Space Vehicle Design

Second Edition

Michael D. Griffin
Oak Hill, Virginia

James R. French
Las Cruces, New Mexico

EDUCATION SERIES
Joseph A. Schetz
Series Editor-in-Chief
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Published by
American Institute of Aeronautics and Astronautics, Inc.
1801 Alexander Bell Drive, Reston, VA 20191-4344

American Institute of Aeronautics and Astronautics, Inc., Reston, Virginia

2 3 4 5

Library of Congress Cataloging-in-Publication Data

Griffin, Michael D. (Michael Douglas), 1949–
Space vehicle design / Michael D. Griffin, James R. French. – 2nd ed.
p. cm. – (AIAA education series)
Includes bibliographical references and index.
ISBN 1-56347-539-1
1. Space vehicles–Design and construction. I. French, James R. II.
Title. III. Series.

TL875.G68 2004
629.47–dc22

2003019167

ISBN 1-56347-539-1

Foreword

This second edition of *Space Vehicle Design* by Michael D. Griffin and James R. French is an updated, thorough treatment of an important and rapidly evolving subject in the aerospace field. The first edition has been a valuable part of the AIAA Education Book Series, and we are very pleased to welcome this new edition to the series. The second edition features the addition of a new chapter on reliability analysis, as well as more and updated technical material and many excercises.

This design textbook is arranged in a logical fashion starting with mission considerations then spacecraft environment, astrodynamics, propulsion, atmospheric entry, attitude control, configuration and structures, subsystems, and finally reliability, so that university courses at different academic levels can be based upon it. In addition, this text can be used as a basis for continuing education short courses or independent self study. The book is divided into 12 chapters and 2 appendices covering more than 600 pages.

The AIAA Education Series aims to cover a broad range of topics in the general aerospace field, including basic theory, applications, and design. A complete list of titles published in the series can be found on the last pages in this volume. The philosophy of the series is to develop textbooks that can be used in a college or university setting, instructional materials for intensive continuing education and professional development courses, and also books that can serve as the basis for independent self study for working professionals in the aerospace field. Suggestions for new topics and authors for the series are always welcome.

Joseph A. Schetz
Editor-in-Chief
AIAA Education Series

Foreword to the Previous Edition

The publication of *Space Vehicle Design* by Michael D. Griffin and James R. French satisfies an urgent need for a comprehensive text on space systems engineering. This new text provides both suitable material for senior-level courses in aerospace engineering and a useful reference for the practicing aerospace engineer. The text incorporates several different engineering disciplines that must be considered concurrently as a part of the integrated design process and optimization. It also gives an excellent description of the design process and its accompanying tradeoffs for subsystems such as propulsion, power sources, guidance and control, and communications.

The text starts with an overall description of the basic mission considerations for spacecraft design, including space environment, astrodynamics, and atmospheric reentry. Then the various subsystems are discussed, and in each case both the theoretical background and the current engineering practice are fully explained. Thus the reader is exposed to the overall systems-engineering process, with its attendant conflicting requirements of individual subsystems.

Space Vehicle Design reflects the authors' long experience with the spacecraft design process. It embodies a wealth of information for designers and research engineers alike. But most importantly, it provides the fundamental knowledge for the space systems engineer to evaluate the overall impact of candidate design concepts on the various component subsystems *and* the integrated system leading to the final design selection.

With the national commitment to space exploration, as evidenced by the continuing support of the Space Station and the National Aero-Space Plane programs, this new text on space system engineering will prove a timely service in support of future space activities.

J. S. PRZEMIENIECKI
Editor-in-Chief
AIAA Education Series
1991

Table of Contents

Preface

We can only smile, more than a bit ironically, when we read the preface to the first edition of this text, which follows. Much has changed, both in the space community and in the larger world, in the 13 years since that edition appeared. Even more has changed in the two decades since the project was originally begun. One thing that has not is the difficulty of shoehorning a book project, even a "mere" revision, into lives dominated by professional careers. We are not unique in that regard; still, we would not have guessed that the production of this second edition would have required twice the time of the first.

Our earlier comments concerning the dearth of texts in the general field of space vehicle systems engineering and design now seem quaint. There are many excellent offerings, as well as in the various allied specialty disciplines. An even greater collection of core knowledge, tutorial material, mathematical "applets," and design data is available on the World Wide Web, which did not even exist when the first edition was published. Why, then, this new edition? Because we hope, and believe, that this text continues to fulfill its original goal, that of linking and integrating the many disciplines relevant to the field of space systems engineering in a way that is impossible when they are considered separately, or even in one text that is the product of many authors.

We have attempted to update the material to make the treatment consistent with current experience and practice in the field. At the same time, there is much that remains relevant from what are now the earlier decades of the space program. We have endeavored to omit nothing of real value merely on the grounds that it is old.

This edition contains a new chapter on reliability analysis, much new technical material in other sections, and many homework problems. As always, we regret that it cannot contain more. We constantly grappled with decisions on what to include and what to omit, both to control the scope of the text and to allow it to be completed—eventually.

Finally, we had to address the issue of how to treat the wealth of material available online. The temptation was strong to use more of it than we did in preparing this edition, and to reference it appropriately in the reference and bibliographic sections at the end of each chapter. As one example among dozens, it seems silly in some respects to include material on RF link analysis, as we have done in Chapter 11, when dozens of such "applets" are available on the web. The same can be said of orbit dynamics calculations, Euler angle visualization tools, and so on, almost literally ad infinitum. In the end, however, we decided against the inclusion of such material, and have included and referenced only that which is accessible through archived references.

We made this choice for the reason that, despite the incredible richness of web-based resources for the modern engineer, it remains true that most websites

and links are exceedingly volatile. We felt that this volatility would likely result in more irritation to the user than if he were left to the good graces of his favorite search engine. Suffice it to say, however, that every topic, and every subtopic, in this text can be explored in full detail online by those with the curiosity to do so. And, there is always the third edition. . . .

Michael D. Griffin
James R. French
November 2003

Preface to the Previous Edition

The idea for this text originated in the early 1980s with a senior-level aerospace engineering course in Spacecraft Design, taught by one of us at the University of Maryland. It was then a very frustrating exercise to provide appropriate reference materials for the students. Space vehicle design being an extraordinarily diverse field, no one text—in fact, no small group of texts—was available to unify the many disciplines of spacecraft systems engineering. As a consequence, in 1983 we decided to collaborate on a unifying text. The structure and academic level of the book followed from our development of a professional seminar series in spacecraft design. To meet the needs of engineers and others attending the seminars, the original academic course notes were radically revised and greatly expanded; when complete, the notes formed the outline for the present textbook.

The book meets, we believe, the needs of an upper-level undergraduate or Master's-level graduate course in aerospace vehicle design, and should likewise prove useful at the professional level. In this regard, our text represents somewhat of a departure from the more conventional academic style; it generally omits first-principle derivations in favor of integrating results from many specialized technical fields as they pertain to vehicle design and engineering tradeoffs at the system level.

It has been a long and torturous path to publication. Writing the manuscript was the easy part; publication was much more difficult. In the mid-1980s various publishers (not AIAA) showed discomfort with a perceived low-volume, "niche" product and backed away from the commitment we wanted. Job changes and the authors' busy schedules forced additional delays. And despite all the time it has taken to obtain the finished product, we both see many changes and improvements we would have liked to have made—but that would doubtless be true no matter how long we had worked.

In any event, the job is done for now. To all who have begun conversations with us in the last several years with, "When is the book coming out?," here it is. We hope you find it worth the wait.

Michael D. Griffin
James R. French
November 1990

1
Introduction

1.1 Introduction

In this book we attempt to treat the major engineering specialty areas involved in space vehicle and mission design from the viewpoint of the systems engineer. To attain this breadth, the depth of coverage in each area is necessarily limited. This is not a book for the specialist in attitude control, propulsion, astrodynamics, heat transfer, structures, etc., who seeks to enhance his knowledge of his own area. It is a book for those who wish to see how their own specialty is incorporated into a final spacecraft design and for those who wish to add to their knowledge of other disciplines.

To this end we have subordinated our desires to include involved analyses, detailed discussions of design and fabrication methods, etc. Equations are rarely derived, and never when they would interfere with the flow of the text; however, we take pains to state the assumptions behind any equations used. We believe that the detailed developments appropriate to each specialty area are well covered in other texts or in the archive journals. We refer the reader to these works where appropriate. Our goal in this work is to show how the knowledge and constraints from various fields are synthesized at the overall system level to obtain a completed design.

We intend this book to be suitable as a text for use in a senior- or graduate-level design course in a typical aerospace engineering curriculum. Very few students emerge from four years of schooling in engineering or physical science feeling comfortable with the larger arena in which they will practice their specialty. This is rarely their fault; academic work by its nature tends to concentrate on that which is known and done, and to educate the student in such techniques. This it does very well, subject of course to the cooperation of the student. What is not taught is how to function in the fact of the unknown, the uncertain, and the not-yet-done. This is where the practicing engineer or scientist must learn to synthesize his knowledge, to combine the specialized concepts he has learned in order to obtain a new and useful result. This does not seem to be a quality that is taught in school.

It is also our intention that this book be useful as a reference tool for the working engineer. With this in mind, we have included as much state-of-the-art material as practicable in the various areas that we treat. Thus, although we discuss the methods by which, say, rocket vehicle performance is analyzed, we are under no illusion that analytical methods produce the final answers in all

cases of interest. We therefore include much more data in tabular and graphic form on the actual performance and construction of various rocket vehicles. We follow the same philosophy for attitude control, guidance, power, telecommunications, and for the other specialty areas and systems discussed here. However, this is not a "cookbook" or a compendium of standard results that can be applied to every problem. No book or course of instruction can serve as a solution manual to all engineering problems. In fact, we take as an article of faith that, in any interesting engineering work, one is paid to solve previously unsolved problems. The most that any text can do is to provide a guide to the fundamentals. This we have tried to do by providing both data and analytical results, with a chain of references leading to appropriate sources.

1.2 Systems Engineering Process

1.2.1 What Is Systems Engineering?

The responses to this question are many and varied. To some who claim to practice systems engineering, the activity seems to mean maintaining detailed lists of vehicle components, mass properties, and the name, number, and pedigree of each conductor that crosses the boundary between any two subsystems. To others it means computer architecture and software with little or no attention to hardware. To still others it means sophisticated computer programs for management and decision making, and so on.

In the opinion of the authors, definitions such as these are too restricted. As with the fabled blind men describing the elephant, each perceives some element of fact, but none fully describes the beast. As an aid to understanding the purpose of this book, we offer the following definition:

> Space systems engineering is the art and science of developing an operable system capable of meeting mission requirements within imposed constraints including (but not restricted to) mass, cost, and schedule.

Clearly, all of the concepts mentioned earlier, plus many more, play a part in such an activity. Some may feel that the definition is too broad. That, however, is precisely the point. Systems engineering, properly done, is perhaps the broadest of engineering disciplines. The space systems engineer has the responsibility of defining a system based on requirements and constraints and overseeing its creation from a variety of technologies and subsystems.

In such a complex environment, conflict is the order of the day. The resolution of such conflict in an effective and productive manner is the goal of systems engineering. For all of today's high technology and sophisticated analytical capability, the solution is not always clear. This, plus the fact that one is dealing with people as much as with hardware or software, accounts for the inclusion of the word "art" in our definition. There will come a time in any system

development when educated human judgment and understanding will be worth more than any amount of computer analysis. This in no way demeans the importance of detailed analysis and the specialists who perform it, but, applied without judgment or conducted in an atmosphere of preconception and prejudice, such analysis can be a road to failure. This truth has been demonstrated more than once, unfortunately, in the history of both military and civilian technical developments. It is the task of the systems engineer to avoid these pitfalls and to make the technical decisions that best serve the achievement of the goal outlined in our definition.

1.2.2 Systems Engineering Requirements

To perform the task, there are certain characteristics that, if not mandatory, are at least desirable in the systems engineer. These are presented and amplified in this section.

The systems engineer must have an understanding of the goals of the project. These may be scientific, military, or commercial. Whatever the case, it is not possible to meet these goals without a full understanding of them. Decisions made without full knowledge of their context are subject to errors that would otherwise be avoided. Not only must the systems engineer understand the goals, but it is incumbent upon him to share this knowledge with his team, so that they too understand the purpose of the effort.

A broad comprehension of the relevant technical issues is mandatory. It is beyond reasonable expectation that the systems engineer be an expert in all disciplines. No single human can aspire to the full breadth and depth of knowledge required in all of the technical specialties relevant to space vehicle design. That is why a broadly capable design team is required for any significant engineering project. However, to make proper use of the resources afforded by such a team, the systems engineer must be sufficiently conversant with each of the relevant technical areas to comprehend the issues and to make appropriate decisions. It is imperative that any technical decision must be evaluated in terms of its effects on the entire system, not just those subsystems most obviously involved. This can be done only if there is a broad understanding of all space vehicle technologies, leading to an appreciation for the unintended, as well as intentional, consequences of a design decision. Ideally the systems engineer should be able to carry out a preliminary analysis in most aerospace disciplines. This, as much as any other single factor, is the primary motivation for this text.

There are individual traits and organizational practices that commend themselves to systems engineering, and others that do not. The university system has a natural tendency to create specialists rather than generalists, especially in advanced degree programs. Initial advancement within any organization is generally accorded to those who make clearly outstanding contributions within their area of responsibility, often rather narrowly defined. It is therefore quite common to find engineers having substantial credentials of education and

experience, who exhibit great depth of knowledge in a given discipline, but who lack the breadth of knowledge required for effective systems engineering.

This combination of successful performance in a specialized area and excellent academic credentials often results in promotion to a position requiring a systems-oriented viewpoint. If this requirement is recognized, and if the selected individual has the ability and natural inclination to pursue a necessarily broader perspective, this can work very well. If, however, the individual inherently prefers to maintain a narrower view, becoming a "specialist in systems engineering clothing," problems will arise from excessive concentration in some areas and neglect of others. This is not to say that the job cannot be done, but it will probably not be done as well or proceed as smoothly as it would otherwise. Effective systems engineering truly requires a different mindset than that appropriate to more specialized disciplines, and there is little available in the way of formal training and practical experience to allow one to prepare for it.

Given that the systems engineer cannot do everything, and requires the assistance of a design team, it follows that an important characteristic of the systems engineer is the ability to make maximum use of the capabilities of others. Part of this involves the difficult-to-define characteristic of "leadership." However one might define it, the manifestation of leadership of interest here involves obtaining maximum productivity from the team. Again, this is a matter of degree. A team of capable people will usually produce an acceptable product even with poor leadership. However, the same team, properly led, is vastly more effective. Participating in such an effort is generally an enjoyable experience for all who are involved. This aspect of the systems engineering task is discussed in more detail later in this chapter.

The essence of the previous paragraph is that the systems engineer must advocate and embrace, to the maximum extent possible, the hackneyed word "teamwork." It is truly appropriate in this instance in that, if the design team does not function as a fully integrated team rather than as a group of individuals, effectiveness will be diminished. The systems engineer has as one of his duties that of fostering the team spirit.

In any complex system, there is normally more than one solution to a problem. Various requirements will often conflict, and requirements and capabilities will not match perfectly. Success requires compromise. Indeed, it often seems that the essence of the task of systems engineering is to effect a series of compromises along the path to project completion. To those who feel that technical decisions should be pure, free of compromise, and always have a clear answer, the real world of engineering, and especially systems engineering, will bring considerable disappointment. Willingness to compromise within reasonable limits is a vital characteristic of the systems engineer.

The key ingredient in successful systems engineering and design, and in effecting the compromises discussed here, is sound engineering judgment. Engineering analysis is an incredibly useful tool, but not everything that is

important to the success of a project can be analyzed, sometimes because the data or tools are not available, and sometimes because of resource limitations.

Moreover, even when analysis is possible, it must be constantly realized that analytical models used in the practice of engineering are just that—models. Engineering models approximate the real world, some more accurately than others, but no model can do so perfectly. Very often the results derived from such models are ambiguous, or can be understood only in a particular context. Also, such results will always be silent with respect to the importance of physical effects not included in the underlying model. The judgment of the team, and ultimately of the systems engineer, must be the final decision mechanism in such cases.

To some degree, judgment is a characteristic with which one is born. However, to be meaningful its use must be grounded in both education and experience.

1.2.3 Managing the Design Team

We have referred repeatedly to the design team and its importance, which we feel can hardly be overemphasized. A competent multidiscipline team is the most powerful tool at the disposal of the systems engineer. The quality of the product is a direct reflection of the capability of the team and the quality of its leadership. Computer-aided design packages and other analytical tools can enhance the productivity of the team and make the task easier, but cannot substitute for human judgment and knowledge, a point that we have made previously and will continue to emphasize throughout this text.

As mentioned, the reason for using a design team is simply that no single person can have sufficient knowledge in all of the technical discipline areas required to carry out a complex engineering task. The protean "mad scientist" of popular fiction who can carry out a complex project (e.g., a rocket to the moon) unaided is indeed purely fictional. This does not seem to preclude people from trying, however. The authors can point to a number of projects, nameless in this volume to protect us from the wrath of the guilty, that were in fact done as a "one-man show" to the extent that a single individual tried, single-handedly, to integrate the inputs of the specialists rather than to lead the team in a coordinated effort. Uniformly, the output is a system of greater complexity and cost, and lesser capability, than it might have been.

A properly run design team is synergistic in that it is greater than the sum of its parts. If all of the same people were used but kept apart, interacting only with the systems engineer, each would obviously be no less intelligent than when part of a team. Yet experience shows that the well-run team outperforms a diverse array of specialists. The authors attribute this to the vigorous interaction between team members, and to the sharing of knowledge, viewpoints, and concerns that often cause a solution to surface that no individual would have conceived when working alone. Often this is serendipitous; the discussion of one problem may suggest a solution for some other apparently unrelated concern. This can only happen in a closely knit team experiencing frequent interaction.

Although there is no cut-and-dried rule, reasonably frequent design meetings are necessary to promote the concept and sense of a team. Meetings should be held sufficiently often to maintain momentum and to reinforce a habit of attendance. They should not be so frequent as to become boring or to waste time. Except in rare instances, formal, full-team meetings should not be held more frequently than once per week. Intervals greater than two weeks are generally undesirable because of the loss of momentum that ensues. Of course, there will be many individual and subgroup interactions on specific topics once the team is accustomed to working together.

As the leader of the team, there are certain responsibilities borne by the systems engineer. He must ensure that all members contribute. Personality differences among design team members often result in meetings being dominated by a few extroverts, to the exclusion of other introverts who have as much to say, but lack the aggressiveness to assert themselves. The systems engineer must ensure that each individual contributes, both because of his responsibility to foster true teamwork, and because it is important to have all ideas available for consideration, not just those belonging to the extroverts. This may require the systems engineer to ask a few leading questions, or to press for an expanded answer, but this is fully a part of the task of leading the team. So, unfortunately, is that of suppressing the excess verbosity of other individuals!

A phenomenon that plagues many meetings is that of digression from the relevant topic prior to its orderly resolution. In any reasonably large group of people, many spurious thoughts will arise that are not germane to the topic at hand. The group can easily be seduced into following the new line of thought, and ignoring the prior topic. It is the duty of the team leader to prevent excessive deviation from the intended subject, and thus to maintain appropriate focus. Of course, in a long, intense meeting, an occasional digression can be refreshing and can ease the tension. This must be allowed, but with—again—judgment, to prevent the waste of time and, importantly, the failure to address all of the relevant matters.

Equally distressing is the tendency of some to ramble at great length, repeating themselves and offering unnecessary detail. The team leader must intervene, with due sensitivity and concern for the feelings of others, when in his judgment the point of useful return is passed. In a similar vein, a few individuals involved in a discussion concerning the fine details of a problem that appears to be below the reasonable level of interest to the team should be directed to arrange a separate meeting. Again, judgment is required as to when the point of productivity for the team has been reached and passed.

1.3 Requirements and Tradeoffs

As noted earlier, the goal of the process led by the systems engineer is to develop a system to meet the requirements of the project. However, it is rarely if ever true that even the highest-level requirements are edicted in complete,

detailed, and unequivocal form. President John F. Kennedy's famously audacious goal, "...before this decade is out, to land a man upon the moon and return him safely to the Earth" stands in its stark simplicity as one of the few so expressed. Indeed, the pithy enunciation of this top-level requirement has been credited by many as being an important factor behind the ultimate success of Project Apollo. However, most engineers, and most engineering projects, do not benefit from goals so succinctly expressed and so clearly motivated. They instead are usually the result of a complex, interactive process involving a variety of factors that may not be obvious. The following sections will discuss requirements derivation in general terms.

1.3.1 Top-Level Requirements

The basic goals and constraints of a given space mission will generally be defined by the user or customer for the resulting system. Such goals will usually be expressed in terms of the target and activity, e.g., "Orbit Mars and observe atmospheric phenomena with particular emphasis on..." or "Develop a geosynchronous communications satellite capable of carrying 24 transponders operating in...." At the same time, various constraints may be levied such as project start date, launch date, total cost, first year cost, etc. The top-level system requirements will then be derived from these goals and constraints.

Inputs for development of top-level requirements may come from a variety of sources. For example, scientific missions will typically have associated a science working group (SWG) composed of specialists in the field. (Usually these individuals will not be potential investigators on the actual mission to prevent any possible conflict of interest.) This group will provide detailed definition of the science goals of the mission in terms of specific observations to be made, types of instruments, sensors that might be used, etc.

The SWG requirements and desires must be evaluated against the constraints and capabilities that otherwise define the mission. This will often be the systems engineer's most difficult task. The various scientific goals are often in conflict with one another, or with the reality of practical engineering. Scientific investigators in single-minded pursuit of a goal often tolerate compromise poorly. Development of an innovative mission and system design to satisfy as many requirements and desires as possible, while simultaneously achieving a suitable compromise among those which conflict, is a major test of both engineering and diplomatic skill. Furthermore, once slain, the dragon will not remain dead, but continues to revive as the mission and system design and the science payload become better defined.

Nonscientific missions usually have a similar source of inputs. This group may go by various names, but can generically be referred to as a user working group, and represents the needs and desires of the user community. As with science requirements, some of these may be in conflict, and resolution and compromise will be required. In many cases, spacecraft may be single-purpose devices, e.g., a

communications relay satellite. In such a case, the problems with resolution of conflicting requirements are greatly reduced.

The study team itself has the primary responsibility for the development of top-level requirements for the system by turning the mission goals and desires into engineering requirements for the spacecraft, to be later converted into specific numerical requirements. As always, this is a process involving design iteration and compromise in order to establish a realistic set of requirements. Interaction between various subsystem and technology areas is essential to understand the impact of requirements on the complete system, and to minimize the likelihood of expensive surprises.

In some cases, particularly when the mission requires operation at the limits of available technology, various expert advisory groups may contribute to the process. Such groups may provide current data or projections of probable direction and degree of development during the course of the project.

1.3.2 Functional Requirements

Once the top-level requirements are defined, the next step is to derive from them the functional requirements defining what the system and the subsystems of which it is composed must accomplish in order to carry out the mission. Functional requirements are derived by converting the top-level requirements into engineering specifications such as velocity change, orbital elements, instrument fields of view, pointing direction, pointing accuracy, available power, operational duty cycle, and a variety of other parameters.

The derivation of the functional requirements must be done within the context of technical capability and constraints on cost and schedule. This is a critical juncture in the project. Unthinking acceptance of unrealistic requirements on a subsystem, or arbitrary assumptions as to the availability of necessary technology, can lead to major problems with schedule and/or cost. As an example, it is very easy to accept a requirement for a given level of pointing accuracy without critically assessing what the requirement may imply in terms of demands on attitude control sensors and effectors, structural fabrication accuracy and rigidity, etc.

Excessively demanding requirements can increase costs, delay schedule, or both. To avoid this, the proposed requirement should first be evaluated as to its necessity. Is the desired accuracy essential to the mission, or was it selected because of prior experience or heritage that might or might not be relevant? Sometimes a demanding requirement will be levied in a deliberate effort to justify use of an exciting new technology. If one of the mission goals is to advance technology, this may be appropriate; if the goal is to obtain observational data at the lowest cost in the least possible time, it may be essential to avoid performance requirements at or close to state-of-the-art limits.

The preliminary version of the functional requirements document is based on the top-level requirements and a preliminary assessment of the intended

spacecraft capability. At this stage, design details will be limited in most areas, and a great many specific requirements will remain "to be determined" (TBD).

The TBDs will be replaced by quantitative values early in the design phase. It is important for the design team to work toward early completion of the functional requirement, and to establish values for the TBD items. Of course, as the design progresses, the functional requirements will evolve and mature. Some requirements will inevitably change; however, striving for early definition helps to accelerate the achievement of requirements stability.

Early definition of functional requirements is desirable, some would even say vital, to program stability and cost control. Probably no single factor has been more to blame for cost and schedule overruns than changing requirements in midprogram. This may happen at the top level or at the functional level. In the former case, the systems engineer has little control, although it is his duty to point out to his management and customers the impact of the change. At the functional requirements level, the systems engineer has substantial control and should exercise it. Absolute inflexibility is, of course, highly undesirable, because circumstances change and some modifications to functional requirements are inevitable. On the other hand, a relaxed attitude in this matter, allowing easy and casual change without adequate coordination and review, is an invitation to disaster.

1.3.3 Functional Block Diagram

The functional block diagram (FBD) is a tool that many people equate with the practice of systems engineering. Indeed, the FBD is a highly useful tool for visualizing relationships between elements of the system. It is applicable at all levels.

The FBD may be used to demonstrate the relationships of major mission and system elements, elements such as spacecraft, ground tracking system, mission operations, facility, user, etc. At the next level, it might be used to indicate the interaction of major subsystems within a system. An example is a diagram showing the relationship between the major spacecraft subsystems that comprise a spacecraft.

The basic concept can be carried to as low a level as desired. A block diagram showing the relationships between the major assemblies within a subsystem, e.g., solar arrays, batteries, and power conditioning and control electronics within the power subsystem, can be most useful. One must be careful not to push it too far, however. Although in principle the FBD could be carried to the point of showing relationships between individual components, this really is not useful; indeed, it can be actively harmful. It must be remembered that once the decision is made to create such documentation, it must be maintained as and when the design changes. If not current, a given document can be not only irrelevant but also damaging. It becomes a source of misinformation, leading to costly and possibly

dangerous errors. Maintaining the accuracy of required program documentation can be a major task.

It is easy today to be seduced into creating overly complex and unnecessary paper systems. There is a multitude of software available to "help" the manager. Once created, these systems seem to take on a life of their own, to expand and propagate. Significant amounts of time and money can be wasted in creating excessive documentation. The systems engineer should think through the documentation requirements for his activity, and implement a plan to meet them. Unnecessary "bells and whistles" that do not contribute to meeting the established requirements should be avoided, or else they will exact a price later.

1.3.4 Tradeoff Analysis

Tradeoff analysis is the essence of mission and system design. The combination of requirements, desires, and capabilities that define a mission and the system that accomplishes it rarely fit together smoothly. The goal of the system designer is to obtain the best compromise among these factors, to meet the requirements as thoroughly as possible, to accommodate various desires, and to do so within the technical, financial, and schedule resources available.

Much has been said and written about how to do tradeoff analyses at the system and subsystem level. At one time it was admittedly a heuristic process, in plainer terms, a "judgment call." Decisions were made through the application of experience and intuition applied to the desires and requirements, the analytical results, and the available test data. More recently, what has become virtually a new industry has arisen to "systemize" (some would say "legitimize") the process. Elaborate mathematical decision-theoretic analyses and the computers to implement them are now commonplace. It is debatable whether better results are achieved in this fashion; without doubt, it has led to greater diffusion of responsibility for decisions. This can hardly be a virtue, since any engineer worthy of the name must be willing to stand behind his work. In the case of the systems engineer, his work consists of the decisions he makes.

What is sometimes overlooked is the fact that, even with the use of computer analyses, engineering decisions are still, at bottom, based on the judgment of individuals or groups who determine the weighting factors, figures of merit, and algorithms that go into the models. Although technical specialists in various subsystems provide the expertise in their particular areas, it is the responsibility of the systems engineer to ensure that all pertinent factors are included and properly weighted. This should not be construed as an argument against the use of computers or any other labor-saving device allowing a more detailed analysis to be done, or a wider range of options to be explored. It is rather to point out that such means are only useful with the proper inputs, and in the hands of one with the knowledge and understanding to evaluate the output intelligently.

It may be instructive to consider some examples of tradeoffs in which a systems engineer might become involved. Note that we do not give the answers

per se, merely the problems and some of the considerations involved in solving them. As we have indicated, there is rarely only one right answer. The answer, a completed system design, will be specific to the circumstances.

1.3.4.1 Spacecraft propulsion trades. Onboard spacecraft propulsion requirements vary widely, ranging from trajectory correction maneuvers of 100 or 200 m/s, to orbit insertion burns requiring a change in velocity (ΔV) on the order of 1000–2000 m/s. Options for meeting these requirements may include solid propulsion, liquid monopropellant or bipropellant, or some form of electric propulsion. Some missions may employ a combination of these.

Solid motors have the virtue of being simple and reliable. The specific impulse (see Chapter 5) is not as high as for most bipropellant systems, but the mass ratio (preburn to postburn mass; again see Chapter 5) is usually better. If the mission requires a single large impulse, a solid may be the best choice. However, relatively high acceleration is typical with such motors, which may not be acceptable for a delicate structure in a deployed configuration.

A requirement for multiple maneuvers usually dictates the use of a liquid-propulsion system. The choice of a monopropellant or bipropellant is not necessarily obvious, however. The specific impulse of monopropellants tends to be one-half to two-thirds that of bipropellants; however, a monopropellant system has half the number of valves and tanks, and operates with a cooler thrust chamber. For a given total impulse, the mass of monopropellant carried must be greater, but the total propulsion system mass, not merely the propellant mass, is the relevant quantity. It will also be true that, if launch vehicle capability allows it, the greater simplicity of a monopropellant system may favor this choice even for relatively large ΔV requirements. Often a solid rocket will provide the major velocity change, whereas a low-thrust mono- or bipropellant system will provide thrust vector control during the solid burn, as well as subsequent orbit maintenance and correction maneuvers.

Electric propulsion offers very low thrust and very high specific impulse. Obviously it is most attractive on vehicles that have considerable electric power available. Applications requiring continuous low thrust for long periods, very high impulse resolution (small "impulse bits"), or minimum propellant consumption may favor these systems. Some examples that have been identified are communications satellites in geosynchronous orbit (see Chapter 2), where long-period, low-impulse stationkeeping requirements exist, and comet rendezvous missions, where the total impulse needed exceeds that available with chemical propulsion systems.

1.3.4.2 Communications system trades. Telecommunications requirements are driven by the amount of information to be transmitted, the time available to do so, and the distance over which it must be sent. Given the required data rate, the tradeoff devolves to one between antenna gain (which, if it is a parabolic dish, translates directly to size) and broadcast power. In the

present discussion, we assume that the antenna is a parabolic dish. For a given data rate and a specified maximum bit error rate with known range and power, the required antenna size is a function of operating frequency.

Antenna size can easily become a problem, because packaging for launch may be difficult or impossible. Antennas that fold for launch and are deployed for operation in space may avoid the packaging difficulty, but introduce cost and reliability problems. Also, such antennas are of necessity usually rather flexible, which, for large sizes, may result in rather poor figure control. Without good figure control, the potential gain of a large antenna cannot be realized. Larger antennas have other problems as well. Increased gain (with any antenna) implies a reduced beamwidth that results in a requirement for more accurate antenna and/or spacecraft pointing knowledge and stability. This can reverberate through the system, often causing overall spacecraft cost and complexity to increase. Orientation accuracy for many spacecraft is driven by the requirements of the communications system.

Higher broadcast power allows use of a smaller antenna, but will naturally have a significant effect on the power subsystem, increasing mass and solar array size. If flight-qualified amplifiers of adequate power do not exist, expensive development and qualification of new systems must be initiated.

Use of higher frequencies (e.g., X-band as opposed to S-band) allows increased data rates for a given antenna size and power, but, because the effective gain of the dish is higher at higher frequencies, again there results a requirement for increased pointing accuracy. Also, if communication with ground stations must be guaranteed, the use of high frequencies can become a problem. Heavy rain can attenuate X-band signals significantly and may obliterate higher frequencies such as Ka- or Ku-band.

In the final analysis, the solution may not lie within the hardware design at all. More sophisticated onboard processing or data encoding can reduce the amount of data that need to be transmitted to achieve the same information transfer (or reduce the bit error rate), to a point compatible with constraints on power, mass, antenna size, and frequency. Of course, this alternative is not free either. More computational capability will be required, and careful (e.g., expensive) prelaunch analysis must be done to ensure that the data are not unacceptably degraded in the process. The cost of developing and qualifying the software for onboard processing is also a factor to be considered.

1.3.4.3 Power system trades. Spacecraft power sources to date have been limited to choices between solar photovoltaic, isotope-heated thermoelectric, and chemical (batteries or fuel cells) sources. Generally speaking, batteries or fuel cells are acceptable as sole power sources only for short-duration missions, measured in terms of days or at most a few weeks. Batteries in particular are restricted to the shorter end of the scale because of limited efficiency and unfavorable power-to-mass ratio. Fuel cells are much more

efficient but are more complex. They have the advantage of producing potable water, which can be an advantage for manned missions.

Solar photovoltaic arrays have powered the majority of spacecraft to date. The simplicity and reliability of these devices make them most attractive. They can be used as close to the sun as the orbit of Mercury, although careful attention to thermal control is required. New technology in materials and fabrication will allow use even closer than the Mercury orbit. Such arrays can provide power as far out as the inner regions of the asteroid belt. With concentrators, they may be useful as far from the sun as the orbit of Jupiter, although the complexity of deployable concentrators has limited interest in these devices until recently. In the future, man-tended assembly or deployment in space may render such concepts more attractive. Batteries are usually required as auxiliary sources when solar arrays are used to provide overload power or power during maneuvers and eclipse periods.

For long missions far from the sun, or for missions requiring full operation during the night on a planetary surface, radioisotope thermoelectric generators (RTG) have been the choice (as with Voyager 1 and 2, the Viking landers, and the Apollo lunar surface experiments packages). These units are long lived, and produce steady power in sunlight or darkness. They tend to be heavy, and the radiation produced can be a problem for electronics and science instruments, especially gamma ray spectrometers.

All of the sources mentioned earlier have difficulty when high power is desired. Deployable solar arrays in the 10–20 kW range are now relatively common, if not cheap, and individual solar arrays for the International Space Station are in the 75-kW range. Larger arrays have been proposed and are probably possible, but present a variety of problems in terms of drag, maneuverability, articulation control, interaction with spacecraft attitude control, etc. Solar dynamic heat engines using Rankine, Brayton, or Stirling cycles driving an electrical generator or alternator have been proposed. These take advantage of the higher efficiency of such thermodynamic cycles as compared to that of solar cells; however, none has yet been flown. As mentioned, all solar power systems suffer from operational constraints due to eclipse periods and distance from the sun.

Nuclear power plants (reactors) offer great promise for the future, offering a combination of high power at moderate weight for long periods. As will be discussed later, however, such units introduce substantial additional complexity into both mission and spacecraft design, not to mention the political problems of obtaining approval for launch. In the final analysis, the spacecraft designer must trade off the characteristics and requirements of all systems to choose the best power source or combination of sources for his mission.

The preceding examples of tradeoff considerations are by no means all that will be encountered in the design of a spacecraft system. They are merely a few examples of high-level trades on major engineering subsystems. The process becomes more complex and convoluted as the system develops, and occurs at

every level in the design. Every technologist in every subsystem area will have his favored approach, often with little regard to its system value. The task of the systems engineer is to evaluate the overall impact of these concepts on all of the other subsystems and upon the integrated system before making a selection.

1.3.4.4 Technology tradeoffs. A difficult area for decisions is that of using new vs existing technology. The systems engineer is often caught between opposing forces in this matter. On one side is program and project management, who, in general, are primarily interested in completing the job on schedule, within budget, and with minimum uncertainty. To this end, management tends to apply pressure to "do what you did last time;" i.e., minimize the introduction of new concepts or technology with their attendant risk and uncertainty.

On the other side is a host of technical specialists responsible for the various spacecraft subsystems. These people are more likely to be interested in applying the most current technology in their field, and will have very little interest in flying the "same old thing" again, particularly if several years have elapsed.

The dichotomy here is real, and the decision may be of profound significance. To maximize capability, remain competitive, encourage new development, etc., it is clearly desirable to apply new technology when possible. Yet one must avoid being seduced by a promise or potential that is not yet real. It is almost axiomatic that any project pushing the state of the art in too many areas will, even if ultimately successful, be both late and expensive.

In a properly managed program it will be the lot of the systems engineer either to make the technology decision or to make recommendations to management so that the issue can be properly decided. Many issues must be considered in this matter; some of these will be discussed in the remainder of this section.

The first question to be addressed is the most basic: "Will the existing technology do the job?" If a well-understood technology embodied in existing systems will do everything required with a comfortable margin, then there is little incentive to do something new merely because it is new. On the other hand, if the task mandates the use of new technology to be accomplished at all, the decision is again obvious. It then becomes the task of the systems engineer to define, as accurately as possible, the effect on cost and schedule and the risks that may be involved, with regard to the total system.

The cost impact of incorporating new technology can be highly variable. Savings may be realized because of higher efficiency, lower mass, lower volume, or all of these. These effects can propagate through the entire system, reducing structural mass, power demands, etc. However, changes such as this usually reduce cost only if the entire system is being designed to incorporate the new approach. If the spacecraft in question is merely one in a long series, and other subsystems are already designed (or even already built), then full realization of the potential advantages is unlikely. Attempting to capture such advantages would require redesign of most of the other subsystems, resulting in what is

effectively a new system design, and in all likelihood actually increasing overall costs.

This example points to a major risk associated with the introduction of new technology and emphasizes the need for the systems engineer to focus on the complete system, and upon the unforeseen ways in which changes in a subsystem may propagate. A subsystem engineer might propose introduction of a new technology item in his subsystem after the design is well advanced. The advantages cited might be higher efficiency, greater capability, or just the fact that it is the latest technology. It will probably be argued that the cost increase within the subsystem will be small or nonexistent. The subsystem engineer's interest (and the depth of his argument) will usually end at that point. The systems engineer must look beyond this, addressing other questions that include, but are not necessarily limited to, some of the following: If ground support and test equipment already exist, will they be compatible with the new change, or will extensive modifications be required? Will new or special test and handling requirements be invoked (e.g., static electricity precautions, inert gas purge, etc.)? Probably the most important questions relate to the effect on other subsystems. Is this change truly transparent to them, or will new requirements (e.g., noise limits, special power requirements or restrictions, etc.) be imposed? Will the new item affect mission planning because of greater radiation sensitivity (or require shielding mass, which negates some of the purported advantages)? Failure to assess these issues early, and to coordinate with the designers of other subsystems during the decision process, can lead to very costly surprises later.

Another area of concern is that of the actual availability of components based on the new technology. Demonstrations in the laboratory, even fabrication of test components, do not correspond to actual production availability. Even if commercial parts are available, the space-qualified units required for most projects may not be. Thus, commitment to the new item could imply that the system engineer's project must pay the cost of establishing a production line or a space-qualification program. This may not only be costly, but may also be incompatible with the project schedule.

Of course, the issue of component availability question has two sides. It may be equally difficult to obtain older components if several years have passed since their previous use in an application. This is especially true in the rapidly evolving electronics component field. A case in point is that of the Voyager spacecraft, in which it was desired to duplicate many electronic subsystems from the Viking Orbiter. To the dismay of project management, it was found that the manufacturer was terminating production of certain critical integrated circuits, and was not interested in keeping the line open in order to produce the relatively small volume of parts needed. Because the redesign necessary to incorporate new components would have been both expensive and late, the project paid to maintain the production line for the required parts. In a more recent example, space shuttle program officials have found it necessary to resort to on-line auctions to identify and procure what are, as of the early twenty-first century, quite outmoded parts.

This issue is not unique to electronic systems. Increasingly restrictive environmental rules or political events may restrict the availability of structural alloys or particular materials that were readily available a few years earlier.

It might be construed from this discussion that the authors are opposed to the use of new technology unless there is no other choice. This is by no means the case; all else being equal, one would almost always choose to implement a proposed new technology. Unfortunately, new technology is often promoted quite optimistically, with little consideration of its possible unintended consequences. All sides of the issue must be assessed in order to make a proper decision, and the person responsible for so doing is the systems engineer, with the support of technical experts.

It must be equally understood that excessive concern with the problems just discussed can cause organizational or program management to adopt a somewhat "bearish" approach to the adoption of new technology. This can result in adherence to old approaches long after newer, safer, more effective capabilities have become available and well proved. It is as much the responsibility of the systems engineer to avoid this trap as it is to avoid prematurely adopting new technology for the reasons discussed. The challenge is to know which approach to follow, and when.

Bibliography

Augustine, N. F., *Augustine's Laws*, 6th ed., AIAA Reston, VA, 1997.

Goldberg, B. E., Everhart, K., Stevens, R., Babbitt, N. III, Clemens, P., and Stout, L., "System Engineering 'Toolbox' for Design-Oriented Engineers," NASA RP-1358, Dec. 1994.

"NASA Systems Engineering Handbook," NASA SP-6105, June 1995.

"Readings in System Engineering," NASA SP-6102, 1993.

Ryan, R. S., "A History of Aerospace Problems, Their Solutions, Their Lessons," NASA TP-3653, Sept. 1996.

Ryan, R. S., Blair, J., Townsend, J., and Verderaime, V., "Working on the Boundaries: Philosophies and Practices of the Design Process," NASA TP-3642, July 1996.

"What Made Apollo a Success?" NASA SP-287, 1971.

2
Mission Design

2.1 Introduction

Space vehicle design requirements do not, except in very basic terms, have an existence that is independent of the mission to be performed. In fact, it is almost trivial to note that the type of mission to be flown and the performance requirements that are imposed define the spacecraft design that results. Just as a wide variety of aircraft exist to satisfy different broad classes of tasks, so may most space missions be categorized as belonging to one or another general type of flight. Missions to near Earth orbit, for example, will impose fundamentally different design requirements than planetary exploration missions, no matter what the end goal in each case. In this chapter we examine a variety of different mission classes, with a view to the high-level considerations that are thus imposed on the vehicle design process.

2.2 Low Earth Orbit

Low Earth orbit (LEO) can be loosely defined as any orbit that is below perhaps 1000 km, or generally below the inner Van Allen radiation belt. By far the majority of space missions flown to date have been to LEO, and it is probable that this trend will continue. Examples of LEO missions include flight tests, Earth observations for scientific, military, meteorological, and other utilitarian purposes, and observations of local or deep space phenomena. Future missions can be expected to have similar goals plus the addition of new classes for purely commercial purposes. Indeed, the first generation of such commercial missions began appearing at the turn of the century, which saw the advent of global voice and data networks in LEO, commercial FM radio broadcasting, and the first purely commercial Earth observation and photoreconnaissance satellites. The fact that none of the business ventures founded on these mission concepts has yet proved profitable has delayed more aggressive efforts to exploit the LEO environment. Nonetheless, it is widely believed that the purely commercial use of near-Earth space can only grow. Further examples of such missions may include delivery service to the International Space Station, space materials processing, and more sophisticated Earth resource survey spacecraft.

2.2.1 Flight Tests

In the early days of orbital flight, every mission was in some sense a flight test, regardless of its primary goals, simply because of the uncertainty in technology and procedures. With increasing technical and operational maturity, however, many missions have become essentially routine. In such cases, flight tests are conducted only for qualification of new vehicles, systems, or techniques.

Flight tests in general are characterized by extensive instrumentation packages devoted to checking vehicle or system performance. Mission profiles are often more complex than for an operational mission because of the desire to verify as many modes of operation as possible. There is a close analogy with aircraft flight testing, where no real payload is carried and the performance envelope is explored to extremes that are not expected to be encountered under ordinary conditions.

An important difference arises in that aircraft testing will involve many hours of operation over many flights, probably with a number of test units. Space systems, on the other hand, are usually restricted to one or very few test units and one flight per operational unit. It is interesting to recall that Apollo 11, the first lunar landing mission, was only the fifth manned flight using the command module, the third to use a lunar module, and in fact only the 21st U.S. manned mission. The space shuttle provides the first instance of multiple flight tests of the same unit. Even in this case, the number of test flights was very low by aircraft standards, with the vehicle having been declared "operational" after only four flights. As this is written, 113 space shuttle missions have been flown, with no single crewmember having been on more than seven flights. One can hardly imagine, for example, Lindbergh having flown the Atlantic on the basis of such limited experience.

Because of the limited number of flight tests usually allowed for space systems, it is essential that a maximum value be obtained from each one. Not only must the mission profile be designed for the fullest possible exercise of the system, but the instrumentation package must provide the maximum return. LEO offers an excellent environment for test missions. The time to reach orbit is short, the energy expenditure is as low as possible for a space mission, communication is nearly instantaneous, and many hours of flight operation may be accumulated by a single launch to orbit.

As indicated earlier, the Apollo manned lunar program is an excellent example of this type of testing. The various vehicles and procedures were put through a series of unmanned and manned exercises in LEO prior to lunar orbit testing and the lunar landing. Even the unmanned first flight of the Saturn 5/Apollo command service module (CSM) illustrates the philosophy of striving for maximum return on each flight. This flight featured an "all-up" test of the three Saturn 5 stages, plus restart of the third stage in Earth orbit, as required for a lunar mission, followed by a reentry test of the Apollo command module. Viewed as a daring (and spectacularly successful) gamble at the time, it is seen in retrospect that little if any additional program risk was incurred. If the first stage had failed, nothing would have been learned about the second and higher stages—exactly

the situation if dummy upper stages had been used until a first stage of proven reliability had been obtained. Moreover, a failure in any higher stage would still have resulted in obtaining more information than would have been the case with dummy upper stages. Of course, the cost of all-up testing can be much higher if repeated failures are incurred. However, even here equipment costs must be traded off against manpower costs incurred when extra flights are included to allow a more graduated testing program. Even if equipment costs alone are considered, one must note that, when testing upper stages, many perfectly good lower stages must be used to provide the correct flight environment.

The systematic flight-test program for Apollo, leading to a lunar landing after a series of manned and unmanned flights, is apparent in Table 2.1. This table is not a complete summary of all Apollo flight tests. Between 1961 and 1966 some 10 Saturn 1 flights were conducted, of which three were used to launch the Pegasus series of scientific missions. Also, two pad-abort and four high-altitude

Table 2.1 Summary of Apollo test missions

Date	Mission	Comments
Feb. 26, 1966	AS-201	Saturn 1B first flight. Suborbital mission testing command service module (CSM) entry systems at Earth orbital speeds. Partial success due to loss of data.
Aug. 25, 1966	AS-202	Successful repeat of AS-201.
July 5, 1966	AS-203	Orbital checkout of S-4B stage. No payload.
Nov. 9, 1967	AS-501 (Apollo 4)	Saturn 5 first flight. Test of Apollo service propulsion system (SPS) restart capability and reentry performance at lunar return speeds.
Jan. 22, 1968	AS-204 (Apollo 5)	Earth orbit test of lunar module (LM) descent and ascent engines.
April 4, 1968	AS-502 (Apollo 6)	Repeat of Apollo 4. Third stage failed to restart. SPS engines used for high-speed reentry tests.
Oct. 11, 1968	AS-205 (Apollo 7)	First manned Apollo flight. Eleven-day checkout of CSM systems.
Dec. 21, 1968	AS-503 (Apollo 8)	First manned lunar orbital flight. Third flight of Saturn 5.
March 3, 1969	AS-504 (Apollo 9)	Earth orbital checkout of lunar module and CSM/LM rendezvous procedures.
May 18, 1969	AS-505 (Apollo 10)	Lunar landing rehearsal; test of all systems and procedures except landing.
July 16, 1969	AS-506 (Apollo 11)	First manned lunar landing. Sixth Saturn 5 flight, fifth manned Apollo flight, third use of lunar module.

tests of the Apollo launch escape system were conducted during this period. However, only "boilerplate" versions of the Apollo spacecraft were used for these missions, and only the first stage of the Saturn 1 was ever employed for a manned flight, and even then its use was not crucial to the program. Adding the third stage of the Saturn 5 (the S-IVB) to an upgraded Saturn 1 first stage resulted in the Saturn 1B mentioned in the table. Table 2.1 summarizes the tests conducted involving major use of flight hardware.

As may be seen in Table 2.1, one class of flight test that does not actually require injection into orbit is entry vehicle testing. There is seldom any advantage to long-term orbital flight for such tests. The entry must be flown in some approximation of real time, and an instrumented range is often desired. Therefore, such tests are usually suborbital ballistic lobs with the goal of placing the entry vehicle on some desired trajectory. Propulsion may be applied on the descending leg to achieve high entry velocity on a relatively short flight. This was, in fact, done on the previously mentioned unmanned Apollo test flights to simulate lunar return conditions. Note that such flight tests may not be required to match precisely the geometry and velocity of a "real-life" mission. If the main parameter of interest is, for example, heat flux into the shield, this may be achieved at lower velocity by flying a lower-altitude profile than would be the case for the actual mission.

Entry flight tests are often performed in the Earth's atmosphere for the purpose of simulating a planetary entry. Typically, it is impossible to simulate the complete entry profile because of atmospheric and other differences; however, critical segments may be simulated by careful selection of parameters. The Viking Mars entry system and the Galileo probe entry system were both tested in this way. The former used a rocket-boosted ballistic flight launched from a balloon, while the latter involved a parachute drop from a balloon to study parachute deployment dynamics.

Launch vehicle tests usually involve flying the mission profile while carrying a dummy payload. In some cases it is possible to minimize range and operational costs by flying a lofted trajectory that does not go full range or into orbit. For example, propulsion performance, staging, and guidance and control for an orbital vehicle can be demonstrated on a suborbital, high-angle, intercontinental ballistic missile (ICBM) like flight.

2.2.2 Earth Observation

Earth observation missions cover the full gamut from purely scientific to completely utilitarian. Both extremes may be concerned with observations of the surface, the atmosphere, the magnetosphere, or the interior of the planet, and of the interactions of these entities among themselves or with their solar system environment.

Missions concerned with direct observation of the surface and atmosphere are generally placed in low circular orbits to minimize the observation distance.

Selecting an orbit altitude is generally a compromise among field of view, ground track spacing, observational swath width, and the need to maintain orbit stability against atmospheric drag without overly frequent propulsive corrections or premature mission termination. In some cases the orbital period may be a factor because of the need for synchronization with a station or event on the surface. In other cases the orbital period may be required to be such that an integral number of orbits occur in a day or a small number of days. This is particularly the case with navigation satellites and photoreconnaissance spacecraft.

Orbital inclination is usually driven by a desire to cover specific latitudes, sometimes compromised by launch vehicle and launch site azimuth constraints. For full global coverage, polar or near-polar orbits are required. Military observation satellites make frequent use of such orbits, often in conjunction with orbit altitudes chosen to produce a period that is a convenient fraction of the day or week, thus producing very regular coverage of the globe. In many cases it is desired to make all observations or photographs at the same local sun angle or time (e.g., under conditions that obtain locally at, say, 1030 hrs). As will be discussed in Chapter 4, orbital precession effects due to the perturbing influence of Earth's equatorial bulge may be utilized to provide this capability. A near-polar, slightly retrograde orbit with the proper altitude will precess at the same angular rate as the Earth revolves about the sun, thus maintaining constant sun angle throughout the year.

The LEO missions having the most impact on everyday life are weather satellites. Low-altitude satellites provide close-up observations, which, in conjunction with global coverage by spacecraft in high orbit, provide the basis for our modern weather forecasting and reporting system. Such spacecraft are placed in the previously mentioned sun-synchronous orbits of sufficient altitude for long-term stability. The Television and Infrared Observation Satellite (TIROS) series has dominated this field since the 1960s, undergoing very substantial technical evolution in that time. These satellites are operated by the National Oceanic and Atmospheric Administration (NOAA). The Department of Defense operates similar satellites under a program called the Defense Meteorological Support Program (DMSP).

Ocean survey satellites, of which SEASAT was an early example, have requirements similar to those of the weather satellites. All of these vehicles aim most of their instruments toward the region directly beneath the spacecraft or near its ground track. Such spacecraft are often referred to as "nadir-pointed."

Many military missions flown for observational purposes are similar in general requirements and characteristics to those discussed earlier. Specific requirements may be quite different, being driven by particular payload and target considerations.

Missions dedicated to observation of the magnetic field, radiation belts, etc., will usually tend to be in elliptical orbits because of the desire to map the given phenomena in terms of distance from the Earth as well as over a wide latitude band. For this reason, substantial orbital eccentricity and a variety of orbital inclinations may be desired. Requirements by the payload range from simple

sensor operation without regard to direction, to tracking particular points or to scanning various regions.

Many satellites require elliptic orbits for other reasons. It may be desired to operate at very low altitudes either to sample the upper atmosphere (as with the Atmospheric Explorer series) or to get as close as possible to a particular point on the Earth for high resolution. In such cases, higher ellipticity is required to obtain orbit stability, because a circular orbit at the desired periapsis altitude might last only a few hours.

2.2.3 Space Observation

Space observation has fully matured with our ability to place advanced scientific payloads in orbit. Gone are the days when the astronomer was restricted essentially to the visible spectrum. From Earth orbit we can examine space and the bodies contained therein across the full spectral range and with resolution no longer severely limited by the atmosphere. (The Mount Palomar telescope has a diffraction-limited resolving power some 20 times better than can be realized in practice because of atmospheric turbulence.) This type of observation took its first steps with balloons and sounding rockets, but came to full maturity with orbital vehicles.

Predictably, our sun was one of the first objects to be studied with space-based instruments, and interest in the subject continues unabated. Spacecraft have ranged from the Orbiting Solar Observatory to the impressive array of solar observation equipment that was carried on the manned Skylab mission. Orbits are generally characterized by the desire that they be high enough that drag and atmospheric effects can be ignored. Inclination is generally not critical, although in some cases it may be desired to orbit in the ecliptic plane. If features on the sun itself are to be studied, fairly accurate pointing requirements are necessary, because the solar disk subtends only 0.5 deg of arc as seen from Earth.

Many space observation satellites are concerned with mapping the sky in various wavelengths, looking for specific sources, and/or the universal background. Satellites have been flown to study spectral regimes from gamma radiation down to infrared wavelengths so low that the detectors are cooled to near absolute zero to allow them to function. An excellent example is the highly successful Cosmic Background Explorer (COBE) spacecraft, with liquid helium at 4.2 K used for cooling. COBE has enabled astronomers to verify the very high degree of uniformity that exists in the 3-K background radiation left over from the "big bang" formation of the universe, and also to identify just enough non-uniformity in that background to account for the formation of the galaxies we observe today. In the x-ray band, the High Energy Astronomical Observatory (HEAO-2) spacecraft succeeded in producing the first high-resolution (comparable to ground-based optical telescopes) pictures of the sky and various sources at these wavelengths. The more sophisticated Chandra spacecraft, operating in a highly elliptic orbit, greatly extends this capability. Although most

such work has concentrated on stellar and galactic sources, there has recently been some interest in applying such observations to bodies in our solar system, e.g., ultraviolet observations of Jupiter or infrared observations of the asteroids.

Despite early problems resulting from a systematic flaw in the manufacture of its primary mirror, the Hubble Space Telescope (HST) represents the first space analog of a full-fledged Earth-based observatory. This device, with its 2.4-m mirror, is a sizeable optical system even by ground-based standards, and offers an impressive capability for deep space and planetary observations of various types. Periodic servicing by the shuttle to conduct repairs, to reboost the spacecraft in its orbit, and to replace outmoded instruments with more advanced versions has made the HST the closest thing yet to a permanent observatory in space. Observations from the HST have extended man's reach to previously unknown depths of space; however, it operates chiefly in the visible band, and so smaller, more specialized observatories will continue to be needed for coverage of gamma, x-ray, and infrared wavelengths.

Radio astronomers also suffer from the attenuating effects of the atmosphere in certain bands, as well as limits on resolution due to the impracticality of large, ground-based dish antennas. Although so far unrealized, there is great potential for radio astronomy observations from space. Antennas can be larger, lighter, and more easily steered. Moreover, the use of extremely high precision atomic clocks allows signals from many different antennas to be combined coherently, resulting in the possibility of space-based antenna apertures of almost unlimited size. Radio observations with such antennas could eventually be made to a precision exceeding even the best optical measurements.

Space observatories are precision instruments featuring severe constraints on structural rigidity and stability, internally generated noise and disturbances, pointing accuracy and stability, etc. Operation is usually complicated by the need to avoid directly looking at the sun or even the Earth and moon. Orbit requirements are not generally severe, but may be constrained by the need for shuttle accessibility while at the same time avoiding unacceptable atmospheric effects, such as excessive drag or interference by the molecules of the upper atmosphere with the observations to be made.

2.2.4 Space-Processing Payloads

As discussed in Chapter 3, the space environment offers certain unique features that are impossible or difficult, and thus extremely expensive, to reproduce on the surface of a planet. Chief among these are weightlessness or microgravity (not the same as absence of gravity; tidal forces will still exist) and nearly unlimited access to hard vacuum. These factors offer the possibility of manufacturing in space many items that cannot easily be produced on the ground. Examples that have been considered include large, essentially perfect crystals for the semiconductor industry, various types of pharmaceuticals, and alloys of metals, which, because of their different densities, are essentially immiscible on Earth.

Space-processing payloads to date have been small and experimental in nature. Such payloads have flown on several Russian missions and on U.S. missions on sounding rockets, Skylab, and the shuttle. The advent of the shuttle, with its more routine access to LEO, has resulted in substantial increases in the number of experiments being planned and flown. The shuttle environment has made it possible for such experiments to be substantially less constrained by spacecraft design considerations than in the past. Furthermore, it is now possible for a "payload specialist" from the sponsoring organization to fly as a shuttle crew member with only minimal training. The International Space Station (ISS) is expected to replace the shuttle as the base for on-orbit experiments. As this is written, fiscal constraints on the ISS are severely eroding crew size and equipment capability, placing the ability of the space station to carry out meaningful experiments in question. In any case, most of the shuttle launch capacity will be consumed in the ISS assembly support for a number of years.

Because manned vehicles, whether space stations or shuttle, are subject to disturbances caused by the presence of the crew, it seems likely that processing stations will evolve into shuttle-deployed free flyers to achieve the efficiency of continuous operation and tighter control over the environment (important for many manufacturing processes) than would be possible in the multi-user shuttle environment. Such stations would require periodic replenishment of feedstock and removal of the products. This might be accomplished with the shuttle or other vehicles as dictated by economics and the current state of the art. In any case, it introduces a concept previously seldom considered in spacecraft design: the transport and handling of bulk cargo. Space processing and manufacturing has not evolved as rapidly as expected. However, the potential is still there and eventual development of such capability seems likely.

Autonomy, low recurrent cost, and reliability will probably be the hallmarks of such delivery systems. The Russian Progress series of resupply vehicles used in the Salyut and Mir space station programs, and now in the resupply of the ISS, may be viewed as early attempts in the design of vehicles of this type. However, the Progress vehicles still depend on the station crew to effect most of the cargo transfer (though liquid fuel was transferred to Mir essentially without crew involvement). It may be desirable for economic reasons to have future resupply operations of this nature carried out by unmanned vehicles. This will add some interesting challenges to the design of spacecraft systems. It seems certain that there will be a strong and growing need for robotics technology and manufacturing methods in astronautics.

In the longer term, the high-energy aspects of the space environment may be as significant as the availability of hard vacuum and $0g$. The sun produces about 1400 W/m^2 at Earth, and this power is essentially uninterrupted for many orbits of possible future interest. The advance of solar energy collection and storage technology cannot fail to have an impact on the economic feasibility of orbital manufacturing operations. In this same vein, it is also clear that the requirement to supply raw material from Earth for space manufacturing processes is a

tremendous economic burden on the viability of the total system. Again, it seems certain that, in the long term, development of unmanned freighter vehicles capable of returning lunar or asteroid materials to Earth orbit will be undertaken. With the advent of this technology, and the use of solar energy, the economic advantage in many manufacturing operations could fall to products manufactured in geosynchronous or other high Earth orbits.

2.3 Medium-Altitude Earth Orbit

In the early days of the space program, most Earth-orbiting spacecraft were either in low Earth orbit or geosynchronous orbit. More recently, however, there has been increasing interest in intermediate orbits, i.e., those with a 12-h period (half-geosynchronous). The Global Positioning System (GPS), an array of satellites supporting the increasingly crucial GPS navigation system, is located in this orbital regime. These orbits avoid the dangerous inner radiation belt but are significantly deeper in the outer belt than geostationary satellites and thus experience a substantially higher electron flux.

2.4 Geosynchronous Earth Orbit

Geosynchronous Earth orbit (GEO), and particularly the specific geosynchronous orbit known as geostationary, is some of the most valuable "property" in space. The brilliance of Arthur Clarke's foresight in suggesting the use of communications satellites in GEO has been amply demonstrated. However, in addition to comsats, weather satellites now occupy numerous slots in GEO.

As the name implies, a spacecraft in GEO is moving in synchrony with the Earth, i.e., the orbit period is that of Earth's day, 24 h (actually the 23 h, 56 m, 4 s sidereal day, as will be discussed in Chapter 4). This does not imply that the satellite appears in a fixed position in the sky from the ground, however. Only in the special case of a 24-h circular equatorial orbit will the satellite appear to hover in one spot over the Earth. Other synchronous orbits will produce ground tracks with average locations that remain over a fixed point; however, there may be considerable variation from this average during the 24-h period. The special case of the 24-h circular equatorial orbit is properly referred to as geostationary.

A 24-h circular orbit with nonzero inclination will appear from the ground to describe a nodding motion in the sky, that is, it will travel north and south each day along the same line of longitude, crossing the equator every 12 h. The latitude excursion will, of course, be equal to the orbital inclination. If the orbit is equatorial and has a 24-h period but is not exactly circular, it will appear to oscillate along the equator, crossing back and forth through lines of longitude. If the orbit is both noncircular and of nonzero inclination (the usual case, to a slight extent, due to various injection and stationkeeping errors), the spacecraft will

appear to describe a figure eight in the sky, oscillating through both latitude and longitude about its average point on the equator. If the orbit is highly inclined or highly elliptic, then the figure eight will become badly distorted. In all cases, however, a true 24-h orbit will appear over the same point on Earth at the same time each day. An orbit with a slightly different period will have a slow, permanent drift across the sky as seen from the ground. Such slightly non-synchronous orbits are used to move spacecraft from one point in GEO to another by means of minor trajectory corrections.

It is also interesting to consider very high orbits that are not synchronous but that have periods that are simply related to a 24-h day. Examples are the 12-h and 48-h orbits. Of interest are the orbits used by the Russian Molniya spacecraft for communications relay. Much of Russia lies at very high latitudes, areas that are poorly served by geostationary comsats. The Molniya spacecraft use highly inclined, highly elliptic orbits with 12-h periods that place them, at the high point of their arc, over Russia twice each day for long periods. Minimum time is spent over the unused southern latitudes. While in view, communications coverage is good, and these orbits are easily reached from the high-latitude launch sites accessible to the Russians. The disadvantage, of course, is that some form of antenna tracking control is required.

The utility of the geostationary or very nearly geostationary orbit is of course that a communications satellite in such an orbit is always over the same point on the ground, thus greatly simplifying antenna tracking and ground–space–ground relay procedures. Nonetheless, as long as the spacecraft drift is not so severe as to take it out of sight of a desired relay point, antenna tracking control is reasonably simple and is not a severe operational constraint, so that near-geostationary orbits are also quite valuable. The same feature is also important with weather satellites; it is generally desired that a given satellite be able to have essentially continuous coverage of a given area on the ground, and it is equally desirable that ground antennas be readily able to find the satellite in the sky.

The economic value of such orbits was abundantly emphasized during the 1979 World Administrative Radio Conference (WARC-79), when large groups of underdeveloped nations, having little immediate prospect of using geostationary orbital slots, nonetheless successfully prosecuted their claims for reservations of these slots for future use. Of concern was the possibility that, by the time these nations were ready to use the appropriate technology, the geostationary orbit would be too crowded to admit further spacecraft. With present-day technology and political realities, this concern is somewhat valid. There are limits on the proximity within which individual satellites may be placed.

The first limitation is antenna beamwidth. With reasonably sized ground antennas, at frequencies now in use (mostly C-band; see Chapter 12), the antenna beamwidth is about 3 deg. To prevent inadvertent commanding of the wrong satellite, international agreements limit geostationary satellite spacing to 3 deg. Competition for desirable spots among nations lying in similar longitude belts has become severe. A trend to higher frequencies and other improvements

(receiver selectivity and the ability to reject signals not of one's own modulation method are factors here) has allowed a reduction to 2-deg spacing, which alleviates but does not eliminate the problem. Political problems also appear, in that each country wants its own autonomous satellite, rather than to be part of a communal platform, a step that could eliminate the problem of inadvertent commands by using a central controller.

There is also the increasing potential of a physical hazard. Older satellites have worn out and, without active stationkeeping, will drift in orbit, posing a hazard to other spacecraft. Also, jettisoned launch stages and other hardware are in near-GEO orbits. All of this drifting hardware constitutes a hazard to operating systems, which is increasing due to the increasing size of newer systems. There is evidence that some collisions have already occurred. Mission designers are sensitive to the problem, and procedures are often implemented, upon retiring a satellite from active use, to lift it out of geostationary orbit prior to shutdown.

2.4.1 Communications Satellites

Of all the facets of space technology, the one that has most obviously affected the everyday life of the average citizen is the communications satellite, so much so that it is now taken for granted. In the early 1950s a tightly scheduled plan involving helicopters and transatlantic aircraft was devised to transport films of the coronation of Queen Elizabeth II so that it could be seen on U.S. television the next day. In contrast, the 1981 wedding of Prince Charles was telecast live all over the world without so much as a comment on the fact of its possibility. Today, most adults cannot recall any other environment. Less spectacular, but having even greater impact, is the ease and reliability of long distance business and private communication by satellite. Gone are the days of "putting in" a transcontinental or transoceanic phone call and waiting for the operator to call back hours later. Today, direct dialing to most developed countries is routine, and we are upset only when the echo-canceling feature does not work properly.

The communications satellites that have brought about this revolution are to the spacecraft designer quite paradoxical, in the sense that in many ways they are quite simple (we exclude, of course, the communications gear itself, which is increasingly capable of feats of signal handling and processing that are truly remarkable). Because, by definition, a communications satellite is always in communication with the ground, such vehicles have required very little in the way of autonomous operational capability. Problems can often be detected early and dealt with by direct ground command. Orbit placement and correction maneuvers can, if desired, be done in an essentially real-time, "fly-by-wire" mode. Most of the complexity (and much of the mass) is in the communications equipment, which is the raison d'être for these vehicles. Given the cost of placing a satellite in orbit and the immense commercial value of every channel, the tendency is to cram the absolute maximum of communications capacity into

every vehicle. Lifespan and reliability are also important, and reliability is usually enhanced by the use of simple designs.

The value of and demand for communications channels, together with the spacing problems discussed earlier, are driving vehicle design in the direction of larger, more complex multipurpose communications platforms. Indeed, economic reality is pushing us toward the very large stations originally envisioned by Clarke for the role, but with capabilities far exceeding anything imagined in those days of vacuum tubes, discrete circuit components, and point-to-point wiring. Also noteworthy is that comsats thus far have been unmanned. This trend will probably continue, although there may be some tendency, once very large GEO stations are built, to allow for temporary manned occupancy for maintenance or other purposes. Pioneering concepts assumed an essential role for man in a communications satellite; as Clarke has said, it was viewed as inconceivable (if it was considered at all) that large, complex circuits and systems could operate reliably and autonomously for years at a time.

A high degree of specialization is already developing in comsat systems, especially in carefully designed antenna patterns that service specific and often irregularly shaped regions on Earth. This trend can be expected to continue in the future. The large communications platforms discussed earlier will essentially (in terms of size, not complexity) be elaborate antenna farms with a variety of specialized antennas operating at different frequencies and aimed at a variety of areas on the Earth and at other satellites.

It will be no surprise that the military services operate comsat systems as well. In a number of cases, such as the latest MILSTAR models, these vehicles have become quite elaborate, with multiple functions and frequencies. Reliability and backup capability are especially important in these applications, as well as provision for secure communications. Of interest to the spacecraft design engineer is the growing trend toward "hardening" of these spacecraft. In the event of war, nuclear or conventional, preservation of communications capability becomes essential. Spacecraft generally are rather vulnerable to intense radiation pulses, whether from nuclear blasts in space (generating electromagnetic pulses as well) or laser radiation from the ground. The use of well-shielded electrical circuits and, where possible, fiberoptic circuits can be expected. There is, in fact, some evidence of "blinding" of U.S. observation satellites during the Cold War years by the then-Soviet Union, using ground-based lasers. Designers can also expect to see requirements for hardening spacecraft against blast and shrapnel from potential "killer" satellites.

2.4.2 Weather Satellites

Weather satellites in GEO are the perfect complement to the LEO vehicles discussed earlier. High-altitude observations can show cloud, thermal, and moisture patterns over roughly one-third of the globe at a glance. This provides

the large-scale context for interpretation of the data from low-altitude satellites, aircraft, and surface observations.

Obviously, it is not necessary for a satellite to be in a geostationary or even a geosynchronous orbit to obtain a wide-area view. But, as discussed, it is still considered very convenient, and operationally desirable, for the spacecraft to stand still in the sky for purposes of continued observation, command, and control. Crowding of weather satellites does not present the problems associated with comsats, however, because entirely different frequency bands can be used for command and control purposes. The only real concern in this case is collision avoidance.

The Geostationary Operational Environmental Satellites (GOES) system is an excellent example of this type of satellite. Even though the purpose is different, many of the requirements of weather and communications satellites are similar, and the idea of combined functions, especially on larger platforms, may well become attractive in the future.

2.4.3 Space Observation

To date, there has been relatively little deep space observation from GEO. Generally speaking, there has been little reason to go to this energetically expensive orbit for observations from deep space. There are some exceptions; the International Ultraviolet Explorer (IUE) observatory satellite used an elliptic geosynchronous orbit with a 24,300-km perigee altitude and a 47,300-km apogee altitude. The previously mentioned Chandra telescope uses a similar orbit. Such orbits allow more viewing time of celestial objects with less interference from Earth's radiation belts than would have been the case for a circular orbit, while still allowing the spacecraft to be in continuous view of the Goddard Space Flight Center tracking stations.

At higher altitudes the Earth subtends a smaller arc, and more of the sky is visible. This can be important for sensitive optical instruments, which often cannot be pointed within many degrees of bright objects like the sun, moon, or Earth, because of the degradation of observations resulting from leakage of stray light into the optics. As more sensitive observatories for different spectral bands proliferate, there may be a desire to place them as far as possible from the radio, thermal, and visible light noise emanating from Earth.

A recent example is the Wilkinson Microwave Anisotropy Probe (WMAP), launched in 2001. This mission is the first to use a "halo" orbit about the Sun-Earth L2 Lagrange point (see Chapter 4) as a permanent observing station. WMAP orbits L2 in an oval pattern every six months, requiring stationkeeping maneuvers every few months to remain in position. This allows a complete WMAP full-sky observation every six months. As this goes to press, WMAP has succeeded in refining the earlier COBE data, allowing the distribution of background radiation in the universe to be mapped to within a few millionths of a Kelvin.

It will be important with the advent of very large antenna arrays (whether for communications or radio astronomy) to minimize gravity-gradient and atmospheric disturbances, and this will imply high orbits.

In this connection, an interesting possibility for the future is the so-called Orbiting Deep Space Relay Satellite (ODSRS), which has been studied on various occasions under different names. This concept would use a very large spacecraft as a replacement or supplement for the existing ground-based Deep Space Network (DSN). The DSN currently consists of large dish-antenna facilities in California, Australia, and Spain, with the placement chosen so as to enable continuous observation and tracking of interplanetary spacecraft irrespective of Earth's rotation.

The ODSRS concept has several advantages. Long-term, continuous tracking of a spacecraft would be possible and would not be limited by Earth's rotation. Usage of higher frequencies would be possible, thus enhancing data rates and narrowing beamwidths. This in turn would allow spacecraft transmitters to use lower power. The atmosphere poses a significant problem to the use of extremely high frequencies from Earth-based antennas. Attenuation in some bands is quite high, and rain can obliterate a signal (X-band signals are attenuated by some 40 dB in the presence of rain). Furthermore, a space-borne receiver can be easily cooled to much lower temperatures than is possible on Earth, improving its signal-to-noise ratio. The ODSRS would receive incoming signals from deep space and relay them to ground at frequencies compatible with atmospheric passage. Between tracking assignments, it could have some utility as a radio telescope.

Spacecraft performing surveys of the atmosphere, radiation belts, magnetic field, etc., around the Earth may be in synchronous, subsynchronous, or supersynchronous orbits that may or may not be circular. This might be done to synchronize the spacecraft with some phenomena related to Earth's rotation, or simply to bring it over the same ground station each day for data transmission or command and control.

As our sophistication in orbit design grows and experimental or other requirements pose new challenges, more complex and subtle orbits involving various types of synchrony as well as perturbations and other phenomena will be seen. We have only scratched the surface in this fascinating area.

2.5 Lunar and Deep Space Missions

Missions to the moon and beyond are often very similar to Earth orbital missions in terms of basic goals and methods. However, because of the higher energy requirements, longer flight times, and infrequent launch opportunities available using current propulsion systems, evolution of these missions from the basic to the more detailed and utilitarian type has been arrested compared to Earth orbital missions. In general, deep space missions fall into one of three categories: inner solar system targets, outer solar system targets, and solar orbital.

2.5.1 Inner Planetary Missions

The target bodies included in this category are those from Mercury to the inner reaches of the asteroid belt. The energy required to reach these extremes from Earth is roughly the same, a vis-viva energy of 30–40 km^2/s^2 (see Chapter 4). Even though the region encompasses a variation in solar radiative and gravitational intensity of about 60, it can be said to be dominated by the sun. Within this range, it is feasible to design solar-powered spacecraft and to use solar orientation as a factor in thermal control. Flight times to the various targets are measured in months, rather than years, for most trajectory designs of interest.

As would be expected, our first efforts to explore another planet were directed toward the nearby moon. Indeed, the first crude efforts by both the United States and the USSR to fly by or even orbit the moon came only months after the first Earth orbiters. Needless to say, there were at first more failures than successes. The first U.S. Pioneer spacecraft were plagued with various problems and were only partly successful. Probably the scientific highlight of this period was the return of the first crude images of the unknown lunar farside by the Soviet Luna 3 spacecraft. The lunar program then settled into what might be considered the classic sequence of events in the exploration of a planetary body. The early Pioneer flybys were followed by the Ranger family, designed to use close-approach photography of a single site followed by destruction on impact. Reconnaissance, via the Lunar Orbiter series, came next, followed by the Surveyor program of soft landers. Finally, manned exploration followed with the Apollo program.

Although omitting the hard landers, the Russian (Soviet at that time) program followed a similar path, and was clearly building toward manned missions until a combination of technical problems and the spectacular Apollo successes terminated the effort. A number of notable successes were achieved, however. Luna 9 made a "soft" (actually a controlled crash, with cameras encased in an airbag sphere for survival) landing on the moon in February 1966, some months prior to Surveyor 1. The propaganda impact of this achievement was somewhat lessened by the early decoding and release of the returned pictures from Jodrell Bank Observatory in England. The Lunokhod series subsequently demonstrated autonomous surface mobility, and some of the later Luna landers returned samples to Earth, though not before the Apollo landings.

Exploration of the other inner planets, so far as it has gone, has followed essentially the scenario previously outlined. Both the United States and Russia have sent flyby and orbital missions to Venus and Mars. The Russians landed a series of Venera spacecraft on Venus (where the survival problems dwarf anything so far found outside the sun or Jupiter), and the United States achieved two spectacularly successful Viking landings (also orbiters) on Mars. Following a 20-year hiatus after Viking, Mars is once again a focus of U.S. exploration with a series of landers, orbiters, and rovers. The holy grail of sample return is still the ultimate goal presently envisioned, with manned flight to Mars consigned to the indefinite future.

The asteroids have not so far been a major target of planetary science, although many mission concepts have been advanced and some preliminary efforts have been made. Both Voyager spacecraft, as well as Galileo and Cassini, have returned data from flybys of main belt asteroids while en route to the outer planets. The first exploration of a near-Earth asteroid was conducted with the Near Earth Asteroid Rendezvous (NEAR) mission to Eros. NEAR became the first spacecraft to orbit an asteroid, and, in a dramatic end-of-life experiment, also executed a series of maneuvers resulting in the first soft landing of a spacecraft on an asteroid. As this is written, Deep Space 1, an experimental solar electric propulsion vehicle, is conducting a series of slow flybys of asteroids.

The innermost planet, Mercury, has so far been the subject only of flybys and even these by only one spacecraft, Mariner 10. The use of a Venus gravity assist (see Chapter 4) to reach Mercury, plus the selection of a resonant solar orbit, allowed Mariner 10 to make three passes of the planet. This mission was one of the first astrodynamically complex missions to be flown, involving as it did a succession of gravity assist maneuvers, and it was also one of the most successful. Mariner 10 provided our first good look at this small, dense, heavily cratered member of the solar system.

Table 2.2 summarizes a few of the key lunar and inner planetary missions to date.

2.5.2 Outer Planetary Missions

As this is written, the outer planets, except for Pluto, have all been visited, though only Jupiter has been the target of an orbiting research satellite, on the Galileo mission. Cassini, launched in October 1997 for a July 2004 injection into a Saturn orbit, will be the second such outer-planet observatory. This mission is planned to deploy the Huygens probe into the atmosphere of Titan, the only planetary moon known to possess an atmosphere (other than possibly Charon, whose status as either a moon of Pluto, or as the smaller of a double-planetary system, is a matter of current debate).

Pioneers 10 and 11 led the way to the outer planets, with Pioneer 10 flying by Jupiter and Pioneer 11 visiting both Jupiter and Saturn. These missions were followed by Voyagers 1 and 2, both of which have flown by both Jupiter and Saturn, surveying both the planets and many of their moons. The rings of Jupiter and several new satellites of Saturn were discovered. All four vehicles acquired sufficient energy from the flybys to exceed solar escape velocity, becoming, in effect, mankind's first emissaries to the stars. The two Pioneers and Voyager 1 will not pass another solid body in the foreseeable future (barring the possibility of an unknown 10th planet or a "brown dwarf" star), but Voyager 2 carried out a Uranus encounter in 1986 and a Neptune flyby in 1989. Achievement of these goals is remarkable, because the spacecraft has far exceeded its four-year design lifetime. Even though the instrumentation designed for Jupiter and Saturn is not optimal at the greater distances of Uranus and Neptune, excellent results were achieved.

Table 2.2 Summary of key lunar and inner planet missions

Date	Mission	Comments
Late 1950s	Luna	Early Soviet missions. First pictures of far side of moon.
Late 1950s	Pioneer	Early U.S. missions to lunar vicinity.
Early 1960s	Luna	Continued Soviet missions. First unmanned lunar landing.
Early 1960s	Ranger	U.S. lunar impact missions. Detailed photos of surface.
1966–1968	Surveyor	U.S. lunar soft lander. Five successful landings.
1966–1968	Lunar Orbiter	U.S. photographic survey of moon.
1968–1972	Apollo	U.S. manned lunar orbiters and landings. First manned landing.
1968	Zond	Soviet unmanned tests of a manned lunar swingby mission.
Late 1960s	Luna	Soviet unmanned lunar sample return.
Early 1970s	Lunakhod	Soviet unmanned teleoperated lunar rover.
1962 and 1965	Mariner 2 and 5	U.S. Venus flyby missions. Mariner 2 first planetary flyby.
1964 and 1969	Mariner 4, 6, 7	U.S. Mars flyby missions.
1971	Mariner 9	U.S. Mars orbiter. First planetary orbiter.
1973	Mariner 10	U.S. Venus/Mercury flyby.
1975	Viking 1 and 2	U.S. Mars orbiter/lander missions.
1990	Magellan	U.S. Venus radar mapper.
1960s, 1970s	Mars	Series of Soviet Mars orbiter/lander missions.
1970s, 1980s	Venera	Long-running series of Soviet Venus featuring orbiters and landers.
1990	Ulysses	Solar polar region exploration enabled via Jupiter gravity assist.
1994	Clementine	Discovery of ice at lunar poles.
1996	NEAR	First asteroid rendezvous and soft landing.
1996	Mars Global Surveyor	High-resolution surface pictures.
1997	Mars Pathfinder	Successful Mars lander with airbag landing; first Mars rover.
1998	Lunar Prospector	Lunar surface chemistry map; confirmation of polar ice.
2001	Mars Odyssey	Mapping of Mars subsurface water.

It is interesting to note that the scientific value of the Pioneers and Voyagers did not end with their last encounter operation. Long-distance tracking data on these spacecraft have been used to obtain information on the possibility, and potential location, of a suspected 10th planet of the solar system. Such

expectations arose because of the inability to reconcile the orbits of the outer planets, particularly Neptune, with the theoretical predictions including all known perturbations. Both Neptune and Pluto (somewhat fortuitously, it now seems) were discovered as a result of such observations. Tracking data from the Pioneers and Voyagers can return more data, and more accurate data, in a few years than in several centuries of planetary observations. Moreover, because these spacecraft are departing the solar system at an angle to the ecliptic, they provide data otherwise totally unobtainable. The Pioneers and the Voyagers were still being tracked (sporadically in the case of Pioneers) in the early 2000s, nearly three decades after launch. Among other things, they are still attempting to discover boundaries of the heliopause, the interface at which the solar wind gives way to the interstellar medium.

By the logical sequence outlined previously, Jupiter would be the next target for an orbiter and an atmospheric probe, as was in fact the case. The Galileo program achieved these goals, as well as conducting many successive flybys of the Jovian moons from its Jupiter orbit. Although delayed by many factors, including the 1986 *Challenger* accident, Galileo was launched in 1989 on a circuitous path involving a Venus flyby and two Earth flybys on route to Jupiter. This complexity is a result of the cancellation of the effort to develop a high-energy Centaur upper stage for the shuttle, and consequent substitution of a lower-energy inertial upper stage (IUS).

The Galileo spacecraft has been severely crippled by the failure of its rib-mesh antenna to deploy fully. As a result, the data rate to Earth, planned to be tens of kilobits per second, was significantly degraded, greatly curtailing the number of images returned. Nevertheless, the mission must be rated a huge success because of the quality of data that has been received.

The Galileo mission was also an astrodynamical tour de force, with a flyby of one satellite used to target the next in a succession of visits to the Jovian satellites, all achieved with minimal use of propellant. In complexity it has far eclipsed the trail-blazing Mariner 10.

As mentioned, Cassini and its Huygens probe follow in the footsteps of the Galileo Jupiter orbiter and probe. Cassini used an even more complex trajectory than Galileo, referred to as a Venus–Venus–Earth–Jupiter gravity assist (VVEJGA) trajectory. Huygens will separate from the Cassini orbiter to enter the atmosphere of Titan, while Cassini is planned to make at least 30 planetary orbits, each optimized for a different set of observations.

The Cassini mission design is particularly interesting in its use of gravity-assist maneuvers to achieve an otherwise unattainable goal. As noted earlier, Cassini's flight time to Saturn is about 6.7 years, which compares very favorably with the Hohmann transfer time of approximately 6 years (see Chapter 4). The Hohmann transfer to Saturn requires a ΔV from Earth parking orbit in excess of 7 km/s, and although this is the minimum possible for a two-impulse maneuver, it is substantially in excess of that capable of being supplied by any existing upper stage. However, the initial ΔV required to effect a Venus flyby for Cassini was

only about half this value, after which subsequent encounters were used to boost the orbital energy to that required for the outer-planet trip. The multiple-gravity-assist Cassini mission design thus provided a reasonable flight time while remaining within the constraints of the available launch vehicle technology.

Spacecraft visiting the outer planets cannot depend on solar energy for electrical power and heating. Use of solar concentrators can extend the range of useful solar power possibly as far as Jupiter, but at the cost of considerable complexity. The spacecraft that have flown to these regions, as well as those that are planned, depend on power obtained by radioactive decay processes. These power units, generally called radioisotope thermoelectric generators (RTG), use banks of thermoelectric elements to convert the heat generated by radioisotope decay into electric power. The sun is no longer a significant factor at this point, and all heat required, for example to keep propellants warm, must be supplied by electricity or by using the waste heat of the RTGs. On the positive side, surfaces designed to radiate heat at modest temperatures, such as electronics boxes, can do so in full sunlight, a convenience for the configuration designer that is not available inside the orbit of Mars.

2.5.3 Small Bodies

Comets and asteroids, the small bodies of the solar system, were largely ignored during the early phases of space exploration, although various mission possibilities were discussed and, as noted, some have come into fruition. Although most of the scientific interest (and public attention) focuses on comets, the asteroids present a subject of great interest also. Not only are they of scientific interest, but, as we have discussed, some may offer great promise as sources of important raw materials for space fabrication and colonization projects.

The main belt asteroids are sufficiently distant from the sun that they are relatively difficult to reach in terms of energy and flight time. Except for the inner regions of the belt, solar power is not really practical. For example, an asteroid at a typical 2.8 AU distance from the sun suffers a decrease in solar energy by a factor of 8.84 compared with that available at the orbit of Earth. RTGs or, in the future, possibly full-scale nuclear reactors will be required.

However, many asteroids have orbits that stray significantly from the main belt, some passing inside the orbit of Earth. These asteroids are generally in elliptic orbits, many of which are significantly inclined to the ecliptic plane. Orbits having high eccentricity and/or large inclinations are quite difficult, in terms of energy, to reach from Earth. However, a few of these bodies are in near-ecliptic orbits with low eccentricity, and are the easiest extraterrestrial bodies to reach after the moon. In fact, if one includes the energy expenditure required for landing, some of these asteroids are easier to reach than the lunar surface. Clearly, these bodies offer the potential of future exploration and exploitation. Relatively few of these Earth-approaching asteroids are known as yet, but analysis indicates

that there should be large numbers of them. Discovery of new asteroids in this class is a relatively frequent event.

Comets generally occupy highly eccentric orbits, often with very high inclination. Some orbits are so eccentric that it is debatable whether they are in fact closed orbits at all. In any case, the orbital periods, if the term is meaningful, are very large for such comets. Some comets are in much shorter but still highly eccentric orbits; the comet Halley, with a period of 76 years, lies at the upper end of this short-period class. The shortest known cometary period is that of Encke, at 3.6 years.

As stated, most comets are in high-inclination orbits, of which Halley's Comet is an extreme example, with an inclination of 160 deg. This means that it circles the sun in a retrograde direction at an angle of 20 deg to the ecliptic. With few exceptions, comet rendezvous (as distinct from intercept) is not possible using chemical propulsion. High-energy solar or nuclear powered electric propulsion or solar sailing can, with reasonable technological advances, allow rendezvous with most comets.

As this goes to press, the first cometary exploration mission will be the NASA Deep Impact probe, scheduled for an early 2004 launch and later intercept with Comet Tempel 1.

2.5.4 Orbit Design Considerations

Although we will consider this topic in more detail in Chapter 4, the field of orbit and trajectory design for planetary missions is so rich in variety that an overview is appropriate at this point. Transfer trajectories to other planets are determined at the most basic level by the phasing of the launch and target planets. Simply put, both must be in the proper place at the proper time. This is not nearly as constraining as it may sound, particularly with modern computational mission design techniques. A wide variety of transfer orbits can usually be found to match launch dates that are proper from other points of view, such as the availability of hardware and funding.

The conventional transfer trajectory is a solar orbit designed around an inferior conjunction (for inner planets) or opposition (for outer planets). Such orbits, although they do not possess the flexibility described earlier, are often the best compromise of minimum energy and minimum flight time. These orbits typically travel an arc of somewhat less than 180 deg (type 1 transfer) or somewhat more than 180 deg (type 2 transfer) about the sun. A special case here is the classical two-impulse, minimum-energy Hohmann transfer. This trajectory is completely specified by specifying a 180 deg arc between the launch and target planets that it is tangent to both the departure and arrival orbits. However, the Hohmann orbit assumes coplanar circular orbits for the two planets, a condition that is in practice never met exactly. Because the final trajectory is rather sensitive to these assumptions, true Hohmann transfers are not used. Furthermore, flight times using such a transfer would be unreasonably long for any planetary target outside the orbit of Mars. Ingenuity in orbit design or added booster power, or both, must be used to obtain acceptable mission durations for flights to the outer planets.

The expenditure of additional launch energy is the obvious approach to reducing flight times. This involves placing the apsis of the transfer orbit well beyond the target orbit, thus causing the vehicle to complete its transfer to the desired planet much more quickly. In the limit, this section can be made to appear as nearly a straight line, but at great energy cost at both departure and arrival. A planetary transfer such as this is beyond present technological capabilities.

The other extreme is to accept longer flight times to obtain minimum energy expenditure. In its simplest form, this involves an orbit of 540 deg of arc. The vehicle flies to the target orbit (the target is elsewhere), back to the launch orbit (the launch planet is elsewhere), then finally back to the target. Such an evolution sometimes saves energy relative to shorter trajectories through more favorable nodal positioning or other factors. This gain must be traded off against other factors such as increased operations cost, budgeting of onboard consumables, failure risk, and utility of the science data.

A more complicated but more commonly used option involves the application of a velocity change sometime during the solar orbit phase. This can be done propulsively or by a suitable target flyby (increasingly the method of choice) of a third body, or by some combination of these. The propulsive ΔV approach is simplest. A substantial impulse applied in deep space may, for example, allow an efficient change in orbital plane, thus reducing total energy requirements. A more exacting technique is to fly past another body in route and use the swingby to gain or lose energy (relative to the sun, not the planet providing the gravity assist). Mariner 10 used this technique at Venus to reach Mercury, and Pioneer 11 and the two Voyagers used it at Jupiter to reach Saturn. Voyager 2, of course, used a second gravity assist at Saturn to continue to Uranus. The Venus and Earth swingbys mentioned in conjunction with the Galileo mission supply both plane change and added energy. The Jupiter satellite flybys perform a similar function in Jupiter orbit.

The gravity-assist technique, now well established, was first used with Mariner 10. In fact, the only means of reaching Mercury with current launch vehicles and a mass sufficient to allow injection into Mercury orbit with chemical propulsion is via a multirevolution transfer orbit with one or more Venus flybys to reduce the energy of the orbit at Mercury arrival to manageable levels. Of course, in planetary exploration, the additional time spent in doing swingbys is hardly a penalty; we have not yet reached the point where so much is known about any planet that an additional swingby is considered a waste of time.

As noted, this is now a mature technique. It was exploited to the fullest during the Galileo mission to Jupiter, where repeated pumping of the spacecraft orbit through gravity assists from its moons was used to raise and lower the orbit and change its inclination. The orbit in fact was never the same twice. These "tours" allowed the maximum data collection about the planet and its satellites, while permitting a thorough survey of the magnetic field and the space environment.

The final class of methods whereby difficult targets can be reached without excessive propulsive capability involves the use of the launch planet itself for

gravity assist maneuvers. The spacecraft is initially launched into a solar orbit synchronized to intercept the launch planet again, usually after one full revolution of the planet, unless a midcourse ΔV is applied. The subsequent flyby can be used to change the energy or inclination of the transfer orbit, or both. It is also possible to apply a propulsive ΔV during the flyby. Such mission profiles have been frequently studied as options for outer planetary missions, and, as discussed, were applied to both Galileo and Cassini.

The orbits into which spacecraft are placed about a target planet are driven by substantially the same criteria as for spacecraft in Earth orbit. For instance, the Viking orbiters were placed in highly elliptic 24.6-h orbits (a "sol," or one Martian day) so that they would arrive over their respective lander vehicles at the same time each day to relay data. Mars geoscience mappers may utilize polar sun-synchronous orbits like those used by similar vehicles at Earth. A possibility for planetary orbiters is that, rather than being synchronized with anything at the target planet, they can be in an orbit with a period synchronized with Earth. For example, the spacecraft might be at periapsis each time a particular tracking station was in view.

Low-thrust planetary trajectories are required for electric and solar sail propulsion and are quite different from the ballistic trajectory designs described thus far, because the thrust is applied constantly over very long arcs in the trajectory. Such trajectories also may make use of planetary flybys to conserve energy or reduce mission duration. The most notable difference is at the departure and target planets. At the former, unless boosted by chemical rockets to escape velocity, the vehicle must spend months spiraling out of the planetary gravity field. In some cases this phase may be as long as the interplanetary flight time. At the target, the reverse occurs.

This situation results from the very low thrust-to-mass ratio of such systems. In one instance where solar-electric propulsion was proposed for a Mars sample return mission, it was found that the solar-electric vehicle did not have time to spiral down to an altitude compatible with the use of a chemically-propelled sample carrier from the surface. To return to Earth, it had to begin spiraling back out before reaching a reasonable rendezvous altitude. Higher thrust-to-mass ratios such as those offered by nuclear-electric propulsion or advanced solar sails would overcome this problem. Solar-electric propulsion and less capable solar sails are most satisfactory for missions not encountering a deep gravity well. Comet and asteroid missions and close-approach or out-of-ecliptic solar missions are examples.

2.6 Advanced Mission Concepts

Thus far we have dealt with mission design criteria and characteristics primarily for space missions that have flown, or are planned for flight in the near future. In a sense, design tasks at all levels for these missions are known

quantities. Though space flight still has not progressed to the level of routine airline-like operations, nonetheless, much experience has been accumulated since Sputnik 1, to the point where spacecraft design for many types of tasks can be very prosaic. In many areas, there is a well-established way to do things, and designs evolve only within narrow limits.

This is not true of missions that are very advanced by today's standards. Such missions include the development of large structures for solar power satellites or antenna farms, construction of permanent space stations, lunar and asteroid mining, propellant manufacture on other planets, and many other activities that cannot be accurately envisioned at present. For these advanced concepts, the designer's imagination is still free to roam, limited only by established principles of sound engineering practice. In this section, we examine some of the possibilities for future space missions that have been advocated in recent years, with attention given to the mission and spacecraft design requirements they will pose.

2.6.1 Large Space Structures

Many of the advanced mission concepts that have surfaced have in common the element of requiring the deployment in Earth orbit of what are, by present standards, extremely large structures. Examples of such systems include solar power satellites, first conceived by Dr. Peter Glaser, and the large, centralized antenna platforms alluded to previously in connection with communications satellites. These structures will have one outstanding difference from Earth-based structures of similar size, and that is their extremely low mass. If erected in a 0-g environment, these platforms need not cope with the stresses of Earth's gravitational field, and need only be designed to offer sufficient rigidity for the task at hand. This fact alone will offer many opportunities for both success and failure in exploiting the capabilities of large space platforms.

Orbit selection for large space structures will in principle be guided by much the same criteria as for smaller systems, that is, the orbit design will be defined by the mission to be performed. However, the potentially extreme size of the vehicles involved will offer some new criteria for optimization. Systems of large area and low mass will be highly susceptible to aerodynamic drag, and will generally need to be in very high orbits to avoid requirements for excessive drag compensation propulsion. For such platforms, solar pressure can become the dominant orbital perturbation. Similarly, systems with very large mass will tend toward low orbits to minimize the expense of construction with materials ferried up from Earth. When the time comes that many large platforms are deployed in high Earth orbit, it is likely that the use of lunar and asteroid materials for construction will become economically attractive. In terms of energy requirements, the moon is closer to geosynchronous orbit than is the surface of the Earth. The consequences of this fact have been explored in a number of studies.

Other characteristics of expected large space systems have also received considerable analytical attention. As mentioned, structures such as very large antennas or solar power satellites will have quite low mass for their size by Earth standards. Yet these structures, particularly antennas, require quite precise shape control to achieve their basic goals. On Earth, this requirement is basically met through the use of sufficient mass to provide the needed rigidity, a requirement that is not usually inconsistent with that for sufficient strength to allow the structure to support itself in Earth's gravitational field. As mentioned, in a $0g$ environment this will not be the case. Very large structures of low mass will have very low characteristic frequencies of vibration, and quite possibly very little damping at these frequencies. Thus, it has been expected that some form of active shape control will often be required, and much effort has been expended in defining the nature of such control schemes.

Translation control requires similar care. For example, it will hardly be sufficient to attach a single engine to the middle of a solar power satellite some tens of square kilometers in size and ignite it. Not much of the structure will remain with the engine. It may be expected that electric or other low-thrust propulsion systems will come into their own with the development of large space platforms.

2.6.2 Space Stations

Concepts for manned space stations have existed since the earliest days of astronautics. Von Braun's 1952 study, published in *Collier's*, remains a classic in this field. The first-generation space stations, the Russian Salyut and American Skylab vehicles, as well as the more sophisticated Russian Mir and even the ISS, fall far short of von Braun's ambitious concepts. This from some points of view is quite surprising; early work in astronautics seems often to have assumed that construction of large, permanent stations would be among the first priorities to be addressed once the necessary space transportation capability was developed. This has not turned out to be the case. Political factors, including the "moon race," have influenced the course of events, but technical reality has also been recognized. Repeated studies have failed to show any single overriding requirement for the deployment of a space station. The consensus that has instead emerged is that, if a permanent station or stations existed, many uses would be found for it that currently require separate satellites, or are simply not done. However, no single utilitarian function for a space station appears, by itself, sufficient to justify the difficulty and expense of building it.

As this is written, and after many years of gestation, the ISS is being assembled in LEO and is inhabited on an essentially permanent basis. It is advertised as being, and many hope it will be, the first true space station. Even now, it is by far the largest and most technically ambitious artifact yet assembled in space. If it can overcome its rocky start and the funding restrictions that seriously diminish its capability, it may yet live up to these hopes. It seems inevitable that, if space utilization is to continue and expand, there will be a variety of large and small manned and

man-tended orbital stations carrying out numerous functions, some now performed by autonomous vehicles while others not currently available will become so.

Selection of space station orbits will be driven by the same factors as for smaller spacecraft, a tradeoff between operational requirements, energy required to achieve orbit, and difficulty of maintaining the desired orbit. For small space stations such as the Salyut series, maneuvering is not especially difficult, and periodic orbit maintenance can be accomplished with thrusters. The large, flexible assemblies proposed for future stations may be more difficult to maneuver and for this reason may tend to favor higher orbits. As mentioned, some type of electric propulsion will probably be required for orbit maintenance in this case, both because of its reduced propellant requirements and its low thrust.

Space stations designed for observation, whether civil or otherwise, will have characteristics similar to their smaller unmanned brethren. They will generally be found in high-inclination low orbits, perhaps sun-synchronous, for close observation, or in high orbits where a more global view is required. On the other hand, stations of the space operations center type, which are used as way stations en route to geosynchronous orbit or planetary missions as well as for scientific purposes, will probably be in fairly low orbits at inclinations compatible with launch site requirements.

Space stations of the von Braun rotary wheel type may never be realized because of the realization that artificial gravity is not necessary for human flight times up to several months' duration. This has been demonstrated by both Russian and American missions, wherein proper crew training and exercise have allowed the maintenance of reasonably satisfactory physical conditioning, albeit with the need for substantial reconditioning time upon return to Earth. By eliminating the need for artificial gravity, the need for a symmetric, rotating design is also eliminated. This greatly simplifies configuration and structural design, observational techniques, and operations, especially flight operations with resupply vehicles.

However, it is clear that long-term exposure to microgravity is quite debilitating, and very long residence times in space will undoubtedly require the provision of artificial gravity. For an interesting visual demonstration of the problems of docking with a rotating structure, the reader is urged to view Stanley Kubrick's classic film *2001: A Space Odyssey*.

The problem of supplying electric power for space station operations is substantial. Skylab, Salyut, Mir, and ISS have used solar panel arrays with batteries for energy storage during eclipse periods. This will probably remain the best choice for stations with power requirements measured in a few tens of kilowatts. As power requirements become large, which history indicates is inevitable, the choice becomes less clear. The large areas of high-power solar arrays pose a major drag and gravity-gradient stabilization problem in LEO, and their intrinsic flimsiness poses severe attitude control problems even in high orbit. The use of dynamic conversion of solar heat to electricity is promising in reducing the collection area but has other problems.

The only presently viable alternative to solar power for a permanent station is a nuclear system, and here we are generally talking about nuclear reactors rather than the RTGs discussed earlier. RTGs do not have a sufficiently high power-to-weight ratio to be acceptable when high power levels are required. Chemical energy systems such as fuel cells are not practical for permanent orbital stations when the reactants must be brought from Earth. This conclusion could change in the short term if a practical means of recovering unused launch vehicle propellant could be devised, and in the long term if use of extraterrestrial materials becomes common. In the meantime, nuclear power offers the only compact, long-lived source of power in the kilowatt to megawatt range.

Nuclear power also raises substantial problems. The high-temperature reactor and thermal radiators, the high level of ionizing radiation, and the difficulty of systems integration caused by these factors present substantial engineering problems. No less serious is public concern with possible environmental effects due to the uncontrolled reentry of a reactor. This first happened with the Russian Cosmos 954 vehicle, which fortunately crashed in a remote region of Canada. The cleanup operations involved were not trivial.

Of similar importance is the environmental control system of the station. The more independent of resupply from the ground it can be, the more economical the permanent operation of the station will become. The ultimate goal of a fully recycled, closed environmental system will be long in coming, but even a reasonably high percentage of water and oxygen recycling will be of significant help. The possibility of an ecological approach to oxygen recycling may allow production of fresh fruits, vegetables, and decorative plants. The latter may be of only small significance to the resupply problem, but may be quite important for crew morale. Similar concern with environmental issues has gone into the design of U.S. Navy nuclear submarines, which spend long periods submerged.

As the construction and operation of the ISS continues, it will be of interest to examine these and other methods by which crew morale is maintained. That the issue is not trivial is shown by the records of more than one U.S. space flight, where both flight crew boredom and overwork have on occasion led to some acrimonious exchanges with ground control. With the greater visibility now available into the Russian manned space program, similar cases have emerged, again reaffirming the importance of crew morale to mission success.

2.6.3 Space Colonies

Long-term-habitability space stations can be expected to provide the initial basis for the design of space colonies or colonies on other planets or asteroids. The borderline between space stations, or research or work stations on other planets, and true colonies is necessarily somewhat blurred, but the use of the term "colonies" is generally taken to imply self-sufficient habitats with residents of all types who expect to live out their lives in the colony. Trade with Earth is presumed, as a colony with no economic basis for its existence probably will not

have one. On the other hand, it seems reasonable that "research stations" or "lunar mining bases" could grow into colonies, given the right circumstances.

The late Gerard K. O'Neill and his co-workers have been the most ardent recent proponents of the utility and viability of space colonies. In the O'Neill concept, the colonies will have as their economic justification the construction of solar power satellites for Earth, using raw materials derived from lunar or asteroid bases. It would seem that other uses for such habitats could be found as well; as mentioned previously, in the very long run it may be that eventually much of Earth's heavy manufacturing is relocated to sites in space to take advantage of the availability of energy and raw materials. In any case, O'Neill envisioned truly extensive space habitats, tens of kilometers in dimension, featuring literally all of the comforts of home, including grass, trees, and houses in picturesque rural settings.

Whether or not these developments ever come to pass (and the authors do not wish to say that they cannot; well-reasoned economic arguments for developing such colonies have been advanced), such concepts would seem to be the near-ultimate in spacecraft design. In every way, construction of such habitats would pose problems that, without doubt, are presently unforeseen. The engineering of space colonies and colonies on other planets will demand the use of every specialty known on Earth today, from agriculture to zoology, and these specialists will have to learn to transfer their knowledge to extraterrestrial conditions. The history of the efforts of Western Europeans simply to colonize other regions of Earth in the sixteenth and seventeenth centuries suggests both that it will be done and that it will not be done easily.

2.6.4 Use of Lunar and Asteroid Materials

Even our limited exploration of the moon has indicated considerable potential for supplying useful material. We have not in our preliminary forays observed rich beds of ore such as can be found on Earth. Some geologists have speculated that such concentrations may not exist on the moon, and it certainly seems reasonable to suppose that they do not exist near the surface, which is a regolith composed of material pulverized and dispersed in countless meteoric impacts. However, the common material of the lunar crust offers a variety of useful materials, most prominently aluminum, oxygen, and titanium, which is surprisingly in relatively large supply in the lunar samples so far seen. A more useful metal for space manufacturing would be hard to find. The metals exist as oxides or in more complex compounds. A variety of processes have been suggested for the production of useful metals and oxygen; which material is the product and which is the by-product depends on the prejudices of the reader.

Because of the cost of refining the material on the moon and transporting it to Earth, it is improbable that such materials would be economically competitive with materials produced here on Earth. An exception would be special alloys made in $0g$ or other substances uniquely depending on the space environment for

their creation. However, extraterrestrial materials may well compete with materials ferried up from Earth for construction in orbit or on the moon itself. This is the primary justification for lunar and asteroid mining, and it seems so strong that it must eventually come to pass, when the necessary base of capital equipment exists in space.

It may well be that products (as opposed to raw materials) manufactured in space will compete successfully with comparable products manufactured on Earth. Early candidates will be goods whose price is high for the mass they possess and whose manufacture is energy intensive, hampered by gravity and/or atmospheric contaminants, and highly suitable for automated production. Semiconductors and integrated circuits, pharmaceuticals, and certain alloys have been identified in this category. Other activities may follow; one can imagine good and sufficient reasons for locating genetic engineering research and development efforts in an isolated space-based laboratory.

With the accumulation in orbit of sufficient capital equipment to allow large-scale use of lunar or other extraterrestrial materials, and the development of effective solar energy collection methods, the growth of heavy manufacturing must follow. As noted, the surface of the moon is much closer to either GEO or LEO in terms of energy expenditure than is the surface of Earth. Any really large projects will probably be more economical with lunar material, even considering the necessary investment in lunar mining bases. Further, some resources are more readily used than others; even relatively modest traffic from LEO to GEO, the moon, or deep space will probably benefit from oxygen generated on the moon and sent down to Earth orbit.

The probability, long theorized and now supported by observational data from the Clementine and Lunar Prospector missions, that water ice is trapped in permanently dark, very cold regions near the lunar poles is of great interest. Water is not only vital for life-support functions (though with closed systems, humans generate water as a by-product of other activities, thus reducing the life-support problem to that of food alone), but it is also useful in a variety of chemical processes, and especially in the production of hydrogen. Thus far it appears that no economically viable supply of hydrogen exists on the moon except in these ice reservoirs. Hydrogen is useful as a propellant and in a variety of chemical reactions. If it cannot be obtained on the moon, it will have to be imported from Earth, at least in the short term. Although its low mass makes importation of hydrogen at least somewhat tolerable, the desirability of finding it on the moon is obvious.

The use of asteroid materials has equally fascinating potential. Taken as a class, asteroids offer an even more interesting spectrum of materials than has so far been identified on the moon. The metallic bodies consist mostly of nickel-iron, which should be a reasonably good structural material as found and would be refinable into a variety of others. The carbonaceous chondrite types seem to contain water, carbon, and organic materials as well as silicates. These would have the obvious advantage of being water and hydrogen sources; indeed, some models of the Martian climate have postulated that such asteroids are the source

of what Martian water exists. The most common, and probably least useful, asteroids are composed mostly of silicate materials; essentially, they are indistinguishable from common inorganic Earth dirt.

Although, as mentioned, most asteroids lie in the main belt between Mars and Jupiter, a modest number lie in orbits near to or crossing that of Earth. Some of these are energetically quite easy to reach, but with the problem that the low round-trip energy requirement is achieved at the cost of travel times on the order of three years or more. Launch windows are restricted to a few weeks every two or three years. Thus, although it is true that some asteroids are easier to reach than the surface of the moon, this must be balanced against the lunar round-trip time of a few days, together with the ability to make the trip nearly any time. Thus, although asteroid materials of either the Earth-approaching or main-belt variety will probably become of substantial importance eventually, it seems likely that lunar materials will do so first, if only because of convenience.

2.6.5 Propellant Manufacturing

Propellant manufacturing is a special case involving the use of resources naturally occurring on the various bodies of the solar system. It was mentioned in passing under the more general subject of lunar and asteroid resources, but it is by no means restricted to these bodies. In the inner solar system, Mars seems to offer the most promise for application of in situ propellant manufacturing technology.

As noted previously, for the manufacture of a full set of propellants (both fuel and oxidizer), water is both necessary and sufficient. However, carbon, which is also in short supply on the moon, is also important. The atmosphere of Mars provides carbon dioxide in abundance, and water is known to exist in the polar ice caps and most probably in the form of permafrost over much of the planet. Propellant manufacturing has been studied both for unmanned sample return missions and for manned missions. The advantages are comparable to those that accrue by refueling airliners at each end of a flight, rather than designing them to carry fuel for a coast-to-coast round-trip.

Because of the difficulty of mining permafrost or low-temperature ice, it has been suggested that the first propellant manufacturing effort might use the atmosphere exclusively. Carbon dioxide can be taken in by compression and then, in a cell using thermal decomposition and an oxygen permeable membrane, split into carbon monoxide and oxygen. The oxygen can then be liquified and burned with a fuel brought from Earth. Methane is the preferred choice, because it has high performance, a high oxidizer-to-fuel ratio (to minimize the mass brought from Earth), and is a good refrigerant. The latter quality contributes to the process of liquifying the oxygen and keeping both propellants liquid until enough oxidizer is accumulated and the launch window opens.

It should be noted that the combination of carbon monoxide and oxygen is a potential propellant combination. The theoretical performance is modest at best, indicating a delivered specific impulse of 260 s at Mars conditions. Tests in 1991

have confirmed the theoretical predictions. This performance might be adequate for short-range vehicles supporting a manned base on Mars, however, and would certainly be convenient. It is even suitable for orbital vehicles although propellant mass is large. A final advantage is that, because the exhaust product is carbon dioxide, there would be no net effect on the Martian atmosphere.

Making use of Martian water broadens the potential options considerably. Besides the obvious hydrogen/oxygen combination, use of both water and carbon dioxide allows the synthesis of other chemicals such as methane. Methane is an excellent fuel and is more easily storable than hydrogen. Methanol can also be created, either as a fuel or for use in other chemical processes. Another possible option is to bring hydrogen from Earth. The required mass is relatively small, although the bulkiness resulting from it is low density and the difficulty of long-term storage may cause problems. From this brief glimpse, it can be seen that water and carbon or carbon dioxide form the basis for propellant manufacturing as well as other chemical processes.

Because carbonaceous chondrites presumably contain both water and carbon compounds, it is probable that these bodies have potential for various types of chemical synthesis as well. The satellites of the outer planets contain considerable water; indeed, some are mostly water. Whether useful carbon-containing compounds are available is less certain, but at least the hydrogen/oxygen propellant combination will be available.

In all propellant manufacturing processes, the key is power. Regardless of the availability of raw materials, substantial energy is required to decompose the water or carbon dioxide. Compression and liquefaction of the products also require energy. The possible sources of energy are solar arrays, nuclear systems using radioisotopic decay, and critical assemblies (reactors). The use of solar energy is only practical in the inner solar system, and then probably only for small production rates.

2.6.6 Nuclear Waste Disposal

Disposal of long-lived highly radioactive waste in space has been discussed for many years. The attraction is obvious; it is the one disposal mode that, properly implemented, has no chance of contaminating the biosphere of Earth because of leakage or natural disaster.

The least demanding technique would be to place the waste into an orbit of Earth that is at sufficient altitude that no conceivable combination of atmospheric drag or orbital perturbations would cause the orbit to decay. Even though this is workable, it is not considered satisfactory by some, because the material is still within the Earth's sphere of influence and thus might somehow come down. A more practical objection is that, as use of near-Earth space increases, it might not be desirable to have one region rendered unsafe.

Another suggestion is to place all of the material on the moon, say, in a particular crater. This generally avoids the orbit stability problem but has the

disadvantage of rendering one area of the moon quite unhealthy. Energy cost would be high as well, because the material would need to be soft landed to avoid scattering on impact.

From an emotional viewpoint at least, interplanetary space seems the most desirable arena for disposal, preferably in an orbit far from that of Earth. One approach would steal a page from the Mariner 10 mission. For a total energy expenditure less than that for a landing on the moon, the material could be sent on a trajectory to fly by Venus. This could move the perihelion of the orbit to a point between Venus and Mercury. A relatively minor velocity change at the perihelion of the orbit would then lower aphelion inside the orbit of Venus. The package would then be in a stable, predictable orbit that would never again come close to Earth.

The major problem with the space disposal of nuclear waste is the emotional fear of a launch failure spreading the material widely over the surface of the Earth. Although a number of concepts could be applied to minimize the risk, it seems doubtful that this concept will become acceptable to the public in the near future.

Bibliography

Baker, D., *The History of Manned Space Flight*, Crown Publishers, New York, 1981.

Burrough, B., *Dragonfly*, HarperCollins, New York, 1998.

Burrows, W. E., *Deep Black*, Random House, New York, 1986.

Burrows, W. E., *This New Ocean*, Random House, New York, 1998.

Clark, P., *The Soviet Manned Space Program*, Orion Books, New York, 1988.

Gatland, K., *The Illustrated Encyclopedia of Space Technology*, 2nd ed., Orion Books, New York, 1989.

Launius, R. D., *Apollo: A Retrospective Analysis*, Monographs in Aerospace History, No. 3, NASA, 1994.

Logsdon, J. M. (ed.), *Exploring the Unknown*, Vols. I–III, NASA SP-4407, 1996.

Mather, J. C., and Boslough, J., *The Very First Light*, Basic Books, New York, 1996.

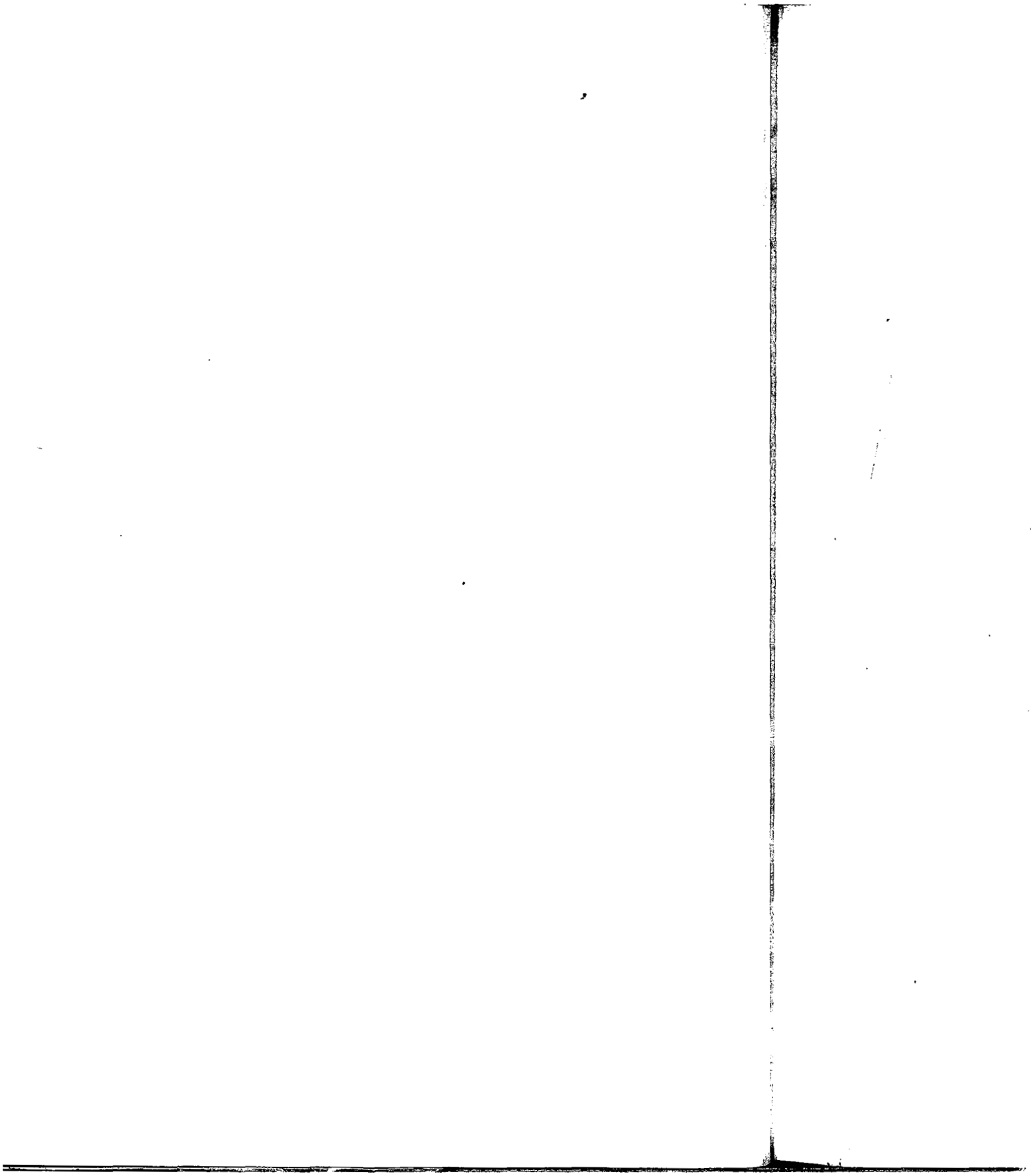
Murray, B., *Journey into Space*, Norton Books, New York, 1989.

Nicogossian, A. E., and Parker, J. F., *Space Physiology and Medicine*, NASA SP-447, 1982.

O'Neill, G. K., *The High Frontier: Human Colonies in Space*, Morrow, New York, 1976.

Von Braun, W., "Man Will Conquer Space Soon," *Colliers*, 1952.

Weissman, P. R., McFadden, L.-A., and Johnson, T. V. (eds.), *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999.



3
Spacecraft Environment

3.1 Introduction

In the broadest sense, the spacecraft environment includes everything to which the spacecraft is exposed from its beginning as raw material to the end of its operating life. This includes the fabrication, assembly, and test environment on Earth, transportation from point to point on Earth, launch, the space environment, and possibly an atmospheric entry and continued operation in a destination environment at another planet.

Both natural and man-made environments are imposed upon the spacecraft. Contrary to the popular view, the rigors of launch and the space environment itself are often not the greatest hazards to the spacecraft. The spacecraft is designed to be launched and to fly in space. If the design is properly done, these environments are not a problem; a spacecraft sometimes seems at greatest risk on Earth in the hands of its creators. Spacecraft are often designed with only the briefest consideration of the need for ground handling, transportation, and test. As a result, these operations and the compromises and accommodations necessary to carry them out may in fact represent a more substantial risk than anything that happens in a normal flight.

However, the preceding comments imply that the spacecraft is designed for proper functioning in flight. To do this it is necessary to know the range of conditions encountered. This includes not only the flight environment but also the qualification test conditions that must be met to demonstrate that the design is correct. To provide confidence that the design will be robust in the face of unexpectedly severe conditions, these tests are typically more stringent than the expected actual environment. In some cases, especially where the rigorous safety standards applied to manned flight are concerned, even the origin of the materials used and the details of the processes by which they are fashioned into spacecraft components may be important to the process of qualifying the spacecraft for flight. Many spacecraft have been lost due to lack of full understanding of the environment.[1]

In this chapter we will discuss the Earth, launch, and space environments, but in somewhat different terms. The launch and flight environments are usually quite well defined for specific launch vehicles and missions. These conditions, and the qualification test levels that are derived from them, will be treated as the actual environment for which the vehicle must be designed. The Earth

environment is assumed to be controllable, within limits, to meet the requirements of a spacecraft, subsystem, or component. Also, the variety of Earth environments, modes of handling and transport, etc., is so great as to preclude a detailed quantitative discussion of them in this volume. Accordingly, the discussion will be of a more general nature when addressing Earth environments.

3.2 Earth Environment

Throughout its tenure on Earth, the spacecraft and its components are subjected to a variety of potentially degrading environments. The atmosphere itself is a primary source of problems. Containing both water and oxygen, the Earth's atmosphere is quite corrosive to a variety of materials, including many of those used in spacecraft, such as lightweight structural alloys. Corrosion of structural materials can cause stress concentration or embrittlement, possibly leading to failure during launch. Corrosion of pins in electrical connectors can lead to excessive circuit resistance and thus unsatisfactory performance. Because of these effects it is desirable to control the relative humidity and in extreme cases to exclude oxygen and moisture entirely by use of a dry nitrogen or helium purge. This is normally required only for individual subsystems such as scientific instruments; in general, the spacecraft can tolerate exposure to the atmosphere if humidity is not excessive. However, too low a relative humidity is also poor practice both from consideration of worker comfort and from a desire to minimize buildup of static electric charge (discussed later in more detail). A relative humidity in the 40–50% range is normally a good compromise.

Another environmental problem arising from the atmosphere is airborne particulate contamination, or dust. Even in a normally clean environment, dust will accumulate on horizontal surfaces fairly rapidly. For some spacecraft a burden of dust particles is not significant; however, in many cases it can have undesirable effects. Dust can cause wear in delicate mechanisms and can plug small orifices. Dislodged dust particles drifting in space, illuminated by the sun, can look very much like stars to a star sensor or tracker on the spacecraft. This confusion can and has caused loss of attitude reference accuracy in operating spacecraft. Finally, dust typically hosts a population of viruses and bacteria that are unacceptable on a spacecraft destined for a visit to a planet on which Earth life might be viable.

Because of the concern for preventing dust contamination, spacecraft and their subsystems are normally assembled and tested in "clean room" environments. Details of how such environments are obtained are not of primary interest here. In general, clean rooms (see Fig. 3.1) require careful control of surfaces in the room to minimize dust generation and supply of conditioned air through high-efficiency particulate filters. In more stringent cases a unidirectional flow of

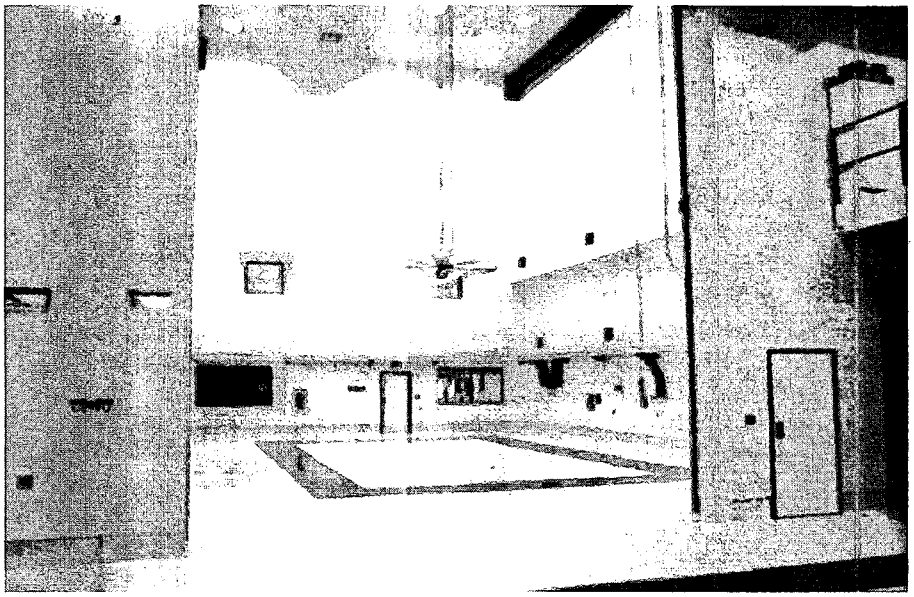


Fig. 3.1 Clean room. (Courtesy of Astrotech Space Operations.)

air is maintained, entering at the ceiling or one wall and exiting at the opposite surface.

The most advanced type of facility is the so-called laminar flow clean room, in which the air is introduced uniformly over the entire surface of a porous ceiling or wall and withdrawn uniformly through the opposing surface or allowed to exit as from a tunnel. Actual laminarity of flow is unlikely, especially in a large facility, but the very uniform flow of clean air does minimize particulate collection. Small component work is done at "clean benches," workbench type facilities where the clean environment is essentially restricted to the benchtop. The airflow exhausts toward the worker seated at the bench, as in Fig. 3.2.

Clean room workers usually must wear special clothing that minimizes particulate production from regular clothing or the body. Clean room garb typically involves gloves, smocks or "bunnysuits," head covering, and foot covering. All this must be lint free. In some cases masks are required as well. Because of the constant airflow and blower noise and the restrictive nature of the clothing, clean room work is often tiring even though it does not involve heavy labor.

Clean facilities are given class ratings such as Class 100,000, Class 1000, or Class 100 facilities. The rating refers to the particulate content of a cubic foot of air for particles between specified upper and lower size limits; thus, lower numbers represent cleaner facilities. Class 100 is the cleanest rating normally discussed and is extremely difficult to maintain in a large facility, especially when

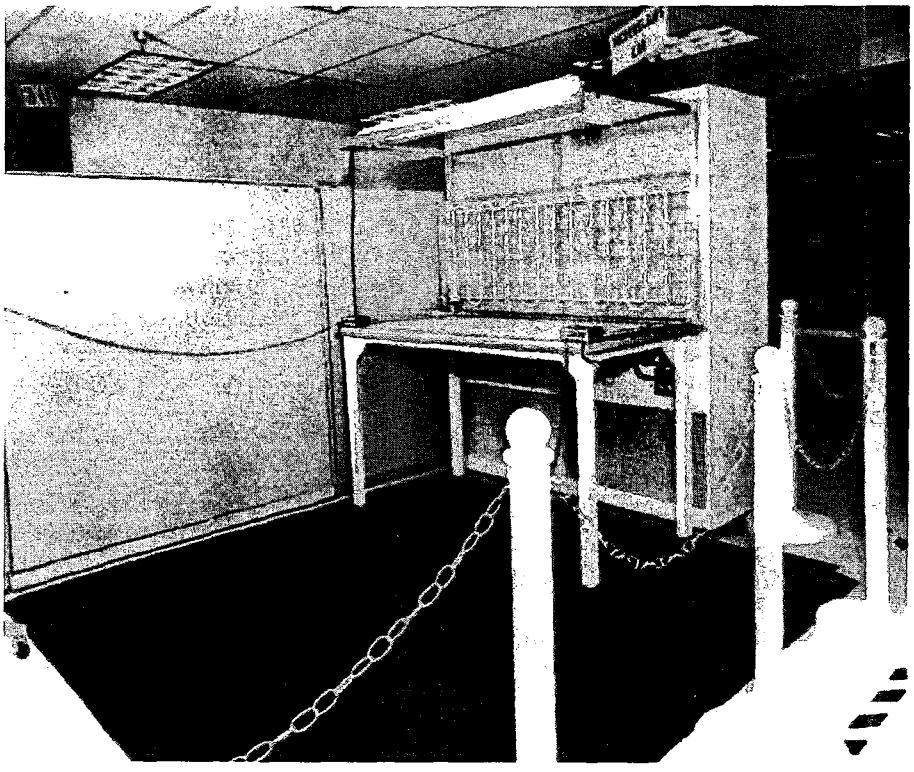


Fig. 3.2 Clean bench. (Courtesy of Ball Aerospace Systems Division.)

any work is in progress. Even Class 1000 is difficult in a facility big enough for a large spacecraft and one in which several persons might be working. A Class 10,000 facility is the best that might normally be achievable under such conditions and represents a typical standard for spacecraft work. Fresh country air would typically yield a rating of approximately Class 300,000. Clean rooms are usually provided with anterooms for dressing and airlocks for entry. Airshowers and sticky floormats or shoe scrubbers provide final cleanup.

A major hazard to many spacecraft components is static electricity. The triboelectric effect can produce very substantial voltages on human skin, plastics, and other surfaces. Some electronic components, in particular, integrated circuits or other components using metal-oxide semiconductor (MOS) technology, are extremely sensitive to high voltage and can easily be damaged by a discharge such as might occur from a technician's fingertip. To prevent such occurrences, clean room workers must be grounded when handling hardware. This is usually done using conductive flooring and conductive shoes or

ankle ground straps. For especially sensitive cases a ground strap on the wrist may be worn.

Because low relative humidity contributes to static charge accumulation, it is desirable that air in spacecraft work areas not be excessively dry. The compromise with the corrosion problem discussed earlier usually results in a chosen relative humidity of about 40–50%. Plastic cases and covers and tightly woven synthetic garments, all favored for low particle generation, tend to build up very high voltages unless treated to prevent it. Special conductive plastics are available, as are fabric treatment techniques. However, the conductive character can be lost over time, and so clean room articles must be constantly monitored.

In theory, with all electronic components mounted and all electrical connections mated, the spacecraft should be safe from static discharge. In practice, however, the precautions discussed earlier are generally observed by anyone touching or handling the spacecraft. The primary risk arises from contact with the circuit that occurs when pins are touched in an unmated connector. Unnecessary contact of this type should be avoided.

Transporting the spacecraft from point to point on Earth may well subject it to more damaging vibration and shock than experienced during launch. Road vibration and shock during ground transportation can be higher than those imposed by launch and the duration is much longer, usually hours or days compared with the few minutes required for launch. For short trips, as from building to building within a facility, the problem can best be handled by moving the spacecraft very slowly over a carefully selected and/or prepared route. For longer trips where higher speed is required, special vehicles employing air cushion suspension are usually required. These vehicles may be specially built for the purpose, or may simply be commercial vans specialized for delicate cargo. Truck or trailer suspensions can deteriorate in service, and it is usually desirable to subject them to instrumented road tests before committing expensive and delicate hardware to a long haul.

Flying is generally preferable to ground transportation for long trips. Jets are preferred to propeller-driven aircraft because of the lower vibration and acoustic levels. High g loads can occur at landing or as a result of turbulence, and the spacecraft must be properly supported to provide protection. The depressurization/pressurization cycle involved in climb and descent can also be a problem. For example, a closed vessel, although designed for several atmospheres of internal pressure, can easily collapse if it bleeds down to an internal pressure equivalent to several thousand feet altitude during flight and then is quickly returned to sea level. This is particularly a problem when transporting propulsion stages having large tanks with relatively thin walls.

When deciding between flight or ground transportation, it should be recalled that it will generally be necessary to transport the spacecraft by road to the airport, load it on the plane, and then reverse the procedure at the other end. For trips of moderate length, a decision should be made as to whether flying, with all

the additional handling involved, is in fact better than completing the entire trip on the ground.

In all cases, whether transporting the space vehicle by ground or air, it is essential that it be properly secured to the carrier vehicle structure. This requires careful design of the handling and support equipment. Furthermore, all delicate structures that could be damaged by continued vibration should be well secured or supported.

For some very large structures, the only practical means of long-range transportation is via water. Barges were used for the lower stages of the Saturn 5 launch vehicle and continue to be used to transport the shuttle external tank from Michoud, Louisiana, to Cape Canaveral, Florida.

The cleanliness, humidity, and other environmental constraints discussed earlier usually must remain in force during transportation. In many cases, as with the shipment by boat of the Hubble Space Telescope from its Sunnyvale, California, fabrication site to Cape Canaveral, this can present a significant logistical challenge.

3.3 Launch Environment

Launch imposes a highly stressful environment on the spacecraft for a relatively brief period. During the few minutes of launch, the spacecraft is subjected to significant axial loads by the accelerating launch vehicle, as well as lateral loads from steering and wind gusts. There will be substantial mechanical vibration and severe acoustic energy input. The latter is especially pronounced just after liftoff as the rocket engine noise is reflected from the ground. Aerodynamic noise also contributes, especially in the vicinity of Mach 1. During the initial phase of launch, atmospheric pressure will drop from essentially sea level to space vacuum. Aerodynamic heating of the spacecraft may impose thermal loads that drive some aspects of the spacecraft design. This initially occurs through heating of the nose fairing during low-altitude ascent, then directly by free molecular heating (see Chapter 6) after fairing jettison. Stage shutdown, fairing jettison, and spacecraft separation will each produce shock transients.

To ensure that the spacecraft is delivered to its desired orbit or trajectory in condition to carry out the mission, it must be designed for and qualified to the expected stress levels, with a margin of safety (see Chapter 8). To facilitate preliminary design, launch vehicle user handbooks specify pertinent parameters such as acoustic, vibration, and shock levels. For vehicles with a well-established flight history, the data are based on actual in-flight measurements. Vehicles in the developmental phase provide estimated or calculated data based on modeling and comparison with similar vehicles.

Environmental data of the type presented in user handbooks are suitable for preliminary analysis in the early phases of spacecraft design and are useful in

establishing initial structural design requirements. Because the spacecraft and launch vehicle interact, however, the actual environment will vary somewhat from one spacecraft payload to another, and the combination of launch vehicle and spacecraft must be analyzed as a coupled system.[2] As a result, the actual environment anticipated for the spacecraft changes with its maturing design and the resulting changes in the total system. Because this in turn affects the spacecraft design, it is clear that an iterative process is required.

The degree of analytical fidelity required in this process is a function of mass margins, fiscal resources, and schedule constraints. For example, structural modeling of the Viking Mars Orbiter/Lander was detailed and thorough because mass margins were tight. On the other hand, the Solar Mesosphere Explorer, a low-budget Earth orbiter that had a very large launch vehicle margin, was subjected to limited analysis. Many structures were made from heavy plate or other material that was so overdesigned that it limited the need for detailed analysis. When schedule is critical, extra mass may well be allocated to the structural design to limit the need for detailed analysis and testing.

Acoustic loads are pervasive within the nose fairing or payload bay, with peaks sometimes occurring at certain locations. Vibration spectra are usually defined at the base of the attach fitting or adapter. Shock inputs are usually defined at the location of the generating device, typically an explosively actuated or mechanically released device.

In many cases the various inputs actually vary somewhat from point to point, especially in the case of shock spectra. For convenience in preliminary design, this is often represented by a single curve that envelops all the individual cases. Examples of this may be seen among the curves presented in this chapter. In general, use of such curves will lead to a conservative design that, at the cost of some extra mass, is well able to withstand the actual flight environment.

To examine launch vehicle data, we present data drawn from user handbooks for some of the various major launch vehicles discussed in Chapter 5. Random vibration data are presented as curves of spectral density in g^2/Hz, essentially a measure of energy vs frequency of vibration.

For the shuttle, data are presented at the main longeron and keel fittings, whereas for the expendable vehicles it is at the spacecraft attachment plane. The first two curves for the shuttle (see Figs. 3.3 and 3.4) represent early predictions, and the third (Fig. 3.5) presents flight data for longeron vibration based on Space Transportation System (STS) flights 1–4. It is instructive to compare Figs. 3.3 and 3.5 and note that the flight data yield higher frequency vibration and higher y-axis levels than predicted. This is not a serious problem, because trunion fitting slippage tends to isolate much of this vibration from the payload. Flight data for the keel fitting (not shown) are very close to the predicted curve (Fig. 3.4).

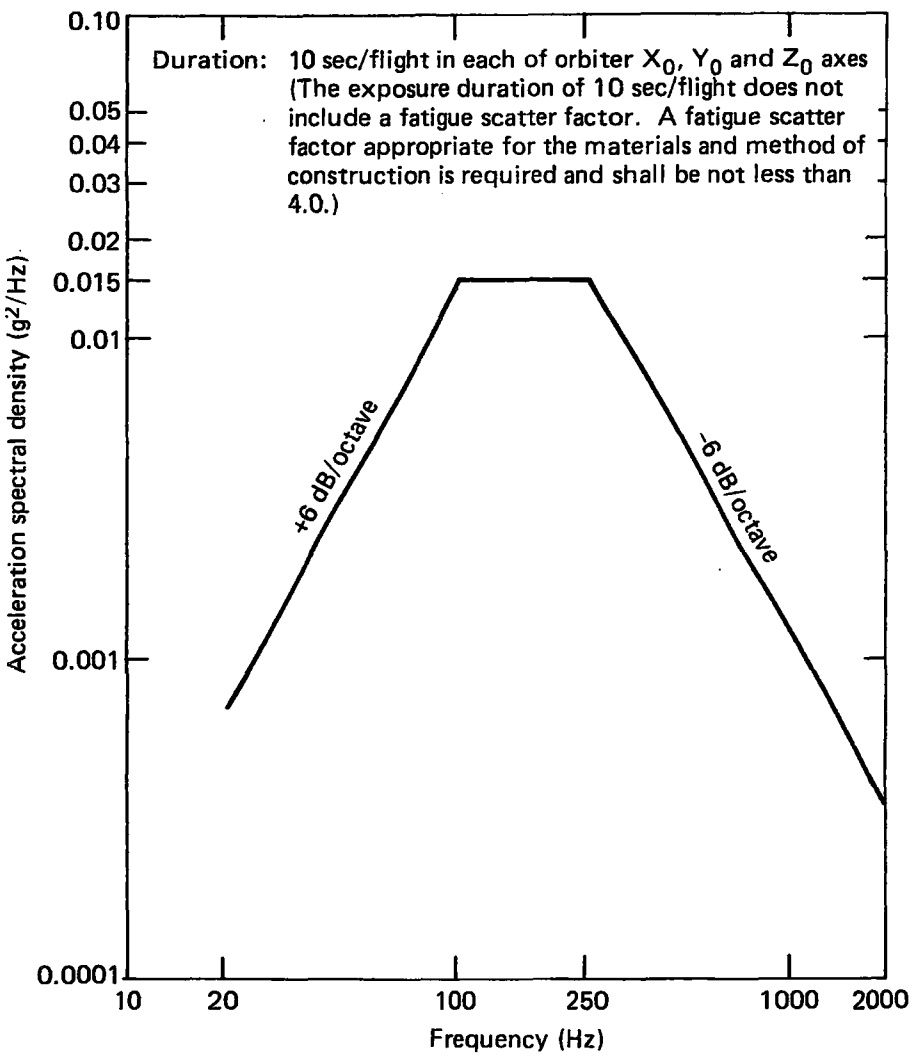


Fig. 3.3 Shuttle vibration environment: unloaded main longeron trunion-fitting vibration.

Provisions for mounting payloads in the shuttle bay are discussed in Chapter 5. These mountings allow for limited motion in certain directions. This helps decouple payloads from orbiter structural vibrations. Furthermore, the presence of the payload mass itself tends to damp the vibration. These effects lead to a vibration attenuation factor *CV*. This is presented in

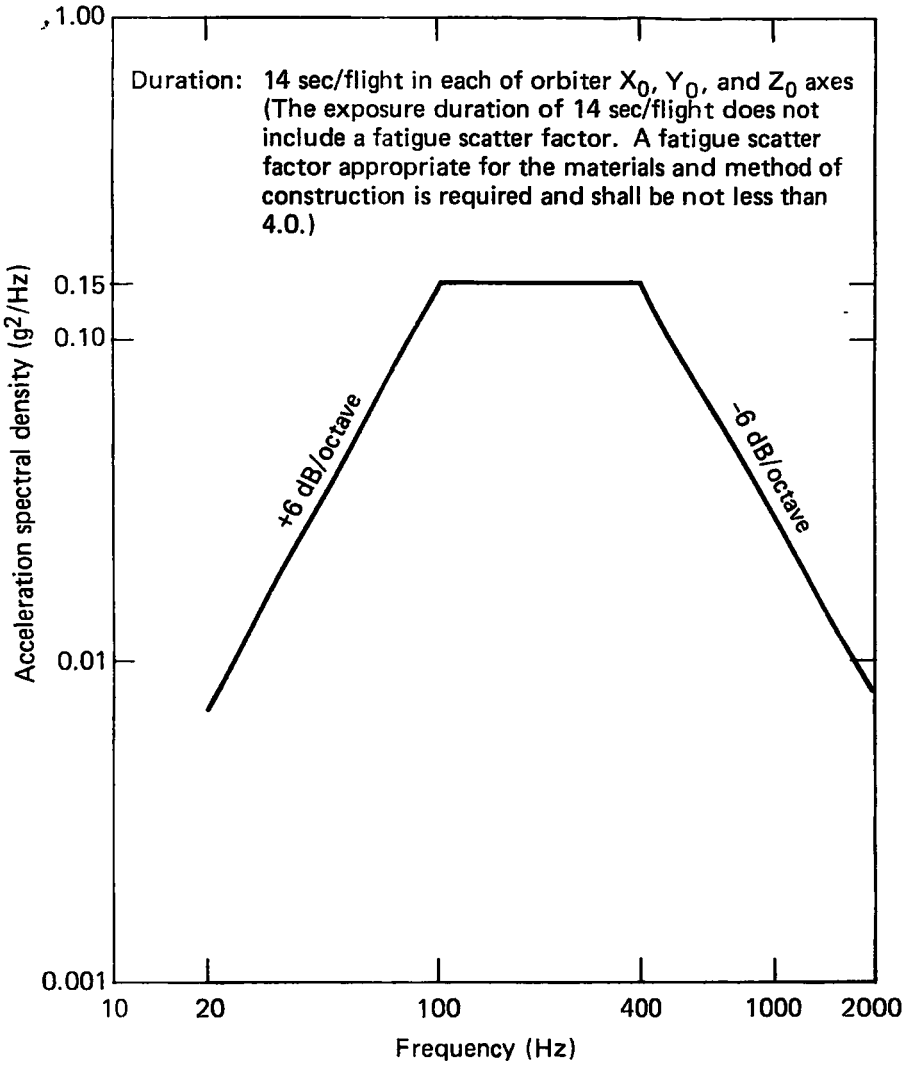


Fig. 3.4 Shuttle vibration environment: unloaded keel trunion fitting vibration.

Fig. 3.6. It is applied as

$$ASD_{\text{payload}} = CV \times ASD_{\text{unloaded orbiter structure}} \tag{3.1}$$

where *ASD* is the acceleration spectral density, i.e., the power spectral density of the vibrational acceleration (see Chapter 12).

Longitudinal vibration is generally caused by thrust buildup and tailoff of the various stages plus such phenomena as the "pogo" effect, which sometimes

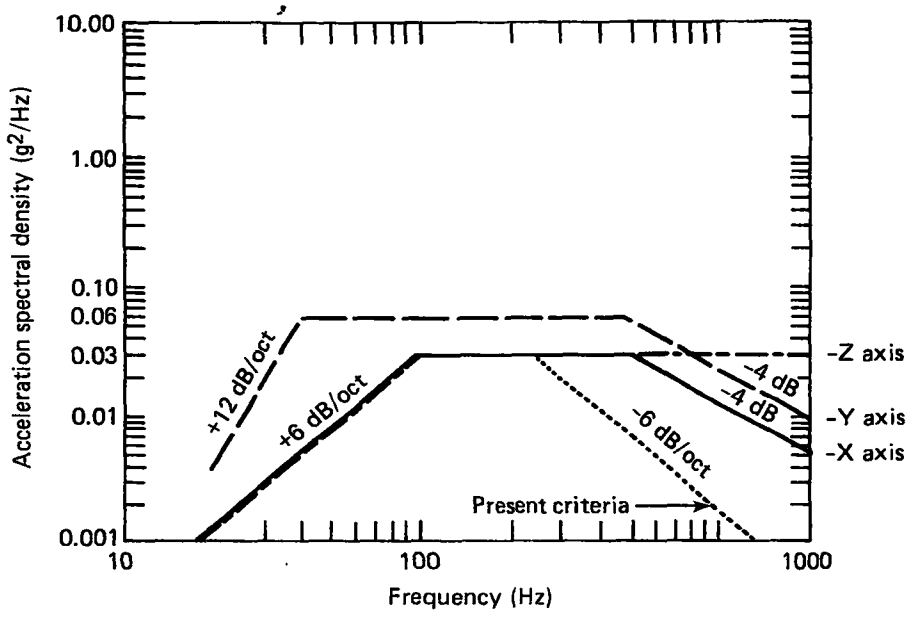


Fig. 3.5 Shuttle vibration environment: Orbiter main longeron random vibration criteria derived from flight data.

plagues liquid-propellant propulsion systems. This is manifested by thrust oscillations generally in the 5–50-Hz range. The phenomenon results from coupling of structural and flow system oscillations and can usually be controlled by a suitably designed gas-loaded damper in the propellant feed lines.

Lateral vibrations usually result from wind gust and steering loads as well as thrust buildup and tailoff.

Expendable vehicle data, presented as longitudinal and lateral sinusoidal vibration data, random vibration, and acoustic and shock spectra, are presented in Tables 3.1 and 3.2 and Figs. 3.7–3.20.

3.4 Atmospheric Environment

By definition, space vehicles are not primarily intended for operation within an atmosphere, whether that of Earth or otherwise. However, flight through an atmosphere, either upon ascent or reentry or both, and possibly at different planets, represents an important operational phase for many space vehicles. Significant portions of Chapter 5, and the entirety of Chapter 6, are devoted to this topic. In this section, we consider in some detail the properties of both the

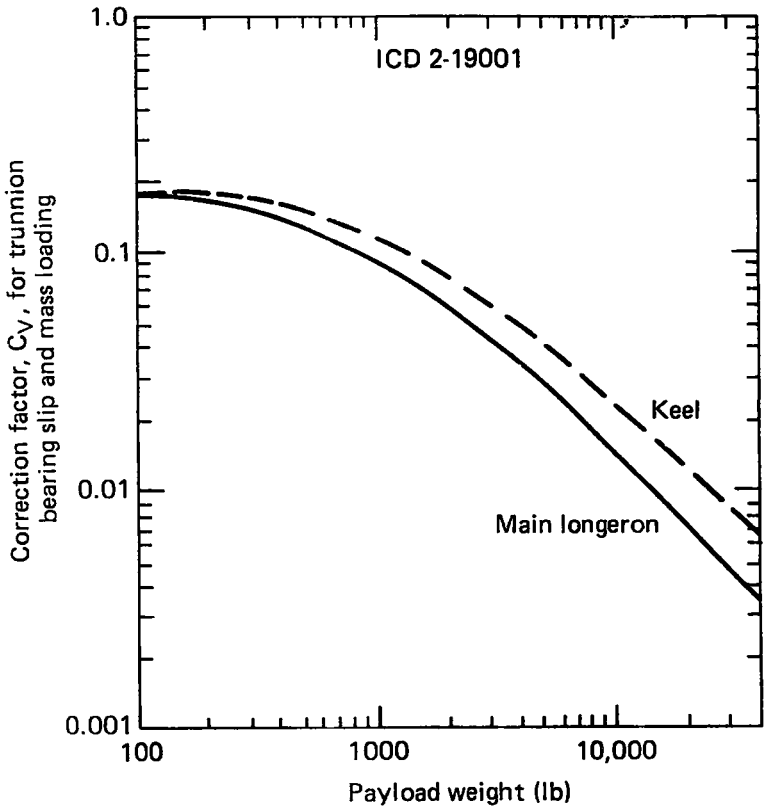


Fig. 3.6 Shuttle vibration environment: vibration attenuation factor.

"standard" Earth atmospheric environment, as well as the effect of some important variations likely to be encountered in practice. The present discussion is restricted to the properties of the atmosphere when viewed as a neutral gas. The upper atmosphere environment, including the effects of partial vacuum and space plasma, are treated in subsequent sections.

Table B.17 and Fig. 3.21 present the current U.S. Standard Atmosphere model,[3] and Fig. 3.22 shows the density of atomic oxygen at low-orbit altitudes, the effects of which are discussed in a later section. It is seen that substantial variation of upper atmosphere properties with the 11-year solar cycle exists. Figure 3.23 shows historical and predicted solar cycle variations[4] as measured by the $F_{10.7}$ flux, i.e., the measured solar intensity at a wavelength of 10.7 μm.

As will be discussed further both here and in Chapters 4 and 7, the solar cycle variation and its effect on the upper atmosphere and space radiation environments can be of great importance in both mission and spacecraft design. Orbital

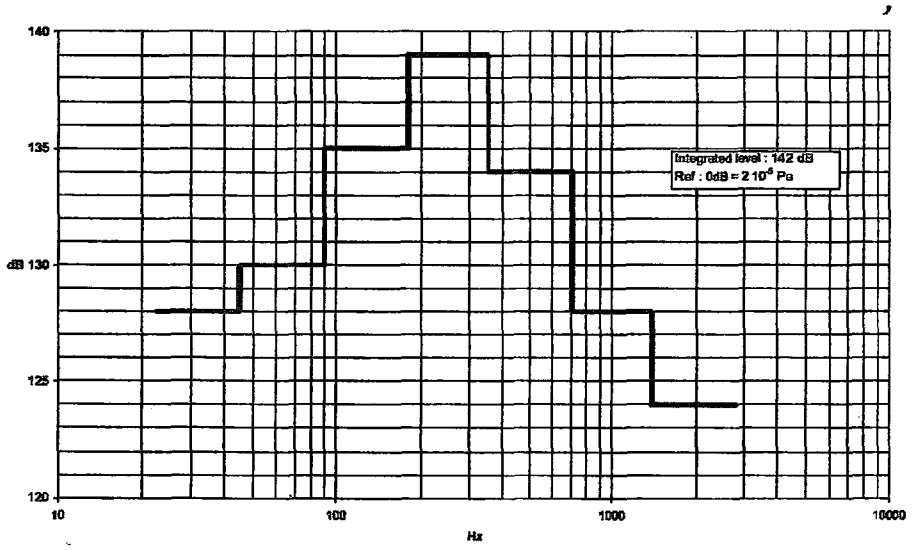


Fig. 3.7 Ariane V payload acoustic environment. (Courtesy Arianespace.)

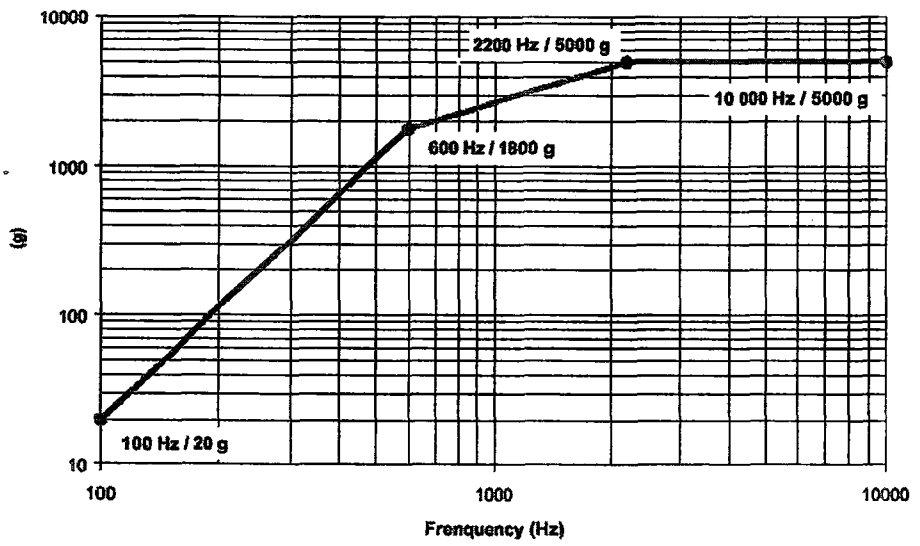


Fig. 3.8 Ariane V shock spectrum envelope at spacecraft separation interface. (Courtesy Arianespace.)

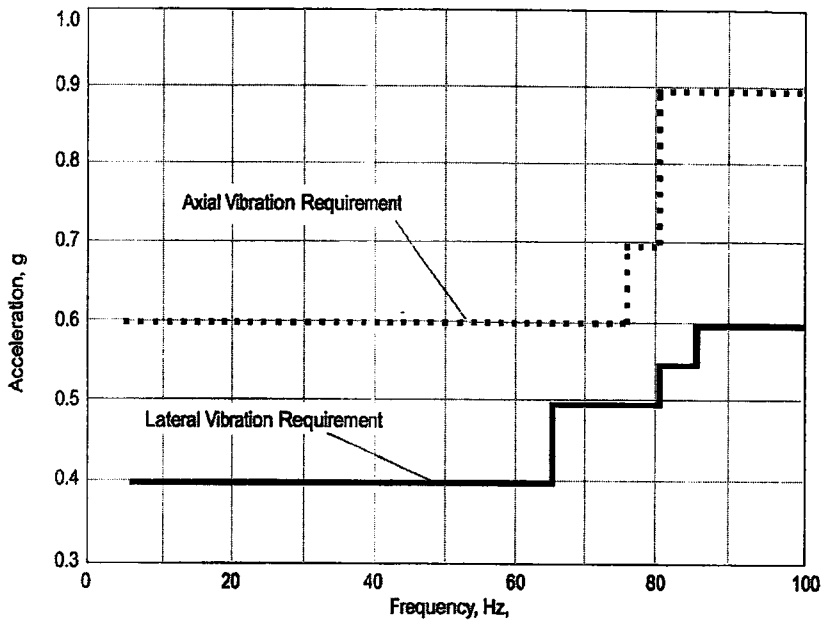


Fig. 3.9 Atlas IIAS, IIIA, IIIB, V-400 sinusoidal vibration requirement. (Courtesy Lockheed Martin.)

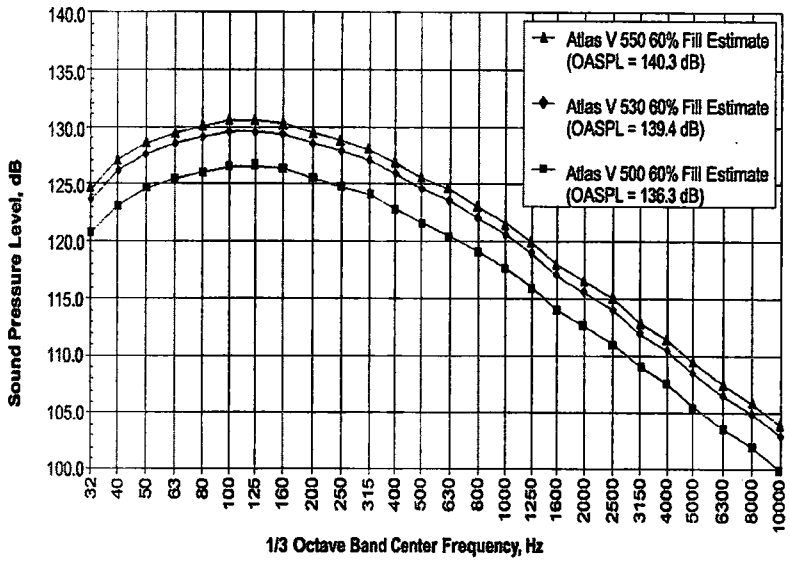


Fig. 3.10 Acoustic environment for Atlas V short payload fairing. (Courtesy Lockheed Martin.)

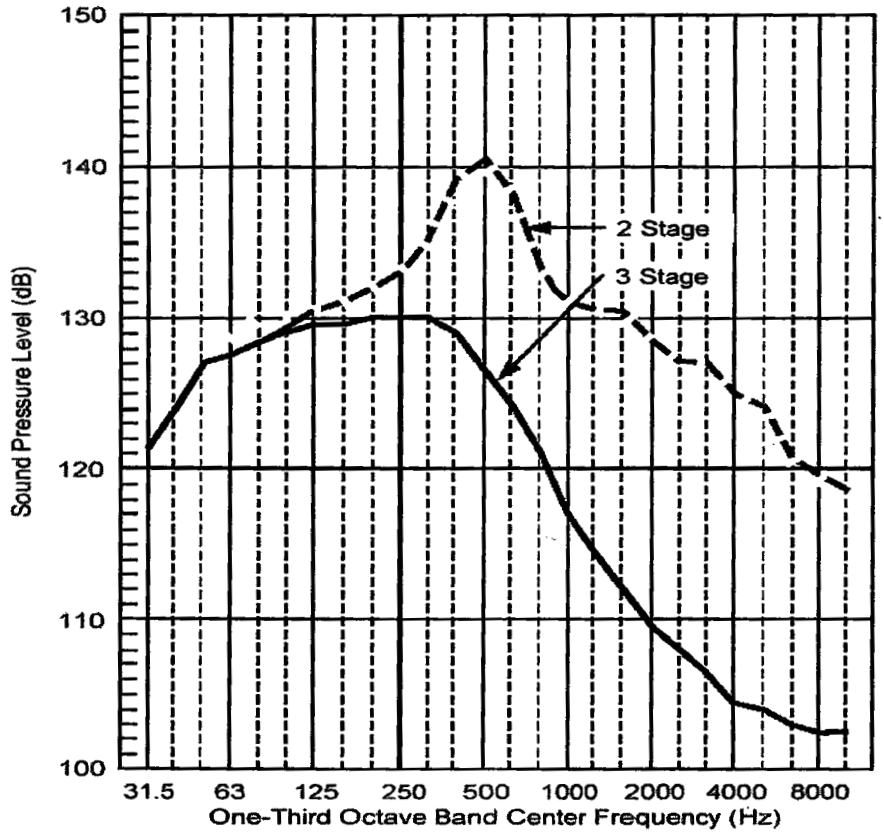


Fig. 3.11 Delta II 7920 and 7925 acoustic environment, 9.5 foot fairing. (Courtesy Boeing.)

operations during periods of greater solar activity, and consequently higher upper atmosphere density, produce both more rapid orbit decay and more severe aerodynamic torques on the spacecraft. This can in turn necessitate a greater mass budget for secondary propulsion requirements for drag makeup and similar compensations in the attitude control system design. The radiation exposure budget must also be assessed with an understanding of the portion of the solar cycle in which the spacecraft is expected to operate.

Other variations in the standard atmosphere are of significance in the design of both launch and entry vehicles. Atmosphere models exhibit smoothly varying properties, representative of average behavior, whereas in nature numerous fairly abrupt boundaries can exist on a transient basis. An important example is that of wind shear, which as the name implies is an abrupt variation of wind speed with altitude.

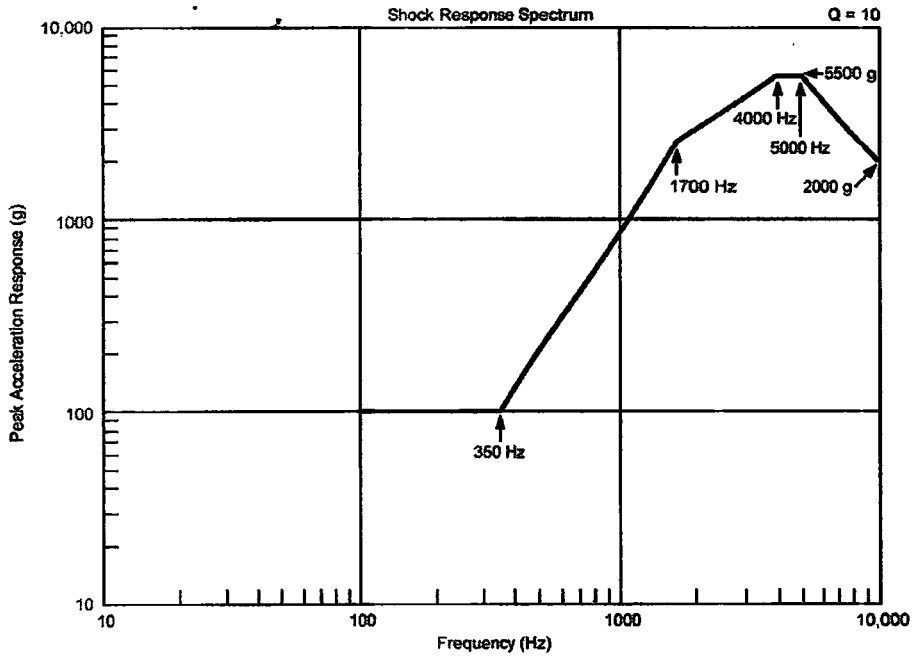
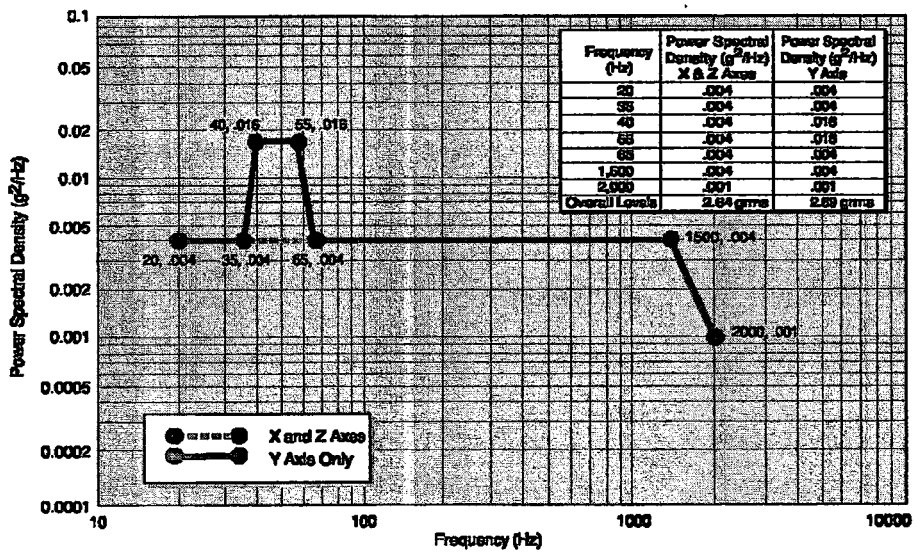


Fig. 3.12 Delta II spacecraft interface shock environment (6019 and 6915 payload attach fitting). (Courtesy Boeing.)



Frequency (Hz)	Power Spectral Density (g^2/Hz) X & Z Axes	Power Spectral Density (g^2/Hz) Y Axis
20	.004	.004
35	.004	.004
40	.004	.016
55	.004	.016
65	.004	.004
1,500	.004	.004
2,000	.001	.001
Overall Levels	2.64 grms	2.59 grms

Fig. 3.13 Pegaus XL random vibration environment. (Courtesy Orbital Sciences Corporation.)

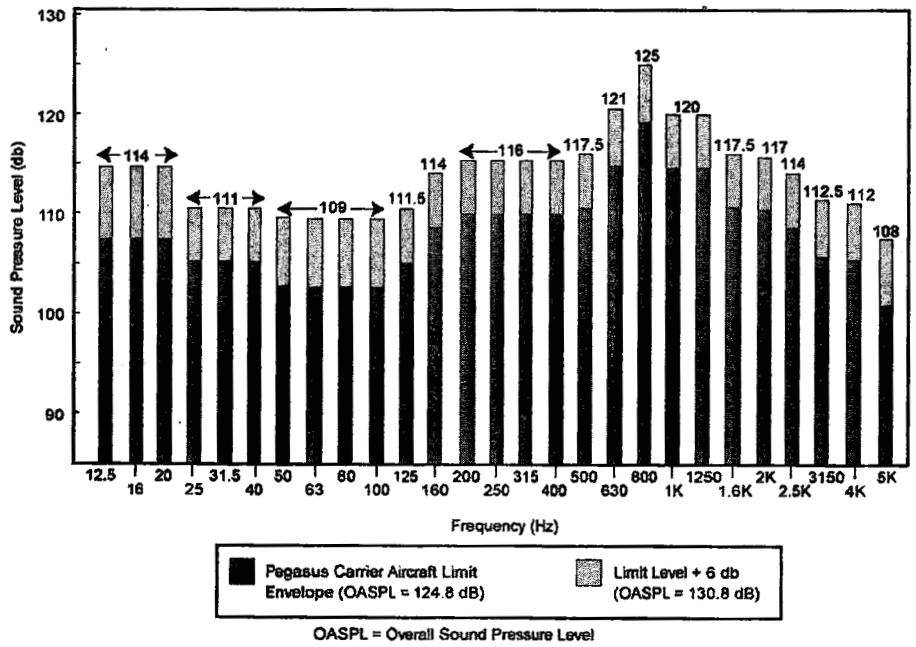


Fig. 3.14 Pegaus XL payload acoustic environment. (Courtesy Orbital Sciences Corporation.)

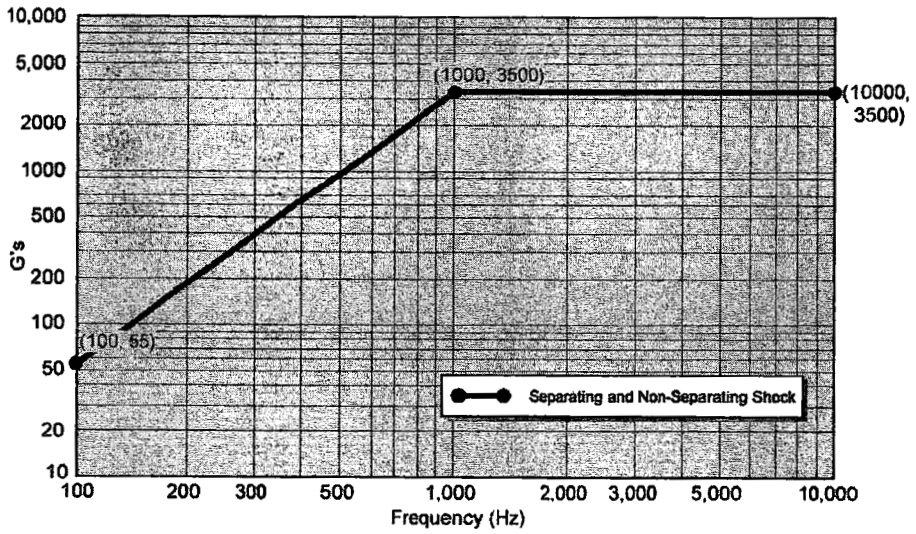


Fig. 3.15 Pegaus XL payload shock environment at separation plane. (Courtesy Orbital Sciences Corporation.)

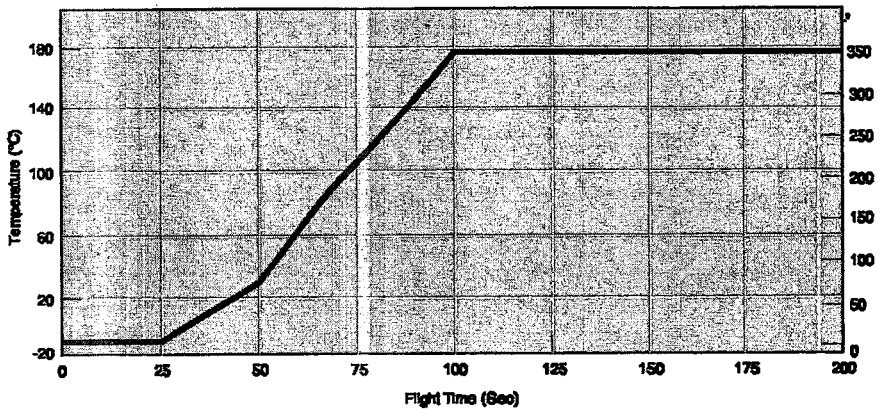


Fig. 3.16 Pegaus XL fairing inner surface temperature for worst-case hot trajectory. (Courtesy Orbital Sciences Corporation.)

Wind shear appears to an ascent vehicle climbing between layers as a sharp gust, effectively increasing the aerodynamic angle of attack and imposing transient loads on the vehicle. Such loads, if excessive, can cause in-flight breakup or, on a lesser scale, violation of payload lateral load constraints. Thus, all launch vehicles will be subject to a wind shear constraint, the magnitude of which depends on the vehicle, as a condition of launch.

For unguided ballistic and semiballistic entry vehicles, the primary effect of unmodeled wind shear is on landing point accuracy. For gliding entry vehicles such as the space shuttle, the threat of excessive wind shear is the same as that for ascent vehicles; excessive transient loads could overstress the vehicle. Also, of course, excessive unmodeled headwinds, whether shear is present or not, reduce the vehicle's kinetic energy. Entry trajectory design and terminal area energy management schemes must incorporate reasonable worst-case headwind predictions, or risk failing to reach the intended runway. Several shuttle missions have reached the terminal area in an unexpectedly low energy state.

Conceptually similar to wind shear is density shear, i.e., a sudden variation in layer density as a function of altitude. Shuttle flight experience has revealed drag—hence atmospheric density—variations of up to 19% over periods of a few seconds.[5] Again, unmodeled drag variations are of concern for gliding entry vehicles, for which energy control is critical. Depending on the vehicle control system design, abrupt drag variations may result in an undesirable autopilot response. The space shuttle, for example, attempts to fly a nominal reference drag profile; differences between flight and reference values result in vehicle attitude adjustments as the autopilot seeks to converge on the nominal drag value. Spurious drag variations result in anomalous fuel

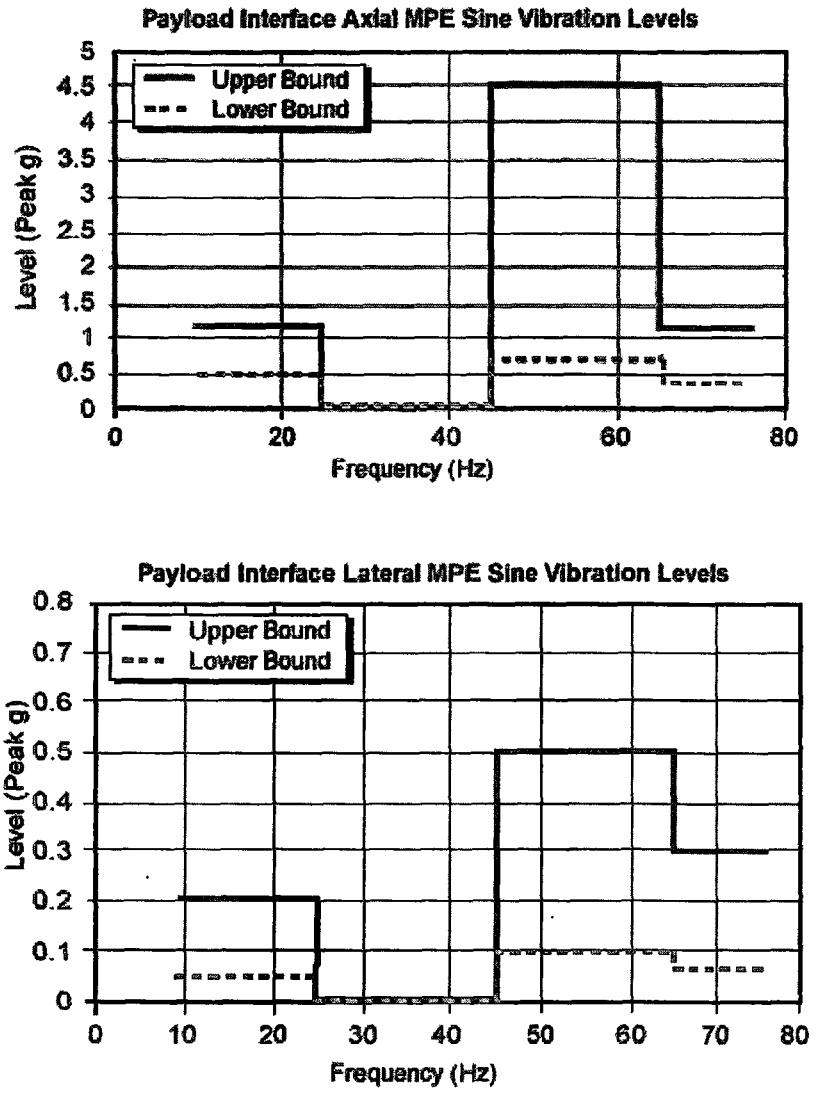


Fig. 3.17 Taurus axial and lateral sine vibration environment. (Courtesy Orbital Sciences Corporation.)

consumption as the attitude is altered to respond to what is effectively just noise in the system.

Not included in standard atmosphere models, but present in reality, are so-called noctilucent or polar mesospheric clouds. These clouds are found at high latitudes, typically above 50°, are comprised of very fine ice crystals averaging

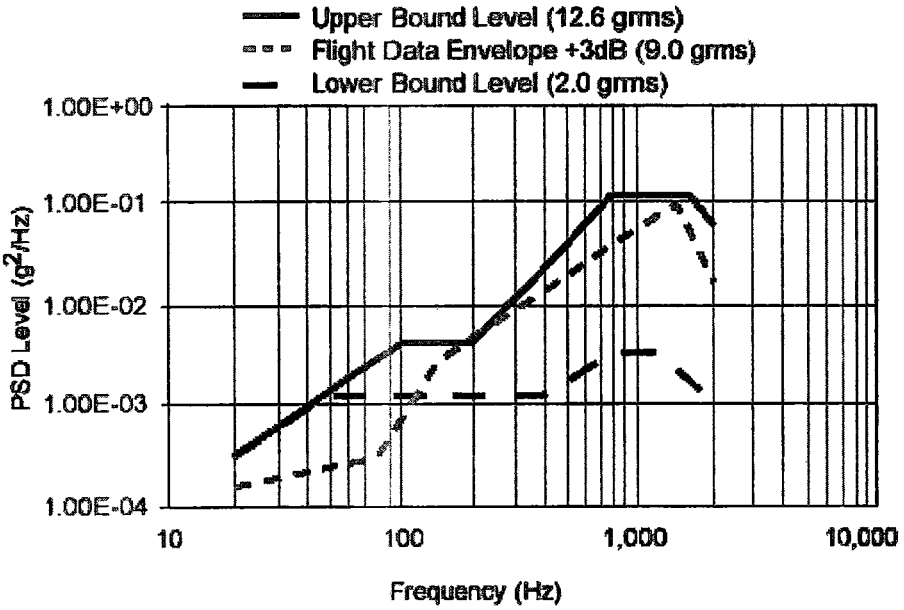


Fig. 3.18 Taurus random vibration environment. (Courtesy Orbital Sciences Corporation.)

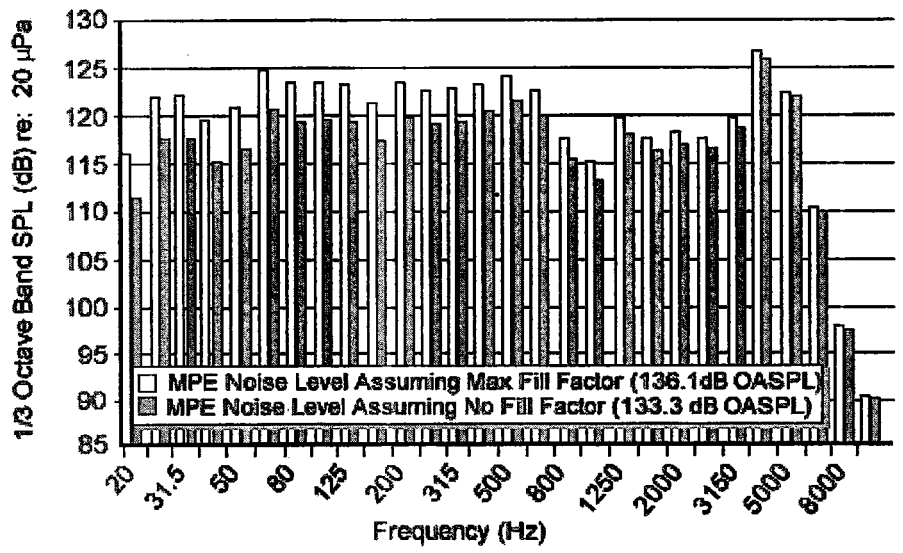


Fig. 3.19 Taurus payload acoustic environment, 63" fairing. (Courtesy Orbital Sciences Corporation.)

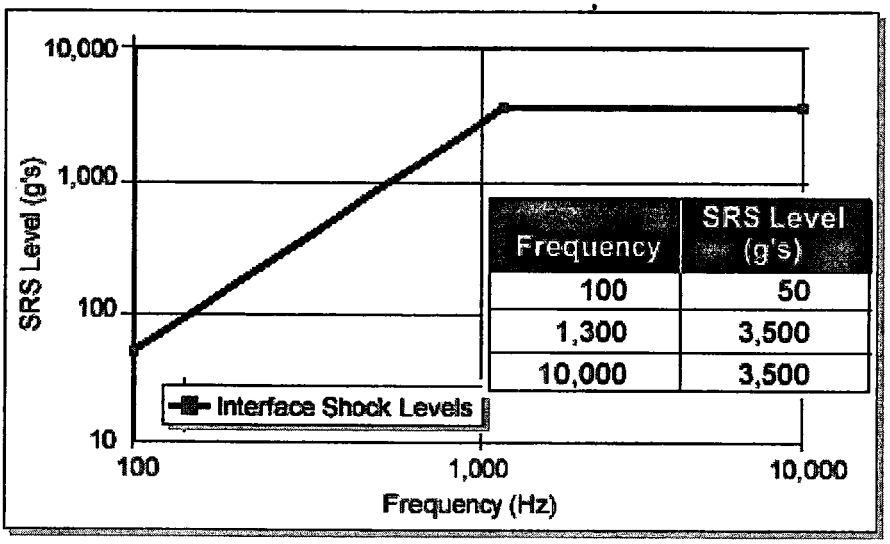


Fig. 3.20 Taurus shock spectrum at payload interface. (Courtesy Orbital Sciences Corporation.)

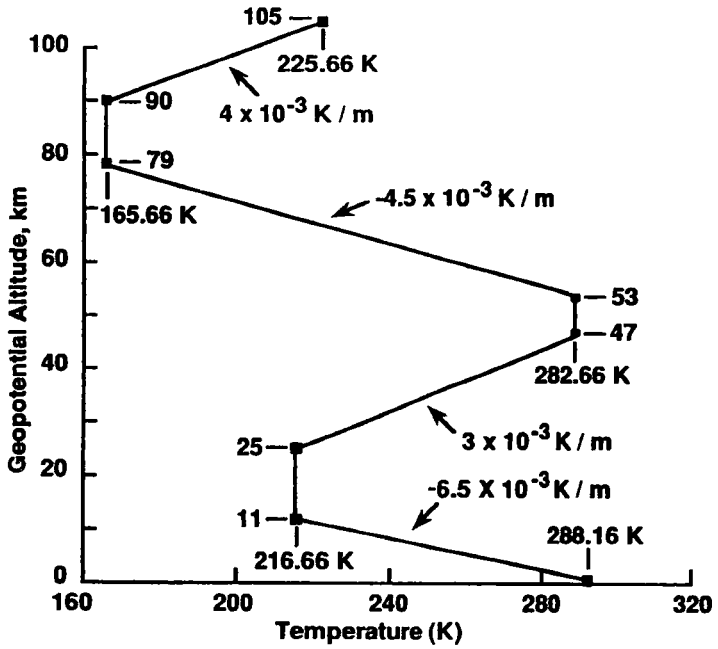


Fig. 3.21 Temperature distribution of standard atmosphere.

Table 3.1 Ariane V load factors at spacecraft separation plane

Events/Axis	Acceleration, *g*
Solid Booster Shutdown	
Axial	+4.5
Lateral	0.25
Core Stage Shutdown	
Axial	+3.5
Lateral	0.25
Upper Stage Shutdown	
Axial	+0.4
Lateral	0.25
Sinusoidal Loads	
Axial, 5–100 Hz	< 1.0
Lateral, 0–25 Hz	< 0.8
Lateral, 25–100 Hz	< 0.6

50 nm in size, and are confined to altitudes of 80–90 km. These clouds have no significant effect on launch vehicles and are too low to be of concern for satellites, but may be of concern for entry vehicles. Because of concerns that such particles could significantly abrade shuttle thermal protection tiles, shuttle entry trajectories are planned to avoid passage through the regions of latitude and altitude where noctilucent clouds can form. This poses a significant constraint, because it requires the avoidance of descending-node reentries for high-inclination flights.[5]

3.5 Space and Upper Atmosphere Environment

The space environment is characterized by a very hard (but not total) vacuum, very low (but not zero) gravitational acceleration, possibly intermittent or impulsive nongravitational accelerations, ionizing radiation, extremes of thermal radiation source and sink temperatures, severe thermal gradients, micrometeoroids, and orbital debris. Some or all of these features may drive various aspects of spacecraft design.

3.5.1 Vacuum

Hard vacuum is of course one of the first properties of interest in designing for the space environment. Many key spacecraft design characteristics and techniques are due to the effects of vacuum on electrical, mechanical,

Table 3.2 Atlas center of gravity limit load factors

Event/Axis	Steady-state (g)		Dynamic (g)	
	Axial	Lateral	Axial	Lateral
Launch				
IIAS, IIIA, IIIB	1.2		± 1.1	± 1.3
V-400	1.2		± 0.5	± 0.8
V-500	1.6		± 2.0	± 2.0
Winds				
IIAS	2.7	± 0.4	± 0.8	± 1.6
IIIA, IIIB	2.7	± 0.4	± 0.3	± 1.6
V-400	2.2	± 0.4	± 0.5	± 1.6
V-500	2.4	± 0.4	± 0.5	± 1.6
SRM Separation				
V-500	3.0		± 0.5	± 0.5
BECO				
V-400, V-500	5.5		± 0.5	± 1.0
(Max Axial)				
IIAS	5.0		± 0.5	± 0.5
IIIA, IIIB	5.5		± 0.5	± 0.5
(Max Lateral)				
IIAS	2.5		± 1.0	± 2.0
IIIA, IIIB	2.5		± 1.0	± 1.5
SECO				
IIAS, IIIA, IIIB	2.0		± 0.4	± 0.3
MECO				
(Max Axial)				
All versions	4.8		± 0.5	± 0.2
(Max Lateral)				
All versions	0.0		± 2.0	± 0.6

Notes: (1) For Atlas IIAS, IIIA, IIIB, the load factors above yield a conservative design envelope for spacecraft in the 1800–4500 kg class, with the first lateral mode above 10 Hz and the first axial mode above 15 Hz.

(2) For Atlas V-400, the load factors provide a conservative design for spacecraft in the 900–9000 kg range with the first lateral and axial modes above 8 Hz and 15 Hz, respectively.

(3) For Atlas V-500, the load factors are conservative for spacecraft in the 4500–19,000 kg range, with first lateral and axial modes above 2.5 Hz and 15 Hz.

Table 3.3 Delta sinusoidal vibration flight environment and test requirements

Event/Axis	Frequency (Hz)	Level	Sweep Rate
Flight			
Thrust	5.0–6.2	1.27 cm DA	
	6.2–100	1.0 *g* (0–peak)	
Lateral	5.0–100	0.7 *g* (0–peak)	
Acceptance Test			
Thrust	5.0–6.2	1.27 cm DA	4 octave/min
	6.2–100	1.0 *g* (0–peak)	4 octave/min
Lateral	5.0–100	0.7 *g* (0–peak)	4 octave/min
Design Qualification Test			
Thrust	5.0–7.4	1.27 cm DA	2 octave/min
	7.4–100	1.4 *g* (0–peak)	2 octave/min
Lateral	5.0–6.2	1.27 cm DA	2 octave/min
	6.2–100	1.0 *g* (0–peak)	2 octave/min
Protoflight Test			
Thrust	5.0–7.4	1.27 cm DA	4 octave/min
	7.4–100	1.4 *g* (0–peak)	4 octave/min
Lateral	5.0–6.2	1.27 cm DA	4 octave/min
	6.2–100	1.0 *g* (0–peak)	4 octave/min

Note: DA = double amplitude.

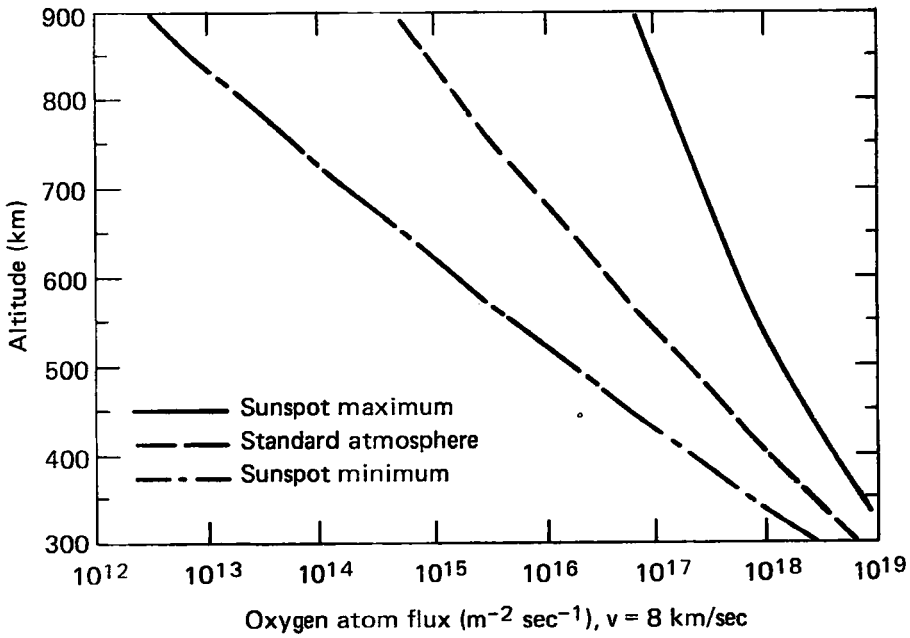


Fig. 3.22 Oxygen atom flux variation with altitude.

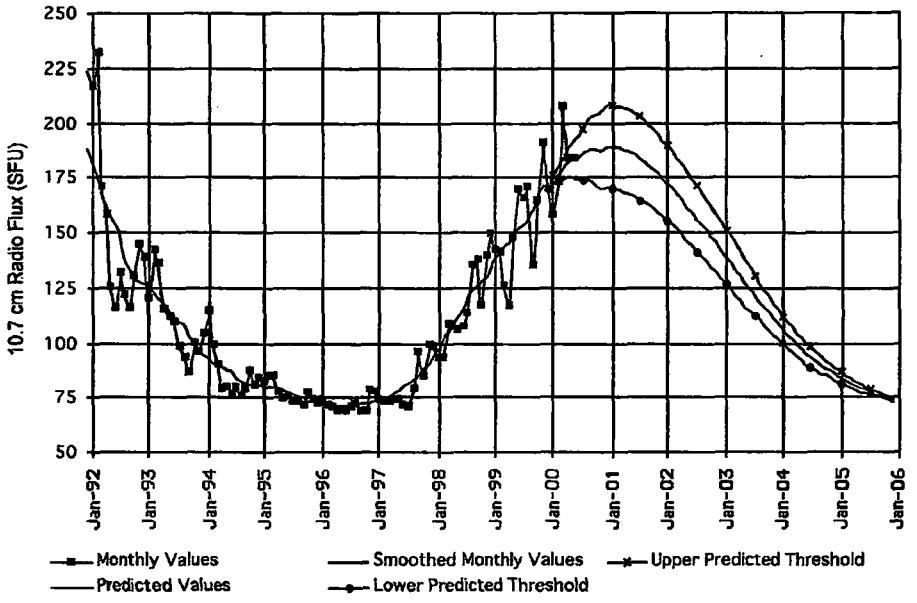


Fig. 3.23 Historical and predicted $F_{10.7}$ solar flux.[4]

and thermal systems. Material selection is crucially affected by its vacuum behavior. Many materials that see routine engineering use for stressful ground engineering applications are inappropriate even for relatively benign spacecraft applications.

Most materials will outgas to at least some extent in a vacuum environment. Metals will usually have an outer layer into which gases have been adsorbed during their tenure on Earth, and which is easily released once in orbit. Polymers and other materials composed of volatile compounds may outgas extensively in vacuum, losing substantial fractions of their initial mass. Some basically nonvolatile materials, such as graphite-epoxy and other composites, are hygroscopic and can absorb considerable water from the air. This water will be released over a period of months once the spacecraft is in orbit. Some plating materials will, when warm, migrate in vacuum to colder areas of the spacecraft when they recondense. Cadmium is notorious in this regard; thus, conventional cadmium-plated fasteners are an anathema in space applications.

Outgassing materials can be a problem for several reasons. In polymeric or other volatile materials, the nature and extent of the outgassing can lead to serious changes in the basic material properties. Even where this does not occur, as in water outgassing from graphite-epoxy, structural distortion can result. Such

composites are often selected because of their high stiffness-to-weight ratio and low coefficient of thermal expansion, for applications where structural alignment is critical. Obviously, it is desirable to preserve on orbit the same structure as was fabricated on the ground. Outgassing is also a problem in that the vapor can recondense on optical or other surfaces where such material depositions would degrade the device performance. Even if the vapor does not condense, it can interfere with the desired measurements. For example, ultraviolet astronomy is effectively impossible in the presence of even trace amounts of water vapor.

Outgassing is usually dealt with by selecting, in advance, those materials where it is less likely to be a problem. In cases where the material is needed because of other desirable properties, it will be "baked out" during a lengthy thermal vacuum session and then wrapped with tape or given some other coating to prevent re-absorption of water and other volatiles. Obviously, other spacecraft instruments and subsystems must be protected while the bake-out procedure is in progress.

Removal of the adsorbed O_2 layer in metals that do not form an oxide layer, such as stainless steel, can result in severe galling, pitting, and cold welding between moving parts where two pieces of metal come into contact. Such problems are usually avoided by not selecting these materials for dynamic applications in the space environment.

Moving parts require lubrication, for which traditional methods are at best problematic in vacuum. Even on the ground, lubricants can degrade with time, and dry out if originally liquid. The difficulty of finding stable lubricants is greatly exacerbated for the spaceflight regime, where we have unattended functional lifetimes measured in years, ambient pressures on the order of $10^{-6}\,N/m^2$ or less, temperatures ranging from 200–350 K or to even greater extremes, and where outgassing or evaporation can pose significant problems for other instruments or subsystems.

Space lubricants must therefore be selected with due consideration for the viscosity, vapor pressure, operating temperature range, and outgassing properties of the material. Of these, outgassing properties, which are treated in standard references,[6] are possibly the most important, because if the material outgasses substantially its other attributes, no matter how desirable, are unlikely to remain stable over time.

3.5.2 Partial Vacuum

Although the vacuum in low Earth orbit, for example at 200 km, is better than anything obtainable on the ground, it is by no means total. At shuttle operating altitudes, enough residual atmosphere remains to interact in a significant fashion with a spacecraft. Drag and orbit decay due to the residual atmosphere are discussed in Chapter 4; it may be necessary to include propulsion for drag compensation to prevent premature reentry and destruction of the

spacecraft. Of greater interest here, however, are the possible chemical interactions between the upper atmosphere atomic and molecular species and spacecraft materials.

It was noted during early shuttle missions that a pronounced blue glow appeared on various external surfaces while in the Earth's shadow. This was ascribed to recombination of atomic oxygen into molecular oxygen on contact with the shuttle skin. Although it presented no problems to the shuttle itself, the background glow is a significant problem for certain scientific observations.

Apart from its role in generating shuttle glow, atomic oxygen is an extremely vigorous oxidizer, and its prevalence in LEO ($\sim 10^{14}$ particles/cm^2/s) dictates the use of non-oxidizing surface coverings for extended missions. Samples returned from the 1984 on-orbit repair of the Solar Maximum Mission spacecraft showed that the Kapton^{TM} thermal blanketing material had been severely eroded by the action of atomic oxygen. It is now known that vulnerable materials such as thin (1 mil) Kapton^{TM} blankets can be destroyed within a few weeks.[6]

The combined effects of thermal extremes and the near-vacuum environment, in combination with solar ultraviolet exposure, may alter the reflective and emissive characteristics of the external spacecraft surfaces. When these surfaces are tailored for a particular energy balance, as is often the case, degradation of the spacecraft thermal control system performance can result. Thus, long-lived spacecraft must have paint or coatings that are "nonyellowing" if changes in the overall thermal balance are to be minimized.

A particularly annoying partial vacuum property is the relative ease with which low-density neutral gases are ionized, a phenomenon known as Paschen breakdown, which provides excellent but unintended conductive paths between points in electronic hardware that are at moderate to high potential differences. This tendency is aggravated by the fact that, at high altitudes, the residual molecular and atomic species are already partly ionized by solar ultraviolet light and various collision processes.

The design of electronic equipment intended for use in launch vehicles is of course strongly affected by this fact, as is the design of spacecraft that are intended for operation in very low orbits. A key point is that, even though a spacecraft system (such as a command receiver or inertial navigation system) is intended for use only when in orbit, it may be turned on during ascent. If this is so, then care needs to be exercised to prevent electrical arcing during certain phases of flight. To this end, spacecraft equipment that must be on during the ascent phase should be operated during the evacuation phase of thermal vacuum chamber testing.

Spacecraft intended for operation on the surface of Mars are also vulnerable to Paschen breakdown effects, as well as to the formation of arcs in the sometimes dusty atmosphere.

3.5.3 Space Plasma and Spacecraft Charging

So far we have discussed the space and upper atmosphere environment as if it were electrically neutral. In fact, it is not, and it should be recognized as a plasma, i.e., a hot, heavily ionized medium often referred to as a "fourth state of matter," after solids, liquids, and gases.[7] The universe is more than 99% plasma by mass; "ordinary" matter is the rare exception. Plasmas are formed whenever there is sufficient energy to dissociate and ionize a gas and to keep it from cooling and recombining into a neutral state. The sheath of hot, ionized gas around a reentry vehicle is one example, the interstellar medium is another, and the interior of a star is yet another.

Interplanetary space is filled with plasma generated by the sun within which the planets, asteroids, comets, etc., move. The magnetic fields of Jupiter, Saturn, and to a lesser extent Earth exert a magnetohydrodynamic effect on the plasma, shaping it into locally toroidal belts of charged particles, called Van Allen belts, in honor of their discoverer, whose radiation counter aboard Explorer 1 provided the first evidence of their existence. Usually these radiation belts have no visible effect; however, during periods of high solar activity, a heavier than normal flow of charged particles into the upper atmosphere can be redirected to the magnetic polar regions, producing the result known as the aurora borealis, or "northern lights."

Motion of the magnetically active planets within the plasma produces an interaction of the local planetary field with the interplanetary medium, creating a "bow shock" very similar to that for a hypersonic entry vehicle in an atmosphere (see Fig. 6.12), but shaped by electromagnetic forces rather than those of continuum fluid dynamics. The motion of the sun through the local interstellar medium produces a similar effect on a much larger scale. One goal of the Voyager missions launched in 1977 was to reach, and thus help define, this solar influence boundary.

The plasma, while essentially neutral as a whole, is populated with moving, electrically charged particles, specifically electrons and positively charged ions, generally having approximately equal kinetic energy. The flow of charge defines an electric current, which is positive by definition if ions are moving, and negative for moving electrons. The lightest possible ion is the single proton, the nucleus of a hydrogen atom, with a mass 1840 times that of the electron. Other ions are even more massive; thus, electrons move at speeds orders of magnitude faster than ions, and even faster relative to any spacecraft.

As the spacecraft moves through the plasma, it preferentially encounters electrons, more of which bombard the spacecraft in a given time than do the slower ions. There is thus a negative current tending to charge the spacecraft. As the resulting negative charge grows, Coulomb forces build, slowing accumulation of electrons and enhancing the attraction

of positively charged ions. Ultimately, the positive and negative currents equilibrate. This will occur with the spacecraft at a "floating potential" somewhat negative relative to that of the surrounding plasma, resulting from the preferential accumulation of the faster electrons, relative to the equally energetic, but more massive and thus slower, ions. This floating potential will depend on the orbit parameters, spacecraft size and geometry, solar cycle, terrestrial season, and other factors.

Spacecraft charging can be "absolute" with respect to the plasma, "differential" with respect to different parts of a spacecraft, or both. If the spacecraft is highly conductive throughout, differential charging cannot occur. At lower altitudes, there is sufficient ion density in the plasma that large charge differences cannot develop even between separate, electrically isolated portions of a spacecraft. At GEO spacecraft altitude, this is not the case. If some portions of such a vehicle are electrically isolated from others, a substantial differential charge buildup can occur. When the point is reached at which the potential difference is sufficient to generate a high-voltage arc, charge equilibration will occur, quite possibly in a destructive manner. This behavior can occur at any time, but is greatly enhanced during periods of high solar activity. Numerous spacecraft have been damaged, or lost, due to this mechanism.[8,9] It is for this reason that it is recommended that conduction paths be provided to all parts of a spacecraft, including especially thermal blankets, solar arrays, etc., as discussed in Chapter 8.

While differential charging is not ordinarily of concern for LEO spacecraft, absolute charging of the spacecraft can cause problems. One effect is sputtering, in which large negative charges attract ions to impact the spacecraft at high speed, physically removing some surface atoms. This alters the thermal properties of the surface and adds to the contamination environment around the spacecraft.

If there are no exposed conductors carrying different voltages, LEO spacecraft will tend to float within a few volts negative of the plasma. However, LEO spacecraft with exposed conductors at differing potential levels will exhibit differential charging, with the same possibilities for damage as for GEO spacecraft. It is found[10] that the spacecraft will equilibrate at a negative potential with respect to the plasma, at roughly 90% of the most negative exposed spacecraft voltage. When all spacecraft operated at low bus voltages, e.g., the 28-V level that was standard for many years, this was not a problem. However, as spacecraft bus voltages have climbed (see Chapter 10), the arcing thresholds of common electrical conductors have been reached (e.g., copper, at around 40 V), with the attendant problems.

A variety of effects can occur. The arcing itself produces electromagnetic interference (EMI) that will generally be considered unacceptable. Such noise is not insignificant; in the case of the shuttle, the EMI environment is dominated by plasma interaction noise. Solar arrays, which depend on maintaining a specified potential difference across the array, can develop arcs between exposed

conductors or into the ambient plasma, degrading array efficiency and possibly damaging array elements or connections. Very large arrays such as on the International Space Station, which are designed to produce 160 V, may require a plasma contactor to keep all parts of the spacecraft below the arcing threshold for copper.

It would seem that using a positive spacecraft ground instead of the conventional negative return line would obviate these problems. However, almost all modern electronic subsystems are designed for positive power input and a negative ground return. LEO spacecraft designers must therefore take care to ensure that conductors carrying medium to high voltages are not exposed to the ambient plasma.

3.5.4 Magnetic Field

A LEO spacecraft spends its operational lifetime in Earth's magnetic field, and planetary spacecraft encountering Jupiter or Saturn will experience similar but stronger fields. Because the primary effect of the magnetic field is on the spacecraft attitude control system, its characteristics are discussed in Chapter 7. However, there can be other effects.

A conductive spacecraft moving in a magnetic field is a generator. For large vehicles the voltage produced can be nontrivial. For example, it has been estimated that the International Space Station may experience as much as a 20-V difference between opposite ends of the vehicle.

This effect is the basis of an interesting concept that has been proposed for generating power in low Earth orbit. A conductive cable several kilometers long would be deployed from a spacecraft and stabilized vertically in a gravity-gradient configuration (see Chapter 7). Motion in Earth's magnetic field would generate a current that could be used by the spacecraft, at the cost of some drag makeup propellant. A preliminary tether experiment was performed from the cargo bay of the space shuttle; however, mechanical problems with the deployment mechanism allowed only limited aspects of the technique to be demonstrated.

3.5.5 Weightlessness and Microgravity

It is common to assume that orbital flight provides a weightless environment for a spacecraft and its contents. To some level of approximation this is true, but as with most absolute statements, it is inexact. A variety of effects result in acceleration levels (i.e., "weight" per unit mass) between 10^{-3} and $10^{-11}g$, where $1g$ is the acceleration due to gravity at the Earth's surface, 9.81 m/s^2.

The acceleration experienced in a particular case will depend on the size of the spacecraft, its configuration, its orbital altitude if in orbit about a planet with an atmosphere, the solar cycle, and residual magnetic moment. Additionally,

the spacecraft will experience periodic impulsive disturbances resulting from attitude or translation control actuators, internal moving parts, or the activities of a human flight crew. If confined to the spacecraft interior, these disturbances may produce no net displacement of the spacecraft center of mass. However, for sensitive payloads such as optical instruments or materials-processing experiments that are fixed to the spacecraft, the result is the same.

The most obvious external sources of perturbing accelerations are environmental influences such as aerodynamic drag and solar radiation pressure, both discussed in Chapter 4. If necessary, these and other nongravitational effects can be removed, to a level of better than $10^{-11}g$, by a disturbance-compensation system to yield essentially drag-free motion. This concept is discussed in Chapter 4 and has been used with navigation satellites, where the ability to remain on a gravitationally determined (thus highly predictable) trajectory is of value.

The disturbance compensation approach referred to has inherently low bandwidth, and so cannot compensate for higher frequency disturbances, which we loosely classify as "vibration." For space microgravity research, reduction of such vibration to very low levels is crucial, and usually requires the implementation of specialized systems to achieve.[11]

A perturbing acceleration that cannot be removed is the so-called gravity-gradient force. Discussed in more detail in Chapter 7, this force results from the fact that only the spacecraft center of mass is truly in a gravitationally determined orbit. Masses on the vehicle that are closer to the center of the earth would, if in a free orbit, drift slowly ahead of those masses located farther away. Because the spacecraft is a more or less rigid structure, this does not happen; the internal elastic forces in the structure balance the orbital dynamic accelerations tending to separate masses orbiting at different altitudes.

Gravity-gradient effects are significant ($10^{-3}g$ or possibly more) over large vehicles such as the shuttle or International Space Station. For most applications this may be unimportant. However, certain materials-processing operations are particularly demanding of low-gravity, low-vibration conditions and thus may need to be conducted in free-flying modules, where they can be located near the center of mass. Higher altitude also diminishes the effect, which follows an inverse-cube force law.

Although we have so far discussed only the departures from the idealized $0g$ environment, it is nonetheless true that the most pronounced and obvious condition associated with space flight is weightlessness. As with other environmental factors, it has both positive and negative effects on space vehicle design and flight operations. The benefits of weightlessness in certain manufacturing and materials-processing applications are in fact a significant practical motivation for the development of a major space operations infrastructure. Here, however, we focus on the effects of $0g$ on the spacecraft functional design.

The 0*g* environment allows the use of relatively light spacecraft structures by comparison with earthbound designs. This is especially true where the structure is actually fabricated in orbit, or is packaged in such a way that it is not actually used or stressed until the transportation phase is complete. The International Space Station is an example of the former approach, while both the Apollo lunar module and the lunar roving vehicle are examples of the latter. A possibly awkward side effect of large, low-mass structures is that they tend to have relatively low damping and hence are susceptible to substantial structural excitation. Readers who have seen the films of the famous Tacoma Narrows Bridge disaster, the classic case in this regard, will be aware of the potential for concern. Less dramatically, attitude stabilization and control of large space vehicles are considerably complicated by structural flexibility. This is discussed in more detail in Chapter 7.

In some cases, the relatively light and fragile mechanical designs appropriate for use in space render ground testing difficult. Booms and other deployable mechanisms may not function properly, or at least the same way, in a 1*g* field if designed for 0*g* or low *g*. Again, a case in point is the Apollo lunar rover. The actual lunar rover, built for one-sixth *g*, could not be used on Earth, and the lunar flight crews trained on a stronger version. In other cases, booms and articulating platforms may need to be tested by deploying them horizontally and supporting them during deployment in Earth's gravity field.

The calibration and mechanical alignment of structures and instruments intended for use in flight can be a problem in that the structure may relax to a different position in the strain-free 0*g* environment. For this and similar reasons, spacecraft structural mass is often dictated by stiffness requirements rather than by concerns over vehicle strength. Critical instrument alignment and orientation procedures are often verified by the simple artifice of making the necessary measurements in a 1*g* field, then inverting the device and repeating the measurements. If significant differences are not observed, the 0*g* behavior is probably adequate.

Weightlessness complicates many fluid and gasdynamic processes, including thermal convection, compared with ground experience. The situation is particularly exacerbated when one is designing for human presence. Effective toilets, showers, and cooking facilities are much harder to develop for use in 0*g*. When convection is required for thermal control or for breathing air circulation, it must be provided by fans or pumps. The same is true of liquids in tanks; if convection is required to maintain thermal or chemical uniformity, it must be explicitly provided. Weightlessness is a further annoyance when liquids must be withdrawn from partially filled tanks, as when a rocket engine is ignited in orbit. Secondary propulsion systems will usually employ special tanks with pressurized bladders or wicking to ensure the presence of fuel in the combustion chamber. Larger engines are usually ignited following an ullage burn of a small thruster to force the propellant to settle in place over the intake lines to the engine.

As mentioned, a significant portion of the concern over spacecraft cleanliness during assembly is due to the desire to avoid problems from floating dust and debris once in orbit. Careful control over assembly operations is necessary to prevent dropped or forgotten bolts, washers, electronic components, tools, and other paraphernalia from causing problems in flight. Again, this may be of particular concern for manned vehicles, where an inhaled foreign object could be deadly. It is for this reason that the shuttle air circulation ports are screened; small objects tend to be drawn by air currents toward the intake screens, where they remain until removed by a crew member.

Weightlessness imposes other design constraints where manned operations are involved. Early attempts at extravehicular operations during the Gemini program of the mid-1960s showed that inordinate and unexpected effort was required to perform even simple tasks in $0g$. Astronaut Gene Cernan on his Gemini 9 flight became so exhausted merely putting on his maneuvering backpack that he was unable to test the unit. Other astronauts experienced difficulty in handling their life-support tethers and in simply shutting the spacecraft hatch upon completion of extravehicular activity (EVA).

These and other problems were in part caused by the bulkiness and limited freedom of movement possible in a spacesuit, but were to a greater extent due to the lack of body restraint normally provided by the combination of friction and the $1g$ Earth environment. With careful attention to the placement of hand and foot restraints, it proved possible to accomplish significant work during EVA without exhausting the astronaut. This was demonstrated by Edwin (Buzz) Aldrin during the flight of Gemini 12 and put into practice "for real" by the Skylab 2 crew of Conrad, Kerwin, and Weitz during the orbital repair of the Skylab workshop. Today, EVA is accepted as a risky and demanding, but still essentially routine, activity when conducted in a disciplined manner and guided by the principles that have been learned. This has been shown during a number of successful retrieval, repair, and assembly operations in the U.S. space shuttle, the Russian Mir, and the International Space Station programs.

3.5.6 Radiation

Naturally occurring radiation from numerous sources at a wide range of wavelengths and particle energies is a fixture of the space environment. The sun is a source of ultraviolet (UV) and soft x-ray radiation and, on occasion, will eject a flux of very high energy protons in what is known as a "solar flare," or more technically as a "solar proton event." The Van Allen radiation belts surrounding Earth, the solar wind, and galactic cosmic rays are all sources of energetic charged particles of differing types. The radiation environment may be a problem for many missions, primarily due to the effect of high-energy charged particles on spacecraft electronic systems, but also in regard to the degradation of paints, coatings, and various polymeric materials as a result of prolonged UV exposure.

Charged-particle effects are of basically two kinds: degradation due to total dose and malfunctions induced by so-called single-event upsets. Fundamentally different mechanisms are involved in these two failure modes.

N- or p-type metal-oxide semiconductors (NMOS or PMOS) are most resistant to radiation effects than CMOS, but require more power. Transistor–transistor logic (TTL) is even more resiliant, but likewise uses more power.

High-energy particulate radiation impacting a semiconductor device will locally alter the carefully tailored crystalline structure of the device. After a sufficient number of such events, the semiconductor is simply no longer the required type of material and ceases to function properly as an electronic device. Total dose effects can be aggravated by the intensity of the radiation; a solar flare can induce failures well below the levels normally tolerated by a given device. At lower dose rates the device will anneal to some extent and "heal" itself, a survival mechanism not available at higher rates.

The other physical effect that occurs when particulate radiation interacts with other matter is localized ionization as the incoming particle slows down and deposits energy in the material. In silicon, for example, one hole-electron pair is produced for each 3.6 eV of energy expended by the incoming particle. Thus, even a relatively low energy cosmic ray of some 10^7 eV will produce about 3×10^6 electrons, or 0.5 pC. This is a significant charge level in modern integrated circuitry and may result in a single-event upset, a state change from a stored "zero" to a "one" in a memory or logic element.

The single-event upset phenomenon has come about as a result of successful efforts to increase speed and sensitivity and reduce power requirements of electronic components by packing more semiconductor devices into a given volume. This is done essentially by increasing the precision of integrated circuit manufacture so that smaller circuits and devices may be used. For example, the mid-1980s state of the art in integrated circuit manufacturing resulted in devices with characteristic feature sizes on the order of 1 μm, while early-2000s designs approach 0.1 μm feature sizes. Ever-smaller circuits and transistor junctions imply operation at lower current and charge levels, obviously a favorable characteristic in most respects. However, beginning in the late 1970s and continuing thereafter, device "critical charge" levels reached the 0.01–1.0-pC range, where a single ionizing particle could produce enough electrons to change a "0" state to a "1," or vice versa. This phenomenon, first observed in ground-based computers, was explained in a classic work by May and Woods.[12] Its potential for harm if the change of state occurs in a critical memory location is obvious.

In practice, the damage potential of the single-event upset may exceed even that due to a serious software malfunction. If complementary metal-oxide semiconductor (CMOS) circuitry is used, the device can "latch up" into a state where it draws excessively high current, destroying itself. This is particularly unfortunate in that CMOS components require very little power for operation and are thus attractive to the spacecraft designer. Latch-up protection is possible,

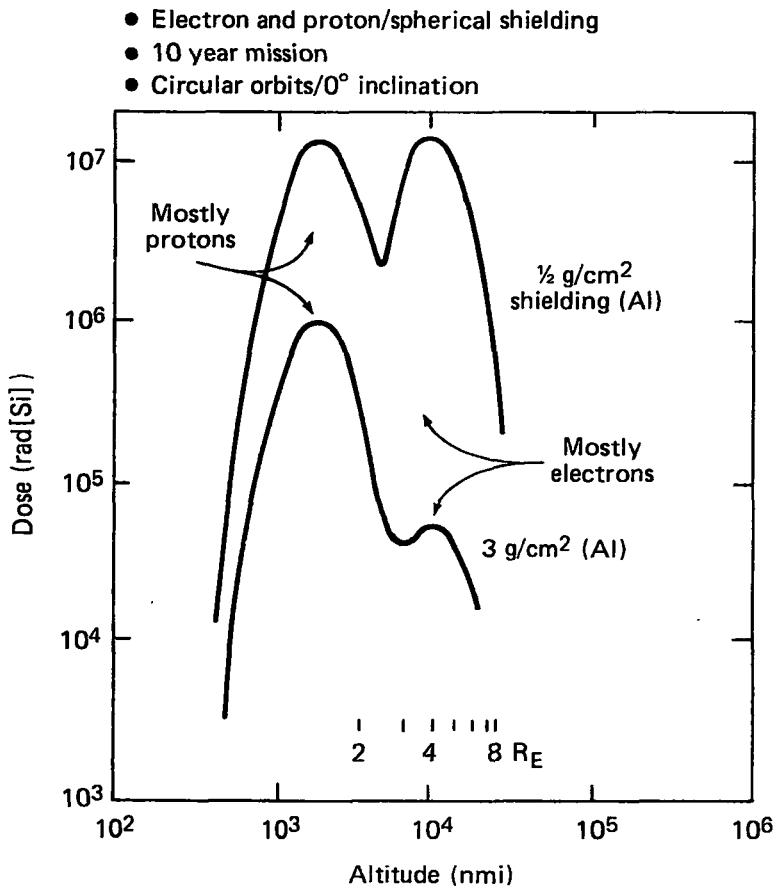


Fig. 3.24 Radiation environment for circular equatorial orbits.

either in the form of external circuitry or built into the device itself. Built-in latch-up protection is characteristic of modern CMOS devices intended for use in high-radiation environments.

The most annoying property of single-event upsets is that, given a device that is susceptible to them, they are statistically guaranteed to occur (this is true even on the ground). One can argue about the rate of such events; however, as noted earlier, even one upset at the wrong time and place could be catastrophic. Protection from total dose effects can be essentially guaranteed with known and usually reasonable amounts of shielding, in combination with careful use of radiation hardened parts. However, there is no reasonable amount of shielding that offers protection against heavy nuclei galactic cosmic rays causing single-event upsets.[13,14]

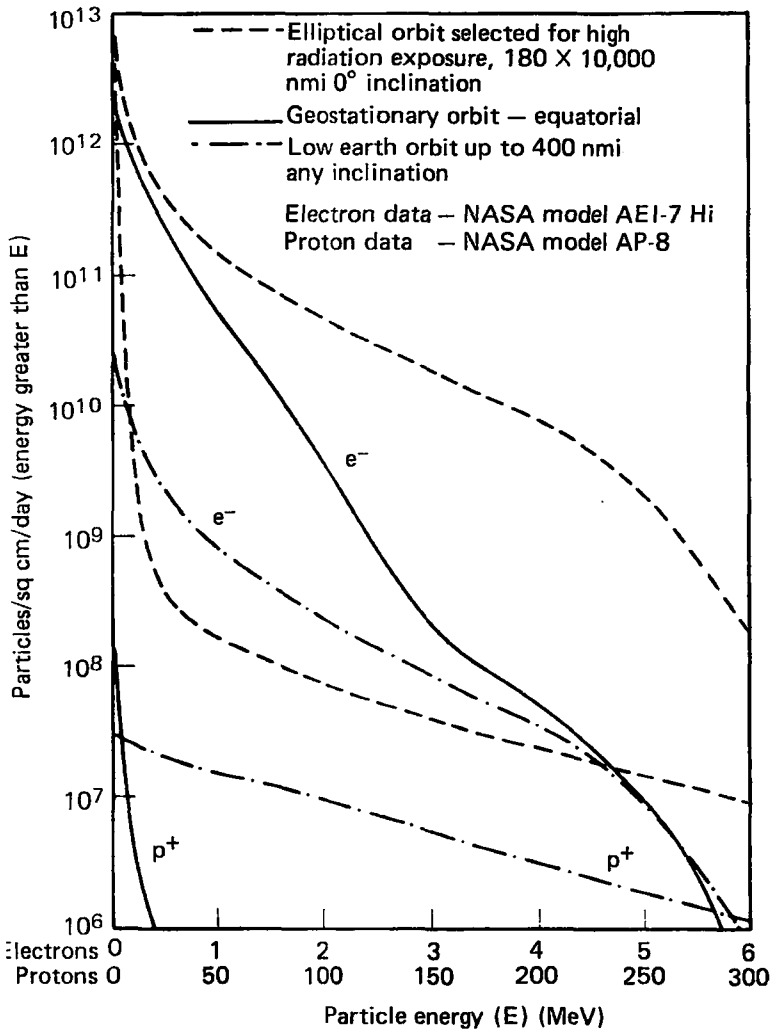


Fig. 3.25 Natural radiation environment.

Upset-resistant parts are available and should be used when analysis indicates the upset rate to be significant. (The level of significance is a debatable matter, with an error rate of 10^{-10}/day a typical standard. Note that, even with such a low rate, several upsets would be expected for a spacecraft with a mere megabit of memory and a projected 10-year lifetime.) As pointed out, shielding will not provide full relief but can be used to advantage to screen out at least the lower energy particles, thus reducing the upset rate. However, in many applications even relatively low error rates cannot be tolerated, and other measures may be required. These basically fall into the category of error detection and correction.

Table 3.4 Radiation hardness levels for semiconductor devices

Technology	Total dose, rads (Si)
CMOS (soft)	10^3–10^4
CMOS (hardened)	5×10^4–10^6
CMOC/SOS (soft)	10^3–10^4
CMOS/SOS (hardened)	$>10^5$
ECL	10^7
I^2L	10^5–4×10^6
Linear IC^2s	5×10^3–10^7
NMOS	7×10^2–7×10^3
PMOS	4×10^3–10^5
TTL/STTL	$>10^6$

Such methods include the use of independent processors with "voting" logic, and the addition of extra bits to the required computer word length to accommodate error detection and correction codes. Other approaches may also be useful in particular cases.

As mentioned, total dose effects are often more tractable because of the more predictable dependence of the dose on the orbit and the mission lifetime. For low-orbit missions, radiation is typically not a major design consideration. For this purpose, low orbit may be defined as less than about 1000-km altitude. At these altitudes, the magnetic field of the Earth deflects most of the incoming solar and galactic charged particle radiation. Because the configuration of the magnetic field does channel some of the particles toward the magnetic poles (the cause of auroral displays), spacecraft in high-inclination orbits will tend to receive somewhat greater exposure than those at lower inclinations. However, because orbital periods are still relatively short and the levels moderate, the expected dosages are not typically a problem, as long as the requirement for some level of radiation hardness is understood.

Figures 3.24 and 3.25 present the natural radiation environment vs altitude for spacecraft in Earth orbit. Figure 3.24 shows the radiation dose accumulated by electronic components over a 10-year mission in circular, equatorial orbits. Because electronic components are normally not exposed directly to space but are contained in a structure, curves are presented for two thicknesses of aluminum structure to account for the shielding effect. The extremely high peaks, of course, correspond to the Van Allen radiation belts, discussed earlier. Note that the shielding is more effective in the outer belt. This reflects the fact that the outer belt is predominantly electrons, whereas protons (heavier by a factor of 1840) dominate the inner belt. Figure 3.25 shows the radiation count vs energy level for selected Earth orbits.

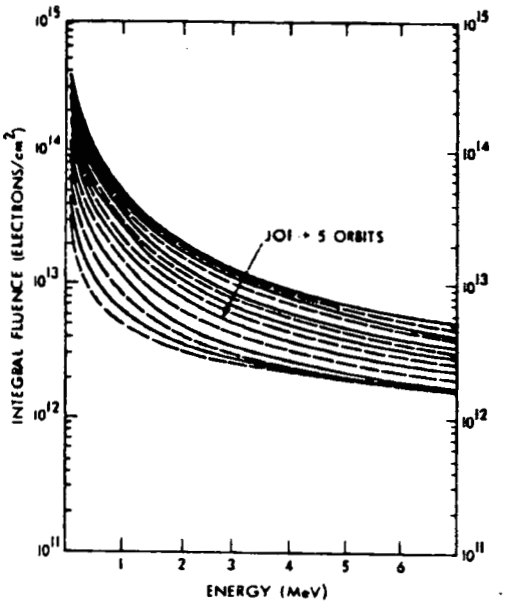
Fortunately for the communications satellite industry, geostationary orbit at about six Earth radii is well beyond the worst of the outer belt and is in a region in which the shielding due to the spacecraft structure alone is quite effective. However, it may be seen that in a 10-year mission a lightly shielded component could accumulate a total dose of 10^6 rad. To put this in perspective, Table 3.4 presents radiation resistance or "hardness" for various classes of electronic components. As this table shows, very few components can sustain this much radiation and survive. The situation becomes worse when one recognizes the need to apply a radiation design margin of the order of two in order to be certain that the components will complete the mission with unimpaired capability. For a dose of 1 Mrad and a design margin of 2, all components must be capable of 2 Mrad. At this level the choices are few, thus mandating increased shielding to guarantee an adequate suite of components for design.

The example discussed earlier is not unreasonable. Most commercial communications satellites are designed for on-orbit lifetime of 5–7 years, and an extended lifetime of 10 years is quite reasonable as a goal. In many cases these vehicles do not recoup the original investment and begin to turn a profit until several years of operation have elapsed.

If the design requirements and operating environment do require shielding beyond that provided by the material thickness needed for structural requirements, it may still be possible to avoid increasing the structural thickness. Spot shielding is very effective for protecting individual sensitive components or circuits. Such shielding may be implemented as a box containing the hardware of interest. Another approach might be to use a potting compound loaded with shielding material. (Obviously, if the shielding substance is electrically conductive, care must be exercised to prevent any detrimental effect on the circuit.) An advantage offered by the nonstructural nature of spot shielding is that it allows for the possibility of using shielding materials, such as tantalum, that are more effective than the normal structural materials. This may allow some saving in mass.

Alterations in the spacecraft configuration may also be used advantageously when certain circuits or components are particularly sensitive to the dose anticipated for a given mission and orbit. Different portions of the spacecraft will receive different dosages according to the amount of self-shielding provided by the configuration. Thus, components placed near rectangular corners may receive as much as 175% of the dose of a component placed equally near the spacecraft skin, but in the middle of a large, thick panel. When some flexibility in the placement of internal electronics packages exists, these and other properties of the configuration may be exploited.

A spacecraft in orbit above the Van Allen belts or in interplanetary space is exposed to solar-generated radiation and galactic cosmic rays. The dose levels from these sources are often negligible, although solar flares can contribute several kilorads when they occur. Galactic cosmic rays, as discussed



a) Integral electron fluences for the Galileo mission (JOI-Jupiter orbit insertion)

b) Electron dose vs aluminum shield thickness for the Galileo mission

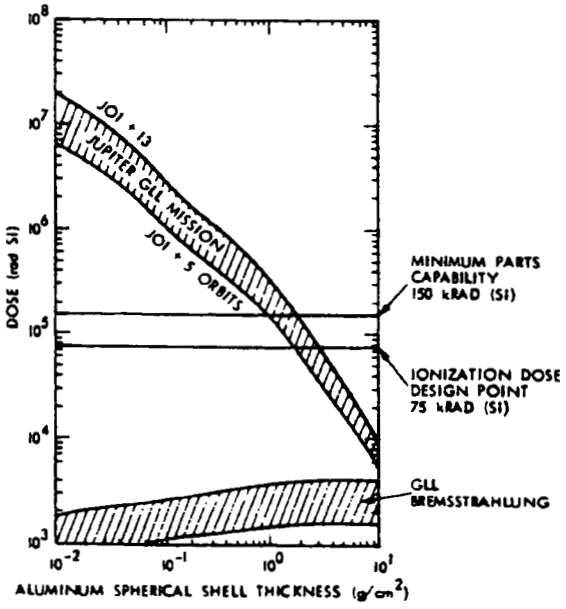


Fig. 3.26 Jupiter radiation environment.

Table 3.5 Radiation tolerance of common space materials

Material	Dose, rads (Si)
Nylon	10^5–10^6
Silver-teflon	10^6–10^7
Neoprene	10^6–10^7
Natural rubber	10^6–10^7
Mylar	10^7–10^8
Polyethylene	10^7–10^8
Sealing compounds	10^8–10^9
Silicone grease	10^8–10^9
Conductive adhesive	10^8–10^9
Kapton®	10^9–10^{10}
Carbon	10^9–10^{10}
Optical glass	5×10^8–5×10^9
Fused glass	10^9–10^{10}
Quartz	10^9–10^{10}

earlier, can produce severe single-event upset problems, because they consist of a greater proportion of high-speed, heavy nuclei against which it is impossible to shield.

Manned flight above the Van Allen belts is a case where solar flares may have a potentially catastrophic effect. The radiation belts provide highly effective shielding against such flares, and in any case a reasonably rapid return to Earth is usually possible for any such close orbit. (This assumption may need to be reexamined for the case of future space station crews.) Once outside the belts, however, the received intensity of solar flare radiation may make it impractical to provide adequate shielding against such an event. For example, although the average flare can be contained, for human physiological purposes, with 2–4 g/cm^2 of shielding, infrequent major events can require up to 40 g/cm^2, an impractical amount unless a vehicle is large enough to have an enclosed, central area to act as a "storm cellar." It is worth noting that the Apollo command module, and certainly the lunar module, did not provide enough shielding to enable crew survival in the presence of a flare of such intensity as that which occurred in August 1972, between the Apollo 16 and 17 missions.

Most of the bodies in the solar system do not have intense magnetic fields and thus have no radiation belts (by the same token, low-altitude orbits and the planetary surface are thus unprotected from solar and galactic radiation). This cannot be said of Jupiter, however. The largest of the planets has a very powerful magnetic field and intense radiation belts. Figure 3.26 indicates the intensity of the Jovian belts.

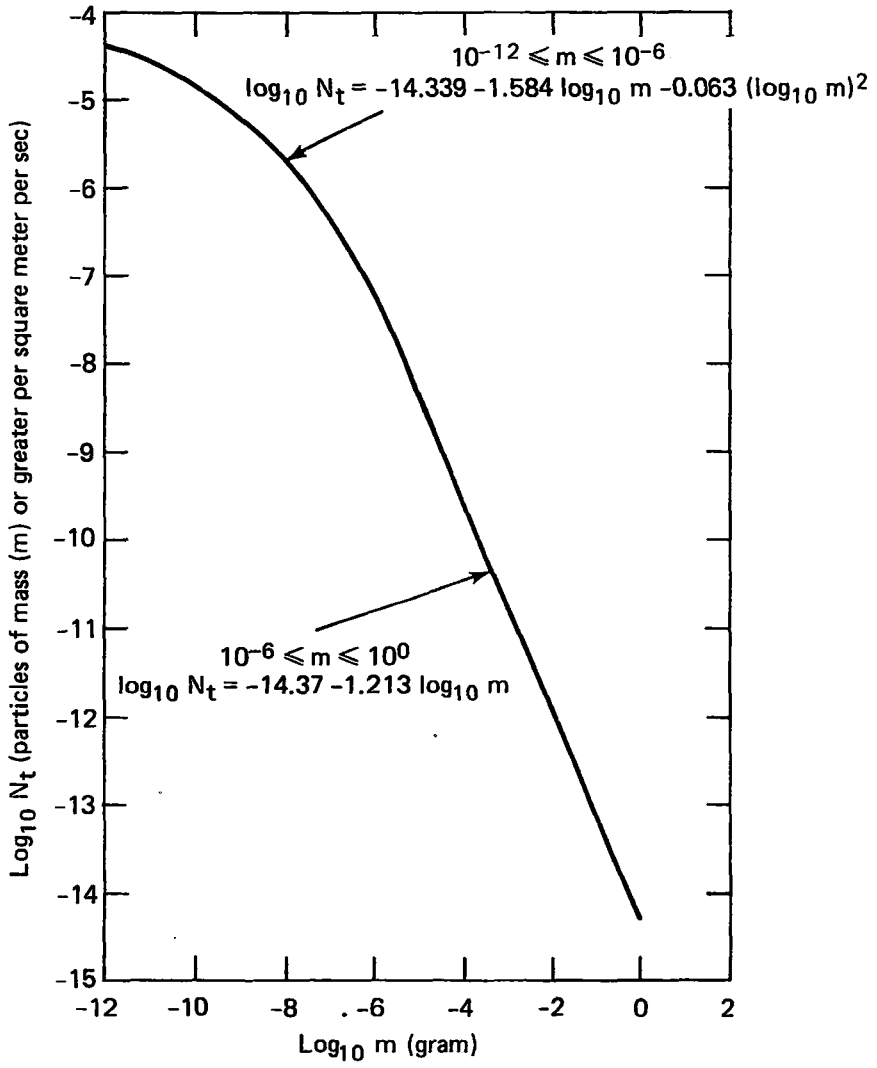


Fig. 3.27 Meteoroid flux vs mass at 1 AU.

Natural radiation sources may not be the only problem for the spacecraft designer. Obviously, military spacecraft for which survival is intended (possibly "hoped for" is the more realistic term) in the event of a nuclear exchange pose special challenges. Less pessimistically, future spacecraft employing nuclear reactors for power generation will require shielding methods not previously employed, at least on U.S. spacecraft. Even relatively low-powered radioisotope thermoelectric generators (RTG), used primarily on planetary spacecraft, can

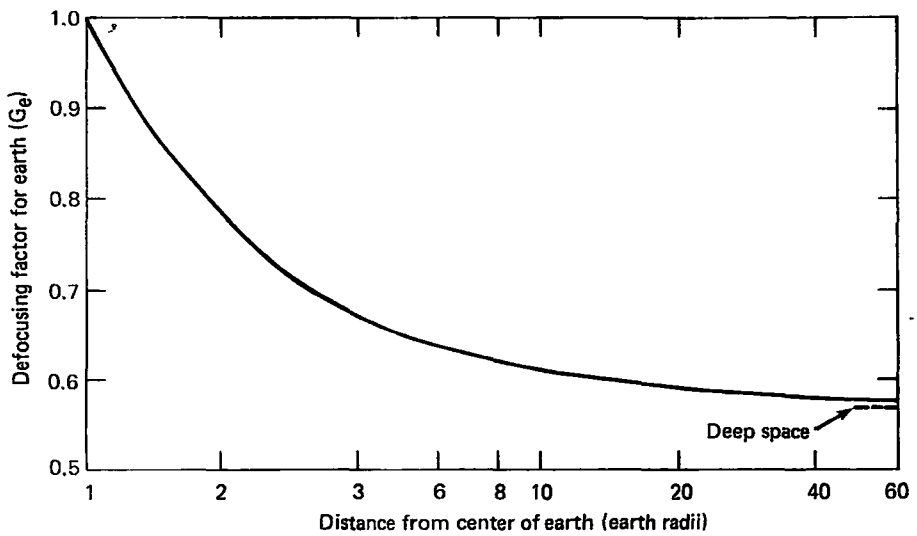


Fig. 3.28 Defocusing factor due to the Earth's gravity for an average meteoroid velocity of 20 km/s.

cause significant design problems. These issues are discussed in more detail in Chapter 10.

Finally, radiation may produce damaging effects on portions of the spacecraft other than its electronic systems. Polymers and other materials formed from organic compounds are known to be radiation sensitive. Such materials, including Teflon® and Delrin®, are not used on external surfaces in high-radiation environments such as Jupiter orbit.[15] Other materials, such as Kevlar® epoxy,

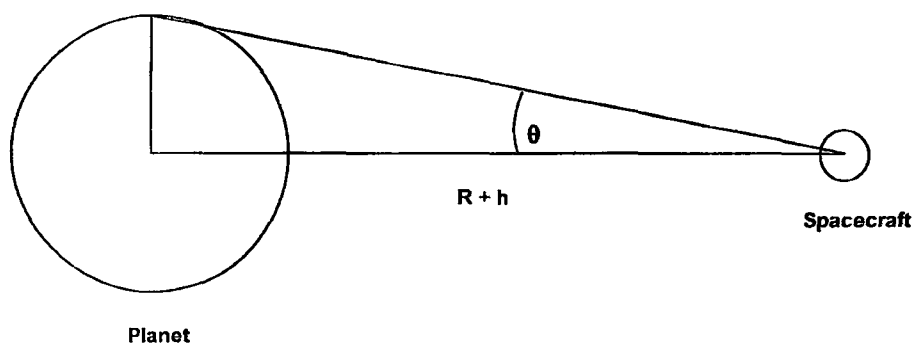


Fig. 3.29 Method for determining body shielding factor for randomly oriented spacecraft.

which may be used in structural or load-bearing members, can suffer a 50–65% reduction in shear strength after exposure to large (3000 Mrad) doses such as those that may be encountered by a permanent space station.[16] Table 3.5 provides order-of-magnitude estimates for radiation tolerance of common materials.

3.5.7 Micrometeoroids

Micrometeoroids are somewhat of a hazard to spacecraft, although substantially less than once imagined. Meteoroid collision events have occurred, but rarely. The two highly probable known cases consist of geostationary spacecraft hit by small objects, probably meteoroids. In one case, the European Space Agency's Olympus satellite was lost as it consumed propellant in an attempt to recover. A Japanese satellite sustained a hit in one solar array, with the only result being a minor loss of power generation capacity.

The standard micrometeoroid model[17] is based on data from numerous sources, included the Pegasus satellites flown in Earth orbit specifically for the purpose of obtaining micrometeoroid flux and penetration data, detectors flown on various lunar and interplanetary spacecraft, and optical and radar observation from Earth. This 1969 model still represents the best source of design information available for near-Earth space. The model approximates near-Earth micrometeoroid flux vs particle mass by

$$\log_{10} N_t > m = -14.339 - 1.584 \log_{10} m - 0.063(\log_{10} m)^2 \qquad (3.2)$$

when the particle mass m is in the range $10^{-12}g < m < 10^{-6}g$. For larger particles such that $10^{-6}g < m < 1g$, the appropriate relation is

$$\log_{10} N_t > m = -14.37 - 1.213 \log_{10} m \qquad (3.3)$$

These relationships are presented graphically in Fig. 3.27. For specific orbital altitudes, gravitational focusing and the shielding effect of the planet must be considered to derive the specific meteoroid flux environment for the orbit in question.

Because of the gravitational attraction of the Earth, more meteoroids are found at low altitudes than farther out. A correction for this focusing effect must be applied when extrapolating near-Earth meteoroid flux data to high orbits or to deep space. Assuming an average meteoroid velocity in deep space of 20 km/s, Fig. 3.28 presents a curve of the defocusing factor that may be used to compute the flux at a given altitude above Earth from the deep-space data of Fig. 3.27.

The increase in particle flux for low-altitude planetary orbits tends to be offset by the shielding factor provided by the planet. The body shielding factor ζ is defined as the ratio of shielded to unshielded flux and is given by

$$\zeta = \frac{(1 + \cos\theta)}{2} \qquad (3.4)$$

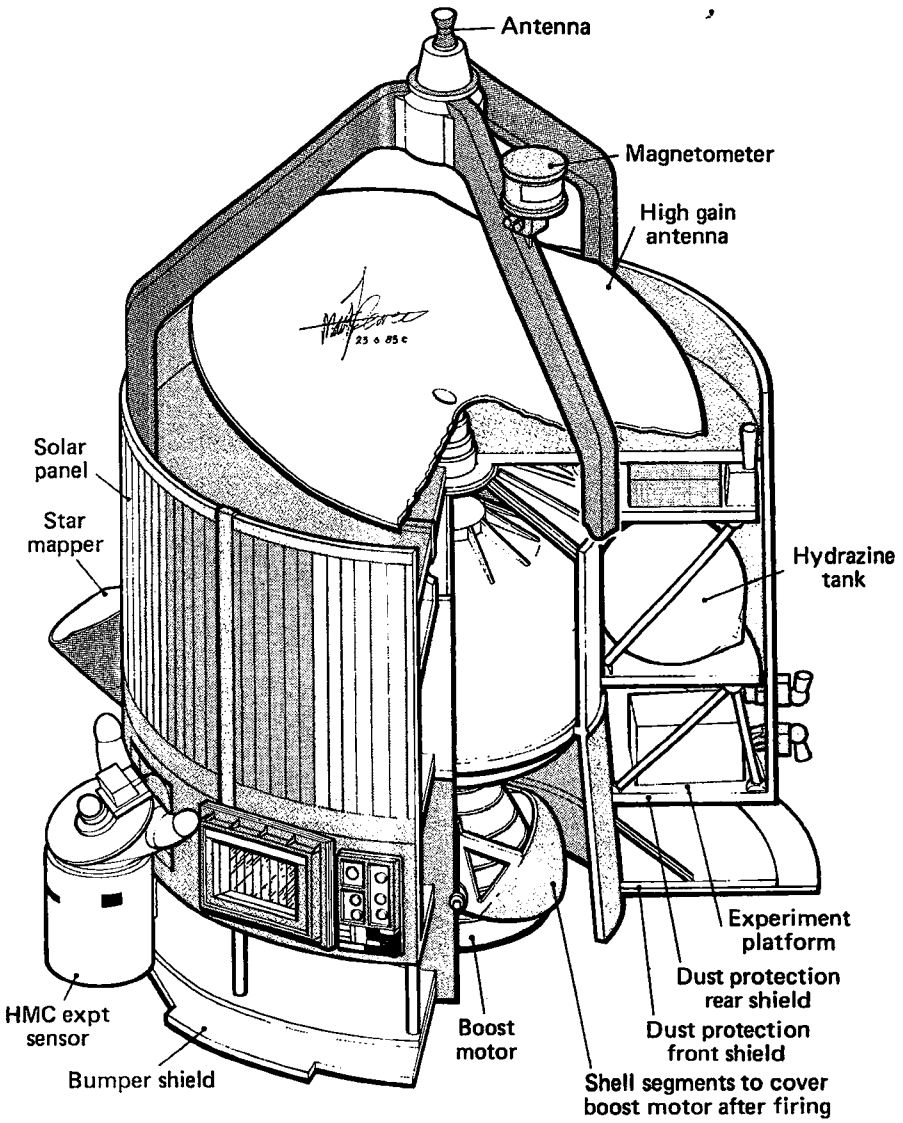


Fig. 3.30 Giotto spacecraft with Whipple meteor bumper.

where

$$\sin\theta = \frac{R}{(R+h)} \tag{3.5}$$

where R is the shielding planet radius and h the spacecraft altitude.

Figure 3.29 shows the geometry for the body shielding factor. Although particles vary considerably in density and velocity, for most purposes a density of 0.5 g/cm^3 and a velocity of 20 km/s are used as average values.

It will be seen that most micrometeoroids are extremely small. To put the threat in perspective, a rule of thumb is that a particle of 1 μg will just penetrate a 0.5-mm-thick sheet of aluminum. For most applications, the spacecraft external structure, thermal blankets, etc., provide adequate protection against particles with any significant probability of impact. For longer missions or more severe environments, additional protection may be needed, as with the Viking Orbiter propulsion system. This presented a fairly large area over a relatively long mission. More significantly, however, micrometeoroid impact on the pressurized tanks was highly undesirable, since, although penetration was extremely unlikely, the stress concentrations caused by the crater could have caused an eventual failure. The problem was dealt with by making the outer layer of the thermal blankets out of Teflon®-impregnated glass cloth.

The kinetic energy of micrometeoroids is typically so high that, upon impact, the impacting body and a similar mass of the impacted surface are vaporized. This leads to the concept of the "meteor bumper" proposed originally by Fred Whipple long before the first orbital flights. Although most spacecraft do not require protection of this magnitude, some very severe environments may dictate use of this concept. The concept involves placing a thin shield (material choice is not highly critical but is preferably metal) to intercept the incoming particle a short distance from the main structure of the pressure vessel. The thickness of the shield is dictated by the anticipated size of the particles. Ideally the shield should be just thick enough to ensure vaporization of the largest particles that have significant probability of being encountered. The spacing between the shield and the main structure is designed to allow the jet of vaporized material, which still has substantial velocity, to spread over a larger area before striking the main structure. The result of such an event is then a hole in the shield and possibly a dent or depression in the inner structure. Without the shield, a particle of sufficient mass and kinetic energy to dictate this type of protection could cause major damage. Even if it did not penetrate, the impact could result in spalling of secondary particles, still quite energetic, off the other side of the structure. Such particles could result in severe vehicle damage or crew injury.

The perceptive reader will see that this ability of an impacting particle to spall larger slower particles off the anti-impact side places a significant constraint on shield design. Any area that is made thicker than the optimum for vaporization, say, for attachment brackets, could become the source of secondary particles. These particles, being more massive than the original and possessing considerable kinetic energy, but not enough to vaporize them on impact, can be very damaging. It is clear from this brief discussion that design of such shields is an exacting task requiring both science and art. An actual flight application of this concept is the European Space Agency's Giotto probe, which flew through the dust cloud of Halley's Comet. In this instance the shield is only required on one

side of the spacecraft. Relative velocity of the dust is 60–70 km/s. Figure 3.30 shows the Giotto configuration.

Cour-Palais[18] provides a very thorough discussion of mechanisms of meteoroid damage. Although a detailed knowledge of the phenomena involved is beyond the usual scope of systems engineering, a general understanding will be useful in assessing protection that may be required for a given spacecraft mission.

3.5.8 Orbital Debris

Naturally occurring particles are not the only or, at some altitudes, the most severe impact hazard. Nearly a half-century of essentially uncontrolled space operations has produced a major hazard in low Earth orbit. As of January 2000, nearly 9000 separate space objects larger than approximately 10 cm were being tracked and catalogued by the U.S. Space Command. Cumulatively, the population of tracked objects was estimated at almost 5×10^6 kg. The number of smaller, but still very dangerous, objects is greater yet. Statistical estimates derived from ground telescope observations indicated the presence of 100,000–150,000 objects larger than 1 cm in diameter as of January 2000.[19] Impact sensors on various spacecraft have demonstrated the presence of literally billions of small particles, consisting mostly of paint flecks and aluminum oxide, in the 0.01–0.5-mm range. In all such cases, the debris level exceeds, and sometimes greatly exceeds, the natural meteoroid background.

This debris cloud has a variety of sources. Hundreds of explosions or other breakups of spacecraft or rocket stages have occurred, with no end immediately in sight. (Nine such events occurred in 1998 and again in 2001. In the latter year, one breakup occurred only 30 km from the International Space Station.) In some cases this has occurred deliberately, or at least with no effort made to prevent it. For example, early Delta second stages were left with fuel tanks in a pressurized state following spacecraft separation, resulting in several on-orbit explosions. These explosions generated a considerable amount of long-lived debris.

The situation is unlikely to improve in the near future. Approximately 2×10^6 kg of spacecraft material resides at altitudes below 2000 km, most in the form of intact vehicles having characteristic dimensions on the order of 3 m. The varying orbital planes of these objects can produce high intersection angles and the potential for high collision velocities. As Kessler and Cour-Palais[20] have shown, such collisions are a statistical certainty and can be expected to contribute to an increasingly dense debris cloud. Routine space operations such as the firing of solid rocket motors, which generate extensive aluminum-oxide particulate debris, will also continue to add to the low-orbit hazard.

In the early years of space operations, such considerations seemed unimportant, because space seemed to be "vast" and "limitless." Although these romantic descriptors are true in general, the volume occupied by moderate altitude, moderate inclination orbits around the Earth is by no means limitless, and in fact becomes somewhat congested when populated by tens of thousands of

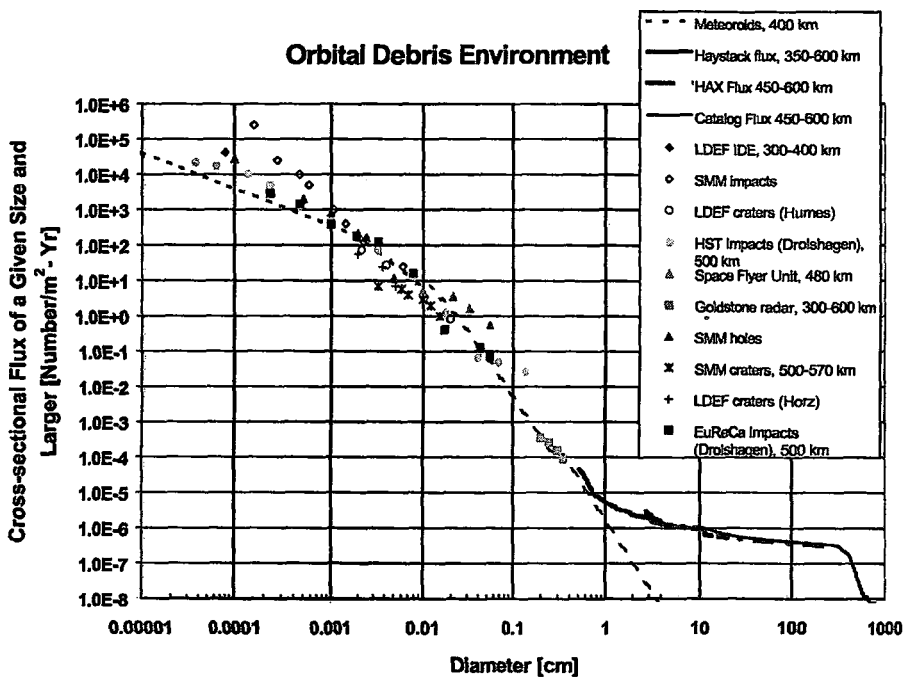


Fig. 3.31 Observed space debris environment. (Courtesy NASA Johnson Space Center, Orbital Debris Program Office.)

particles moving at 8–10 km/s. The debris density is most severe at medium altitudes. The debris flux appears to the worst in the altitude range of 600–1100 km. Below 200–300 km, atmospheric drag causes the debris orbits to decay rapidly into the atmosphere. Above 1100 km, the flux tapers off because of the increasing volume of space and because operations in these orbits have been more limited.

While geosynchronous orbit is becoming crowded, the debris problem has not reached the severity of the lower altitude environment. This is in part because the large, potentially explosive booster stages that have contributed substantially to the low-orbit debris cloud do not reach geosynchronous altitude. However, it is also true that the communications satellite community was among the first to recognize that measures to minimize orbital clutter should be routinely employed. To this end, it has become standard practice in the industry to lift outmoded or nonfunctional satellites out of the geostationary ring, with fuel for this purpose included in the satellite design budget.

Because of the high flux of particles in certain orbits, the probability of a debris strike on a spacecraft can be quite high. Worse still, there is a chance that the strike could involve a "large" particle of a few millimeters diameter. Such an

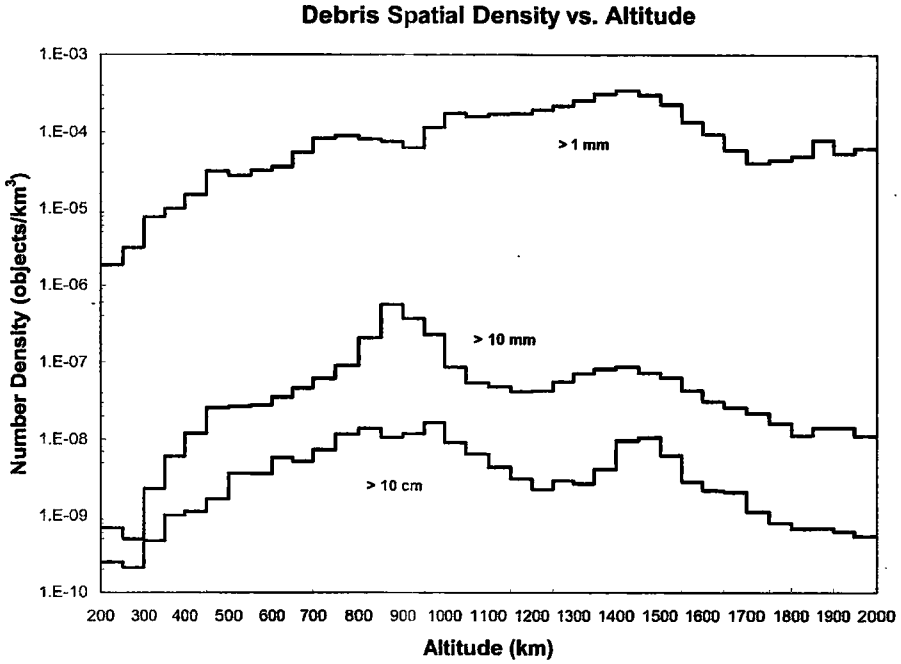


Fig. 3.32 Cumulative spatial debris density. (Courtesy NASA Johnson Space Center, Orbital Debris Program Office.)

impact could well be catastrophic. For example, NASA models of debris hazards for manned orbital operations assume fatal space suit damage from particles in the 0.3–0.5-mm range, and catastrophic shuttle damage from a 4-mm particle. Particles in the 1-mm range could cause a mission abort in some cases, such as impact with the large shuttle thermal radiators in the payload bay doors. As of early 2003, the only known accidental collision between catalogued satellites occurred in July 1996, between a fragment of an Ariane upper stage (which had exploded 10 years earlier) and the French CERISE satellite. The collision severed the spacecraft's gravity gradient attitude stabilization boom. The satellite was able to resume operations after the attitude control system software was modified. Only one new piece of catalogued debris was produced, the upper half of the gravity gradient boom.

Many other incidents of lesser significance have occurred. The first known example was on the STS-7 mission, during which the outer layer of a windshield on the space shuttle *Challenger* was cracked by what, upon postflight analysis, proved to be a fleck of paint. Many small impacts were observed in samples of thermal blanketing returned from the Solar Max spacecraft following its 1984 on-orbit repair. Most shuttle missions now return with some evidence of debris impact seen on the thermal protection system tiles. The recovery and return of the

Long Duration Exposure Facility (LDEF) in 1990, after nearly six years in low Earth orbit (initially 510 km, decaying to about 325 km by the time of its recovery), provided extensive further data on the number and size distribution of particulate debris.

These and other experiences have led to continuing efforts to update standard orbital debris models to reflect changing conditions. Figures 3.31 and 3.32 present NASA results from the ORDEM2000 model, widely regarded as the current standard.[21] ORDEM2000 describes the debris environment in low Earth orbit between 200 and 2000 km altitude. The model is intended to provide engineering solutions estimates of the orbital debris environment (spatial density, flux, etc.). ORDEM2000 incorporates considerable observational data for object sizes from 10 mm to 10 m into its database, and uses a maximum likelihood estimator to convert these observations into debris population probability distribution functions.

ORDEM2000 also performs orbital lifetime calculations based on the orbital parameters and ballistic coefficient of specified objects. The topic of orbital lifetime calculations is treated more fully in Chapter 4.

Even a perfect debris model is not of much help to the spacecraft designer having the task of protecting his vehicle from hypervelocity particle impacts. Consisting primarily of particles of spacecraft and booster structural material, the debris has a much higher density than comet-derived meteoroid particles. For particles smaller than 1 cm, the density is taken to be 2.8 g/cm^3 on average. For large particles, the particle density ρ is found to be approximately

$$\rho = \frac{2.8}{D^{0.074}} \text{ g/cm}^3 \tag{3.6}$$

where D is the average diameter in centimeters. The average relative velocity is usually assumed to be 10 km/s. The requirement to withstand such impacts is obviously very challenging.

Although local shielding of certain critical components or areas is possible, as is done, for example, on the International Space Station, completely armoring a spacecraft is not practical from a mass standpoint and in some cases may not even be possible. At this point, the most practical strategy may be to avoid high-probability orbits. As mentioned, the problem is expected to increase in severity for some years before greater awareness and increased use of various mitigation strategies begins to reverse the trend. A spate of antisatellite (ASAT) vehicle tests of the type conducted by the USSR on several occasions, and by the United States in September 1985, could greatly aggravate the problem.

As an illustrative example of the effects of hypervelocity impact on orbital clutter, the September 1985 test, in which the P78-1 SOLWIND satellite was destroyed by an air-launched ASAT rocket, was estimated to have created approximately 10^6 fragments between 1 mm and 1 cm in diameter. This event alone thus produced, at an altitude sufficient to yield long-lived orbits, a debris environment in excess of the natural micrometeoroid background.

It is possible to conduct such tests in a more suitable fashion. In September 1986, the U.S. Department of Defense (DoD) Strategic Defense Initiative Organization conducted a boost-phase intercept test involving a collision between an experimental interceptor and a Delta 3920 second-stage rocket in powered flight. A direct hit at a relative velocity of approximately 3 km/s ensued. The chosen intercept altitude of about 220 km, which then became the highest possible perigee point of any collision debris, ensured that the residue from the collision remained in orbit for at most a few months.[22]

Numerous national and international efforts have been undertaken to increase the level of awareness of the space debris problem and to develop and promulgate mitigation strategies for the future.[19] Among the recommended approaches are (1) cessation of deliberate spacecraft breakups producing debris in long-lived orbits; (2) minimization of mission-related debris generation; (3) passivation of spacecraft and rocket bodies remaining in orbit after mission completion, i.e., expending residual propellants, discharging batteries, venting tanks, etc.; (4) selection of transfer orbit parameters to ensure reentry of spent transfer stages within 25 years; and (5) boosting separated apogee kick motors, other transfer stages used for geostationary spacecraft circularization, and defunct geostationary satellites to an altitude at least 300 km above the geostationary ring.

Mitigation measures such as these obviously place an additional burden on space vehicle design and operation not present in earlier years. For this reason, while international cooperation over debris mitigation has increased in recent years, full compliance continues to elude the space community. Space operations and plans must increasingly take into account strategies for avoiding, or coping with, orbital debris. For example, in the five years between 1989 and 1994, the space shuttle received four collision-avoidance warnings and acted upon three of them.[23] It has been estimated that the International Space Station can expect to receive about 10 collision avoidance warnings per year of sufficient concern that an avoidance maneuver could be required.[24]

3.5.9 Thermal Environment

Space flight presents both a varied and extreme thermal environment to the space vehicle designer. Spacecraft thermal control is an important topic in its own right, and will be treated in more detail in Chapter 9. However, it is appropriate in this section to survey some of the environmental conditions that must be addressed in the thermal design.

The space vacuum environment essentially allows only one means of energy transport to and from the spacecraft, that of radiative heat transfer. The overall energy balance is therefore completely defined by the solar and planetary heat input, internally generated heat, and the radiative energy transfer properties that are determined by the spacecraft configuration and

materials. The source and sink temperatures (from the sun with a characteristic blackbody temperature of 5780 K and dark space at 3 K, respectively) for radiative transfer are extreme.

Under these conditions, extremes of both temperature and temperature gradient are common. Thermally isolated portions of an Earth-orbiting spacecraft can experience temperature variations from roughly 200 K during darkness to about 350 K in direct sunlight. One has only to consider such everyday experiences as the difficulty of starting a car in very cold weather, with battery and lubrication problems, or very hot weather (which may cause carburetor vapor lock) to appreciate that most machinery functions best at approximately the same temperatures as do humans.

If appropriate internal conduction paths are not provided, temperature differences between the sunlit and dark sides of a spacecraft can be almost as severe as the extremes cited earlier. This results in the possibility of damage or misalignment due to differential expansion in the material. Space vehicles are sometimes rolled slowly about an axis normal to the sun line to minimize this effect. When this is impractical, and other means to minimize thermal gradients are not suitable, special materials having a very low coefficient of thermal expansion (such as Invar® or graphite-epoxy) may need to be employed.

The fatiguing effect on materials of repeated thermal cycling between such extremes is also a problem and has resulted in many spacecraft component and subsystem failures. One relevant example was that of LANDSAT-D, where the solar cell harness connections were made overly tight and pulled loose after repeated thermal cycling, ultimately disabling the spacecraft.

Thermal system design in vacuum is further complicated by the need for special care in ensuring good contact between bolted or riveted joints. Good thermal conductivity under such conditions is difficult to obtain, hard to quantify, and inconsistent in its properties. Use of a special thermal contact grease or pad is required to obtain consistently good conductive heat transfer.

The lack of free convection has been mentioned in connection with the 0*g* environment; it is, of course, equally impossible in vacuum. Heat transfer internal to a spacecraft is therefore by means of conduction and radiation, in contrast to ground applications in which major energy transport is typically due to both free and forced convection. This results in the need for careful equipment design to ensure appropriate conduction paths away from all internal hot spots and detailed analytical verification of the intended design. This may sometimes be avoided by hermetically sealing an individual package or, as is common for Russian spacecraft, by sealing the whole vehicle. The disadvantage here is obviously that a single leak can result in loss of the mission.

The atmospheric entry thermal environment is the most severe normally encountered by a spacecraft, and vehicles designed for this purpose employ a host of special features to achieve the required protection. This is discussed in more detail in Chapters 6 and 9.

3.5.10 Planetary Environments

Interplanetary spacecraft designers face environmental problems that may be unique even in what is, after all, a rather specialized field. Flyby spacecraft, such as Pioneers 10 and 11 and Voyagers 1 and 2, may encounter radiation environments greatly exceeding those in near-Earth space. The Mariner 10 mission to Mercury required the capability to cope with a factor of 10 increase in solar heating compared to Earth orbit, whereas Voyager 2 at Neptune received only about 0.25% of the illumination at Earth. In addition to these considerations, planetary landers face possible hazards such as sulfuric acid in the Venusian atmosphere and finely ground windblown dust on Mars. Spacecraft intended for operation on the lunar surface must be designed to withstand alternating hot and cold soaks of two weeks duration and a range of 200 K.

It is well beyond the scope of this text to discuss in detail the environment of each extraterrestrial body, even where appropriate data exist. Spacecraft system designers involved in missions where such data are required must familiarize themselves with what is known. Because the desired body of knowledge is often lacking, ample safety margins must usually be included in all design calculations.

References

[1]Bedingfield, K. L., Leach, R. D., and Alexander, M. B., "Spacecraft System Failures and Anomalies Attributed to the Natural Space Environment," NASA Ref. Pub. 1390, Aug. 1996.

[2]Engels, R. C., Craig, R. R., and Harcrow, H. W., "A Survey of Payload Integration Methods," *Journal of Spacecraft and Rockets*, Vol. 21, 1984, pp. 417–424.

[3]U.S. Standard Atmosphere, National Oceanic and Atmospheric Administration, NOAA S/T 76-1562, U.S. Government Printing Office, Washington, DC, 1976.

[4]Slobin, S. D., "Atmospheric and Environmental Effects," DSMS Telecommunications Link Design Handbook, Doc. 810-005, Rev. E, Jet Propulsion Lab., Pasadena, CA, Jan. 2001.

[5]Hale, N. W., Lamotte, N. O., and Garner, T. W., "Operational Experience with Hypersonic Flight of the Space Shuttle," AIAA Paper 2002-5259, Oct. 2002.

[6]Campbell, W. A., Marriott, R. S., and Park, J. J., "Outgassing Data for Selecting Spacecraft Materials," NASA Ref. Pub. 1124, 1990.

[7]Baumjohann, W., and Treumann, R. A., *Basic Space Plasma Physics*, Imperial College Press, London, 1986.

[8]Frezet, M., Daly, E. J., Granger, J. P., and Hamelin, J., "Assessment of Electrostatic Charging of Satellites in the Geostationary Environment," *ESA Journal*, Vol. 13, 1989, p. 91.

[9]Leach, R. D., and Alexander, M. B., "Failures and Anomalies Attributed to Spacecraft Charging," NASA Ref. Pub. 1375, Aug. 1995.

[10]Ferguson, D. C., "Interactions Between Spacecraft and Their Environments," National Aeronautics and Space Administration, Glenn Research Center, Cleveland, OH, 1993; also Proceedings, AIAA Aerospace Sciences Meeting, Reno, NV, January 1993.

[11]Whorton, M. S., Eldridge, J. T., Ferebee, R. C., Lassiter, J. O., and Redmon, J. W., Jr., "Damping Mechanisms for Microgravity Vibration Isolation," NASA TM-1998-206953, Jan. 1998.

[12]May, T. C., and Woods, M. H., "Alpha-Particle-Induced Soft Errors in Dynamic Memories," *IEEE Transactions on Electron Devices*, Vol. ED-26, No. 1, 1979, pp. 2–9.

[13]Cunningham, S. S., "Cosmic Rays, Single Event Upsets and Things That Go Bump in the Night," *Proceedings of the AAS Rocky Mountain Guidance and Control Conference*, Paper AAS-84-05, 1984.

[14]Cunningham, S. S., Banasiak, J. A., and Von Flowtow, C. S., "Living with Things That Go Bump in the Night," *Proceedings of the AAS Rocky Mountain Guidance and Control Conference*, Paper AAS-85-056, 1985.

[15]Bouquet, F. L., and Koprowski, K. F., "Radiation Effects on Spacecraft Materials for Jupiter and Near-Earth Orbiters," *IEEE Transactions on Nuclear Science*, Vol. NS-29, No. 6, 1982, pp. 1629–1632.

[16]Frisch, B., "Composites and the Hard Knocks of Space," *Astronautics and Aeronautics*, Vol. ?, pp. 33–38.

[17]"Meteoroid Environment Model-1969," NASA SP-8013.

[18]Cour-Palais, B., "Hypervelocity Impact in Metals, Glass, and Composites," *International Journal of Impact Engineering*, Vol. 5, 1987, pp. 221–237.

[19]International Academy of Astronautics, "Position Paper on Orbital Debris," Paris, France, Nov. 2001.

[20]Kessler, D. J., and Cour-Palais, B. G., "Collision Frequency of Artificial Satellites: The Creation of a Debris Belt," *Journal of Geophysical Research*, Vol. 83, No. A6, 1978, pp. -.

[21]Liou, J., Matney, M. J., Anz-Meador, P. D., Kessler, D. J., Jansen, M., and Theall, J. R., "The New NASA Orbital Debris Engineering Model ORDEM 2000," NASA/TP-2002-210780, May 2002.

[22]Tan, A., and Zhang, D., "Analysis and Interpretation of the Delta 180 Collision Experiment in Space," *Journal of the Astronautical Sciences*, Vol. 49, Oct.–Dec. 2001, pp. 585–599.

[23]National Research Council, *Orbital Debris: A Technical Assessment*, National Academy Press, Washington, DC, 1995.

[24]National Research Council, "Protecting the Space Station from Meteoroids and Orbital Debris," Washington, DC, Jan. 1997.

Problems

3.1 At its atmospheric entry interface of $h = 122$ km altitude, the space shuttle air-relative velocity is about 7.9 km/s. The angle of attack at that time is typically about 40 deg, and the planform area is 367 m^2. What is the drag acceleration (see Chapter 4) at the entry interface under standard atmosphere conditions?

3.2 On a particular day at Cape Canaveral, the air pressure and temperature are measured and found to be 101,000 N/m^2 and 298 K, respectively. What is the density, and what is the density altitude? Assume $R_{gas} = 287.05$ J/kg · K for air (see Chapter 6).

3.3 What is the expected number of impacts on the space shuttle during a two-week mission at 400 km circular orbit altitude and 51.6 deg inclination by debris particles greater than 4 mm in size? For particles greater than 1 cm? Assume the planform area of 367 m^2 to be the relevant target area.

3.4 How much flight time should the space shuttle fleet expect to accumulate before experiencing an impact by a micrometeoroid of 0.1 g or greater mass, assuming an average orbit of 400-km altitude?

3.5 The Global Positioning System (GPS) satellite constellation operates in 63-deg inclination orbits at approximately 11,000 n mile altitude. Give a rough estimate of the expected total radiation dose from protons and electrons for these satellites assuming a nominal ten-year mission.

3.6 Consider a plot of acceleration spectral density (ASD) such as in Fig. 3.17. Note that this is a graph of $\log_{10}$ ASD vs $\log 10_{10} f_{Hz}$. Assuming simple harmonic oscillation, what is the slope (dB/octave) of a curve of constant displacement on such a plot?

3.7 Calculate the average acceleration loading due to random vibration, g_{rms}, for the curve of Fig. 3.17.

4
Astrodynamics

4.1 Introduction

Astrodynamics is the study of the motion of man-made objects in space subject to both natural and artificially induced forces. It is the latter factor that lends a design element to astrodynamics that is lacking in its parent science, celestial mechanics. The function of the astrodynamicist is to synthesize trajectories that, within the limits imposed by physics and launch vehicle performance, accomplish desired mission goals. Experience gained since the dawn of the space age in 1957, together with the tremendous growth in the speed and sophistication of computer analyses, have allowed the implementation of mission designs not foreseen by early pioneers in astronautics. This trend was discussed briefly in Chapter 2 and shows every sign of continuing. The use of halo orbits[1] for the International Sun-Earth Explorer (ISEE) and Wilkinson Microwave Anisotropy Probe missions, the development of space colony concepts using the Earth-moon Lagrangian points,[2] together with the analysis by Heppenheimer[3] of "achromatic" trajectories to reach these points from the moon, and the extensive modern use of gravity-assist maneuvers[4] for interplanetary missions are but a few examples.

Astrodynamics, through its links to classical astronomy, has its roots in the very origins of the scientific revolution. The mathematical elegance of the field exceeds that found in any other area of astronautics. Problems posed in celestial mechanics have been a spur to the development of both pure and applied mathematics since Newton's development of the calculus (which he used, among other things, to derive Kepler's laws of planetary motion and to show that a spherically symmetric body acts gravitationally as if its mass were concentrated at a point at its center). Hoyle[5] comments on this point and notes that it has not been entirely beneficial; the precomputer emphasis on analytical solutions led to the development of many involved methods and "tricks" useful in the solution of celestial mechanics problems. Many of these methods persist to the present as an established part of mathematics education, despite having little relevance in an era of computational sophistication.

Because of the basic simplicity of the phenomena involved, it is possible in astrodynamics to make measurements and predictions to a level of accuracy exceeded in few fields. For example, it is not unusual to measure the position of an interplanetary spacecraft (relative to its tracking stations) to an accuracy of

less than a kilometer. Planets such as Mars, Venus, and Jupiter, which have been orbited by spacecraft, can now be located to within several tens of meters out of hundreds of millions of kilometers. Such precision is attained only at the price of extensive data processing and considerable care in modeling the solar system environment.

We shall not engage in detailed consideration of the methods by which the highest possible degree of precision is attained. Although it is true that the most accurate methods of orbit prediction and determination are desirable in the actual execution of a mission, such accuracy is rarely needed at the levels of mission definition appropriate to spacecraft design. What is required is familiarity with basic orbit dynamics and an understanding of when and why more complex calculations are in order. We take the view that the spacecraft system engineer requires a level of competence in astrodynamics approximately defined by the range of methods suitable for solution via pocket calculator. Any analysis absolutely requiring a computer for its completion is in general the province of specialists.

Of course, this threshold is a moving target. "Pocket" calculators available as this goes to press (2003) offer a level of capability substantially exceeding that of desktop personal computers of the mid-1980s, allowing many formerly prohibitive computations to be completed with ease, even during preliminary design.

4.2 Fundamentals of Orbital Mechanics

4.2.1 Two-Body Motion

The basis of astrodynamics is Newton's law of universal gravitation:

$$F = -\frac{GMm}{r^2} \tag{4.1}$$

which yields the force between two point masses M and m separated by a distance r and directed along the vector $\boldsymbol{r}$ (see Fig. 4.1) between them. G is the universal gravitation constant, a fundamental (and very difficult to determine) constant of nature. It is a sophomore-level exercise in physics[6] to show that M and m may be extended spherically symmetric bodies without affecting the validity of Eq. (4.1).

A necessary and sufficient condition that F be a conservative (nondissipative, path-independent) force is that it be derivable as the negative (by convention) gradient of a scalar potential. This is the case for the gravitational force law, as seen by differentiating the potential function (per unit mass of m) given by

$$U = -\frac{GM}{r} \tag{4.2}$$

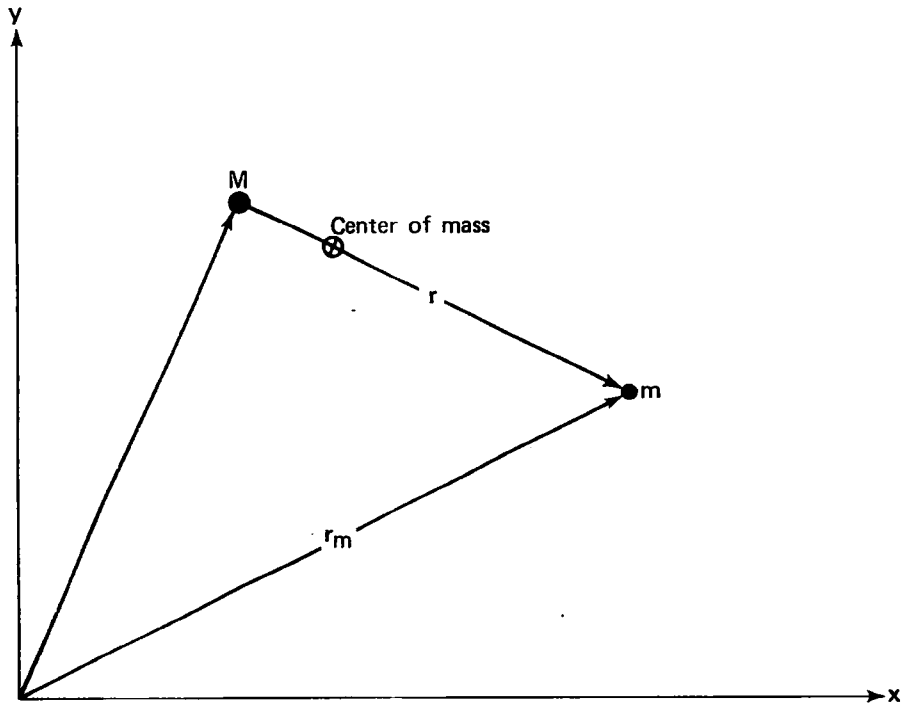


Fig. 4.1 Two-body motion in inertial space.

U has dimensions of energy per unit mass and is thus the potential energy of mass m due to its position relative to mass M. Note that the singularity at the origin is excluded from consideration, because M and m cannot be coincident, and that the potential energy is taken as zero at infinity, an arbitrary choice, because Eq. (4.2) could include an additive constant with no change in the force law. With the negative-gradient sign convention indicated earlier, the potential energy is always negative.

Using Newton's second law,

$$F = ma \tag{4.3}$$

in an inertial frame, and equating the mutual force of each body on the other leads to the familiar inverse square law equation of motion

$$\frac{d^2\boldsymbol{r}}{dr^2} + \frac{G(M+m)}{r^3}\boldsymbol{r} = 0 \tag{4.4}$$

where r is defined as shown in Fig. 4.1. Several key results may be obtained[7] for a universe consisting of only the two masses M and m:

1) The center of mass of the two-body system is unaccelerated and thus may serve as the origin of an inertial reference frame.

2) The angular momentum of the system is constant; as a result, the motion is in a plane normal to the angular momentum vector.

3) The masses M and m follow paths that are conic sections with their center of mass as one focus; thus, the possible orbits are a circle, an ellipse, a parabola, or a hyperbola.

We note that the two-body motion described by Eq. (4.4) is mathematically identical to the motion of a particle of reduced mass:

$$m_r = \frac{Mm}{M+m} \tag{4.5}$$

subject to a radially directed force field of magnitude GMm/r^2. The two-body and central-force formulations are thus equivalent, which leads to the practice of writing Eq. (4.4) as

$$\frac{d^2\boldsymbol{r}}{dt^2} + \frac{\mu}{r^3}\boldsymbol{r} = 0 \tag{4.6}$$

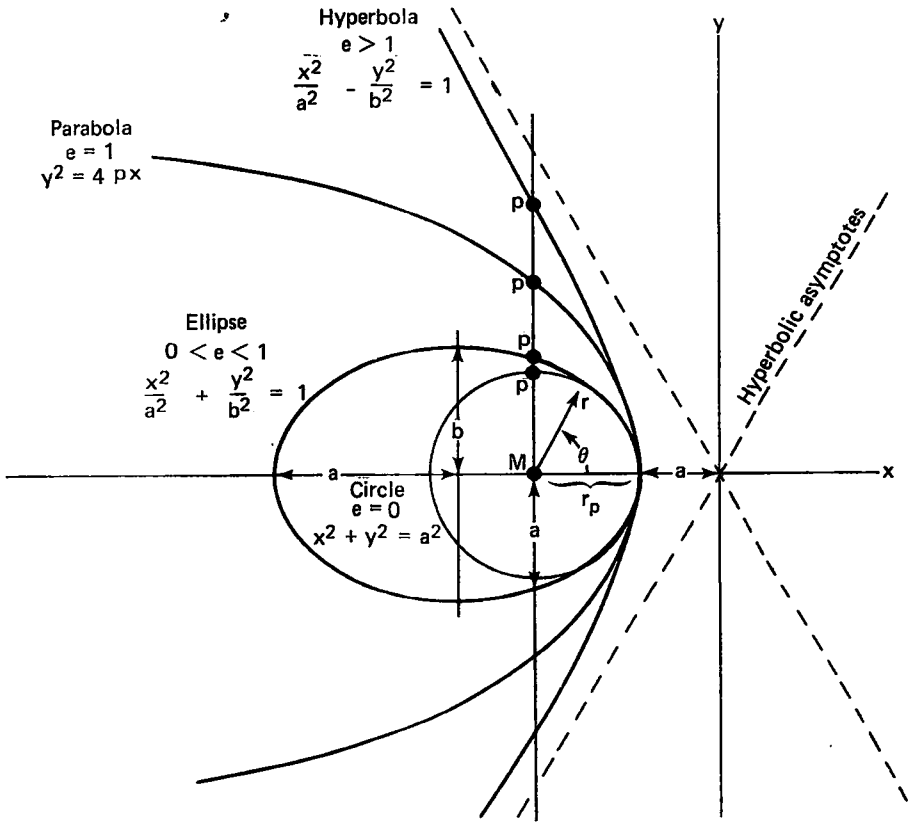
where $\mu = G(M+m)$. In nearly all cases of interest in astrodynamics, $m \ll M$, which leads to $m_r \simeq m$, $\mu \simeq GM$, and a blurring of the physical distinction between two-body and central-force motion. The system center of mass is then in fact the center of mass of the primary body M. For example, a satellite in Earth orbit has no measurable effect on the motion of the Earth, which appears as the generator of the central force. This approximation is still quite valid in the description of planetary motion, although careful measurements can detect the motion of the sun about the mutual center, or barycenter, of the solar system. Interestingly enough, the Earth–moon pair provides one of the few examples in the solar system where the barycenter of the two masses is sufficiently displaced from the center of the "primary" to be readily observable.

The formulation of Eq. (4.6) is especially convenient in that μ is determinable to high accuracy through observation of planetary or spacecraft trajectories, whereas G is itself extremely difficult to measure accurately. As an aside, recent theoretical and experimental work[8] suggests that G may not be a constant, but decreases gradually over cosmologically significant time scales.

We now proceed to quantify the results cited earlier. Figure 4.2 depicts the possible orbits, together with the parameters that define their geometric properties. Note that the different conic sections are distinguished by a single parameter, the eccentricity e, which is related to the parameters a and b or a and p as shown in Fig. 4.2. It is also clear from Fig. 4.2 that a polar coordinate representation provides the most natural description of conic orbits, as a single equation,

$$r = \frac{p}{1 + e\cos\theta} \tag{4.7}$$

accounts for all possible orbits. The angle θ is the true anomaly (often ν in the classical celestial mechanics literature) measured from periapsis, the point of



Circle	Ellipse	Parabola	Hyperbola
$a = b > 0$	$a > b > 0$	$a = \infty$	$a < 0$
$e = 0$	$e^2 = 1 - \frac{b^2}{a^2}$	$e = 1$	$e^2 = 1 + \frac{b^2}{a^2}$
$r = \frac{p}{1 + e\cos\theta} = \frac{r_p(1+e)}{1 + e\cos\theta} = \frac{a(1-e^2)}{1 + e\cos\theta}$			

e = Eccentricity
a = Semi-major or semi-transverse axis
b = Semi-minor or semi-conjugate axis
p = Parameter or semi-latus rectum

Fig. 4.2 Conic section parameters.

closest approach of M and m, as shown in Fig. 4.2. The parameter, or semilatus rectum p, is given by

$$p = a(1 - e^2) \tag{4.8}$$

It may be useful to combine Eqs. (4.7) and (4.8) to yield

$$r = \frac{r_p(1 + e)}{1 + e\cos\theta} \tag{4.9}$$

where

$$r_p = a(1 - e) \tag{4.10}$$

is the periapsis radius, obtained at $\theta = 0$ in Eq. (4.7). Elliptic orbits also have a well-defined maximum or apoapsis radius at $\theta = \pi$ given by

$$r_a = a(1 + e) \tag{4.11}$$

Combining Eq. (4.10) and (4.11) yields the following useful relationships for elliptic orbits only:

$$a = \frac{r_a + r_p}{2} \tag{4.12}$$

and

$$e = \frac{r_a - r_p}{r_a + r_p} \tag{4.13}$$

It remains only to specify the relationships between the geometric parameters a, e, and p, and the physical variables energy and angular momentum. The solution of Eq. (4.6) establishes the required connection; this solution is given in a variety of texts[9] and will not be repeated here. We summarize the results in the following.

The total energy is simply the sum of kinetic and potential energy for each mass. Because the two-body center of mass is unaccelerated and we are assuming that $m \ll M$, the energy of the larger body is negligible; thus, the orbital energy is due to body m alone and is

$$E_t = T + U = \frac{V^2}{2} - \frac{\mu}{r} \tag{4.14}$$

where T is the kinetic energy per unit mass. In polar coordinates, with velocity components V_r and V_θ given by r and $r\,\mathrm{d}\theta/\mathrm{d}t$, respectively,

$$E_t = \frac{r^2 + (r\,\mathrm{d}\theta/\mathrm{d}t)^2}{2} - \frac{\mu}{r} \tag{4.15}$$

which is constant due to the previously discussed conservative property of the force law.

The polar coordinate frame in which V_r and V_θ are defined is referred to as the perifocal system. The Z axis of this system is perpendicular to the orbit plane with the positive direction defined such that the body m orbits counterclockwise about Z when viewed from the $+Z$ direction. The origin of coordinates lies at the barycenter of the system, and the X axis is positive in the direction of periapsis. The Y axis is chosen to form a conventional right-handed set. As discussed earlier, this axis frame is inertially fixed; however, it should not be confused with other inertial frames to be discussed in Sec. 4.2.7 (Coordinate Frames).

One of the more elegant features of the solution for central-force motion is the result that

$$E_t = -\frac{\mu}{2a} \tag{4.16}$$

i.e., the specific energy of the orbit (energy per unit mass of the satellite) depends only on its semimajor (or semitransverse) axis. From Eqs. (4.14) and (4.16), we obtain

$$V^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right) \tag{4.17}$$

which is known as the vis-viva or energy equation.

The orbital angular momentum per unit mass of body m is

$$\boldsymbol{h} = \boldsymbol{r} \times \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \boldsymbol{r} \times \boldsymbol{V} \tag{4.18}$$

with magnitude given in terms of polar velocity components by

$$h = rV_\theta = \frac{r^2\,\mathrm{d}\theta}{\mathrm{d}t} \tag{4.19}$$

and is constant for the orbit, as previously discussed. This is a consequence of the radially directed force law; a force normal to the radius vector is required for a torque to exist, and in the absence of such a torque, angular momentum must be conserved. From the solution of Eq. (4.6), it is found that

$$h^2 = \mu p \tag{4.20}$$

Thus, the orbital angular momentum depends only on the parameter, or semilatus rectum, p. It is also readily shown that the angular rate of the radius vector from the focus to the body m is

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{h}{r^2} = \frac{h(1 + e\cos\theta)^2}{p^2} \tag{4.21}$$

Equations (4.16) and (4.20) may be combined with the geometric result (4.8) to yield the eccentricity in terms of the orbital energy and angular momentum,

$$e^2 = 1 + 2E_t\left(\frac{h}{\mu}\right)^2 \tag{4.22}$$

This completes the summary of results from two-body theory that are applicable to all possible orbits. In subsequent sections, we consider specialized aspects of motion in particular orbits.

4.2.2 Circular and Escape Velocity

From Eq. (4.17), and noting that for a circular orbit $r = a$, we find that the required velocity at radius r,

$$V_{cir} = \sqrt{\frac{\mu}{r}} \tag{4.23}$$

where V_{cir} is circular velocity. If $E_t = 0$, we have the condition for a parabolic orbit, which is the minimum-energy escape orbit. From Eq. (4.14),

$$V_{esc} = \sqrt{\frac{2\mu}{r}} = \sqrt{2}V_{cir} \tag{4.24}$$

where V_{esc} is escape velocity. Of course, circular velocity can have no radial component, whereas escape velocity may be in any direction not intersecting the central body.

Circular and parabolic orbits are interesting limiting cases corresponding to particular values of eccentricity. Such exact values cannot be expected in practice; thus, in reality all orbits are either elliptic or hyperbolic, with $E_t < 0$ or $E_t > 0$. Nonetheless, circular or parabolic orbits may be used as reference trajectories for the actual motion, which is seen as a perturbation of the reference orbit. We will consider this topic in more detail later; for the present, we examine the features of motion in elliptic and hyperbolic orbits.

4.2.3 Motion in Elliptic Orbits

Figure 4.3 defines the parameters of interest in elliptic orbit motion. The conic section results given earlier are sufficient to describe the size and shape of the orbit, but do not provide the position of body m as a function of time. Because it is awkward to attempt a direct solution of Eq. (4.21) to yield θ (and hence r) as a function of time, the auxiliary variable E, the eccentric anomaly, is introduced. The transformation between true and eccentric anomaly is

$$\tan\left(\frac{\theta}{2}\right) = \left(\frac{1+e}{1-e}\right)^{1/2} \tan\left(\frac{E}{2}\right) \tag{4.25}$$

It is found[10] that E obeys the transcendental equation

$$f(E) = E - e\sin E - n(t - t_p) = 0 \tag{4.26}$$

where

n = mean motion = $\sqrt{\mu/a^3}$
t_p = time of periapsis passage

which is known as Kepler's equation. The mean motion n is the average orbital rate, or the orbital rate for a circular orbit having the same semimajor axis as the

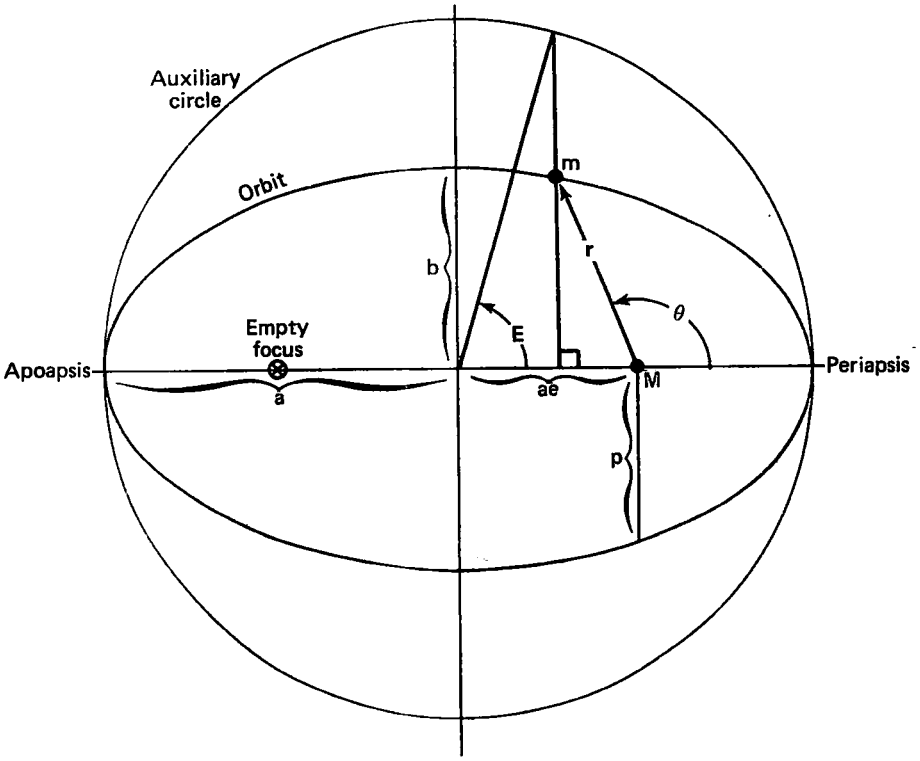


Fig. 4.3 Elliptic orbit parameters.

given elliptic orbit. The mean anomaly

$$M \equiv n(t - t_p) \tag{4.27}$$

is thus an average orbital angular position and has no physical significance unless the orbit is circular, in which case $n \equiv \dot{\theta} \equiv d\theta/dt$ exactly, and $E = \theta = M$ at all times.

When E is obtained, it is often desired to know r directly, without the inconvenience of computing θ and solving the orbit equation. In such a case, the result

$$r = a(1 - e\cos E) \tag{4.28}$$

is useful. The radial velocity in the orbit plane is

$$rV_r = na^2 e \sin E = e(\mu a)^{1/2} \sin E \tag{4.29}$$

The tangential velocity V_θ is found from

$$V_\theta = \frac{r\,d\theta}{dt} = \frac{h}{r} \tag{4.30}$$

If $E = \theta = 2\pi$, then $(t - t_p) = \tau$, the orbital period. Equation (4.26) then gives

$$\tau = 2\pi\sqrt{\frac{a^3}{\mu}} \tag{4.31}$$

which is Kepler's third law.

The question of extracting E as a function of time in an efficient manner is of some interest, especially prior to the modern era with its surfeit of computational capability. A numerical approach is required because no closed-form solution for $E(M)$ exists. At the same time, Eq. (4.26) is not a particularly difficult specimen; existence and uniqueness of a solution are easy to show.[9] Any common root-finding method such as Newton's method or the modified false position method[11] will serve. All such methods are based on solving the equation in the "easy" direction (i.e., guessing E, computing M, and comparing with the known value) and employing a more or less sophisticated procedure to choose updated estimates for E. The question of choosing starting values for E to speed convergence to the solution has received considerable attention.[12] However, the availability of programmable calculators, including some with built-in root finders, renders this question somewhat less important than in the past, at least for the types of applications stressed in this book.

When the orbit is nearly circular, $e \simeq 0$, and approximate solutions of adequate accuracy are available that yield θ directly in terms of mean anomaly. By expanding in powers of e, Eqs. (4.25) and (4.26) can be reduced to the result[10]

$$\theta \simeq M + 2e\sin M + \left(\frac{5e^2}{4}\right)\sin 2M + \cdots \tag{4.32}$$

In many cases of interest for orbital operations, the orbits will be nearly circular, and Eq. (4.32) can be used to advantage. For example, an Earth orbit of 200×1000 km, quite lopsided by parking orbit standards, has an eccentricity of 0.0573, which implies that, in using Eq. (4.32), terms of order 2×10^{-4} are being neglected. For many purposes, such an error is unimportant.

When the orbit is nearly parabolic, numerical difficulties are encountered in the use of Kepler's equation and its associated auxiliary relations. This can be seen from consideration of Eq. (4.25) for $e \simeq 1$, where there is considerable loss of numerical accuracy in relating eccentric to true anomaly. The difficulty is also seen in the use of Kepler's equation near periapsis, where E and $e\sin E$ will be almost equal for near-parabolic orbits. Battin[13] and others have developed universal formulas that avoid the difficulties in time-of-flight computations for Keplerian orbits. However, in spacecraft design the problems of numerical inaccuracy for nearly parabolic orbits are more theoretical than practical and will not concern us here.

4.2.4 Motion in Hyperbolic Orbits

The study of hyperbolic orbits is accorded substantially more attention in astrodynamics than it traditionally receives in celestial mechanics. In celestial mechanics, only comets pursue escape orbits, and these are generally almost parabolic; hence, orbit prediction and determination methods tend to center around perturbations to parabolic trajectories. In contrast, all interplanetary missions follow hyperbolic orbits, both for Earth departure and upon arrival at possible target planets. Also, study of the gravity-assist maneuvers mentioned earlier requires detailed analysis of hyperbolic trajectories.

Figure 4.4 shows the parameters of interest for hyperbolic orbits. Because a hyperbolic orbit of mass m and body M is a one-time event (possibly terminated by a direct atmospheric entry or a propulsive or atmospheric braking maneuver to effect orbital capture), the encounter is often referred to as hyperbolic passage. Although described by the same basic conic equation as for an elliptic orbit, hyperbolic passage presents some significant features not found with closed orbits.

In this section, we consider only the encounter between m and M, i.e., the two-body problem; hence, motion of M is ignored. Thus, if m is a spacecraft and M a planetary flyby target, then the separate motion of both m and M in solar orbit is neglected. This is equivalent to regarding the influence of M as dominating the encounter and ignoring that of the sun. This is, of course, the same approximation we have used in the preceding discussions, and, when m and M are relatively close, it is not a major source of error in the analysis of interest here. For example, the gravitational influence of the sun on a spacecraft in low Earth orbit will not usually be of significance in preliminary mission design and analysis.

Hyperbolic passage is fundamentally different. Although the actual encounter can indeed be modeled as a two-body phenomenon to the same fidelity as before, the complete passage must usually be examined in the context of the larger reference frame in which it takes place. This external frame provides the "infinity conditions" and orientation for the hyperbolic passage, and it is in this frame that gravity-assisted velocity changes must be analyzed. For the present, however, we consider only the actual two-body encounter. The results will be useful within a so-called sphere of influence (actually not a sphere and not sharply defined) about the target body M. Determination of spheres of influence will be addressed in a subsequent section.

When m is "infinitely" distant from M, the orbit equation may be solved with $r = \infty$ to yield the true anomaly of the asymptotes. From Eq. (4.7),

$$\theta_a = \cos^{-1}\left(\frac{-1}{e}\right) \tag{4.33}$$

and due to the even symmetry of the cosine function, asymptotes at $\pm\theta_a$ are obtained, as required for a full hyperbola. Since $r > 0$, values of true anomaly for

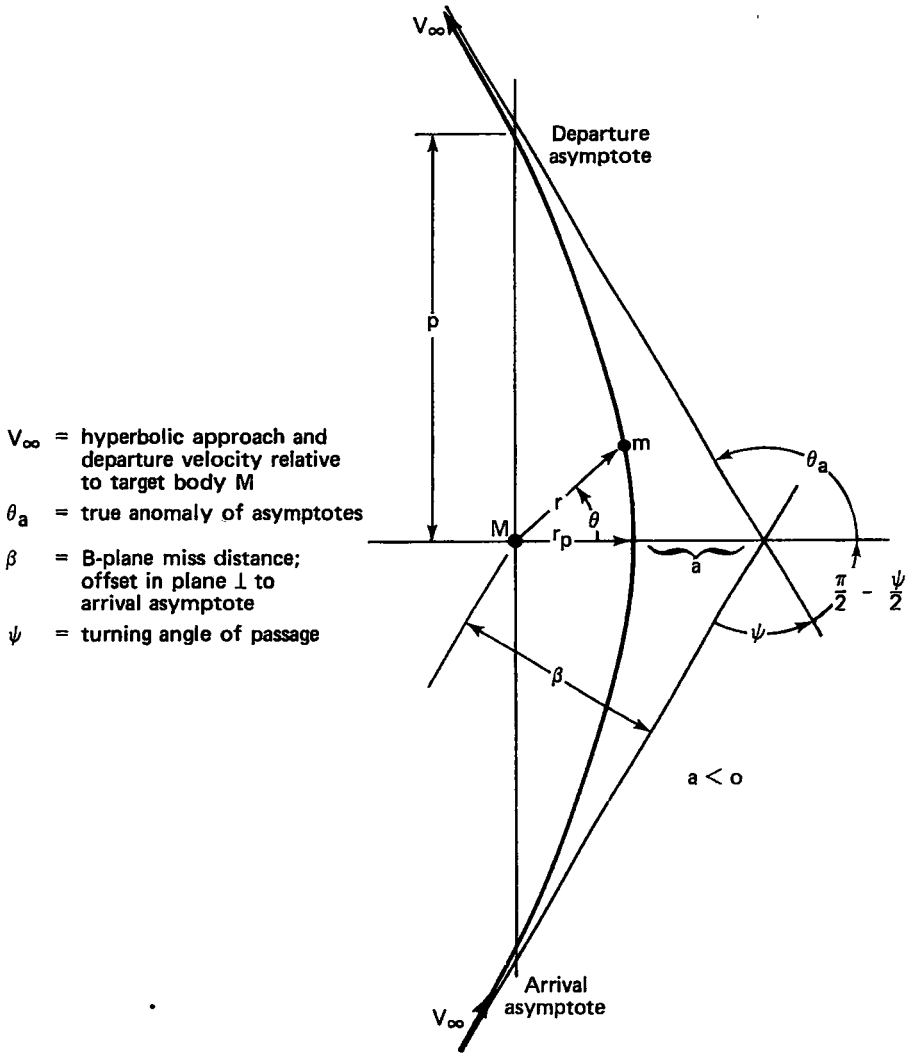


Fig. 4.4 Hyperbolic orbit geometry.

finite r are restricted to the range $[-\theta_a, \theta_a]$, and so the orbit is concave toward the focus occupied by M.

Knowledge of θ_a serves to orient the orbit in the external frame discussed earlier. The hyperbolic arrival or departure velocity V_∞ is the vector difference, in the external frame, between the velocity of m and that of M. The condition that it must lie along an asymptote determines the orientation of the hyperbolic passage with respect to the external frame. This topic will be considered in additional detail in a subsequent section.

The magnitude of the hyperbolic velocity V_∞ is found from the vis-viva equation with $r = \infty$ to be

$$V_\infty^2 = -\frac{\mu}{a} = 2E_t \tag{4.34}$$

where it is noted that the semimajor axis a is negative for hyperbolas. Since V_∞ is usually known for the passage from the infinity conditions, in practice one generally uses Eq. (4.34) to solve for a. Conservation of energy in the two-body frame requires V_∞ to be the same on both the arrival and departure asymptotes. However, the vector velocity V_∞ is altered by the encounter due to the change in its direction. This alteration of V_∞ is fundamental to hyperbolic passage and is the basis of gravity-assist maneuvers.

The change in direction of V_∞ is denoted by Ψ, the turning angle of the passage. If the motion of m were unperturbed by M, the departure asymptote would have a true anomaly of $-\theta_a + \pi$, whereas due to the influence of M the departure is in fact at a true anomaly:

$$\theta_a = -\theta_a + \pi + \Psi \tag{4.35}$$

Hence,

$$\frac{\Psi}{2} = \theta_a - \frac{\pi}{2} = \sin^{-1}\left(\frac{1}{e}\right) \tag{4.36}$$

where the second equality follows from Eq. (4.33). The velocity change seen in the external frame due to the turning angle of passage is, as seen from Fig. 4.5,

$$\Delta V = 2V_\infty \sin\frac{\Psi}{2} = \frac{2V_\infty}{e} \tag{4.37}$$

The eccentricity may be found from Eq. (4.10) with the semimajor axis a given by Eq. (4.34), yielding

$$e = 1 + \frac{V_\infty^2 r_p}{\mu} \tag{4.38}$$

This result is useful in the calculation of a hyperbolic departure from an initial parking orbit, or when, as is often the case for interplanetary exploration missions, the periapsis radius at a target plane is specified. However, Eq. (4.38) is inappropriate for use in pre-encounter trajectory correction maneuvers, which are conventionally referred to as the so-called B-plane, the plane normal to the arrival asymptote. Pre-encounter trajectory corrections will generally be applied at an effectively "infinite" distance from the target body and will thus, almost by definition, alter the placement and magnitude of V_∞ relative to M. The B-plane is therefore a convenient reference frame for such maneuvers.

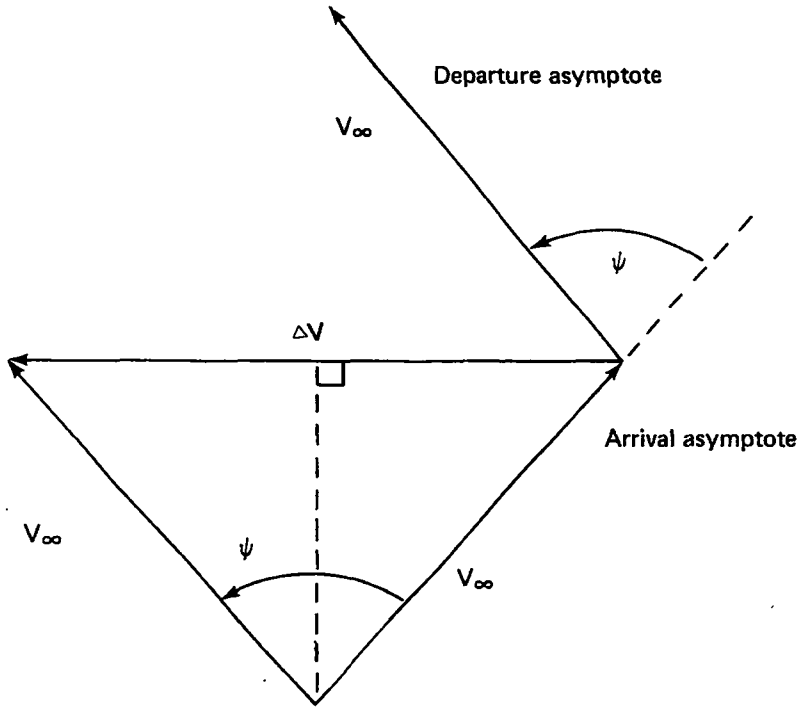


Fig. 4.5 Velocity vector change during hyperbolic passage.

The orbital angular momentum is easily evaluated in terms of the B-plane miss distance β, yielding

$$h = \beta V_\infty \tag{4.39}$$

which follows readily from the basic vector definition of Eq. (4.18) applied at infinity in rectangular coordinates. Using this result plus Eq. (4.34) in Eq. (4.22), we find

$$e^2 = 1 + \left(\frac{\beta V_\infty^2}{\mu}\right)^2 \tag{4.40}$$

From Eqs. (4.38) and (4.40), the periapsis radius is given directly in terms of the approach parameters β and V_∞ as

$$\frac{r_p}{\beta} = -\left(\frac{\mu}{\beta V_\infty^2}\right) + \sqrt{1 + \left(\frac{\mu}{\beta V_\infty^2}\right)^2} \tag{4.41}$$

while the B-plane offset required to obtain a desired periapsis is

$$\beta = r_p\left[1 + \frac{2\mu}{r_p V_\infty^2}\right]^{1/2} \tag{4.42}$$

Equations (4.33–4.42), together with the basic conic section results given in Sec. 4.2.1 (Two-Body Motion), suffice to describe the spatial properties of hyperbolic passage. It remains to discuss the evolution of the orbit in time. As with elliptic orbits, the motion is most easily described via a Kepler equation,

$$f(F) = e\sinh F - F - n(t - t_p) = 0 \tag{4.43}$$

where

n = mean motion = $[\mu/(-a)^3]^{1/2}$
t_p = time of periapsis passage

and again it is recalled that $a < 0$ for hyperbolic orbits. The hyperbolic anomaly F, as with the eccentric anomaly E, is an auxiliary variable defined in relation to a reference geometric figure, in this case an equilateral hyperbola tangent to the actual orbit at periapsis.[14] The details are not of particular interest here, because the analysis has even less physical significance than was the case for the eccentric anomaly. The transformation between true and hyperbolic anomaly is given by

$$\tan\left(\frac{\theta}{2}\right) = \left[\frac{e+1}{e-1}\right]^{1/2}\tanh\left(\frac{F}{2}\right) \tag{4.44}$$

Analogously to Eqs. (4.28) and (4.29), it is found that

$$r = a(1 - e\cosh F) \tag{4.45}$$

and

$$rV_r = na^2 e\sinh F = e(-a\mu)^{1/2}\sinh F \tag{4.46}$$

with the tangential velocity again given by Eq. (4.30).

4.2.5 Motion in Parabolic Orbits

Parabolic orbits may be viewed as a limiting case of either elliptic or hyperbolic orbits as eccentricity approaches unity. This results in some mathematical awkwardness, as seen from Eq. (4.8),

$$a = \lim_{e\to 1}\left[\frac{p}{1-e^2}\right] = \infty \tag{4.47}$$

The semimajor axis is thus undefined for parabolic orbits. The result is of somewhat limited concern, however, and serves mainly to indicate the desirability of using Eq. (4.9), from which the semimajor axis has been

eliminated, for parabolic orbits. Thus,

$$r = \frac{2r_p}{1 + \cos\theta} \tag{4.48}$$

and by comparison with Eq. (4.8), it is seen that

$$p = 2r_p \tag{4.49}$$

It may be shown that the motion in time is given by

$$\frac{D^3}{6} + D = M = n(t - t_p) \tag{4.50}$$

The parabolic anomaly D is an auxiliary variable defined as

$$D = \sqrt{2}\tan\frac{\theta}{2} \tag{4.51}$$

and the mean motion is

$$n = \sqrt{\frac{\mu}{r_p^3}} \tag{4.52}$$

As with hyperbolic orbits, n has no particular physical significance. In terms of parabolic anomaly,

$$r = r_p\left(1 + \frac{D^2}{2}\right) \tag{4.53}$$

while

$$rV_r = (\mu r_p)^{1/2} D = n r_p^2 D \tag{4.54}$$

As always, the tangential velocity V_θ is given by Eq. (4.30). It should be noted that the exact definition of D varies considerably in the literature, as does the form of Kepler's equation [Eq. (4.50)]. Care should be taken in using analytical results from different sources for parabolic orbits.

4.2.6 Keplerian Orbital Elements

Orbital motion subject to Newtonian laws of motion and gravitational force results in a description of the trajectory in terms of second-order ordinary differential equations, as exemplified by Eq. (4.6). Six independent constants are thus required to determine a unique solution for an orbit; in conventional analysis, these could be the initial conditions consisting of the position and velocity vectors $\boldsymbol{r}$ and $\boldsymbol{V}$ at some specified initial time t_0, often taken as zero for convenience. In fact, however, any six independent constants will serve, with the physical nature of the problem usually dictating the choice.

In classical celestial mechanics, position and velocity information are never directly attainable. Only the angular coordinates (right ascension and declination) of objects on the celestial sphere are directly observable. Classical orbit determination is essentially the process of specifying orbital position and velocity given a time history of angular coordinate measurements. Direct measurement of $\boldsymbol{r}$ and $\boldsymbol{V}$ (or their relatively simple calculation from given data) is possible in astrodynamics, where ground tracking stations and/or onboard guidance systems may, for example, supply position and velocity vector estimates on a nearly continuous basis.

However, even when $\boldsymbol{r}$ and $\boldsymbol{V}$ are obtained, information in this form is of mathematical utility only, because it conveys no physical "feel" for or geometric "picture" of the orbit. It is thus customary to describe the orbit in terms of six other quantities plus an epoch, a time t_0 at which they apply. These quantities, chosen to provide a more direct representation of the motion, are the Keplerian orbital elements, defined graphically in Fig. 4.6 and listed in the following:

a or p = semimajor axis or semilatus rectum
e = eccentricity
i = inclination of orbit plane relative to a defined reference plane
Ω = longitude or right ascension of ascending node, measured in the reference plane from a defined reference meridian
ω = argument of periapsis, measured counterclockwise from the ascending node in the orbit plane
θ_0, M_0, or t_p = true or mean anomaly at epoch, or time of relevant periapsis passage

As seen, there is no completely standardized set of elements in use, a circumstance in part due to physical necessity. For example, the semimajor axis a is undefined for parabolic orbits and observationally meaningless for hyperbolic orbits, requiring use of a more fundamental quantity, the semilatus rectum p, or occasionally the angular momentum h. Nonetheless, the geometric significance and convenience of the semimajor axis for circular and elliptic orbits will not be denied, and it is always employed when physically meaningful.

Other problems occur as well. For nearly circular orbits, ω is ill defined, as is Ω for orbits with near-zero inclination. In such cases, various convenient alternate procedures are used to establish a well-defined set of orbital elements. For example, when the orbit is nominally circular, $\omega \equiv 0$ is often adopted by convention. When $i \simeq 0$, accurate specification of Ω and hence ω is difficult, and the parameter Π, the so-called longitude of periapsis, is often used. Here, $\Pi = \Omega + \omega$, with Ω measured in the X–Y plane and ω measured in the orbital plane. In such cases Π may be accurately known even though Ω and ω are each poorly specified.

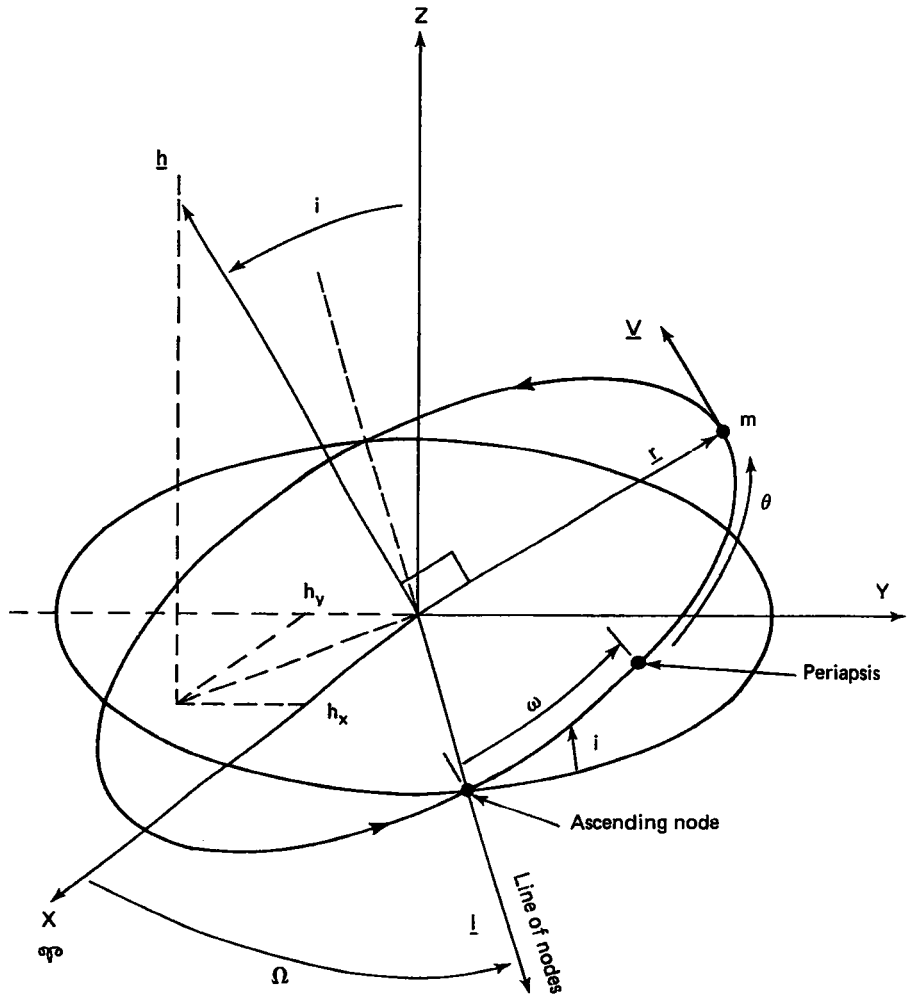


Fig. 4.6 Orbital elements.

It may be seen that a (or p) and e together specify the size and shape or, equivalently, the energy and angular momentum, of the orbit, and i, Ω, and ω provide the three independent quantities necessary to describe the orientation of the orbit with respect to some external inertial reference frame. The final required parameter serves to specify the position of body m in its orbit at a particular time. As indicated, the true or mean anomaly at epoch, or the time of an appropriate periapsis passage, may also be used. For spacecraft orbits, conditions at injection into orbit or following a midcourse maneuver may also be employed, as may the true or mean anomaly at some particular time other than the epoch.

In a simple two-body universe, the orbital elements are constant, and their specification determines the motion for all time. In the real world, additional

influences or perturbations are always present and result in a departure from purely Keplerian motion; hence, the orbital elements are not constant. Perturbing influences can be of both gravitational and nongravitational origin and may include aerodynamic drag, solar radiation, and solar wind; the presence of a third body; a nonspherically symmetric mass distribution in an attracting body; or, in certain very special cases, relativistic effects. One or more of these effects will be important in all detailed analyses, such as are necessary for the actual execution of a mission, and in many cases for preliminary analysis and mission design as well. Indeed, it is common practice in mission design to make use of certain special perturbations to achieve desired orbital coverage, as discussed briefly in Chapter 2 and in Sec. 4.3.3 (Aspherical Mass Distribution).

It commonly happens that a spacecraft or planetary trajectory is predominantly Keplerian, but that there exist perturbations that are significant at some level of mission design or analysis. When approximate analyses of such cases are carried out, it may be found that the perturbing influences alter various elements or combinations of elements in a periodic manner, or in a secular fashion with a time constant that is small compared to the orbit period. Such analyses may be used to provide relatively simple corrections to a given set of orbital elements describing the average motion, or to a set of elements accurately defined at some epoch.

The result of procedures such as those just described is a description of the orbit in terms of a set of osculating elements, which are time varying and describe a Keplerian orbit that is instantaneously tangent to the true trajectory. In this way increased accuracy can be obtained while still retaining a description of the motion in terms of orbital elements, i.e., without resorting to a numerical solution. These topics will be considered in more detail in a later section.

4.2.7 Coordinate Frames

Within the fixed orbit plane of two-body Keplerian motion, the coordinate system of choice is the polar coordinate system depicted in Fig. 4.2. The position of the object is given by the coordinates (r, θ), with the true anomaly θ measured from periapsis. This system of perifocal or orbit plane coordinates is both natural and sufficient as long as the orientation of the orbit in space need not be considered.

However, we have seen in the preceding discussion that a particular orbit is defined through its elements in relation to a known inertial reference frame, as shown in Fig. 4.6. There are two major inertial reference frames of interest.

The X axis in all cases of interest in the solar system is defined in the direction of the vernal equinox, the position of the sun against the fixed stars on (presently) March 21, the first day of spring. More precisely, the X axis is the line from the center of the Earth to the center of the sun when the sun crosses the Earth's equatorial plane from the southern to northern hemisphere.

For orbits about or observations from Earth (or, when appropriate, any other planet), the natural reference plane is the planetary equator. The orbital inclination i is measured with respect to the equatorial plane, and the longitude of the ascending node Ω is measured in this plane. The positive Z direction is taken normal to the reference plane in the northerly direction (i.e., approximately toward the North Star, Polaris, for Earth). The Y axis is taken to form a right-handed set and thus lies in the direction of the winter solstice, the position of the sun as seen from Earth on the first day of winter.

The coordinate frame thus defined is referred to as the geocentric inertial (GCI) system. Though fixed in the Earth, it does not rotate with the planet. It is seen that, in labeling the frame as "inertial," the angular velocity of the Earth about the sun is ignored. Because the frame is defined with respect to the "infinitely" distant stars, the translational offsets of the frame throughout the year are also irrelevant, and any axis set parallel to the defined set is equally valid.

It will often be necessary to transform vectors ($\boldsymbol{r}$, $\boldsymbol{V}$) in orbit plane coordinates to their equivalents in inertial space. From Fig. 4.6, it is clear that this can be accomplished in three steps, starting with the assumption that the orbit plane is coincident with the inertial X–Y plane, and that the abscissas are co-aligned:

1) Rotate the orbit plane by angle Ω about the inertial Z axis (colinear with the angular momentum vector $\boldsymbol{h}$).
2) Rotate the orbit plane about the new line of nodes by inclination i.
3) Perform a final rotation about the new angular momentum vector (also the new Z axis) by angle ω, the argument of periapsis.

In the terminology of rotational transformations (see also Chapter 7), this is a 3-1-3 Euler angle rotation sequence composed of elementary rotation matrices:

$$T_{P\to I} = \begin{bmatrix} C\omega & -S\omega & 0 \\ S\omega & C\omega & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & Ci & -Si \\ 0 & Si & Ci \end{bmatrix} \begin{bmatrix} C\Omega & -S\Omega & 0 \\ S\Omega & C\Omega & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.55}$$

This yields the rotation matrix that transforms a vector in perifocal coordinates into a vector in the inertial frame. In combined form, with $S\theta$ and $C\theta$ representing $\sin\theta$ and $\cos\theta$, we have

$$T_{P\to I} = \begin{bmatrix} C\Omega C\omega - S\Omega S\omega Ci & -C\Omega S\omega - S\Omega C\omega Ci & S\Omega Si \\ S\Omega C\omega + C\Omega S\omega Ci & -S\Omega S\omega + C\Omega C\omega Ci & -C\Omega Si \\ S\omega Si & C\omega Si & Ci \end{bmatrix} \tag{4.56}$$

Transformation matrices are orthonormal, and so the inverse transformation (in this case, from inertial to orbit plane coordinates) is found by transposing Eq. (4.56):

$$T_{I\to P} = T_{P\to I}^{-1} = T_{P\to I}^{t} \tag{4.57}$$

For heliocentric calculations, planetary equators are not suitable reference planes, and another choice is required. It is customary to define the Earth's orbital

plane about the sun, the ecliptic plane, as the reference plane for the solar system. The Earth's orbit thus has zero inclination, by definition, whereas all other solar orbiting objects have some nonzero inclination. The Z axis of this heliocentric inertial (HCI) system is again normal to the reference plane in the (roughly) northern direction, and the Y axis again is taken to form a right-handed set. The Earth's polar axis is inclined at approximately 23.5° relative to the ecliptic, and so the transformation from GCI to HCI is accomplished via a coordinate rotation of 23.5° about the X axis. The relationship between these two frames is shown in Fig. 4.7.

In either system, a variety of coordinate representations are possible in addition to the basic Cartesian (X, Y, Z) frame. The choice will depend in part on the type of equipment and observations employed. For example, a radar or other radiometric tracking system will produce information in the form of range to the spacecraft, azimuth angle measured from due North, and elevation angle above the horizon. Given knowledge of the tracking station location, such information is readily converted to standard spherical coordinates (r, θ, ϕ) and therefore to (X, Y, Z) or other coordinates.

When optical observations are made, as in classical orbit determination or when seeking to determine the position of an object against the background of fixed stars, range is not a suitable parameter. All objects appear to be located at the same distance and are said to be projected onto the celestial sphere. In this case, only angular information is available, and a celestial longitude–latitude system similar to that used for navigation on Earth is adopted. Longitude and latitude are replaced by right ascension and declination (α, δ). Right ascension is measured in the conventional trigonometric sense in the equatorial plane (about the Z axis), with 0° at the X axis. Declination is positive above and negative below the equatorial $(X-Y)$ plane, with a range of $\pm 90°$. Figure 4.8 shows the relationships between Cartesian, celestial, and spherical coordinates. Useful transformations are

$$X = r\sin\theta\cos\phi = r\cos\delta\cos\alpha \tag{4.58a}$$

$$Y = r\sin\theta\sin\phi = r\cos\delta\sin\alpha \tag{4.58b}$$

$$Z = r\cos\theta = r\sin\delta \tag{4.58c}$$

$$r^2 = X^2 + Y^2 + Z^2 \tag{4.58d}$$

$$\theta = \cos^{-1}\left[\frac{Z}{(X^2+Y^2+Z^2)^{1/2}}\right] \tag{4.58e}$$

$$\phi = \alpha = \tan^{-1}\left(\frac{Y}{X}\right) \tag{4.58f}$$

$$\delta = \sin^{-1}\left[\frac{Z}{(X^2+Y^2+Z^2)^{1/2}}\right] \tag{4.58g}$$

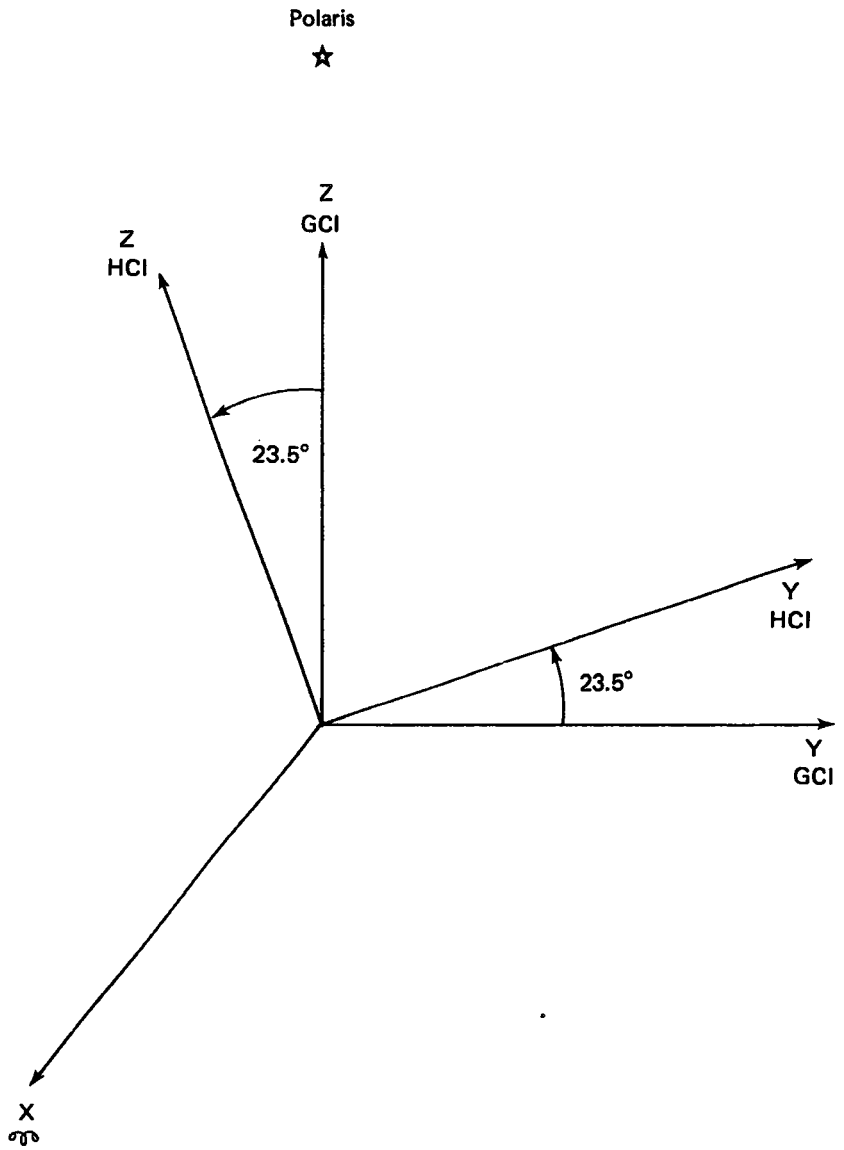


Fig. 4.7 Relation of GCI and HCI coordinates.

where line-of-sight vectors only are obtained, as with optical sightings, and r is assumed to be of unit length in Eqs. (4.58).

The Earth's spin axis is not fixed in space but precesses in a circle with a period of about 26,000 yr. This effect is due to the fact that the Earth is not spherically symmetric but has (to the first order) an equatorial bulge upon which solar and lunar gravitation act to produce a perturbing torque. The vernal equinox

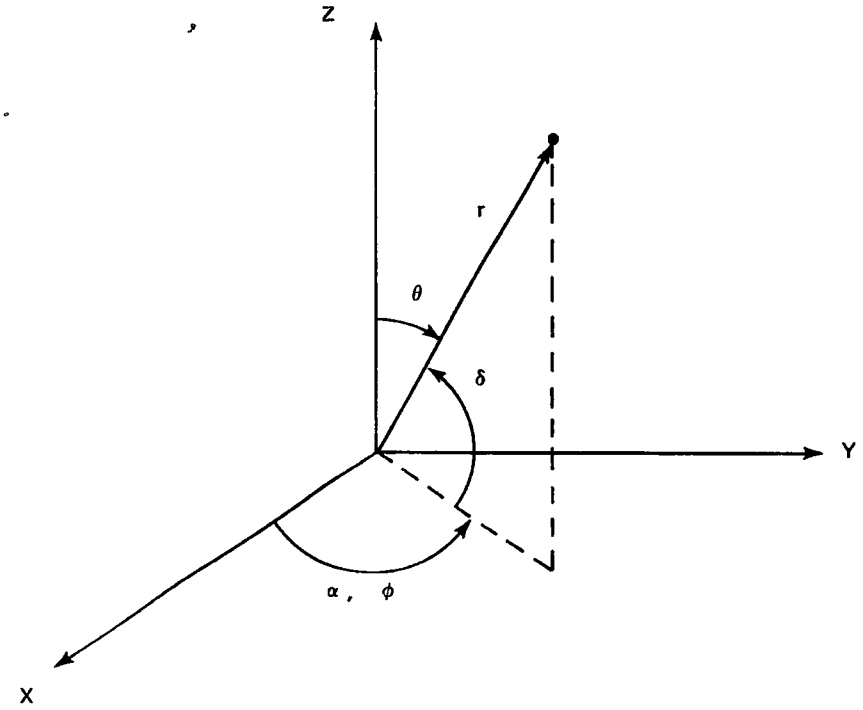


Fig. 4.8 Relationship between Cartesian, spherical, and celestial coordinates.

of course processes at the same rate, with the result that very precise or very long-term observations or calculations must account for the change in the "inertial" frames that are referred to the equinox. There is also a small deviation or about 9 arc-seconds over a 19-yr period due to lunar orbit precession that may in some cases need to be included. Specification of celestial coordinates for precise work thus includes a date or epoch (2000 is in common current use) that allows the exact orientation of the reference frame with respect to the "fixed" stars (which themselves have measurable proper motion) to be computed.

In many spacecraft applications, corrections over the mission lifetime are small with respect to those relative to the year 2000. For this reason, space missions are commonly defined with reference to true of date (TOD) coordinates, which have an epoch defined in a manner convenient to a particular mission.

4.2.8 Orbital Elements from Position and Velocity

As mentioned, knowledge of position and velocity at any single point and time in the orbit is sufficient to allow computation of all Keplerian elements. As an important example, it is possible given the position and velocity of the ascent vehicle at burnout to determine the various orbit injection parameters. A similar

calculation would be required following a midcourse maneuver in an interplanetary mission.

As a matter of engineering practice, the single-measurement errors in vehicle position and velocity are of such magnitude as to result in a rather crude estimate of the orbit from one observation, and so rather elaborate filtering and estimation algorithms are employed in actual mission operations to obtain accurate results. However, for mission design and analysis such issues are unimportant, and it is of interest to know the orbital elements in terms of nominal position and velocity vectors. This information is also of use in sensitivity studies, in which the orbit dispersions that result from specified launch vehicle injection errors (see Chapter 5) are examined.

It is assumed that $\boldsymbol{r}$ and $\boldsymbol{V}$ are known in a coordinate system of interest, such as GCI for Earth orbital missions or HCI for planetary missions. In practice, one or more coordinate transformations must be performed to obtain data in the required form, because direct measurements will be made in coordinates appropriate to a ground-based tracking station or network. It is assumed here that $\boldsymbol{i}, \boldsymbol{j}$, and $\boldsymbol{k}$ are the unit vectors in the (X, Y, Z) directions for the appropriate coordinate system.

Given $\boldsymbol{r}$ and $\boldsymbol{V}$ in a desired coordinate system, the angular momentum is, from Eq. (4.18),

$$\boldsymbol{h} = \boldsymbol{r} \times \boldsymbol{V} \tag{4.59}$$

with magnitude

$$h = rv \sin\left(\frac{\pi}{2} - \gamma\right) = rV \cos\gamma \tag{4.60}$$

where γ, the flight-path angle relative to the local horizon, is defined in Fig. 4.9. Thus,

$$\sin\gamma = \frac{rV}{rV} \tag{4.61}$$

Eccentricity may be found from Eq. (4.22), with E_t given by Eq. (4.14). Alternatively, it may be shown[9] that, in terms of flight-path angle γ,

$$e^2 = \left(\frac{rV^2}{\mu} - 1\right)^2 \cos^2\gamma + \sin^2\gamma \tag{4.62}$$

or

$$e^2 = \left[\left(\frac{rV^2}{\mu} - 1\right)^2 - 1\right] \cos^2\gamma + 1 \tag{4.63}$$

and

$$\tan\theta = \frac{(rV^2/\mu)\sin\gamma\cos\gamma}{(rV^2/\mu)\cos^2\gamma - 1} \tag{4.64}$$

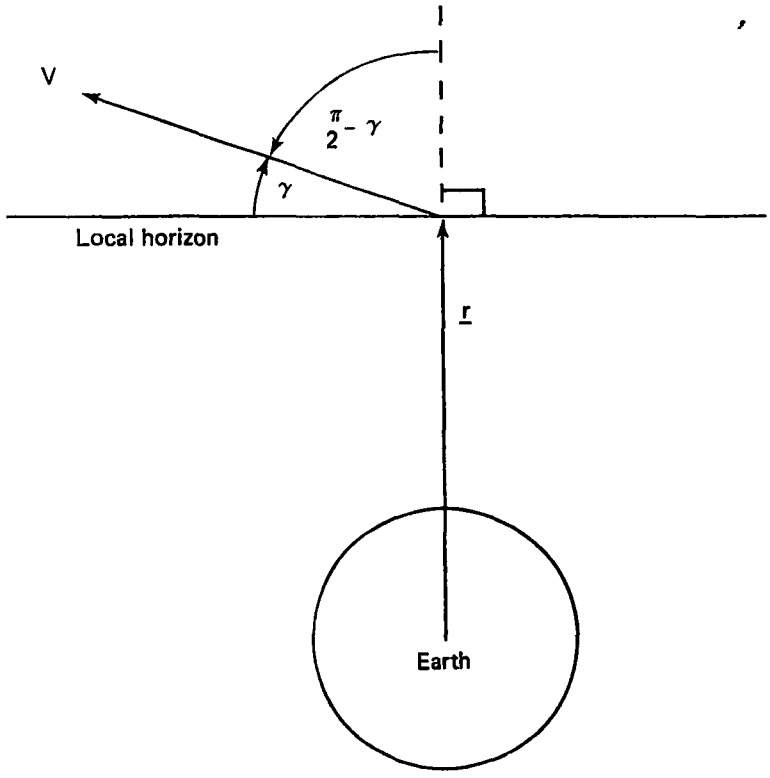


Fig. 4.9 Motion in orbit plane showing flight path angle.

Equation (4.62) avoids the ambiguity in true anomaly inherent in the use of the inverse cosine when Eq. (4.7) is used. Note that $\gamma = 0$ implies $\theta = 0$ if $rV^2/\mu > 1$, and $\theta = \pi$, if $rV^2/\mu < 1$. Thus, the spacecraft moves horizontally only at perigee or apogee if the orbit is elliptic. If $rV^2/\mu = 1$, the orbit is circular, and Eq. (4.61) will yield $\gamma = 0$, which implies that θ is undefined in Eq. (4.64). As discussed in Section 4.2.6, this difficulty is due to the fact that ω, and hence θ, are undefined for circular orbits. Defining $\omega = 0$ in this case will resolve the problem.

If defined, the semimajor axis is found from Eq. (4.8):

$$a = \frac{p}{1 - e^2} \tag{4.65}$$

or, if the orbit is nearly parabolic, we may use Eq. (4.20):

$$p = \frac{h^2}{u} \tag{4.66}$$

Since i is defined as the angle between $\boldsymbol{h}$ and $\boldsymbol{k}$,

$$\boldsymbol{h} \cdot \boldsymbol{k} = h \cos i = h_z \tag{4.67}$$

and

$$\cos i = \frac{h_z}{h} \tag{4.68}$$

where it is noted that $0° \leq i \leq 180°$.

The node vector $\boldsymbol{l}$ lies along the line of nodes between the equatorial and orbit planes and is positive in the direction of the ascending node. Thus,

$$\boldsymbol{l} = \boldsymbol{k} \times \boldsymbol{h} \tag{4.69}$$

Since Ω, the right ascension of the ascending node, is defined as the angle between $\boldsymbol{i}$ and $\boldsymbol{l}$,

$$\boldsymbol{l} \cdot \boldsymbol{i} = l \cos \Omega = l_x \tag{4.70}$$

and

$$\cos \Omega = \frac{l_x}{l} \tag{4.71}$$

where $l_y < 0$ implies $\Omega > 180°$. It may also be noted that

$$\tan \Omega = \frac{h_x}{-h_y} \tag{4.72}$$

which avoids the cosine ambiguity in Eq. (4.71).

Finally, it is noted that ω is the angle from the node vector $\boldsymbol{l}$ to perigee, in the orbit plane, whereas θ is measured from perigee to $\boldsymbol{r}$, also in the orbit plane. Thus,

$$\boldsymbol{l} \cdot \boldsymbol{r} = lr \cos(\omega + \theta) \tag{4.73}$$

Hence, the argument of perigee is

$$\omega = \cos^{-1}\left(\frac{\boldsymbol{l} \cdot \boldsymbol{r}}{lr}\right) - \theta \tag{4.74}$$

In Eq. (4.74), $\boldsymbol{k} \cdot \boldsymbol{r} > 0$ implies $0° < \omega + \theta < 180°$, and $\boldsymbol{k} \cdot \boldsymbol{r} < 0$ implies $180° < \omega + \theta < 360°$.

If desired, the periapsis time t_p may be found from the Kepler time-of-flight relations (4.26), (4.43), or (4.50), plus the auxiliary equations relating E, F, or D to true anomaly θ

The location of the launch site (or, more accurately, the location and timing of actual booster thrust termination) is a determining factor in the specification of some orbital elements. Of these, the possible range of orbital inclinations is the most important. Qualitatively, it is clear that not all inclinations are accessible from a given launch site. For example, if injection into orbit does not occur

precisely over the equator, an equatorial orbit is impossible, because the orbit plane must include the injection point. This is shown quantitatively by Bate et al.[15]:

$$\cos i = \sin\phi\cos\lambda \tag{4.75}$$

where

ϕ = injection azimuth (North = 0°)
λ = injection latitude.

If the boost phase is completed quickly so that injection is relatively close to the launch site, conditions for ϕ and λ as determined for the initial launch azimuth and latitude approximate those at vehicle burnout.

Equation (4.75) implies that direct orbits ($i < 90°$) require $0° \leq \phi \leq 180°$, and furthermore that the orbital inclination is restricted to the range $|i| \geq |\lambda|$.

The longitude of the ascending node Ω also depends on the injection conditions, in this case the injection time. This is because launch is from a rotating planet, whereas Ω is defined relative to the fixed vernal equinox. When the choice of Ω_0 is important, as for a sun-synchronous orbit (see Sec. 4.3.3, Aspherical Mass Distribution), the allowable launch window can become quite small.

4.2.9 State Vector Propagation from Initial Conditions

In practical work it is often desired to compute the orbital state vector ($\boldsymbol{r}$, $\boldsymbol{V}$) at time t given initial conditions ($\boldsymbol{r}_0$, $\boldsymbol{V}_0$) at time t_0. The material presented thus far allows us to do so as follows:

1) From ($\boldsymbol{r}_0$, $\boldsymbol{V}_0$), compute the orbital elements using Eqs. (4.59–4.74).

2) Use the appropriate form of Kepler's equation, depending on eccentricity, to find true anomaly $\theta(t)$ given $\theta_0 = \theta_0(t_0)$.

3) Determine r from Eq. (4.7), the orbit equation.

4) Use (for example) Eqs. (4.29–4.30) to find V_r, V_θ, hence ($\boldsymbol{r}$, $\boldsymbol{V}$) in orbit-plane coordinates.

5) Apply the appropriate coordinate transformation Eq. (4.56) to convert ($\boldsymbol{r}$, $\boldsymbol{V}$) from orbit plane coordinates to GCI or HCI.

This process, while conceptually clear, is undeniably awkward. Some improvement may be realized after $\theta(t)$ is obtained in step 2 through the use of the Lagrangian coefficients,[10,13,15] where the dependence on initial conditions is expressed as

$$\boldsymbol{r} = f\boldsymbol{r}_0 + g\boldsymbol{V}_0 \tag{4.76a}$$

$$\boldsymbol{V} = f\boldsymbol{r}_0 + \dot{g}\boldsymbol{V}_0 \tag{4.76b}$$

This relationship is mandated by the fact that, because ($\boldsymbol{r}_0$, $\boldsymbol{V}_0$) are coplanar but

not colinear, any other vector in the orbit plane can be written as a linear combination of (r_0, V_0). Various forms of f and g exist depending on whether the independent variable is taken to be true anomaly θ or eccentric anomaly D, E, or F. For example, with $\Delta\theta = (\theta - \theta_0)$,

$$f = 1 - (1 - \cos\Delta\theta)\frac{r}{p} \tag{4.77a}$$

$$g = \frac{rr_0 \sin\Delta\theta}{\mu p} \tag{4.77b}$$

$$\dot{f} = \sqrt{\frac{\mu}{p}}\left[\frac{1 - \cos\Delta\theta}{p} - \frac{1}{r} - \frac{1}{r_0}\right]\tan\left(\frac{\Delta\theta}{2}\right) \tag{4.77c}$$

$$\dot{g} = 1 - (1 - \cos\Delta\theta)\frac{r_0}{p} \tag{4.77d}$$

Equations (4.77) are independent of eccentricity, but of course it is necessary to use the appropriate form of Kepler's equation, depending on eccentricity, to obtain $\theta(t)$ given θ_0 and $(t - t_0)$. This is inconvenient in many analytical applications, an issue to which we have alluded previously.

As mentioned earlier, Battin[13] pioneered the development and application of "universal" formulas to eliminate this inconvenience. Numerous formulations exist; we cite here a particularly elegant approach (without its rather tedious derivation), easily implemented on a programmable calculator. Beginning with the end result, it is found that

$$f = 1 - \frac{U_2(\chi, \alpha)}{r_0} \tag{4.78a}$$

$$g = \frac{r_0 U_1(\chi, \alpha) + \sigma_0 U_2(\chi, \alpha)}{\sqrt{\mu}} \tag{4.78b}$$

$$\dot{f} = -\frac{\sqrt{\mu} U_1(\chi, \alpha)}{rr_0} \tag{4.78c}$$

$$\dot{g} = 1 - \frac{U_2(\chi, \alpha)}{r} \tag{4.78d}$$

where

$$\alpha = \frac{1}{a} = \frac{2}{r_0} - \frac{V_0^2}{\mu} \tag{4.79}$$

$$\sigma_0 = \frac{r_0 \cdot V_0}{\sqrt{\mu}} \tag{4.80}$$

and $U_i(\chi, \alpha)$ is the universal function of ith order. The universal functions, which are analogous to the trigonometric functions, satisfy

$$U_0(\chi, \alpha) = 1 - \frac{\alpha\chi^2}{2!} + \frac{(\alpha\chi^2)^2}{4!} - \cdots \tag{4.81}$$

$$U_1(\chi, \alpha) = \chi\left[1 - \frac{\alpha\chi^2}{3!} + \frac{(\alpha\chi^2)^2}{5!} - \cdots\right] \tag{4.82}$$

and satisfy the recursion relation

$$U_{n+2}(\chi, \alpha) = \frac{\chi^n}{\alpha n!} - \frac{U_n(\chi, \alpha)}{\alpha} \tag{4.83}$$

The universal Kepler equation is

$$\sqrt{\mu}(t - t_0) = r_0 U_1(\chi, \alpha) + \sigma_0 U_2(\chi, \alpha) + U_3(\chi, \alpha) \tag{4.84}$$

The variable χ is the universal form of the eccentric anomaly and can be shown to be

$$\chi = (\sigma - \sigma_0) + \alpha\sqrt{\mu}(t - t_0) \tag{4.85}$$

although usually this is unknown and is to be found via iteration.

The application of these results, as opposed to their derivation, is straightforward. Assuming that the functions $U_0(\chi, \alpha)$ and $U_1(\chi, \alpha)$ can be programmed as executable subroutines, in the manner of standard trigonometric functions, the procedure is as follows:

1) Given initial conditions (r_0, V_0), compute α and σ_0.

2) Given $(t - t_0)$, solve the Kepler equation for χ, i.e., guess χ, compute the $U_i(\chi, \alpha)$, and apply conventional root-finding techniques to iterate to convergence.

3) Given χ, compute U_0, U_1, U_2, and U_3.

4) With $U_0(\chi, \alpha)$, $U_1(\chi, \alpha)$, $U_2(\chi, \alpha)$, and $U_3(\chi, \alpha)$ now known, compute the Lagrangian coefficients.

5) Use f, g, $\dot{f}$, and $\dot{g}$ to find the new state vector ($\boldsymbol{r}$, $\boldsymbol{V}$) at time t.

Note that the orbital eccentricity is not used. This approach, of course, fails to convey the visual picture of the orbit offered by use of the classical elements. However, it is unrivaled in its computational efficiency and utility.

4.2.10 Orbit Determination

In earlier sections we have given procedures for obtaining orbital elements when $\boldsymbol{r}$ and $\boldsymbol{V}$ are known in an inertial frame such as GCI or HCI. The problem of orbit determination then essentially consists of obtaining accurate values of $\boldsymbol{r}$ and $\boldsymbol{V}$ at a known time t. Doppler radar systems and certain other types of equipment allow this to be done directly (ignoring for the moment any coordinate

transformations required to relate the observing site location to the inertial frame), whereas other types of radar or passive optical observations do not. Indeed, classical orbit determination may be thought of as the process of determining r and V given angular position (α, δ) observations in an inertial frame at known times.

Classical orbit determination remains important today. Radar observations are not possible for all targets and are seldom as accurate on any given measurement as are optical sightings. Given adequate observation time and sophisticated filtering algorithms (two days of Doppler tracking followed by several hours of computer processing are required to provide the ephemeris data, accurate to within 10 m at epoch, used in the Navy Transit navigation system), radiometric measurement techniques are the method of choice in modern astrodynamics. However, for accurate preliminary orbit determination, optical tracking remains unsurpassed. This is evident from the continued use and expansion of such systems, as for example in the Ground-Based Electro-Optical Deep Space Surveillance (GEODSS) program.[16]

As discussed, six independent pieces of time-tagged information are required to obtain the six orbital elements. Several types of observations have been used to supply these data:

1) Three position vectors $r(t)$ may be obtained at different known times. This case is applicable to radar systems that do not permit Doppler tracking to obtain velocity.

2) Three line-of-sight vectors (angular measurements) may be known at successive times. A solution due to Laplace gives position and velocity at the intermediate time.

3) Two position vectors $r(t)$ may be known and the flight time between them used to determine position and velocity. This is known as the Gauss problem (or the Lambert problem, when viewed from the trajectory designer's perspective), and remains useful today in part for its application to ballistic missile trajectory analysis and other intercept problems.

We note in passing that in few cases would an orbit be computed from only the minimum number of sightings. In practice, all data would be used, with the multiple solutions for the elements filtered or smoothed appropriately to allow a best solution to be obtained. However, practical details are beyond the scope of this text. Many references are available, e.g., Bate et al.,[15] Gelb,[17] Nahi,[18] and Wertz.[19] An excellent survey of the development of Kalman filtering for aerospace applications is given by Schmidt.[20]

We will not consider the details of orbit determination further in this text. Adequate references (e.g., Danby[10] and Bate et al.[15]) exist if required; however, such work is not generally a part of mission and spacecraft design.

4.2.11 Timekeeping Systems

Accurate measurement of time is crucial in astrodynamics. The analytic results derived from the laws of motion and presented here allow predictions to

be made for the position of celestial objects at specific past or future times. Discrepancies between predictions and observations may, in the absence of a priori information, legitimately be attributed either to insufficient fidelity in the dynamic model or to inaccuracy in the measurement of time. Obviously, it is desirable to reduce uncertainties imposed by timekeeping errors, so that differences between measured and predicted motion can be taken as evidence for, and a guide to, needed improvements in the theoretical model.

The concept of time as used here is that of absolute time in the Newtonian or nonrelativistic sense. In this model, time flows forward at a uniform rate for all observers and provides, along with absolute or inertial space, the framework in which physical events occur. In astrodynamics and celestial mechanics, this absolute time is referred to as "ephemeris time," to be discussed later in more detail.

Special relativity theory shows this concept of time (and space) to be fundamentally in error. Perception of time is found to be dependent on the motion of the observer, and nonlocal observers are shown to be incapable of agreeing on the exact timing of events. General relativity shows further that the measurement of time is altered by the gravity field in which such measurements occur. These results notwithstanding, relativistic effects are important in astrodynamics only in very special cases, and certainly not for the topics of interest in this book. For our purposes, Newtonian concepts of absolute space and time are preferable to Einsteinian theories of space-time.

4.2.11.1 Measurement of time. Measurement of absolute time is essentially a counting process in which the fundamental counting unit is derived from some observable, periodic phenomenon. Many such phenomena have been used throughout history as timekeeping standards, always with a progression toward the phenomena that demonstrate greater precision in their periodicity. In this sense, precision is obtained when the fundamental period is relatively insensitive to changes in the physical environment. Thus, pendulum clocks allowed a vast improvement in timekeeping standards compared to sundials, water clocks, etc. However, the period of a pendulum is a function of the local gravitational acceleration, which has a significant variation over the Earth's surface. Thus, a timekeeping standard based on the pendulum clock is truly valid only at one point on Earth. Moreover, the clock must be oriented vertically, and hence is useless, even on an approximate basis, for shipboard applications where accurate timekeeping is essential to navigation.

Until 1 January, 1958, the Earth's motion as measured relative to the fixed stars formed the basis for timekeeping in physics. By the end of the nineteenth century, otherwise unexplainable differences between the observed and predicted position of solar system bodies led to the suspicion that the Earth's day and year were not of constant length. Although this was not conclusively demonstrated until the mid-20th century, the standard or ephemeris second was nonetheless taken as 1/31,556,925.9747 of the tropical year (time between successive vernal equinoxes) 1900.

Since 1 January, 1958, the basis for timekeeping has been international atomic time (TAI), defined in the international system (SI) of units[21] as 9,292,631,770 periods of the hyperfine transition time for the outer electron of the Ce^{133} atom in the ground state. Atomic clocks measure the frequency of the microwave energy absorbed or emitted during such transitions and can yield accuracy of better than one part in 10^{14}. (For comparison, good quartz crystal oscillators may be stable to a few parts in 10^{13}, and good pendulum clocks and commonly available wristwatches may be accurate to one part in 10^6). The SI second was chosen to agree with the previously defined ephemeris second to the precision available in the latter as of the transition date.

4.2.11.2 Calendar time. Calendar time, measured in years, months, days, hours, minutes, and seconds, is computationally inconvenient but remains firmly in place as the basis for civil timekeeping in conventional human activities. Because spacecraft are launched and controlled according to calendar time, it is necessary to relate ephemeris time, based on atomic processes, to calendar time, which is forced by human conventions to be synchronous with the Earth's rotation.

The basic unit of convenience in human time measurement is the day, the period of time between successive appearances of the sun over a given meridian. The mean solar day (as opposed to the apparent solar day) is the average length of the day as measured over a year and is employed to remove variations in the day due to the eccentricity of the Earth's orbit and its inclination relative to the sun's equator.

Because "noon" is a local definition, a reference meridian is necessary in the specification of a planet-wide timekeeping (and navigation) system. The reference meridian, 0° longitude, runs through a particular mark at the former site of the Royal Observatory in Greenwich, England. Differences in apparent solar time between Greenwich and other locations on Earth basically reflect longitude differences between the two points, the principle that is the basis for navigation on Earth's surface. Twenty-four local time zones, each nominally 15° longitude in width, are defined relative to the prime meridian through Greenwich and are employed so that local noon corresponds roughly to the time when the sun is at zenith.

The mean solar time at the prime meridian is defined as universal time (UT), also called Greenwich mean time (GMT) or Zulu time (Z). A variety of poorly understood and essentially unpredictable effects (e.g., variations in the accumulation of polar ice from year to year) alter the Earth's rotation period; thus, UT does not exactly match any given rotation period, or day. Coordinated universal time (UTC) includes these corrections and is the time customarily broadcast over the ratio. "Leap seconds" are inserted or deleted from UTC as needed to keep it in synchrony with Earth's rotation as measured relative to the fixed stars. Past corrections, as well as extrapolations of such corrections into the future, are available in standard astronomical tables and almanacs.

4.2.11.3 Ephemeris time. From earlier discussions, it is clear that ephemeris time is the smoothly flowing, monotonically increasing time measure relevant in the analysis of dynamic systems. In contrast, universal time is based on average solar position as seen against the stars and thus includes variations due to a number of dynamic effects between the sun, Earth, and moon. The resulting variation in the length of the day must be accounted for in computing ephemeris time from universal time. The required correction is

$$\mathrm{ET} = \mathrm{UT} + \Delta T \simeq \mathrm{UT} + \Delta T(A) \tag{4.86}$$

where ΔT is a measured (for the past) or extrapolated (for the future) correction, and $\Delta T(A)$ is an approximation to ΔT given by

$$\Delta T(A) = \mathrm{TAI} - \mathrm{UT} + 32.184\ \mathrm{s} \tag{4.87}$$

Hence,

$$\mathrm{ET} \simeq \mathrm{TAI} + 32.184\ \mathrm{s} \tag{4.88}$$

where $\Delta T = 32.184$ s was the correction to UT on 1 January, 1958, the epoch for international atomic time. For comparison, $\Delta T(A) = 50.54$ s was the correction for 1 January, 1980. The 18.36-s difference that accumulated between TAI (which is merely a running total of SI seconds) and UT from 1958 to 1980 is indicative of the corrections required.

4.2.11.4 Julian dates. Because addition and subtraction of calendar time units are inconvenient, the use of ephemeris time is universal in astronomy and astrodynamics. The origin for ephemeris time is noon on 1 January, 4713 B.C., with time measured in days since that epoch. This is the so-called Julian day (JD). For example, the Julian day for 31 December, 1984 (also, by definition, 0 January, 1985) at 0 hrs is 2,446,065.5, and at noon on that day it is 2,446,066. Noon on 1 January, 1985, is then JD 2,446,067, etc. Tables of Julian date are given in standard astronomical tables and almanacs, as well as in navigation handbooks. Fliegel and Van Flandern[22] published a clever and widely implemented equation for the determination of any Julian date that is suited for use in any computer language (e.g., FORTRAN or PL/1) with integer-truncation arithmetic:

$$\begin{aligned}\mathrm{JD} = K - 32,075 + 1461 * \frac{[I + 4800 + (J - 14)/12]}{4} \\ + 367 * \frac{\{J - 2 - [(J - 14)/12] * 12\}}{12} \\ - 3 * \frac{\{[I + 4900 + (J - 14)/12]/100\}}{4}\end{aligned} \tag{4.89}$$

where

I = year
J = month
K = day of month

The Julian day system has the advantage that practically all times of astronomical interest are given in positive units. However, in astrodynamics and spacecraft work in general, times prior to 1957 are of little interest, and the large positive numbers associated with the basic JD system are somewhat cumbersome. Accordingly, the Julian day for space (JDS) system is defined with an epoch of 0 hrs UT on 17 September, 1957. Thus,

$$\text{JDS} = \text{JD} - 2{,}436{,}099.5 \tag{4.90}$$

Similarly, the modified Julian day (JDM) system has an epoch defined at 0 hrs UT, 1 January, 1950. These systems have the additional advantage for practical spacecraft work of starting at 0 hrs rather than at noon, which is convenient in astronomy for avoiding a change in dates during nighttime observations.

4.2.11.5 Sidereal time. Aside from ephemeris time, the systems discussed thus far are all based on the solar day, the interval between successive noons or solar zenith appearances. It is this period that is the basic 24-h day. However, because of the motion of the Earth in its orbit, this "day" is not the true rotational period of the Earth as measured against the stars, a period known as the sidereal day, 23 hrs 56 min 4 s. There is exactly one extra sidereal day per year.

Sidereal time is of no interest for civil timekeeping but is important in astronautics for both attitude determination and control and astrodynamics, where the orientation of a spacecraft against the stars is considered. For example, it is the sidereal day that must be used to compute the period for a geosynchronous satellite orbit. Sidereal time is also needed to establish the instantaneous relationship of a ground-based observation station or launch site to the GCI or HCI frame. The local sidereal time is given by

$$\theta = \theta_G + \Omega_E \tag{4.91}$$

where

θ = local sidereal time (angular measure)
θ_G = Greenwich sidereal time
Ω_E = east longitude of observing site

The Greenwich sidereal time is given in terms of its value at a defined epoch, t_0, by

$$\theta_G = \theta_{G_0} + \omega_e(t - t_0) \tag{4.92}$$

where ω_e is the inertial rotation rate of Earth and was 7.292116×10^{-5} rad/s for 1980. Again, tables of θ_{G_0} for convenient choices of epoch are provided in standard astronomical tables and almanacs.

4.3 Non-Keplerian Motion

We have, to this point, reviewed the essential aspects of two-body orbital mechanics, alluding only briefly to the existence of perturbing influences that can invalidate Keplerian results. Such influences are always present and can often be ignored in preliminary design. However, this is not always the case, nor is it indeed always desirable; mission design often involves deliberate use of non-Keplerian effects. Examples include the use of atmospheric braking to effect orbital maneuvers and the use of the Earth's oblateness to specify desired (often sun-synchronous) rates of orbital precession. These and other aspects of non-Keplerian orbital dynamics are discussed in the following sections.

4.3.1 Sphere of Influence

Of the various possible perturbations to basic two-body motion, the most obvious are those due to the presence of additional bodies. Such bodies are always present and cannot be easily included in an analysis, particularly at elementary levels. It is then necessary to determine criteria for the validity of Keplerian approximations to real orbits when more than two bodies are present.

If we consider a spacecraft in transit between two planets, it is clear that when close to the departure planet, its orbit is primarily subject to the influence of that planet. Far away from any planet, the trajectory is essentially a heliocentric orbit, whereas near the arrival planet, the new body will dominate the motion. There will clearly be transition regions where two bodies will both have significant influence on the spacecraft motion. The location of these transition regions is determined by the so-called sphere of influence of each body relative to the other, a concept originated by Laplace.

For any two bodies, such as the sun and a planet, the sphere of influence is defined by the locus of points at which the sun's and the planet's gravitational fields have equal influence on the spacecraft. The term "sphere of influence" is somewhat misleading; every body's gravitational field extends to infinity, and in any case, the appropriate equal-influence boundary is not exactly spherical. Nonetheless, the concept is a useful one in preliminary design.

Although the relative regions of primary influence can be readily calculated for any two bodies, the concept is most useful when, as mentioned earlier, one of the masses is much greater than the other. In such a case, the so-called classical

sphere of influence about the small body has the approximate radius

$$r \simeq R\left(\frac{m}{M}\right)^{2/5} \tag{4.93}$$

where

r = sphere of influence radius
R = distance between primary bodies m and M
m = mass of small body
M = mass of large body

A spacecraft at a distance less than r from body m will be dominated by that body, but at greater distances, it will be dominated by body M. Table 4.1 gives sphere-of-influence radii from Eq. (4.81) for the planets relative to the sun and the moon relative to the Earth.

Other sphere-of-influence definitions are possible. A popular alternate to Eq. (4.93) is the Bayliss sphere of influence, which replaces the $2/5$ power with $1/3$. This defines the boundary where the direct acceleration due to the small body equals the perturbing acceleration due to the gradient of the larger body's gravitational field.

Away from the sphere-of-influence boundary, Keplerian orbits are reasonably valid. Near the boundary, a two-body analysis is invalid, and alternate methods are required. In some cases, it may be necessary to know only the general characteristics of the motion in this region. The results of restricted three-body analysis, discussed later, may then be useful. Generally, however, more detailed information is required. If so, the mission analyst has two choices. Accurate trajectory results may be obtained by a variety of methods, all requiring a computer for their practical implementation. Less accurate, preliminary results may be obtained by the method of patched conics, discussed in Sec. 4.5. The use

Table 4.1 Planetary spheres of influence

Planet	Mass ratio (sun planet)	Sphere of influence, km
Mercury	6.0236×10^6	1.12×10^5
Venus	4.0852×10^5	6.16×10^5
Earth	3.3295×10^5	9.25×10^5
Mars	3.0987×10^6	5.77×10^5
Jupiter	1.0474×10^3	4.88×10^7
Saturn	3.4985×10^3	5.46×10^7
Uranus	2.2869×10^4	5.18×10^7
Neptune	1.9314×10^4	8.68×10^7
Pluto	3×10^6	1.51×10^7
Moon[a]	81.30	66,200

[a]Relative to Earth.

of this method is consistent with the basic sphere-of-influence calculation of Eq. (4.93) and thus does not properly account for motion in the transition region between two primary bodies.

4.3.2 Restricted Three-Body Problem

The general case of the motion of three massive bodies under their mutual gravitational attraction has never been solved in closed form. Sundman[23] developed a general power series solution; however, useful results are obtained only for special cases or by direct numerical integration of the basic equations. A special case of particular interest involves the motion of a body of negligible mass (i.e., a spacecraft) in the presence of two more massive bodies. This is the restricted three-body problem, first analyzed by Lagrange, and is applicable to many situations of interest in astrodynamics.

Earlier it was found[10] that, in contrast to the simple Keplerian potential function of Eq. (4.2), the restricted three-body problem yields a complex potential function with multiple peaks and valleys. This is shown in Fig. 4.10, where contours of equal potential energy are plotted. Along these contours, particles may move with zero relative velocity.

The five classical Lagrangian points are shown in Fig. 4.10. L_1, L_2, and L_3 are saddle points, i.e., positions of unstable equilibrium. Objects occupying these positions will remain stationary only if they are completely unperturbed. Any disturbances will result in greater displacement from the initial position.

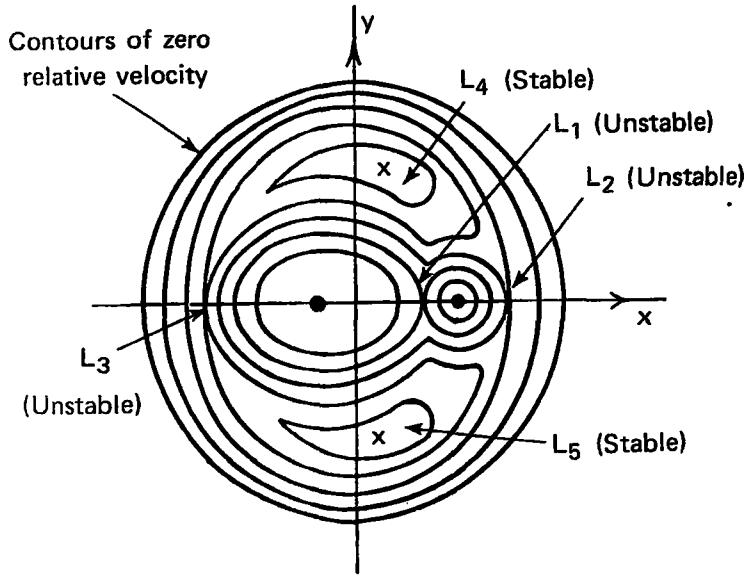


Fig. 4.10 Lagrange points for restricted three-body problem.

Spacecraft can occupy these positions for extended periods, but only if a supply of stationkeeping fuel is provided to overcome perturbing forces.

L_4 and L_5 are positions of stable equilibrium; objects displaced slightly from these positions will experience a restoring force toward them. Small bodies can occupy stable orbits about L_4 or L_5, a fact that is observationally confirmed by the presence of the Trojan asteroids that occupy the stable Lagrange points 60° ahead of and behind Jupiter in its orbit about the sun. The properties of L_4 and L_5 have led to considerable analysis of their use as sites, in the Earth-moon system, for permanent space colonies and manufacturing sites that would be supplied with raw materials form lunar mining sites.[3]

4.3.3 Aspherical Mass Distribution

As indicated in Sec. 4.2.1 (Two-Body Motion), an extended body such as the Earth acts gravitationally as a point mass provided its mass distribution is spherically symmetric. That is, the density function $\rho(r, \theta, \phi)$ in spherical coordinates must reduce to a function $\rho = \rho(r)$; there can be no dependence on latitude θ or longitude ϕ. Actually, the Earth is not spherically symmetric, but more closely resembles an oblate spheroid with a polar radius of 6357 km and an equatorial radius of 6378 km. This deformation is due to the angular acceleration produced by its spin rate (and is quite severe for the large, mostly gaseous outer planets such as Jupiter and Saturn), but numerous higher order variations exist and are significant for most Earth orbital missions.

The Earth is approximately spherically symmetric; thus, it is customary[24] to describe its gravitational potential in spherical coordinates as a perturbation to the basic form of Eq. (4.2), e.g.,

$$U(r, \theta, \phi) = -\frac{\mu}{r} + B(r, \theta, \phi) \tag{4.94}$$

where $B(r, \theta, \phi)$ is a spherical harmonic expansion in Legendre polynomials $P_{nm}(\cos \theta)$. The complete solution[25] includes expansion coefficients dependent on both latitude and longitude. Both are indeed necessary to represent the observable variations in the Earth's field. However, low-orbiting spacecraft having short orbital periods are not sensitive to the longitudinal, or tesseral, variations because, being periodic, they tend to average to zero. Spacecraft in high orbits with periods too slow to justify such an averaging assumption are too high to be influenced significantly by what are, after all, very small effects. Thus, for the analysis of the most significant orbital perturbations, only the latitude-dependent, or zonal, coefficients are important. (An exception is a satellite in a geostationary orbit; because it hovers over a particular region on Earth, its orbit is subjected to a continual perturbing force in the same direction, which will be important.) If the longitude-dependent variations are ignored, the gravitational

potential is[25]

$$U = -\left(\frac{\mu}{r}\right)\left[1 - \sum_{n=2}^{\infty}\left(\frac{R_e}{r}\right)^n J_n P_n(\cos\theta)\right] \tag{4.95}$$

where

R_e = radius of Earth
r = radius vector to spacecraft
J_n = nth zonal harmonic coefficient
$P_n(x) = P_{n0}(x)$ = nth Legendre polynomial

Table 4.2 shows that, for Earth, J_2 dominates the higher order J_n, which themselves are of comparable size, by several orders of magnitude. This is to be expected, because J_2 accounts for the basic oblateness effect, which is the single most significant aspherical deformation in the Earth's figure. Furthermore, since $(R_e/r) < 1$, it is to be expected that the second-order term in Eq. (4.95) will produce by far the most significant effects on the spacecraft orbit.

This is in fact the case. The major effect of Earth's aspherical mass distribution is a secular variation in the argument of perigee ω and the longitude of the ascending node Ω. Both of these depend to the first order only on the Earth's oblateness, quantitatively specified by J_2. The results are

$$\frac{d\omega}{dt} \simeq \left(\frac{3}{4}\right) n J_2 \left(\frac{R_e}{a}\right)^2 \frac{4 - 5\sin^2 i}{(1-e^2)^2} \tag{4.96}$$

$$\frac{d\Omega}{dt} \simeq -\left(\frac{3}{2}\right) n J_2 \left(\frac{R_e}{a}\right)^2 \frac{\cos i}{(1-e^2)^2} \tag{4.97}$$

where

n = mean motion = $\sqrt{\mu/a^3}$
a = semimajor axis
i = inclination
e = eccentricity

Variations in a, e, i, and n also occur but are periodic with zero mean and small amplitude. In Eqs. (4.96) and (4.97) we have approximated these parameters by their values for a Keplerian orbit, whereas in fact they depend on J_2. Errors implicit in this approximation may be as much as 0.1%.[19]

Table 4.2 Zonal harmonics for Earth

J_2	1.082×10^{-3}
J_3	-2.54×10^{-6}
J_4	-1.61×10^{-6}

Of interest here is the fact that, for direct ($i < 90°$) orbits, $d\Omega/dt < 0$, whereas for retrograde orbits it is positive. When the rotation rate equals 360°/yr, the orbit is said to be sun-synchronous, because its plane in inertial space will precess to remain fixed with respect to the sun. Such orbits are frequently used to allow photography or other Earth observations to take place under relatively fixed viewing and lighting conditions. Practical sun-synchronous orbits are usually nearly circular, slightly retrograde ($96° < i < 100°$) and have a mean altitude in the range of 200–1000 km.

From Eq. (4.96) we note that $d\omega/dt = 0$ for $\sin^2 i = 4/5$, or $i = 63.435°$; thus, in this case there is no perturbation to the argument of perigee. For $i < 63.435°$, the rotation of the line of apsides is in the direction of the orbit, whereas for larger inclinations, it rotates in the direction opposite to the motion.

The computed perturbations due to the Earth's oblateness and the higher order geopotential variations are most important for relatively low orbits and may not be the dominant gravitational disturbances for higher orbits. For example, considering only the J_2 term in Eq. (4.95), it is easily found that, at $r \simeq 15{,}000$ km, the maximum possible perturbing acceleration is about 10^{-3} m/s^2. This is comparable to or exceeded by the perturbing acceleration produced by the sun on a spacecraft in such an orbit. At higher altitudes, oblateness effects are even smaller and may well rank behind lunar perturbations (typically two orders of magnitude smaller than solar effects) in significance. The exception, again, can be satellites in geostationary orbits, which are subject to essentially constant perturbing forces due to geopotential variations.

The results of this section are of course not restricted to Earth-orbiting spacecraft and can be used to establish orbits with particular characteristics about other planets. During the Viking missions to Mars, for example, the Viking orbiters were inserted into highly elliptic orbits with a periapsis of a few hundred kilometers and a period of 24.68 h (1 Martian day). This was initially done to optimize data relay from the landers. However, this orbit minimized fuel usage during the injection maneuver, allowed both high- and low-altitude photography, and, because of the rotation of the line of apsides, allowed the periapsis to occur over a substantial range of latitudes. This allowed most of the planet to be photographed at both small and large scale.

The results presented in this section may be used in reverse order; that is, orbital tracking may be used to establish the values of the J_n coefficients (and, in the full-blown expansion, the values of the tesseral harmonic coefficients as well).[26] Tracking of the early Vanguard satellites was used to establish the pear-shaped nature of the Earth's figure, shown by the fact that J_3 is nonzero. Today satellite geodesy is the method that produces the most accurate determination of the harmonic coefficients.

Satellite geodesy imposes unique spacecraft design requirements for its accomplishment. Low-orbiting spacecraft potentially yield the most useful results, yet at low altitudes atmospheric drag dwarfs the minor perturbations that

it can be desired to measure. Great care is required to develop analytical models for these effects so that the biases they introduce can be removed from the data.

The problem can also be circumvented by the use of an onboard drag-compensation system. The system operates by enclosing a small free-floating proof mass inside the spacecraft (see Fig. 4.11). The proof mass is maintained in a fixed position relative to the spacecraft body by means of very small thrusters operating in a closed-loop control system.[27] This drag-compensation system removes all nongravitational forces on the proof mass, and hence the spacecraft, and thus forces it to follow a purely gravitational trajectory. The measured departures from simple Keplerian motion are then due to perturbations in the Earth's potential field.

This approach has advantages for navigation satellites as well as for geodetic research. First implemented in the U.S. Navy TRIAD program in 1972, the drag-free [also called Disturbance Compensation System (DISCOS)] concept[27] has allowed the measurement of along-track gravitational perturbations down to a level of $5 \times 10^{-12} g$. Its use has been proposed for other applications where removal of all nongravitational effects is important.

In recent years the constellation of some two dozen GPS satellites (see also Chapter 11) has been used to refine the geodetic model.[28] The relatively high orbit of the constellation makes it difficult to detect the higher order field harmonics; however, this is compensated by the advantages that result from having numerous satellites in stable orbits available to be tracked for decades. Proposals have been advanced for placing two spacecraft in the same relatively low orbit and using their differential acceleration as an even more sensitive indicator of variations in the Earth's figure[29] (see Table 4.2).

The standard model for Earth geodesy for many years was the World Geodetic System 1984 (WGS-84) model, which provides expansion coefficients for Eq. (4.94) up to $n = m = 180$. A newer version, Earth Gravity Model 1996 (EGM-96) provides coefficient data through $n = m = 360$. These models are maintained by the U.S. National Imagery and Mapping Agency (NIMA), and may be downloaded from the NIMA website.

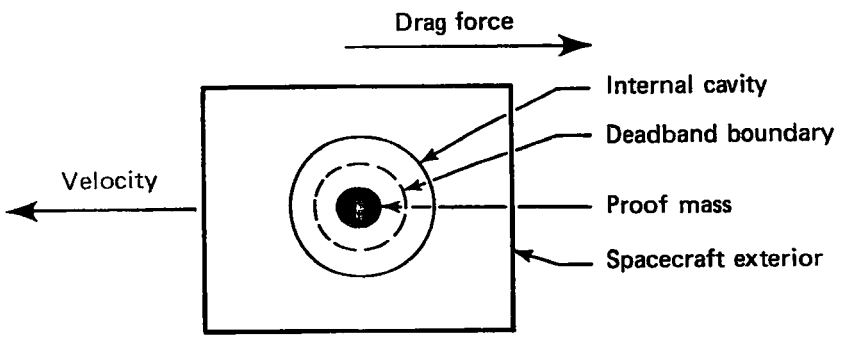


Fig. 4.11 Drag compensation concept.

The use of the potential function expansion approach relies for its computational practicality on the assumption that the Earth is nearly spherical and that its potential may be approximated by small perturbations to the simple point-mass form. For the Earth this is true, but this is not always so elsewhere in the solar system. The moon is both substantially ellipsoidal and possessed of regions of significant mass concentration, or mascons, discovered during the Apollo era by tracking lunar orbiting spacecraft. In such cases, the oblateness term may not be the most significant source of orbital perturbations.

4.3.4 Atmospheric Drag

The influence of atmospheric drag is important for all spacecraft in low Earth orbit, both for attitude control (see Chapter 7) and in astrodynamics. Such spacecraft will eventually reenter the atmosphere due to the cumulative effect of atmospheric drag unless provided with an onboard propulsion system for periodic reboost. The period of time required for this to occur is called the orbital lifetime of the spacecraft, and depends on the mass and aerodynamic properties of the vehicle, its orbital altitude and eccentricity, and the density of the atmosphere.

An example of the effect of atmospheric drag on a spacecraft in low Earth orbit is given in Fig. 4.12, which shows the predicted decay for the Space Shuttle as a function of orbital height and vehicle attitude.[30]

Because of the difficulty of modeling both the environmental and spacecraft aerodynamic properties in this flight regime, quantitative results for orbital perturbations due to atmospheric drag are difficult to obtain. In the following sections we discuss the approximate results that can be obtained and supply guidelines for more detailed modeling effects where appropriate.

4.3.4.1 Atmosphere models. As will be seen, satellite orbital lifetime depends strongly on the variation of the upper atmosphere density with altitude. Although the gross behavior of the atmospheric density is well established, exact properties are difficult to determine and highly variable. It is this factor, more than any other, which makes the determination of satellite lifetimes so uncertain, and renders even the most sophisticated analysis of more academic than practical interest. "Permanent" satellites will have lifetimes measured in years, with an uncertainty measured at least in months and quite possibly also in years.

As a spacecraft approaches reentry, this uncertainty decreases; however, predicted lifetimes may be in error by one or more orbital periods even on the day of reentry. Errors of about 10% of the remaining lifetime represent the best obtainable precision for orbit decay analyses. Difficulties in orbital decay analysis were graphically illustrated during the final months and days prior to the reentry of Skylab in 1979, and again for Mir in 2001. Kaplan et al.[31] present an excellent summary of the analytical and operational effort expended in a largely unsuccessful attempt to effect a controlled reentry of the Skylab vehicle.

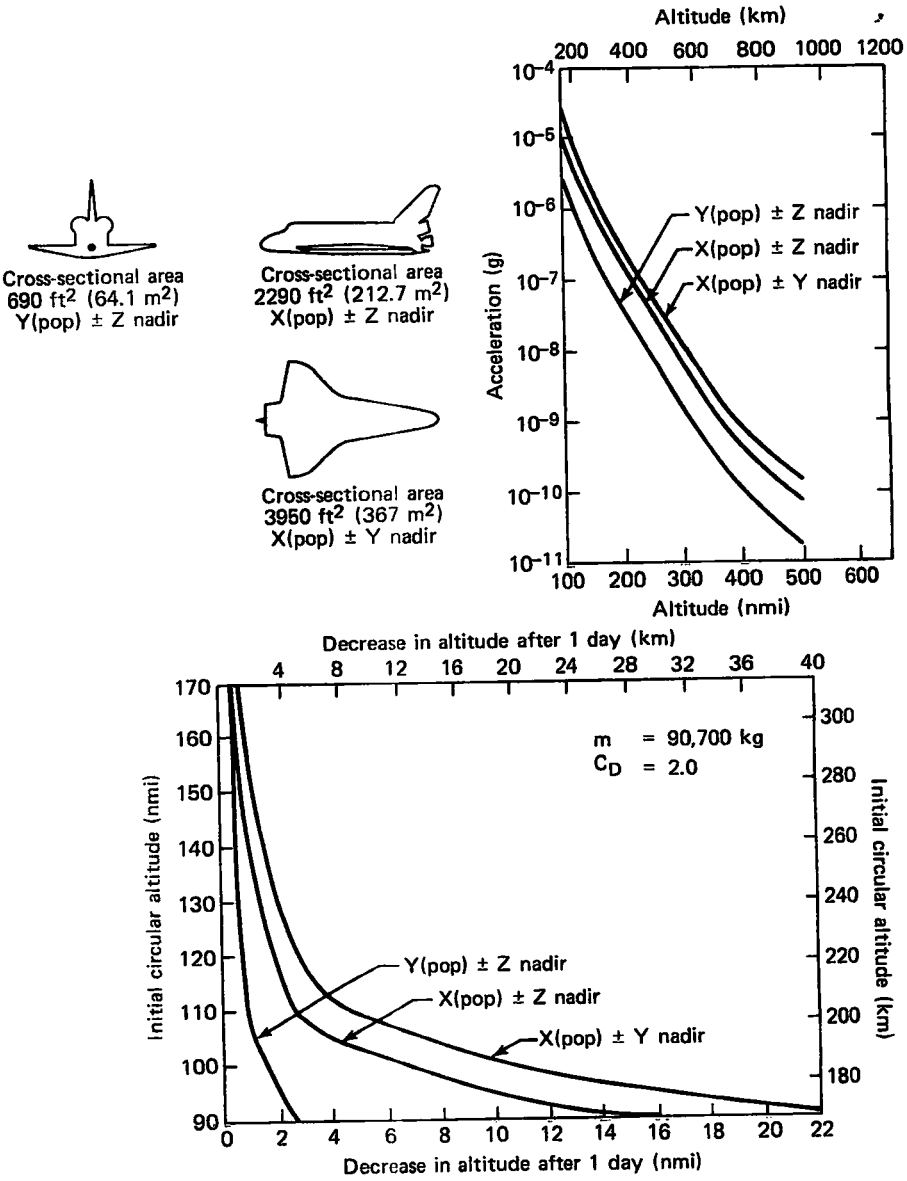


Fig. 4.12 Effects of atmospheric drag on the shuttle orbiter.

Similar concerns existed in connection with the reentry of the Russian Mir space station after 15 years of service. However, at the time of its de-orbit, substantial attitude and propulsion capability still remained, and a controlled reentry was performed on 23 March 2001, with impact in the ocean.

Standard models exist[32] for the variations in average atmospheric properties with altitude (see Chapter 3). For properties at altitudes above 100 km, which are of principal interest in spacecraft dynamics, these models are based primarily on the work of Jacchia.[33] These models are maintained by the National Space Science Data Center (NSSDC), at the NASA Goddard Space Flight Center, and may be downloaded from the NSSDC Web site.

As indicated in Chapter 5, the basic form of the Earth's atmospheric density profile is obtained by requiring hydrostatic equilibrium in conjunction with a specified temperature profile, determined via a combination of analytical and empirical means. The temperature profile is modeled as a sequence of layers having either a constant temperature or a constant temperature gradient (see Fig. 3.21).

These assumptions result in a density profile having the form of Eqs. (5.18a) or (5.18b) in each layer, i.e., an exponential or power law dependence. In practice, there may be little difference in the density as predicted by each of these forms, and it is common and analytically convenient to assume an exponential form:

$$\rho = \rho_0 e^{-\beta(r-r_0)} \tag{4.98}$$

where

$1/\beta$ = scale height
ρ_0 = density at reference altitude
r_0 = reference altitude
r = radius vector magnitude

The reference altitude may be the top or bottom of a specified layer. If a single layer is assumed, the atmosphere is referred to as strictly exponential, and the reference level may be sea level. It may also be desirable to use a reference level at some altitude, such as 100 km, appropriate to the analysis. This allows accuracy to be maintained at orbital altitudes, at the possible sacrifice of sea level results, which are irrelevant in this situation.

The exact form of β is[34]

$$\beta = \frac{g/R_{\text{gas}} + dT/dr}{T} \tag{4.99}$$

where

R_{gas} = specific gas constant
T = temperature
g = gravitational acceleration

The ability to obtain the integrated result of Eq. (4.98), and consequently its accuracy, depends on the assumption of constant scale height, and thus on the assumption that the g and T are both constant. Clearly this is not strictly true. Gravitational acceleration varies by about 4% over the altitude range from sea level to 120 km.

Rather large local temperature gradients, on the order of 10 K/km at 120 km, may exist. Scale heights for Earth below about 120 km range from 5 to 15 km,

with a mean value of about 7.1 km. In the worst case, when $1/\beta = 5\,\text{km}$, examination of temperature profile data for Earth indicates that the gravitational term is approximately $0.14\,\text{km}^{-1}$ and the temperature gradient term is approximately $0.04\,\text{km}^{-1}$. However, for the relatively small range of altitudes of importance in satellite aerodynamic drag analyses, it is often acceptable to assume β to be a constant given by its mean value in the range of interest. Shanklin et al.[35] studied the sensitivity of near-term orbit ephemeris predictions to differences among various atmosphere models and concluded that the simple exponential model discussed earlier yielded, for the four cases studied, results indistinguishable from those using more complex models.

Standard atmosphere data are presented in Figs. 3.21 and 3.22 and Appendix B.17. Extensive variations from average atmosphere properties exist. Fluctuations are observed on a daily, 27-day, seasonal, yearly, and 11-yr basis, as well as with latitude. Of these, the 11-year cycle is the most pronounced, due to the variations in solar flux with the sunspot cycle. Density fluctuations of a factor of 10 at 350 km and a factor of 5 at 800 km are observed.

Higher solar flux produces greater atmospheric density at a given altitude. Solar flux is commonly reported in units of 10^4 Janskys ($1\,\text{Jansky} = 10^{-26}\,\text{W} \cdot \text{m}^{-2} \cdot \text{Hz}^{-1}$), with typical 11-yr minima of 80 and maxima of 150, with peaks of 250 not unrealistic. Jacchia[36] shows the effects of such peaks. For example, at 500 km with a solar flux of 125, a nominal density of $1.25 \times 10^{-12}\,\text{kg/m}^3$ is observed, whereas a solar flux of 160 will produce the same density at 600 km.

4.3.4.2 Effects of drag on orbital parameters. Drag is defined as the component of force antiparallel to the spacecraft velocity vector relative to the atmosphere. We may often ignore the velocity component due to the rotation of the atmosphere, because it is small compared to the spacecraft orbital velocity. There is then no component of force normal to the orbit plane, and, to a first approximation, atmospheric drag thus has no effect on the elements ω, Ω, and i, which determine the orientation of the orbit in space.

As discussed earlier, atmospheric density varies exponentially with height above the Earth, with a scale height on the order of tens of kilometers, and is always small at orbital altitudes. Thus, if the orbit is even slightly elliptic, the principal drag occurs at perigee and may be modeled approximately as an impulsive reduction in velocity. It will be seen in Section 4.4 that the result of such an impulsive ΔV maneuver is to reduce the orbital apogee while leaving the perigee altitude essentially unchanged. Thus, atmospheric drag reduces the semimajor axis and the eccentricity of the orbit, tending to circularize it. A low orbit that is nearly circular experiences a significant continuing drag force and may have a lifetime of only days or hours.

This model is of course only approximate, and aerodynamic drag does cause variations in all orbital elements.[37] However, more detailed analysis does confirm the principal features of this model, namely, that drag tends to circularize the orbit at an altitude very near that of the original perigee. Analysis of the motion of

a satellite in the upper regions of a planetary atmosphere is extremely complex. The classical treatment in this area is that of King-Hele,[38] whereas that of Vinh et al.[34] provides an excellent more recent text. We include here some results of importance in preliminary mission design.

4.3.4.3 Decay of elliptical orbits. We begin by considering the decay of an initially elliptic orbit with $e_0 > 0$ to $e = 0$. This is most conveniently done by specifying eccentricity as the independent variable. In this way, the geometric properties of the decaying orbit can be specified without reference to the spacecraft dynamic characteristics. In non-dimensional form, it is found that the semimajor axis decays as

$$\frac{a}{a_0} = 1 + \varepsilon h_1(\alpha) + \varepsilon^2 h_2(\alpha) + \varepsilon^3 h_3(\alpha) + \varepsilon^4 h_4(\alpha) + \cdots \tag{4.100}$$

where the subscript 0 implies initial orbit conditions and

$$\alpha = \beta a_0 e$$

$$\varepsilon = \frac{1}{\beta a_0}$$

$$h_1 = B - B_0$$

$$h_2 = 2(A - A_0) + (A - 3)(B - B_0)$$

$$h_3 = \frac{7[\alpha^2 - (\beta a_0 e_0)^2]}{2} - \frac{(4A_0 - 13)(A - A_0)}{2} + (2\alpha^2 - 4A - A^2 + 13)(B - B_0) + \frac{(\alpha^2 + A - A^2 + 3)(B - B_0)^2}{2}$$

$$h_4 = \frac{[\alpha^2 - (\beta a_0 e_0)^2](7A - 4A_0 - 35)}{2} + \frac{(A - A_0)(12\alpha^2 + 16A_0^2 + 42A_0 + 4A_0 A - 9A - 8A^2 + 213)}{6} + \frac{(B - B_0)(2A^3 + 7A^2 + 46A - 25\alpha^2 - 138)}{2} + \frac{(B - B_0)^2(A^3 + 6A^2 - \alpha^2 A - 7\alpha^2 - 35)}{2} + 2(A - A_0)(B - B_0)(\alpha^2 + A - A^2 + 3) + (B - B_0)^3 \frac{(2A^3 - 3A^2 - 2\alpha^2 A + \alpha^2 - 6)}{6}$$

$$A = \alpha I_0(\alpha)\alpha I_1(\alpha)$$

$$B = \ell n[\alpha I_1(\alpha)]$$

where I_n is the imaginary Bessel function, nth order, first kind.[25,39]

The decay in semimajor axis a/a_0 found through Eq. (4.100) is then used to determine the changes in other parameters. It is found that the perigee radius decays as

$$\frac{r_p}{r_{p_0}} = \frac{(1 - \varepsilon\alpha)(a/a_0)}{(1 - e_0)} \tag{4.101}$$

while the apogee radius is found from

$$\frac{r_a}{r_{a_0}} = \frac{(1 + \varepsilon\alpha)(a/a_0)}{(1 + e_0)} \tag{4.102}$$

and the decrease in orbital period is

$$\frac{\tau}{\tau_0} = \left(\frac{a}{a_0}\right)^{3/2} \tag{4.103}$$

The preceding results are used to predict the evolving shape of the orbit during the decay process:

1) Given the initial orbit and the atmosphere model, find ε.
2) Choose e in the range $0 < e < e_0$ and compute α.
3) Compute the h_i and solve Eqs. (4.100–4.103).
4) Repeat steps 2 and 3 as needed for a range of eccentricities.

In spacecraft design applications it will commonly be desired to study the evolution of the orbit in time. To do this, the orbital eccentricity must be known as a function of time. King-Hele[38] finds, and the higher order analysis of Vinh et al.[34] confirms, that eccentricity varies with time as

$$\left(\frac{e}{e_0}\right)^2 = 1 - \frac{T}{T_L} \tag{4.104}$$

where T is the nondimensional time and T_L the nondimensional orbital lifetime given in terms of the initial orbit parameters by

$$T_L = \left(\frac{e_0}{\varepsilon}\right)^2 \frac{(1 - 5e_0/6 + 23e_0^2/48 + 7\varepsilon/8 + \varepsilon e_0/6 + 9\varepsilon^2/16e_0)}{2} \tag{4.105}$$

If the orbit is initially nearly circular ($e_0 < 0.02$), T_L is given more accurately by

$$T_L = \left[1 - \frac{(9\beta^2 a_0^2 e_0^2/20 - 1)\varepsilon}{2}\right]\left(\frac{e_0^2}{2\varepsilon^2}\right) \tag{4.106}$$

T contains all of the spacecraft parameters and is defined as

$$T = 2\pi\left(\frac{SC_D}{m}\right)\rho_{p_0} f^2 \beta^2 a_0^3 e_0 I_1(\beta a_0 e_0) \exp(-\beta a_0 e_0)\frac{t^*}{\tau_0} \tag{4.107}$$

where ρ_{p_0} is the initial periapsis atmospheric density and the drag model is

$$D = \tfrac{1}{2}\rho V^2 f^2 \frac{SC_D}{m} \tag{4.108}$$

where

D = drag acceleration of satellite
C_D = drag coefficient
S = projected area normal to flight path
ρ = atmospheric density
m = satellite mass

The parameter f provides a latitude correction, usually less than 10%, to the orbital speed as it appears in the drag model, and is given by

$$f = \left[1 - \left(\frac{\omega_e r_{p_0}}{V_{p_0}}\right)\cos i\right] \tag{4.109}$$

where

ω_e = angular velocity of the Earth (7.292×10^{-5} rad/s)
V_{p_0} = initial periapsis velocity

The term (m/SC_D) in Eq. (4.107) is the ballistic coefficient and is a measure of the ability of an object to overcome fluid resistance (see also Chapter 6). Typical values for spacecraft will be on the order of 10–100 kg/m^2. The projected area can be complicated to compute for other than simple spacecraft shapes, but this presents no fundamental difficulties. Determination of the drag coefficient is more complex. This issue is addressed subsequently.

If only the time to circularize the orbit is required, then Eq. (4.105) or (4.106) alone is sufficient, and the more complex procedure for analyzing the evolution of the orbit due to drag is unnecessary. This is often the case in preliminary design work.

4.3.4.4 Circular orbit lifetime. The lifetime remaining for a circular orbit is easily estimated from first principles. Orbital energy is dissipated by drag, and since the semi-major axis a is solely a function of the orbital energy, it will gradually decay. The rate at which orbital energy is lost must be matched by the power (force multiplied by velocity) dissipated by the satellite due to aerodynamic drag. Thus, from Eqs. (4.16) and (4.108),

$$-\tfrac{1}{2}\rho V^3 f^3 \frac{SC_D}{m} = \frac{dE_t}{dt} = -\frac{d(\mu/2a)}{dt} = \left(\frac{\mu}{2a^2}\right)\frac{da}{dt} \tag{4.110}$$

Because the orbit is circular, $r = a$ and

$$V = V_{\text{cir}} = \frac{\sqrt{\mu}}{a} \tag{4.111}$$

After some straightforward algebraic manipulation, we obtain for the nondimensional circular orbit lifetime

$$\frac{t}{\tau_0} = \int_0^{t/\tau_0} \mathrm{d}t = \frac{1}{2\pi}\left(\frac{1}{\rho_0 a_0^{3/2}}\right)\left(\frac{1}{f^3}\right)\left(\frac{m}{SC_D}\right)\int_{a_{\min}}^{a} \frac{\rho_0}{\rho}\frac{\mathrm{d}a}{a^{1/2}} \tag{4.112}$$

where

$\tau_0 = 2\pi\sqrt{a_0^3/\mu}$ = initial orbit period
$\rho_0/\rho = e^{\beta(a-a_0)}$ = exponential atmosphere model
ρ_0 = initial orbit atmospheric density

and $a_{\min}$ is the value of the semimajor axis below which reentry is assumed to be imminent. For example, as mentioned in Chapter 6, the space shuttle entry interface altitude is taken, partly by convention, to be 122 km or 400,000 ft. Certainly this would represent a very conservative lower bound for orbit decay; in most cases one might consider the circular orbit lifetime to be terminated at an altitude of 150 km.

Equation (4.112) may be integrated numerically, using the exponential atmosphere model of Eq. (4.98) as indicated previously, standard atmosphere tables for $\rho_0/\rho(r)$ given in Appendix B, or nonstandard values representative of a high solar flux, etc. A closed-form result expressed in terms of Dawson's integral[39] may be obtained, but for typical values of β and a_0, the actual computation is quite ill-posed, and of little value. A useful closed-form approximation can be developed by letting

$$a = a_0 + \Delta \tag{4.113}$$

where we assume $|\Delta/a_0| \ll 1$, hence

$$\frac{\mathrm{d}a}{a^{1/2}} \cong \left(1 + \frac{\Delta}{2a_0}\right)\frac{\mathrm{d}\Delta}{a_0^{1/2}} \tag{4.114}$$

Using these results in Eq. (4.112) and dropping terms of order Δ/a_0, we obtain for the circular orbit lifetime

$$\frac{t}{\tau_0} = \frac{1}{4\pi}\left(\frac{2\beta a_0 + 1}{\rho_0 \beta^2 a_0^3}\right)\left(\frac{1}{f^3}\right)\left(\frac{m}{SC_D}\right)(1 - e^{\beta\Delta}) \tag{4.115}$$

and the reader is reminded that for the case of orbit decay, $\Delta < 0$

The results of this section are certainly feasible for use in preliminary design, but are undeniably tedious. A variety of computational alternatives exist. For

example, the NASA ORDEM2000 model,[40] discussed in Chapter 3 in connection with orbital debris hazard assessment, can also be used to estimate the orbital lifetime of a space object, given its orbital parameters and ballistic coefficient. The model can be downloaded from the NASA Johnson Space Center.

4.3.4.5 Drag coefficient. As seen, the orbit lifetime depends directly on C_D. For analysis of orbit decay, we are concerned with the so-called free molecular flow regime. In this regime, the flow loses its continuum nature and appears to a first approximation as a stream of independent particles (i.e., Knudsen number $\gg \infty 1$). Furthermore, the flow is at very high speed, about Mach 25 or more than 7.5 km/s in most cases, and thus is hypersonic. This implies that any pressure forces produced by random thermal motion are small in comparison to those due to the directed motion of the spacecraft through the upper atmosphere.

The high-speed, rarefied flow regime permits the use of Newtonian flow theory, in which the component of momentum flux normal to a body is assumed to be transferred to the body by means of elastic collisions with the gas molecules. The tangential component is assumed unchanged. The net momentum flux normal to the body surface produces a pressure force that, when integrated over the body, yields the drag on the body. Geometric shadowing of the flow is assumed, which implies that only the projected area of the spacecraft can contribute to the drag force. Subject to these assumptions, and assuming Mach $25 \cong$ Mach ∞, the pressure coefficient

$$C_p \equiv \frac{(p - p_\infty)}{\frac{1}{2}\rho V^2} \tag{4.116}$$

is given by

$$C_p = 1.84 \sin^2 \theta \tag{4.117}$$

where θ is the local body angle relative to the flow velocity vector.

The drag coefficient is obtained by integrating the streamwise component of the pressure coefficient over the known body contour. For simple shapes the required integrations can be performed analytically, yielding useful results for preliminary design and analysis. Table 4.3 gives values of C_D for spheres, cylinders, flat plates, and cones subject to the preceding assumptions.

Newtonian flow theory allows only approximate results at best. Even at the high speeds involved in orbital decay analysis, random thermal motion is important, as in the exact, non-elastic nature of the interaction of the gas molecules with the body surface.

Table 4.3 Newtonian flow drag coefficients

Body	C_D
Sphere	1.0
Circular cylinder	1.3
Flat plate at angle α	$1.8 \sin^3 \alpha$
Cone of half-angle δ	$2 \sin^2 \delta$

A more accurate treatment results from the assumption that atmospheric molecules striking the spacecraft are in Maxwellian equilibrium, having both random and directed velocity components. Some of the molecules that strike the surface are assumed to be re-emitted inelastically, with a Maxwellian distribution characteristic of the wall temperature T_w. This model, due to Shaaf and Chambre,[41] results in a pressure force on the wall of

$$p = 2p_i - \sigma_n(p_i - p_w) \tag{4.118}$$

where

p_i = pressure due to incident molecular flux
p_w = pressure due to wall re-emissions
σ_n = normal momentum accommodation coefficient

With Maxwellian distributions assumed, p_i and p_w may be computed and the net pressure force obtained. A similar analysis by Fredo and Kaplan[42] yields the shear force and introduces a dependence on the tangential momentum accommodation coefficient σ_t.

The accommodation coefficients σ_n and σ_t characterize the type of interaction the gas particles make with the surface. The equation $\sigma_n = \sigma_t = 0$ implies specular reflection, as in Newtonian flow, whereas $\sigma_n = \sigma_t = 1$ implies diffuse reflection, i.e., total accommodation ("sticking") of the particles to the surface followed by subsequent Maxwellian re-emission at the wall temperature. It is traditional in orbit decay studies to assume total accommodation as an improvement on the known deficiencies of the Newtonian model. In practice this is never true, with $\sigma_n = 0.9$ about the maximum value observed, and σ_t somewhat less. Furthermore, both coefficients are strongly dependent on incidence angle.[43] The work of Fredo and Kaplan[42] shows these effects to be important, particularly for complex, asymmetric shapes. However, this approach requires considerable computational sophistication to implement, and is of limited usefulness in preliminary design calculations.

4.3.5 Solar Radiation Pressure

Observed solar radiation intensity at the Earth's orbit about the sun is closely approximated (to within 0.3%) by

$$I_s = \frac{1358\ \text{W/m}^2}{(1.0004 + 0.0334\cos D)} \tag{4.119}$$

where

I_s = integrated intensity (in W/m^2) on the area normal to the sun
D = phase of year [$D = 0$ on July 4 (aphelion)]

given by Smith and Gottlieb.[44] Note that 1358 W/m^2 is the mean intensity observed at a distance of 1 A.U.; solar radiation intensity for planets at other distances from the sun is computed from the inverse square law. For practical purposes in spacecraft design, the observed intensity of Eq. (4.119) is essentially that due to an ideal blackbody at 5780 K.[45] This assumption is especially convenient when it is necessary to consider analytically the spectral distribution of radiated solar power. It is of course erroneous in that it does not account for the many absorption lines in the solar spectrum due to the presence of various elements in the sun's atmosphere.

We note that intensity has dimensions of power per unit area, and that power is the product of force and the velocity at which the force is applied. Solar radiation thus produces an effective force per unit area, or pressure, given by

$$p_s = \frac{I_s}{c} = 4.5 \times 10^{-6}\ \text{N/m}^2 \tag{4.120}$$

where c is the speed of light in vacuum.

As we have seen, the passage of a satellite through the upper atmosphere produces a drag force resulting from the dynamic pressure, $\frac{1}{2}\rho V^2$, due to the upper atmosphere density ρ and the orbital velocity V. With $V \simeq 7.6$ km/s, it is found that a density of $\rho \simeq 1.55 \times 10^{-13}\ \text{kg/m}^3$ will produce a dynamic pressure equal to the solar radiation pressure from Eq. (4.118). In the standard atmosphere model, this density is found at an altitude of roughly 500 km. At 1000 km, aerodynamic drag produces only about 10% of the force due to solar radiation pressure. Thus, at many altitudes of interest for Earth orbital missions, solar radiation pressure exerts a perturbing force comparable to or greater than atmospheric drag.

Solar radiation pressure is fundamentally different from aerodynamic drag in that the force produced is in the antisolar direction, rather than always opposite the spacecraft velocity vector. The resulting effects may for some orbits average

nearly to zero over the course of an orbit but are generally not confined to variations in the eccentricity and semimajor axis, as with aerodynamic drag. Depending on the orbit and the orientation and symmetry properties of the spacecraft, changes in all orbital elements are possible. Of course, solar radiation pressure does not act on a spacecraft during periods of solar occultation by the Earth or other bodies.

The perturbing effects of solar radiation pressure can be deliberately enhanced by building a spacecraft with a large area-to-mass ratio. This is the so-called solar sail concept. If the center of mass of the vehicle is maintained near to or in the plane of the sail to minimize "weathervane" or "parachute" tendencies, then the sail can provide significant propulsive force normal to the sun vector, allowing relatively sophisticated orbital maneuvers. Solar sails have been proposed by many authors for use in planetary exploration.[46] The solar sail concept enables some missions, such as a Mercury Orbiter, to be performed in a much better fashion than with currently available or projected chemical boosters. Optimal use of solar radiation pressure to maximize orbital energy and angular momentum has been studied by Van der Ha and Modi.[47]

Satellites in high orbits, such as geosynchronous communications or weather satellites, must be provided with stationkeeping fuel to overcome the long-term perturbations induced by solar radiation pressure. Planetary missions must similarly take these effects into account for detailed trajectory calculations. And, as will be seen in Chapter 7, the force produced by solar radiation can be important in attitude control system design.

Solar radiation produces a total force upon a spacecraft given by

$$F_s = KSp_s \tag{4.121}$$

where

K = accommodation coefficient
S = projected area normal to sun

K is an accommodation coefficient characterizing the interaction of the incident photons with the spacecraft surface. It is in the range $1 \leq K \leq 2$, where $K = 1$ implies total absorption of the radiation (ideal blackbody), and $K = 2$ implies total specular reflection back along the sun line (ideal mirror).

Solar radiation is to be distinguished from the solar wind, which is a continuous stream of particles emanating from the sun. The momentum flux in the solar wind is small compared with that due to solar radiation.

4.4 Basic Orbital Maneuvers

Many spacecraft, especially those intended for unmanned low Earth orbital missions, pose only loose requirements on orbit insertion accuracy and need no

orbit adjustments during the mission. In other cases, nominal launch vehicle insertion accuracies are inadequate, or the desired final orbit cannot be achieved with only a single boost phase, and postinjection orbital maneuvers will be required. Still other missions will involve frequent orbit adjustments to fulfill basic objectives. In this section, we consider simple orbital maneuvers including plane changes, one- and two-impulse transfers, and combined maneuvers.

In the following we assume impulsive transfers, i.e., the maneuver occurs in a time interval that is short with respect to the orbital period. Because orbit adjustment maneuvers typically consume a few minutes at most and orbital periods are 100 min or longer, this is generally a valid approximation. In such cases, the total impulse (change in momentum) per unit mass is simply equal to the change in velocity ΔV.

This quantity is the appropriate measure of maneuvering capability for spacecraft that typically are fuel limited. To see this, note that, if a thruster produces a constant force F on a spacecraft with mass m for time interval Δt, we obtain upon integrating Newton's second law

$$\Delta V = \left(\frac{F}{m}\right)\Delta t \tag{4.122}$$

Since F/m will be essentially constant during small maneuvers, and since the total thruster on-time is limited by available fuel, the total ΔV available for spacecraft maneuvers is fixed and is a measure of vehicle maneuver performance capability.

4.4.1 Plane Changes

Most missions, in cases where any orbit adjustment at all is required, will require some adjustment of the orbital plane. This plane (see Fig. 4.13) is perpendicular to the angular momentum vector $\boldsymbol{h}$, which for a Keplerian orbit is permanently fixed in space. A pure plane rotation alters $\boldsymbol{h}$ in direction but not in magnitude and thus requires an applied torque normal to $\boldsymbol{h}$. This in turn requires the application of a force on the spacecraft, e.g., a thruster firing, parallel to $\boldsymbol{h}$. For pure plane rotation we see from the geometry of Fig. 4.13 and the law of cosines that the change in angular momentum is related to the rotation angle ν by

$$\Delta h = h[2(1 - \cos\nu)]^{1/2} = 2h\sin\left(\frac{\nu}{2}\right) \tag{4.123}$$

and since $h = rV_\theta$, the required impulse is

$$\Delta V = 2V_\theta \sin\left(\frac{\nu}{2}\right) \simeq V_\theta \nu \tag{4.124}$$

where the last equality is valid for small ν.

This impulse is applied perpendicular to the initial orbit plane. There results a node line between the initial and final orbits running through the point where the

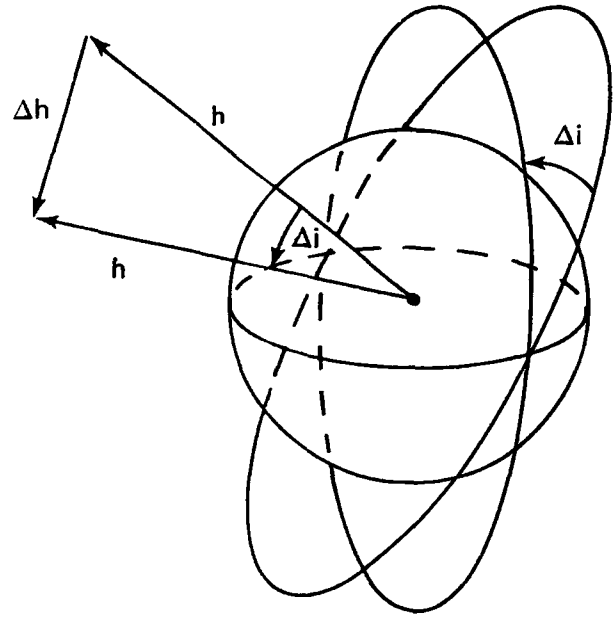


Fig. 4.13 Plane rotation.

thrust is applied. Because, in fact, the location of this node line is usually determined by mission requirements, the timing of the maneuver is often fixed. Maximum spatial separation between the initial and final orbits occurs $\pm 90°$ away from the point of thrust application.

Note that, if the impulse is applied at the line of nodes of the original orbit (i.e., in the equatorial plane of the coordinate system in use), then the plane rotation will result in a change of orbital inclination only, with $\Delta i = \nu$. If the impulse is applied at $\omega + \theta = 90°$, i.e., at a point in the orbit where the radius vector is perpendicular to the line of nodes, the orbit will precess by the amount $\Delta\Omega = \nu$ without altering its inclination. Maneuvers initiated at other points will alter both i and Ω.

In practice, adjustments for both i and Ω may be required. Adjustments to Ω are required to compensate for timing errors in orbit insertion (which may not be errors at all and may be, as for interplanetary missions, unavoidable due to the existing planetary configuration). Alterations to inclination angle are required to compensate for azimuthal heading (see Sec. 4.2.8) errors at injection. Adjustments to Ω and i can be done separately, but the typical maneuver is a single plane rotation executed at the node line between the initial and desired orbits. Figure 4.14 shows the spherical triangle that is applicable to this case. From the spherical triangle law of cosines, the required plane rotation is

$$\cos \nu = \cos i_1 \cos i_2 + \sin i_1 \sin i_2 \cos (\Omega_2 - \Omega_1) \qquad (4.125)$$

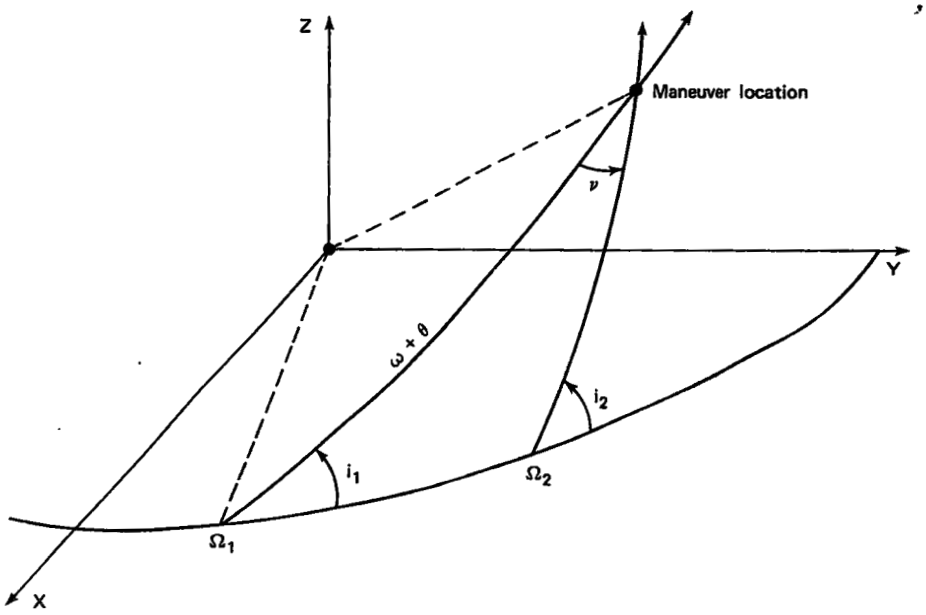


Fig. 4.14 General plane change maneuver.

and from the spherical triangle law of sines, the maneuver is performed at a true anomaly in the initial orbit found from

$$\sin(\omega + \theta) = \frac{\sin i_2 \sin(\Omega_2 - \Omega_1)}{\sin \nu} \tag{4.126}$$

Two locations in true anomaly Δ are possible, corresponding to the choice of $\pm \nu$ at the two nodes. To minimize ΔV requirements, the maneuver should be executed at the node with the largest radius vector.

Interestingly, a plane rotation executed with a single impulsive burn at the line of nodes is not always a minimum-energy maneuver. Lang[48] shows that if the desired node line and the orbital eccentricity are such that

$$e > |\cos \omega^*| \tag{4.127}$$

where ω^* is the central angle from the desired node location to periapsis in the initial orbit, then a two-impulse transfer is best, with the maneuvers occurring at the minor axis points. In this case, the total impulse is given by

$$\Delta V = \nu |\sin \omega^*| \sqrt{\frac{\mu}{a}} \tag{4.128}$$

applied in the ratio

$$\frac{\Delta V_1}{\Delta V_2} = \frac{-\sin(\omega^* + \theta_2)}{\sin(\omega^* + \theta_1)} \tag{4.129}$$

Differences between the one- and two-impulse transfers can be significant if eccentricity is large. Lang shows that, for $e = 0.7$ and $\omega^* = 90$ or 270°, a 29% savings is realized using the optimal maneuver. For $e < 0.1$, savings are always less than 10%.

Equation (4.124) shows that plane changes are expensive; a 0.1-rad rotation for a 200-km circular parking orbit requires a ΔV of approximately 0.78 km/s. The relative expense of plane changes compared with, for example, perigee adjustments, has produced considerable interest in the use of aerodynamic maneuvers in the upper atmosphere to effect plane rotations.[49] Such possibilities are clearly enhanced for vehicles such as the space shuttle, which have a significant lift vector that can be rotated out of the plane of the atmospheric entry trajectory.

Mission requirements specifying node location may well be in conflict with the desire to minimize fuel expenditure. It may be noted that ΔV requirements are minimized by executing the maneuver when V_θ is smallest, i.e., at the apoapsis of an elliptic orbit. Unless the required node line location coincides with the line of apsides of the initial orbit, this minimum-impulse maneuver cannot be achieved.

A simple example is found in the deployment of a communications satellite into a geostationary, hence equatorial, orbit. As seen in Sec. 4.2.8, the minimum inclination orbit for launch from Cape Canaveral is 28.5°. The satellite will either be injected directly into a highly elliptic transfer orbit with apoapsis at geostationary altitude, or into a nearly circular parking orbit and then later into the transfer orbit. Plane rotation must be done over the equator, and it is highly desirable that it be done at apogee. Thus, the initial launch (or the maneuver into the transfer orbit) must be timed to cause the apogee to be so placed. This is most easily accomplished from a circular parking orbit, which is one reason why initial injection into such an orbit is typically a part of more complex mission sequences.

However, such freedom is not always available. Interplanetary missions provide a ready example. Although most of the planets lie close to the plane of Earth's orbit (ecliptic plane), none is exactly in it; and hence, heliocentric orbit plane changes are always required for interplanetary transfer, unless the mission can be timed to allow the target planet to be intercepted when it is at its heliocentric line of nodes (i.e., in the ecliptic plane). This is not often the case.

As will be seen in Sec. 4.5, a spacecraft on an outer planetary mission will at the time of intercept be at, or at least closer to, the transfer orbit apoapsis than it was at departure from the Earth. As discussed, this is the more desirable position for a plane change. However, since the plane rotation maneuver must occur 90° prior to intercept (if maximum effect is to be gained from the maneuver), the node line is thus specified and will not be at the optimal aphelion point. The magnitude

of the required plane change is equal (if the maneuver occurs 90° prior to intercept) to the ecliptic latitude of the target planet at the time of the encounter. Figure 4.15 shows the geometry. A similar situation exists for inner planetary missions, except that it is desirable to change planes as near to Earth as possible, when heliocentric velocity is least.

The plane change may be done closer to the target planet than the optimal point 90° prior to encounter. In such a case, a larger rotation angle is required because there will not be time prior to encounter for the maneuver to take maximum effect. It may be necessary to balance this loss against a gain due to performing the maneuver farther from the sun, where the tangential velocity V_θ is smaller. Appropriate use of these "broken plane" maneuvers can yield acceptable planetary transfers when no direct transfer is possible.[50]

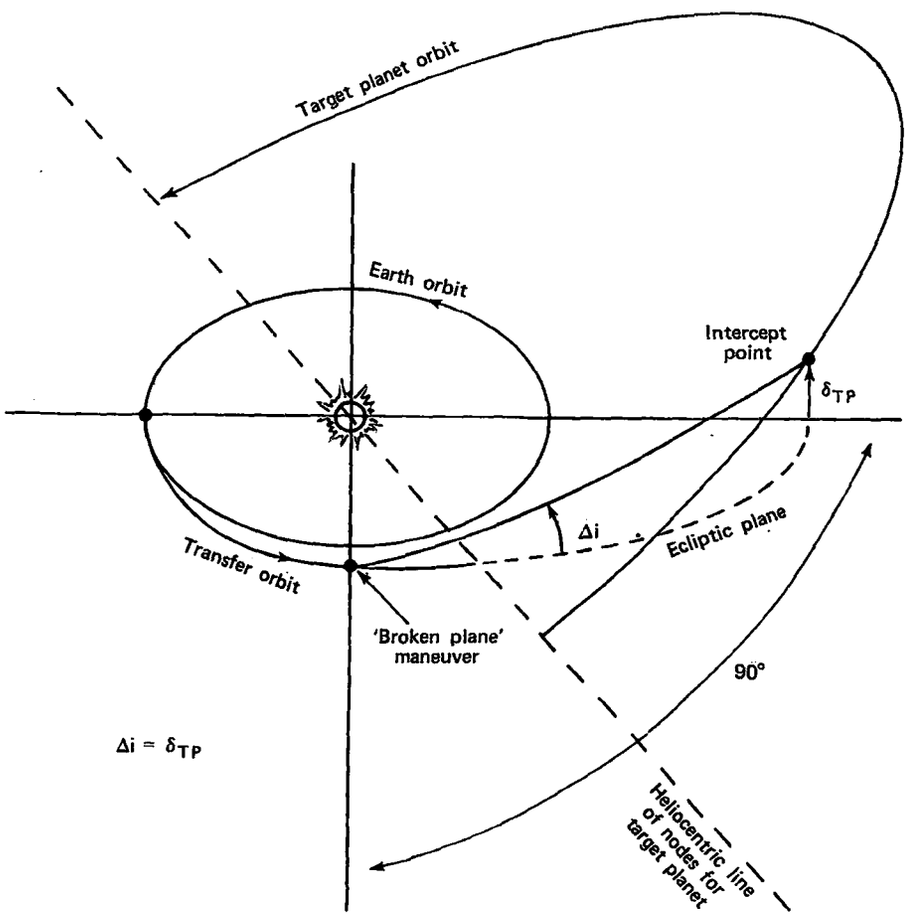


Fig. 4.15 Noncoplanar interplanetary transfer.

4.4.2 Coplanar Transfers

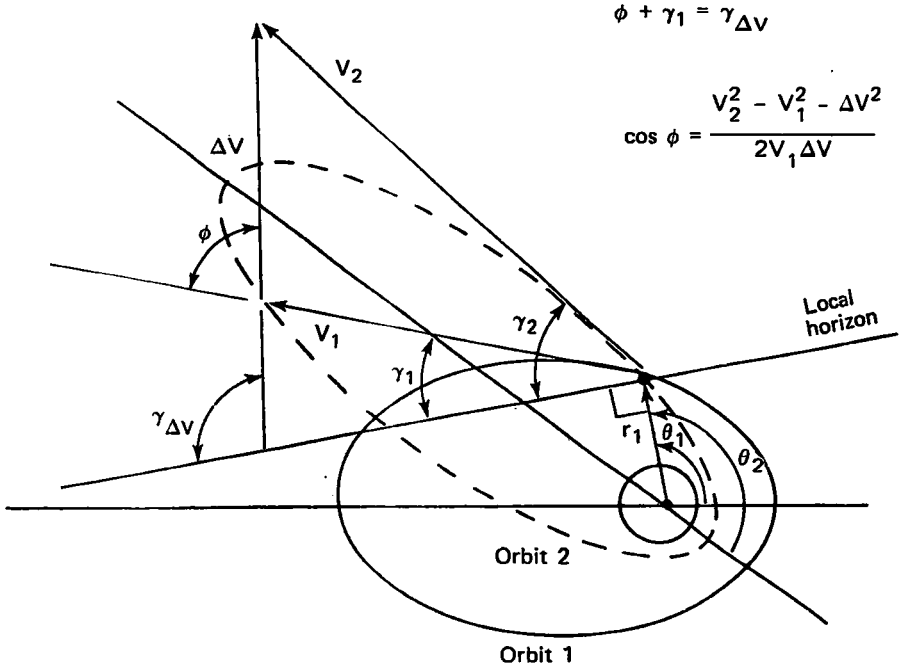
We now consider maneuvers that leave the orientation of the orbit unchanged but that may alter the elements a, e, and ω and the period τ. Because the direction of $\boldsymbol{h}$ is to remain fixed, all maneuvers must produce torques parallel to $\boldsymbol{h}$ and are thus confined to the orbit plane.

4.4.2.1 Single-impulse transfer. The geometry of a general single-impulse orbit transfer is shown in Fig. 4.16. An impulsive maneuver is executed at some position (r_1, θ_1) in the orbit plane, with velocity and flight-path angle (V_1, γ_1) yielding a new orbit (which must in the single-impulse case always intersect the old) with velocity and flight-path angle (V_2, γ_2) and true anomaly θ_2. Because the maneuver is impulsive, the radius vector $\boldsymbol{r}_1$ cannot change during its execution.

Orbit 1 is assumed to be known. In typical situations of interest it may be required to 1) determine the ΔV and heading for the maneuver, given the desired new orbit and the specified transfer point; or 2) determine V_2 and all other characteristics of orbit 2, given the impulse ΔV and the heading for the maneuver. The heading may be the pitch angle ϕ relative to the orbital velocity vector V_1, or

$$\phi + \gamma_1 = \gamma_{\Delta V}$$

$$\cos\phi = \frac{V_2^2 - V_1^2 - \Delta V^2}{2V_1\Delta V}$$



$$\Delta V^2 = V_1^2 + V_2^2 - 2V_1V_2\cos(\gamma_2 - \gamma_1)$$

Fig. 4.16 Single-impulse transfer between two intersecting orbits.

it may be specified by the flight-path angle $\gamma_{\Delta V}$ relative to the local horizontal. In either case, the maneuver satisfies the vector equation

$$V_2 = V_1 + \Delta V \tag{4.130}$$

However, for preliminary calculations it may be more convenient to use the law of cosines with the velocity vector diagram of Fig. 4.16 to determine the required information.

In case 1, where the new orbit is known and the transfer point (r_1, θ_1) is specified, V_2 is immediately found from Eq. (4.17), and γ_2 is found from Eq. (4.60). Then the impulse and heading are

$$\Delta V^2 = V_1^2 + V_2^2 - 2V_1V_2\cos(\gamma_2 - \gamma_1) \tag{4.131}$$

$$\cos\phi = \frac{(V_2^2 - V_1^2 - \Delta V^2)}{2V_1\Delta V} \tag{4.132}$$

or, from the law of sines,

$$\phi = \pi - \sin^{-1}\left[\left(\frac{V_2}{\Delta V}\right)\sin(\gamma_2 - \gamma_1)\right] \tag{4.133}$$

with

$$\phi = \gamma_1 - \gamma_{\Delta V} \tag{4.134}$$

In case 2, the new orbit is to be found given ϕ or $\gamma_{\Delta V}$. With ϕ computed from Eq. (4.134) if need be, Eq. (4.132) is used to solve for V_2, and then Eq. (4.133) is used to find γ_2. The results of Sec. 4.2.8 (Orbital Elements from Position and Velocity) are then applied to find h, p, e, a, θ, and ω.

The most important special case for a single-impulse transfer occurs when the maneuver ΔV is applied tangent to the existing velocity vector V_1. Then $\phi = 0$ or π, and Eq. (4.132) yields $V_2 = V_1 \pm \Delta V$. The tangential ΔV application thus allows the maximum possible change in orbital energy for a given fuel expenditure, adding to or subtracting from the existing velocity in a scalar fashion. Moreover, from Eq. (4.131), $\gamma_2 = \gamma_1$; hence, the flight-path angle is not altered at the point of maneuver execution.

At apogee or perigee, $V_r = 0$, $V_\theta = V_1$, and $\gamma_1 = 0$. A tangentially applied impulse alters only V_θ, leaving $V_r = 0$, and does not change the perigee or apogee radius. Because angular momentum is conserved,

$$h = rV_\theta = r_a h_a = r_p V_p \tag{4.135}$$

an increase or decrease in V_p with r_p constant must result in an increase or decrease in r_a. Thus, a tangential maneuver at one of the apsides takes effect at the opposite apsis. Because γ is not altered at the maneuver point, the line of apsides does not rotate. An example is shown in Fig. 4.17, which illustrates an apogee-raising maneuver for a satellite initially in a low circular parking orbit

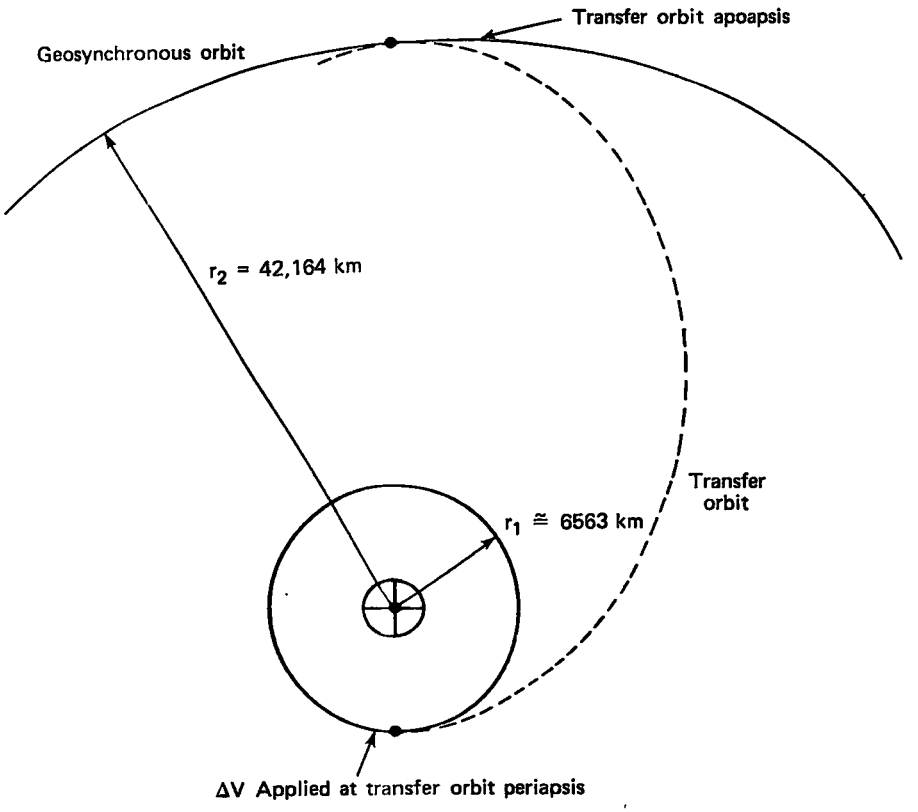


Fig. 4.17 Apogee raising maneuver from circular parking orbit.

and intended for a circular geostationary orbit. The high orbit is attained through injection into an intermediate, highly elliptic geosynchronous transfer orbit.

4.4.2.2 Two-impulse transfer. Two maneuvering impulses are required for transfer between two nonintersecting orbits. Figure 4.18 shows the required geometry. Analysis of this case requires two successive applications of the results in the previous section, first for a maneuver from orbit 1 to the transfer orbit, and then for a maneuver into orbit 2 from the transfer orbit. Again, it may be required to assess the results of particular maneuvers, or to determine the maneuvers required for a particular transfer.

Maneuvers defined by two (or more) impulses involve no particular analytical difficulty. From the trajectory design viewpoint, however, an entirely new order of complexity, and thus flexibility, is introduced. In the second case, the transfer orbit must be known to conduct the maneuver analysis. What, indeed, should the transfer orbit be to perform a particular mission? The problem of determining a suitable transfer orbit between specified end conditions, subject to appropriate constraints, is the essential problem of astrodynamics. As mentioned, the design

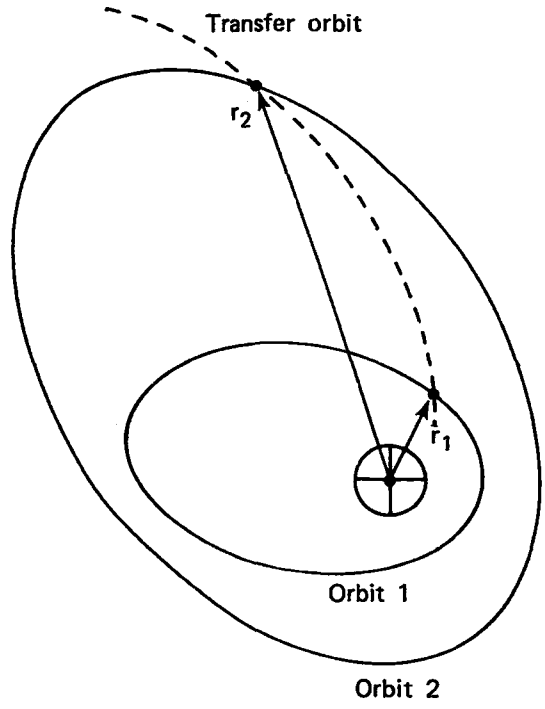


Fig. 4.18 Two-impulse transfer between nonintersecting orbits.

element is what distinguishes astrodynamics from its parent field, classical celestial mechanics. It is this activity that is the main concern of the professional astrodynamicist.

It is worth noting that, conceptually, trajectory design and orbit determination may be viewed as essentially the same problem. If the specified end conditions for the transfer are taken to be observations of an orbiting body, then determination of the "transfer orbit," if it is unique, between these positions is precisely the task of orbit determination. If the given sightings do not uniquely determine the orbit (i.e., determine $\boldsymbol{r}$ and $\boldsymbol{V}$ at a known epoch), then additional information must be obtained. To the trajectory designer, the possible lack of uniqueness between endpoints offers the freedom to apply other constraints, such as fuel usage or transfer time.

4.4.2.3 Lambert problem. The classical two-impulse trajectory design problem is the so-called Lambert, Gauss, or "time-of-flight" problem. The Gauss problem was discussed briefly in Sec. 4.2.10 (Orbit Determination). It was pointed out that the specification of two position vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ together with the flight time between them is sufficient for orbit determination. To the trajectory designer, this is equivalent to the statement that, for fixed endpoints, the possible

trajectories are parameterized according to time of flight. Typically a range of solutions having different energy requirements is available, with the shorter flight times generally (but not always) associated with higher energy requirements.

A property of conic trajectories referred to as Lambert's theorem states more specifically that the transfer time is a function of the form

$$t = t(r_1 + r_2, c, a) \tag{4.136}$$

where c is the chord length between the position vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ and a is the transfer orbit semimajor axis. Because a and E_t are related by Eq. (4.16), the rationale for the statements in the previous paragraph is clear.

Kaplan[9] and Bate et al.[15] give excellent introductory discussions of the time-of-flight problem. Modern practical work in the field is oriented toward trajectory design and is primarily the work of Battin[51] and his co-workers.[52,53]

4.4.2.4 Hohmann transfer. Figure 4.17 shows an important special case for transfer between two coplanar nonintersecting orbits. The transfer orbit is shown with an apogee just tangent to the desired geosynchronous orbit, whereas the perigee is tangent to the initial circular parking orbit. From the results of the previous section, any higher transfer orbit apogee would also allow the geosynchronous orbit to be reached. However, the transfer orbit that is tangent to both the arrival and departure orbits, called the Hohmann transfer, has the property that it is the minimum-energy, two-impulse transfer between two coplanar circular orbits.

In practice Hohmann transfers are seldom used, in part because given departure and arrival orbits are rarely both circular and coplanar. Also, Hohmann orbits are slow, a factor that may be significant for interplanetary missions.

Because of the physical constraints defining the Hohmann transfer, its ΔV requirements are easily computed. ΔV_1 for departure from a circular orbit at $\boldsymbol{r}_1$ is

$$\Delta V_1 = V_{p_{\mathrm{TO}}} - \sqrt{\frac{\mu}{r_1}} \tag{4.137}$$

(where subscript TO denotes transfer orbit), whereas upon arrival at the circular orbit at $\boldsymbol{r}_2$,

$$\Delta V_2 = \sqrt{\frac{\mu}{r_2}} - V_{a_{\mathrm{TO}}} \tag{4.138}$$

and

$$\Delta V = \Delta V_1 + \Delta V_2 \tag{4.139}$$

The transfer orbit properties are easily found; since

$$a_{\mathrm{TO}} = \frac{r_1 + r_2}{2} \tag{4.140}$$

the vis-viva equation yields

$$V_{p_{TO}}^2 = \mu\left(\frac{2}{r_1} - \frac{1}{a_{TO}}\right) \tag{4.141}$$

and

$$V_{a_{TO}}^2 = \mu\left(\frac{2}{r_2} - \frac{1}{a_{TO}}\right) \tag{4.142}$$

or by noting from Eq. (4.135) that

$$V_{a_{TO}} = V_{p_{TO}}\left(\frac{r_1}{r_2}\right) \tag{4.143}$$

Inbound and outbound transfers are symmetric; thus, no loss of generality is incurred by considering only one case. A simple example serves to illustrate the method.

Example 4.1

Compute the mission ΔV for a Hohmann transfer from a 185-km circular space shuttle parking orbit to a geosynchronous orbit at an altitude of 35,786 km above the Earth.

Solution. For Earth

$$\mu = 3.986 \times 10^5 \text{ km}^3/\text{s}^2$$
$$R_e = 6378 \text{ km}$$

Thus,

$$r_1 = 6563 \text{ km}$$
$$r_2 = 42{,}164 \text{ km}$$
$$a_{TO} = 24{,}364 \text{ km}$$

and from Eq. (4.137),

$$\Delta V_1 = 10.252 \text{ km/s} - 7.793 \text{ km/s} = 2.459 \text{ km/s}$$

Similarly, from Eq. (4.138),

$$\Delta V_2 = 3.075 \text{ km/s} - 1.596 \text{ km/s} = 1.479 \text{ km/s}$$

Hence, for the mission

$$\Delta V = 3.938 \text{ km/s}$$

4.4.3 Combined Maneuvers

It is often possible to combine a required in-plane and out-of-plane maneuver and effect a fuel savings. A practical example is that of the previous section, in which the ΔV requirement for a Hohmann transfer from a parking orbit to geosynchronous orbit was computed. To attain a geostationary orbit, a plane change of 28.5° (assuming a due-east launch from Cape Canaveral) is required.

The situation is as shown in Fig. 4.19. As in previous sections, V_1 is the existing velocity vector, V_2 is the desired vector, and

$$\Delta V = V_1^2 + V_2^2 - 2V_1V_2 \cos \Delta i \tag{4.144}$$

gives the magnitude of the velocity increment required for the transfer to effect the combined plane change and alteration of the in-plane elements a, e, and ω.

In the preceding example, $V_1 = 1.596$ km/s, $V_2 = 3.075$ km/s, and $\Delta i = 28.5$ deg; thus we find $\Delta V = 1.838$ km/s. If the maneuvers are performed separately, with the plane change first, Eq. (4.118) yields $\Delta V_1 = 0.786$ km/s, and the circularization maneuver requires $\Delta V_2 = 1.479$ km/s, as before. This produces a total mission $\Delta V = 2.265$ km/s, clearly a less efficient approach.

More general combinations of maneuvers are possible that have concomitant fuel savings over sequential ΔV applications. Investigation of optimal orbit transfers is a perennial topic of research in astrodynamics. Small[54] and Hulkower et al.[55] provide useful results in this area.

4.5 Interplanetary Transfer

The results of the previous sections are sufficient for the analysis of all basic orbital transfers, including those for interplanetary missions. However, trajectory design for such missions is sufficiently complex to justify separate discussion.

Although trajectory design for advanced mission analysis or actual mission execution will invariably be accomplished using direct numerical integration of the equations of motion, such procedures are inappropriate for preliminary assessments. Not only are the methods time-consuming and highly specialized,

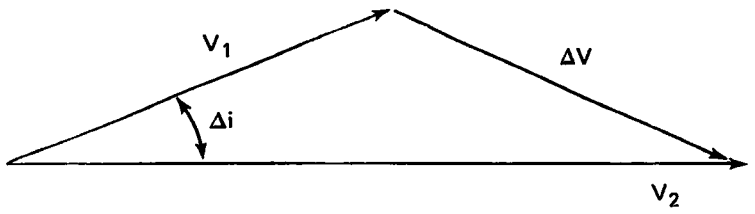


Fig. 4.19 Vector diagram for combined plane change and orbit energy adjustment.

but also, without a preliminary analytical solution, they offer no way to eliminate the many cases that are not at all close to the actual solution of interest.

For initial mission design and feasibility assessments, the so-called method of patched conics is universally employed. We consider the application of the method in some detail.

4.5.1 Method of Patched Conics

As the name implies, this approach uses a series of Keplerian orbits to define the trajectory. Each separate conic section is assumed to be solely due to the influence of the dominant body for that portion of the mission. The different segments are "patched" at the sphere of influence boundaries between different bodies given by Eq. (4.93). Thus, a spacecraft trajectory between Earth and Mars will be modeled for departures as a geocentric escape hyperbola, which at great distances from Earth becomes a heliocentric elliptic orbit, followed again by a hyperbolic approach to Mars under the influence of that planet's gravitational field.

Patched-conic trajectory designs are accomplished in three well-defined steps:

1) The heliocentric orbit from the departure planet to the target planet is computed, ignoring the planet at either end of the arc.

2) A hyperbolic orbit at the departure planet is computed to provide the "infinity" (sphere of influence boundary) conditions required for the departure end of the heliocentric orbit in step 1.

3) A hyperbolic approach to the target is computed from the infinity conditions specified by the heliocentric orbit arrival.

The statement of the patched-conic procedure shows that it cannot yield truly accurate results. A spacecraft in an interplanetary trajectory is always under the influence of more than one body, and especially so near the sphere of influence boundaries. Transitions from one region of dominance to another are gradual and do not occur at sharply defined boundaries. Keplerian orbit assumptions in these regions are incorrect, yet conveniently applicable analytical results do not exist, even for the restricted three-body problem. Other perturbations, such as solar radiation pressure, also occur and can have significant long-term effects.

The patched-conic method yields good estimates of mission ΔV requirements and thus allows quick feasibility assessments. Flight times are less well predicted, being in error by hours, days, or even weeks for lengthy interplanetary missions. Such errors are of no consequence for preliminary mission design but are unacceptable for mission execution. An encounter at the target planet must occur within seconds of the predicted time if a flyby or orbit injection maneuver is to be properly performed. For example, the heliocentric velocity of Mars in its orbit is roughly 24 km/s. If an orbit injection were planned to occur at a 500 km periapsis height, a spacecraft arriving even 10 s late at Mars would likely enter the atmosphere.

Patched-conic techniques are useful at the preliminary design level for hand calculation or for implementation in a relatively simple computer program. As stated, actual mission design and execution must employ the most accurate possible numerical integration techniques. The difference in accuracy obtainable from these two approaches can be a source of difficulty, even at the preliminary design level, for modern interplanetary missions involving application of multiple ΔV or planetary flybys and the imposition of targeting constraints and limitations on total maneuver ΔV. In such cases, the errors implicit in patched-conic approximations during early phases of the trajectory may invalidate subsequent results, and detailed numerical calculations are too cumbersome even on the fastest machines for use in preliminary analysis. Recent applications of constrained parameter optimization theory to the multiple encounter problem have resulted in relatively fast, efficient techniques for trajectory design that eliminate 90–99% of the error of simple conic methods.[50]

4.5.1.1 Heliocentric trajectory. This portion of the interplanetary transfer will usually be computed first, unless certain specific conditions required at an encounter with the target planet should require a particular value of V_∞ for the hyperbolic approach. As stated earlier, the calculation ignores the planet at each end of the transfer and thus gives the ΔV to go from the orbit of the departure planet to the orbit of the arrival planet. To be strictly correct, the departure and arrival should begin and end at the sphere of influence boundary for each planet; however, these regions are typically quite small with respect to the dimensions of the heliocentric transfer and are often ignored. Of course, calculations for Earth-moon missions cannot justifiably employ this assumption.

The heliocentric segment is not restricted to Hohmann transfers or even to coplanar transfers, though these are common assumptions in preliminary design. The assumption of coplanarity may cause serious errors in ΔV computations and should be avoided. However, given the overall accuracy of the method, the assumption of circular orbits at the departure and arrival planets is often reasonable and because of its convenience is used where possible. This assumption may not be justified for missions to planets with substantially elliptic orbits, e.g., Mercury or possibly Mars or, in the extreme case, Pluto.

The heliocentric trajectory design will usually be constrained by available launch energy, desired travel time, or both. When both are important, appropriate tradeoffs must be made, with the realization, however, that energy savings achieved through the use of near-Hohmann trajectories can be nullified—by the increased mass and/or redundant systems required because of the longer flight times. Again, minimum-energy orbits (Hohmann-type doubly tangent transfer plus a heliocentric plane change if required) are often assumed initially because of the computational convenience. If the flight times are unacceptable, a faster transfer must be used, with consequently higher ΔV requirements.

If a doubly tangent transfer orbit is assumed, then Eqs. (4.140–4.143) may be used to determine the transfer orbit characteristics. Equation (4.124) is used to

compute any additional ΔV required to match the heliocentric declination of the target planet at encounter. This cannot be done until the actual timing of the mission is determined.

It will be required to know the arrival velocity of the spacecraft relative to the target planet. This is simply the vector difference

$$V_\infty = V_s - V_p \tag{4.145}$$

between the heliocentric transfer orbit velocity and the velocity of the target planet in its heliocentric orbit, as shown in Fig. 4.20. If the minimum-energy transfer is used, this reduces to the difference in their scalar speeds. If a faster transfer is required, the heliocentric velocity vector will not be tangent to the target planet orbit at encounter. The arrival velocity relative to the target planet will then be given by the methods of Sec. 4.4.2 (Coplanar Transfers). Equations (4.131–4.134) become

$$V_\infty^2 = V_P^2 + V_S^2 - 2V_P V_S \cos(\gamma_s - \gamma_p) \tag{4.146}$$

$$\cos\phi = (V_S^2 - V_P^2 - V_\infty^2)/2V_P V_\infty \tag{4.147}$$

or, from the law of sines,

$$\phi = \pi - \sin^{-1}\left[\left(\frac{V_S}{V_\infty}\right)\sin(\gamma_s - \gamma_P)\right] \tag{4.148}$$

with

$$\gamma = \gamma_P - \phi \tag{4.149}$$

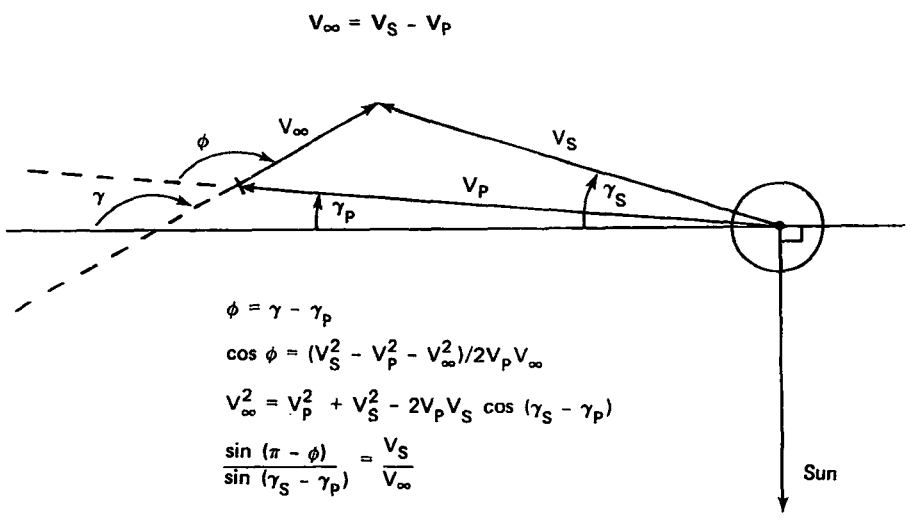


Fig. 4.20 Approach velocity at target planet.

It will always be advantageous for the transfer orbit to be tangent to that of the departure planet. This may not be possible, however, when gravity-assist maneuvers are used at intermediate planets between the departure planet and the ultimate target. The departure conditions from the intermediate planet are determined by the hyperbolic encounter with that planet, as will be seen in Sec. 4.5.2 (Gravity-Assist Trajectories).

Once a trial orbit has been assumed and the heliocentric transfer time computed, it is necessary and possible to consider the phasing or relative angular position of the departure and arrival planets for the mission. The geometric situation is shown in Fig. 4.21. Clearly, departure must occur when the relative planetary positions are located such that, as the spacecraft approaches the target planet orbit in the transfer trajectory, the planet is there also. Assuming the transfer time has been found, the difference in true anomaly between departure and arrival planets at launch is

$$\Delta\theta = (\theta_{TO_A} - \theta_{TO_D}) - (\theta_{TP_A} - \theta_{TP_D}) \tag{4.150}$$

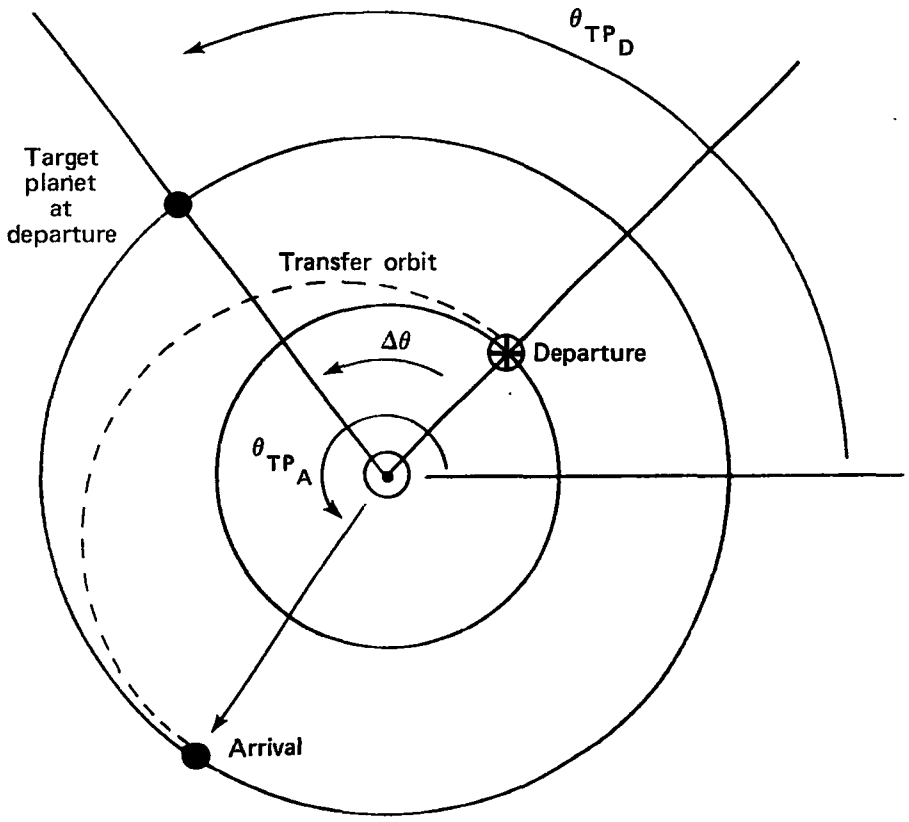


Fig. 4.21 Phasing for interplanetary transfer.

where, from Eq. (4.7), the true anomaly is

$$\theta = \cos^{-1}\left[\frac{(p/r - 1)}{e}\right] \tag{4.151}$$

for any conic orbit. If the departure or arrival planet orbit is circular, $\theta(t) = n(t - t_0)$, whereas if near-circularity can be assumed, $\theta(t)$ may be obtained from Eq. (4.32).

If coplanar circular planetary orbits are assumed, then no difference in ΔV requirements is found between missions executed at different calendar times. In fact, substantial advantages exist for missions that can be timed to encounter the target planet near its heliocentric line of nodes (implying minimum plane change requirements) or when the combination of Earth and target planet perihelion and aphelion phasing is such as to minimize the semimajor axis of the required transfer orbit. For example, an Earth–Mars minimum-energy transfer orbit can have a semimajor axis from 1.12 A.U. to 1.32 A.U., resulting in a ΔV difference at Earth departure of about 500 m/s.

4.5.1.2 Departure hyperbola. When the heliocentric transfer has been computed, the required spacecraft velocity in the neighborhood of the departure planet is found from the vis-viva equation. This velocity is in heliocentric inertial coordinates; the departure planet will itself possess a considerable velocity in the same frame. In the patched-conic method, the heliocentric transfer is assumed to begin at the sphere of influence boundary between the departure planet and the sun. This boundary (see Table 4.1) may be assumed to be at infinity with respect to the planet. As indicated by Eq. (4.145), the planetary departure hyperbola must therefore be designed to supply $\boldsymbol{V}_\infty$, the vector difference between the spacecraft transfer orbit velocity $\boldsymbol{V}_s$ and the planetary velocity $\boldsymbol{V}_p$. Again, when $\boldsymbol{V}_s$ is parallel to $\boldsymbol{V}_p$, $\boldsymbol{V}_\infty$ is their simple scalar difference.

We assume for convenience in this discussion that departure is from Earth. If the departure maneuver is executed with zero geocentric flight-path angle γ, the maneuver execution point will define the periapsis location for the outbound half of a hyperbolic passage, discussed in Sec. 4.2.4 (Motion in Hyperbolic Orbits).

The geocentric hyperbola must be tangent at infinity to the heliocentric transfer orbit; hence, the orientation of the departure asymptote is known in the heliocentric frame. The excess hyperbolic velocity $\boldsymbol{V}_\infty$ is also known from Eq. (4.145), and r_p is usually fixed by parking orbit requirements (r_p is typically about 6600 km for Earth). Therefore, θ_a, e, ΔV and the departure location in the heliocentric frame are specified by Eqs. (4.36–4.38). Equation (4.39) allows the offset β for the passage to be computed if desired. Figure 4.22 shows the geometry for hyperbolic departure.

4.5.1.3 Encounter hyperbola. The determination of $\boldsymbol{V}_\infty$ relative to the target planet from the heliocentric transfer orbit has been discussed. The

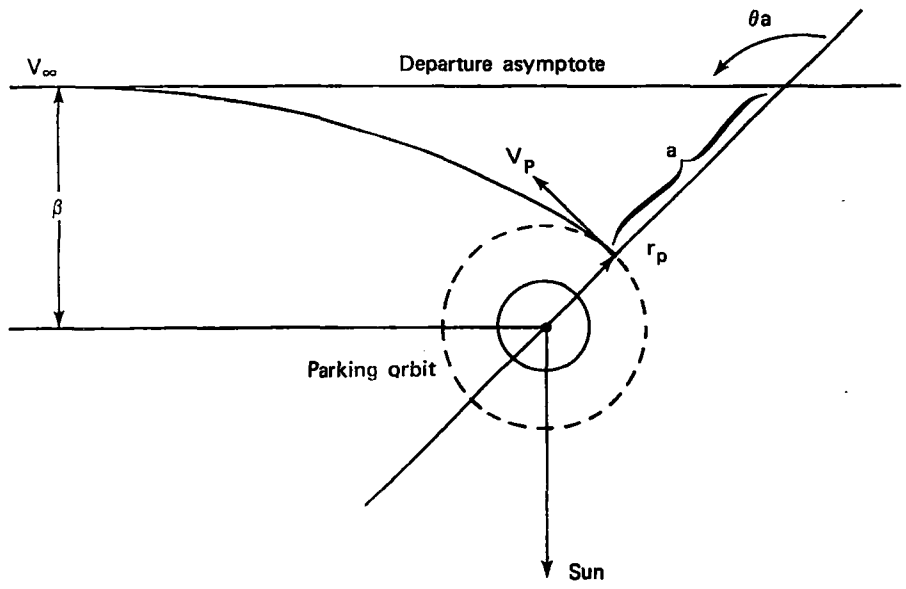


Fig. 4.22 Hyperbolic departure geometry.

encounter orbit at the target planet is shown in Fig. 4.23 in a frame centered in the planet. The results of Sec. 4.2.4 (Motion in Hyperbolic Orbits) again allow the required parameters to be found.

Operational requirements for the encounter will usually differ somewhat from those for departure. The periapsis radius is found from the approach parameters β and V_∞ from Eq. (4.41). This will be of interest for flyby and orbital-injection missions, where a particular periapsis altitude may be appropriate for photography or to attain a desirable orbit about the planet. For passages where a gravity assist is required to allow a continuation of the mission to another planet or moon, the turning angle Ψ of the passage will be critical, and β and r_p will be adjusted accordingly.

Note that impact is achieved for $r_p \leq R$, the planetary radius. From Eq. (4.42), the B-plane offset for impact is then given by

$$\beta_{\text{impact}} \leq R\left(\frac{1+2\mu}{RV_\infty^2}\right)^{1/2} \tag{4.152}$$

The term in parentheses will be somewhat greater than unity; thus, a large targeting area at "infinity" funnels down to a considerably smaller planetary area. This is of course due to the attractive potential of the planet and is referred to as its collision cross section.

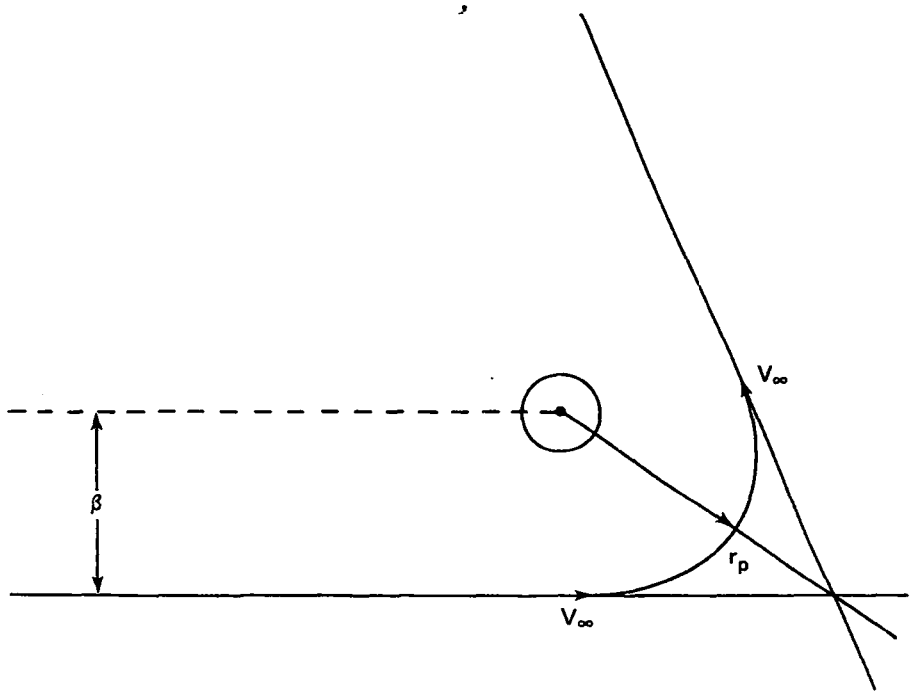


Fig. 4.23 Hyperbolic encounter in target planet frame of reference.

Atmospheric entry and braking without direct planetary impact will require targeting for a small annulus in the atmosphere above the planet, such that

$$R + h_{\min} < r_p < R + h_{\max} \tag{4.153}$$

The minimum acceptable entry height $h_{\min}$ is usually determined by the maximum acceptable dynamic loading due to atmospheric deceleration. The maximum limit $h_{\max}$ will often be fixed by the requirement to avoid "skip out," or by entry heating constraints, or both. These topics are discussed more fully in Chapter 6, but for our purposes here it is sufficient to recognize that an entry corridor of width Δr_p centered at periapsis radius r_p will exist. Targeting for this corridor must be done in the B-plane, at "infinite" distance from the planet.

The B-plane annular width mapping into an annular region near periapsis is given by differentiating Eq. (4.42), yielding

$$\Delta\beta = \left(\frac{\beta}{r_p} - \frac{\mu}{\beta V_\infty^2}\right)\Delta r_p \tag{4.154}$$

Equation (4.154) may be used to determine the B-plane targeting requirement for the encounter to ensure hitting the entry corridor of Eq. (4.153).

4.5.2 Gravity-Assist Trajectories

Upon completion of the hyperbolic passage at a target planet, the approach velocity V_∞ relative to the planet will have been turned through an angle Ψ. In the heliocentric inertial frame, the encounter thus produces the result of Fig. 4.24. As seen by applying Eq. (4.145) on arrival and departure, V_p and V_∞ do not change during the passage, but V_∞ changes because it is turned through angle Ψ, with the result that the spacecraft velocity V_{S_D} in the inertial frame is altered with respect to V_{S_A}.

It will be required to know the values of V_{S_D} and γ_{S_D} to compute V_{S_D} upon exit from the hyperbolic passage. Examination of Fig. 4.24, with ΔV and Ψ known from Eq. (4.37), yields

$$V_{S_D}^2 = V_{S_A}^2 + \Delta V^2 - 2V_{S_A}\Delta V \cos \nu \tag{4.155}$$

$$\nu = \frac{3\pi}{2} + \frac{\Psi}{2} - \phi_A + \gamma_{S_A} - \gamma_P \tag{4.156}$$

$$\phi_A = \pi - \sin^{-1}\left[\left(\frac{V_{S_A}}{V_\infty}\right)\sin(\gamma_{S_A} - \gamma_P)\right] \tag{4.157}$$

$$\gamma_{S_D} = \gamma_{S_A} + \sin^{-1}\left[\left(\frac{\Delta V}{V_{S_D}}\right)\sin \nu\right] \tag{4.158}$$

Figure 4.24 depicts a situation in which the heliocentric energy of the spacecraft is increased (at the infinitesimal expense of that of the planet) as a result of the hyperbolic passage. This would be applicable to missions such as Voyager 1 and 2, in which encounters at Jupiter were used to direct the two spacecraft toward Saturn (and, for Voyager 2, subsequently to Uranus and Neptune) much more efficiently than by direct transfer from Earth. Use of this technique enabled the Galileo mission both in reaching Jupiter and in the many satellite encounters that followed. Deprived, primarily because of political considerations, of an Earth departure stage that could send it directly to Jupiter, Galileo was launched by a two-stage IUS toward Venus. One Venus flyby and two Earth flybys endowed the spacecraft with sufficient energy to reach Jupiter, albeit at the cost of greatly increased flight time. Once in elliptic orbit around Jupiter, each encounter of a satellite was used not only to take data but also to set up another satellite encounter on the next orbit. This allowed for more encounters than would have been possible using a propulsive system alone. The once-exotic gravity-assist technique was thus reduced to a routine flight operations tool.

Energy-loss cases are equally possible. One application is to inner planetary missions such as Mariner 10, which reached Mercury via a pioneering gravity-assist maneuver performed at Venus. Similarly, the Ulysses spacecraft was directed toward the sun by means of an energy-loss maneuver at Jupiter. The same maneuver tipped the heliocentric orbit inclination sufficiently far out of the ecliptic plan that Ulysses obtained good views of the previously unseen polar

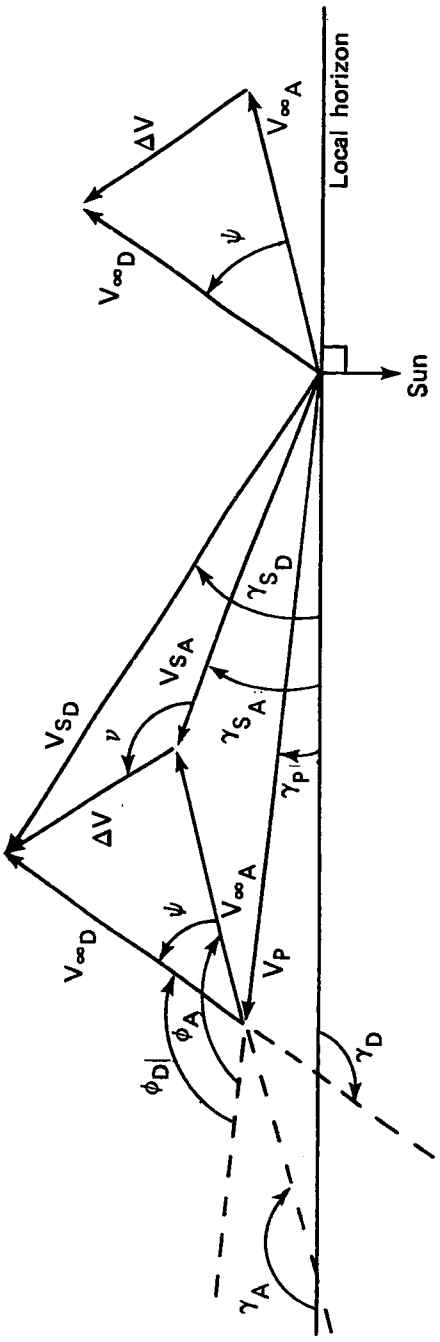


Fig. 4.24 Spacecraft energy gain in inertial frame during hyperbolic passage.

regions of the sun. Figure 4.25 shows a typical situation involving heliocentric energy loss following the encounter.

It will be noted that in Fig. 4.24 the planetary-relative approach vector $V_{\infty A}$ is rotated through angle Ψ in a counterclockwise or positive sense to obtain $V_{\infty D}$, whereas the opposite is true in Fig. 4.25. These are typical, though not required, situations producing energy gain at outer planets and energy loss at inner planets. Consideration of the encounter geometry will show that clockwise rotation of Ψ occurs for spacecraft passage between the target body and its primary (e.g., between a planet and the sun), whereas counterclockwise or positive rotation results from passage behind the target body as seen from its primary. This is shown conceptually in Fig. 4.26. Spacecraft heliocentric energy gain or loss as a result of the encounter depends on the orientation of v_{S_A} and the rotation angle Ψ of the relative approach velocity $V_{\infty A}$ during encounter.

Equation (4.37) shows that the maximum-energy gain or loss occurs if $\Psi = 180°$; in this case scalar addition of V_∞ to V_P results. This is an idealized situation requiring $r_p = 0$ for its implementation. Actually, the maximum heliocentric ΔV obtainable from an encounter occurs for a grazing passage with $r_P = R$. Note that a Hohmann trajectory yields a transfer orbit that is tangent to the target body orbit; V_p and V_{S_A} are thus colinear. Examination of Fig. 4.24 or 4.25 shows that in this case energy can only be gained (for outbound transfers) or lost (for inbound transfers), regardless of the sign of the rotation angle Ψ. Non-Hohmann transfers allow a wider range of encounter maneuvers.

The gravity-assist maneuver for planetary exploration has arrived as a mature technique since its initial use on the Mariner 10 mission. The Galileo Jupiter

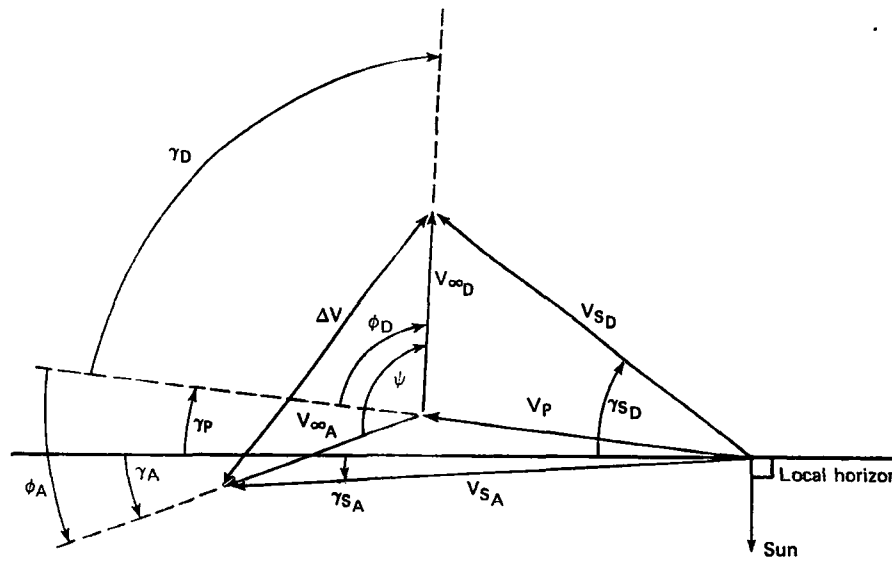


Fig. 4.25 Spacecraft energy loss in inertial frame during hyperbolic passage.

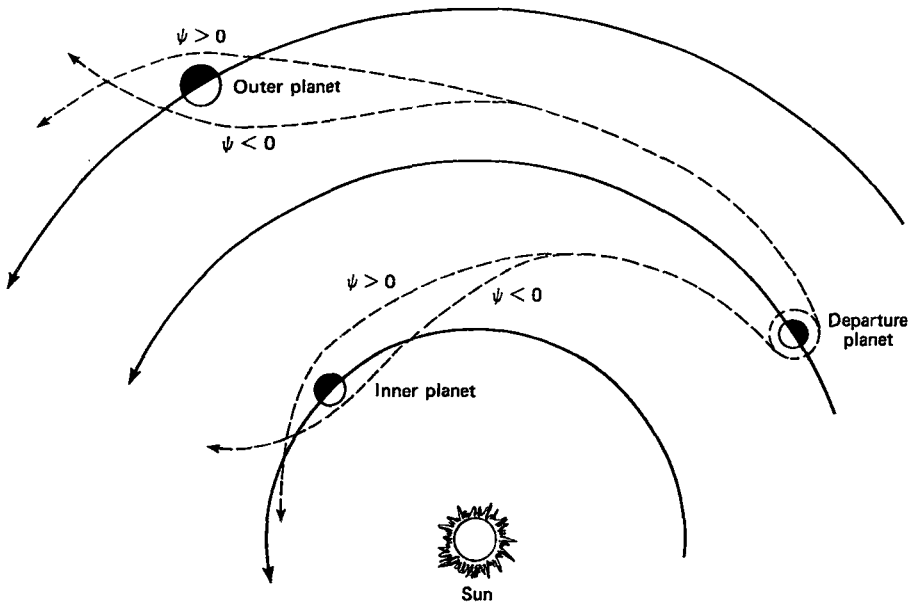


Fig. 4.26 Schematic for hyperbolic passage.

orbital mission made extensive use of gravity assists in the Jovian system to allow the spacecraft to be directed from one moon to another. The first spacecraft to encounter the tail of a comet, ISEE-3 (renamed the International Cometary Explorer), was directed to the comet Giacobini-Zinner in late 1983 via a gravity assist from the moon.

4.5.3 Lunar Transfer

The problem of calculating lunar transfer trajectories is conceptually similar to that of interplanetary transfer analysis, and, indeed, the method of patched conics can be used for preliminary assessment of mission requirements. However, the results obtained are considerably less satisfactory than for interplanetary transfers due to a number of complicating factors.

The masses of the Earth and moon are more nearly equal than for any other primary and satellite (excluding Pluto and Charon) in the solar system. The moon's sphere of influence is therefore large with respect to the Earth-moon separation, and a spacecraft in transit between the two spends much of its time close to the sphere-of-influence boundary. Also, the sun's influence on the trajectory is significant. For these reasons, the patched-conic method is less accurate than for heliocentric transfers and yields truly useful results only for outbound ΔV calculations.

Accurate results are also obtained with considerably more trouble than for interplanetary trajectories, because of the size of the lunar sphere of influence in relation to the transfer orbit dimensions. This implies that the location of the point where the spacecraft crosses the boundary is important in determining the characteristics of the transfer orbit, a fact that adds considerable complexity to the numerical procedures.

We do not include an extended treatment of lunar transfer calculations here. Cursory mission requirements can be assessed by the methods of Sec. 4.5.1 (Method of Patched Conics); more detailed analysis must be done via numerical integration of the equations of motion, possibly using the patched-conic solution as an initial guess. Bate et al.[15] give an excellent discussion of the use of patched-conic techniques in lunar transfer calculations. Their treatment includes non-coplanar transfer analyses, important in this case because of the relatively large lunar orbit inclination (which is in fact not constant, but varies between 18.2° and 28.5° over an 18.6-yr period) in the GCI frame.

4.6 Perturbation Methods

We have on several occasions mentioned that truly Keplerian orbits are essentially nonexistent and have given methods for analyzing some of the perturbations to Keplerian orbits that are important in spacecraft and mission design. Perturbation theory forms an elaborate structure in astrodynamics and classical celestial mechanics and, indeed, comprises much of the current literature in the latter subject. Such topics are completely beyond the scope of this text. However, many of the results cited earlier are due to perturbation theory, and a brief outline of this topic is in order.

Perturbation methods are broadly divided into special and general theories. Special perturbation theory is ultimately characterized by the direct numerical integration of the equations of motion due to a dominant acceleration and one or more small perturbing accelerations. As with all numerical analysis, results are unique to the given case, and it is often unclear how to extrapolate the results of one situation to another case of interest.

General perturbation analysis, historically the first approach to be developed, proceeds as given earlier, except that the perturbing accelerations are integrated analytically, at least to some given order of accuracy. Because closed-form integration of given perturbing accelerations will rarely be possible, series expansion to a desired order of accuracy is used to represent the perturbation, and the series integrated term by term. Analytical results are thus available, and broader applications and more general conclusions are possible. Nearly all important results have been obtained through general perturbation methods; on the other hand, special perturbation techniques are more applicable to practical mission design and execution.

Common special perturbation techniques are the methods of Cowell and Encke and the method of variation of parameters. Cowell's method is conceptually the simplest, at least in the era of digital computers, and consists of directly integrating the equations of motion, with all desired perturbing accelerations included, in some inertial frame. The method is uncomplicated and readily amenable to the inclusion of additional perturbations if a given analysis proves incomplete. The primary pitfalls are those associated with the use of numerical integration schemes by the unwary. Reference to appropriate numerical analysis texts and other sources[56] is recommended even if standard library procedures are to be used. Cowell's method is relatively slow, a factor that is in the modern era often irrelevant. The speed of the method is increased substantially, with only slight complexity, by employing spherical coordinates (r, θ, ϕ) instead of Cartesian coordinates.

Encke's method antedates Cowell's, which is not surprising because the latter posed formidable implementation requirements in the precomputer era. Encke's method also employs numerical integration techniques, but proceeds by integrating the difference between a given reference orbit (often called the osculating or tangent orbit) and the true orbit due to the perturbing acceleration. Because the perturbation is assumed small (a possible pitfall in the application of Encke's method), the difference between the true and reference orbits is presumably small, and larger integration step sizes can be used for much of the orbit. Encke's method, depending on the situation, executes from 3 to 10 times faster than Cowell's.

The method of variation of parameters is conceptually identical to that of general perturbation analysis, with the exception that the final step of series expansion and term by term integration is skipped in favor of direct numerical integration. In this sense, it is something of a compromise method between special and general perturbations. For example, the effects due to nonspherical primary mass distributions discussed in Sec. 4.3 are analyzed by the variation of parameters method. The results yield analytical rather than numerical forms for the variation of the parameters or elements (Ω and ω in this case) by obtaining $d\Omega/dt$, $d\omega/dt$, etc. This allows more interesting general conclusions to be drawn than with a purely numerical approach; however, complete analysis of the final effects must still be done numerically.

4.7 Orbital Rendezvous

Orbital rendezvous and docking operations are essential to the execution of many missions, particularly those involving manned spaceflight. First proven during the manned Gemini flights of 1965 and 1966, rendezvous and docking was a required technique for the Apollo lunar landing missions and the Skylab program. It is essential for space shuttle missions involving satellite retrieval, inspection, or repair as well as assembly and support missions to the International

Space Station. Unmanned, ground-controlled rendezvous and docking procedures have been demonstrated on many Russian flights and have been proposed as an efficient technique for an unmanned Mars sample return mission.[57] In this section, we discuss rendezvous orbit dynamics and procedures.

4.7.1 Equations of Relative Motion

Preliminary rendezvous maneuvers, often called phasing maneuvers, may well be analyzed in an inertial frame such as GCI and carried out using the methods of Sec. 4.4. However, the terminal phase of rendezvous involves the closure of two vehicles separated by distances that are small (e.g., tens or hundreds of kilometers) relative to the dimensions of the orbit. It is then expected that the difference in acceleration experienced by the two vehicles is relatively small and thus that their differential motion might easily be obscured by their gross orbital motion. Also, guidance algorithms are generally described in terms of the position and velocity of one vehicle relative to another. For these reasons, a description of the orbital motion and maneuvers in a planetary-centered reference frame is often inappropriate for rendezvous analysis. Instead, it is customary to define a target vehicle (TV) and a chase vehicle (CV) and to describe the motion of the chase vehicle in a noninertial coordinate frame fixed in the target vehicle. In this way, one obtains the equations of relative motion between the vehicles.

The coordinate frame for the analysis is shown in Fig. 4.27. It is assumed that the two orbits are in some sense "close," having similar values of the elements a, e, i, and Ω. $\boldsymbol{R}_T$ and $\boldsymbol{R}_C$ are the inertial vectors to the target and chase vehicles, respectively, and $\boldsymbol{r}$ is their separation vector,

$$\boldsymbol{r} = \boldsymbol{R}_C - \boldsymbol{R}_T \tag{4.159}$$

initially assumed to be small. The frame in which $\boldsymbol{r}$ is expressed is centered in the TV, and it is convenient to use the rotating local vertical system (r, s, z) shown in Fig. 4.27, where r is parallel to the TV radius vector $\boldsymbol{R}_T$, s is normal to r in the orbit plane, and z is perpendicular to the TV orbit plane. In this system, the vector equations of motion for a CV maneuvering with acceleration $\boldsymbol{a}$ and a nonmaneuvering TV are, from Eq. (4.6),

$$\frac{\mathrm{d}^2\boldsymbol{R}_C}{\mathrm{d}t^2} + \left(\frac{\mu}{R_C^3}\right)\boldsymbol{R}_C = \boldsymbol{a} \tag{4.160}$$

$$\frac{\mathrm{d}^2\boldsymbol{R}_T}{\mathrm{d}t^2} + \left(\frac{\mu}{R_T^3}\right)\boldsymbol{R}_T = 0 \tag{4.161}$$

Equations (4.160) and (4.161) are then differenced and combined with Eq. (4.159), using several simplifying vector identities that assume small r, to yield the equation of motion of the CV in the TV frame. A variety of linearized equations may then be obtained, depending on the simplifying assumptions used.

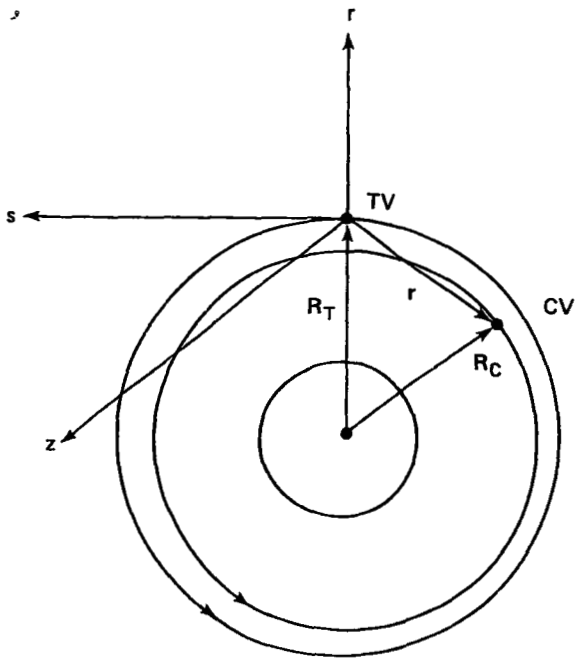


Fig. 4.27 Rotating local vertical coordinates.

If the TV and CV orbits are both nearly circular with similar semimajor axes and orbital inclinations, the basic relative motion equations first given by Hill in 1878 and subsequently rediscovered by Clohessy and Wiltshire[58] apply:

$$\frac{d^2r}{dt^2} - 2n\frac{ds}{dt} - 3n^2r = a_r$$

$$\frac{d^2s}{dt^2} + 2n\frac{dr}{dt} = a_s$$

$$\frac{d^2z}{dt^2} + n^2z = a_z \tag{4.162}$$

where n is the mean TV orbital rate and $n \approx d\theta/dt$ by assumption. Note that, although small separation was initially assumed, the downtrack range s does not explicitly appear and is thus not restricted. In the circular orbit case, the important criteria for orbital separation are the radial and out-of-plane components. If orbits of nonzero eccentricity are allowed, as below, restrictions on downtrack separation will again appear. Note also that the out-of-plane component decouples from the other two; for small inclinations, the motion normal to the orbit plane is a simple sinusoid.

The circular orbit approximation is common and often realistic, because many rendezvous operations can be arranged to occur, at least in the final stages,

between CV and TV in nominally circular orbits. Also, as discussed in Sec. 4.2.3 (Motion in Elliptic Orbits), most practical parking orbits are of nearly zero eccentricity. However, Jones[59] has shown that both zero eccentricity and small eccentricity approximations can yield significant errors (see Fig. 4.28) in some cases compared with results obtained using Stern's equations.[60] These equations are linear and thus retain the assumption of small displacements between TV and CV but are valid for arbitrary eccentricity. We have

$$\frac{d^2r}{dt^2} - 2\left(\frac{d\theta}{dt}\right)\left(\frac{ds}{dt}\right) - \left[\left(\frac{d\theta}{dt}\right)^2 + \frac{2\mu}{R_T^3}\right]r - \left(\frac{d^2\theta}{dt^2}\right)s = a_r \qquad (4.163a)$$

$$\frac{d^2s}{dt^2} + 2\left(\frac{d\theta}{dt}\right)\left(\frac{dr}{dt}\right) - \left[\left(\frac{d\theta}{dt}\right)^2 - \frac{\mu}{R_T^3}\right]s + \left(\frac{d^2\theta}{dt^2}\right)r = a_s \qquad (4.163b)$$

$$\frac{d^2z}{dt^2} + \left(\frac{\mu}{R_T^3}\right)z = a_z \qquad (4.163c)$$

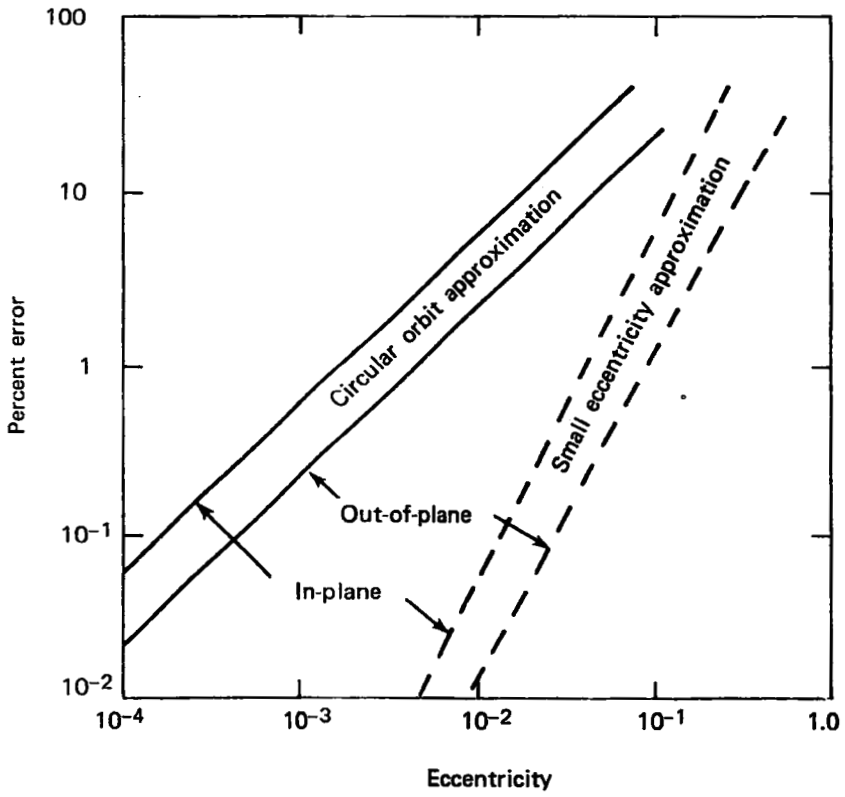


Fig. 4.28 Approximation errors in relative motion equations.

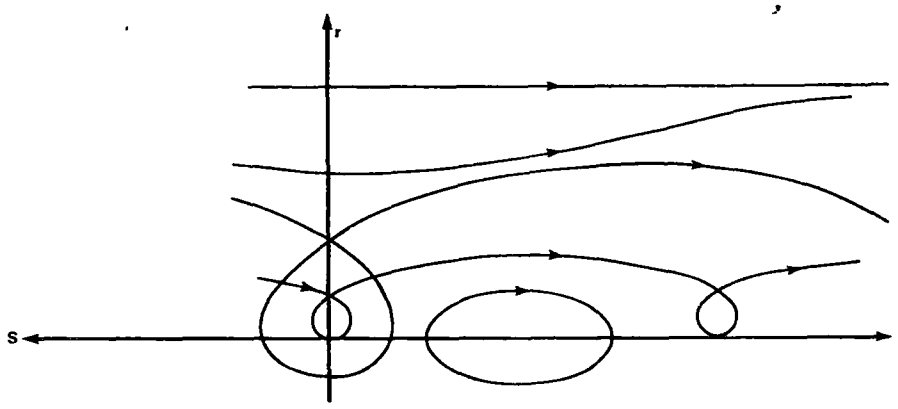


Fig. 4.29 Relative motion trajectories.

where θ and R_T are the true anomaly and radius of the TV in the inertial frame.

It is seen that Eq. (4.163) reduces to Eq. (4.162) when the TV orbit is circular. Dunning[61] gives equations of intermediate complexity between the above two sets, in which the second-derivative terms in true anomaly θ are omitted. These and equivalent results obtained by Jones[59] give first-order corrections for eccentricity compared with the Clohessy-Wiltshire equations.[58]

Care should be exercised in the choice of formulation used. Jones[59] finds approximately 5% error using the Hill equations compared to results obtained using Stern's equations for $e = 0.01$, and 10% error for $e = 0.05$. However, error estimates are only approximate and depend on the actual case of interest, and both sets of equations contain linearization errors that can be expected to dominate at longer CV to TV ranges. Classical guidance algorithms[62] for terminal rendezvous implicitly invoke the same assumptions as for the Hill equations; thus, when doubt exists, it is wise to study the sensitivity of the results obtained to the choice of orbital eccentricity assumed and the dynamics model employed.

The solution to the Hill equations in the case of unforced motion is easily obtained.[9] The rotating coordinate system used in the analysis produces what at first glance appear to be rather unusual trajectories. Figure 4.29 shows typical CV motion[63] for cases where it is above and below the orbit of the TV. Clearly, maneuvers to achieve rendezvous are facilitated if the CV is initially above and ahead or initially below and behind the TV. Rendezvous procedures are structured so as to attain this geometry prior to the initiation of the terminal phase closing maneuvers.

4.7.2 Rendezvous Procedures

A variety of rendezvous procedures have been implemented in the U.S. manned flight program, and others have been proposed for unmanned vehicles

such as would be required for a planetary sample return program.[57] We consider here the basic operational scenario for U.S. manned rendezvous missions. All such missions have utilized essentially circular target vehicle orbits. The baseline procedure that was developed during the Gemini program[64] and implemented operationally on Apollo[65] is the so-called concentric flight plan (CFP) approach. The CFP procedure involves five basic steps:

1) Any out-of-plane component in the CV orbit is removed by waiting until $z = 0$, in the notation of Eq. (4.162), and thrusting with acceleration a_z to yield $dz/dt = 0$. In the formulation of the Hill equations, this is equivalent to a small plane change at the line of nodes, as given by Eq. (4.124). This maneuver was not required to be, and typically was not, the first in the sequence and in operational cases often was not needed at all.

2) A waiting period is needed to allow proper phasing to develop between the two vehicles, as discussed for interplanetary transfer in Section 4.5.1.1 (Heliocentric trajectory). The "above and ahead" or "below and behind" geometry must be attained, together with the requirement imposed by Eq. (4.150). An adjustment to the CV orbit (typically a perigee raising maneuver for the usual case of the CV below the TV) is made to achieve the desired phasing prior to the next step.

3) Upon attainment of proper vehicle phasing, the so-called constant differential height (CDH) or coelliptic maneuver was performed. Assuming the TV to occupy the higher orbit, this maneuver is done at the CV orbit apogee and places the chase craft in an orbit that is concentric (or coelliptic) with that of the target but several tens of kilometers lower. A coelliptic separation of 15 n mile was used for Apollo lunar rendezvous. A strong advantage of the CFP approach is that the CDH phase allows the next, or terminal, phase to be relatively insensitive to the timing and execution of earlier operations.

4) When proper vehicle-to-vehicle phasing is obtained, the terminal phase initiation maneuver is executed. In Gemini and Apollo, the nominal final transfer maneuver was a two-impulse, non-Hohmann trajectory requiring a transfer angle of 130°. This value was selected on the basis of simulations showing a relative lack of sensitivity of the arrival conditions to errors in terminal phase initiation (TPI) timing and impulse magnitude. Also, this transfer trajectory was shown to result in a minimal rotation rate in the CV-to-TV inertial line of sight during the closure process, a feature that is useful both in the design of guidance algorithms and as a piloting aid. Other final transfer trajectories are possible; Jezewski and Donaldson[66] have studied optimal maneuver strategies using the Clohessy-Wiltshire equations.[58]

5) Regardless of the terminal phase maneuver selected, as the range is reduced, a closed-loop terminal guidance scheme will be used to control the reduction of range and range rate to zero. Some form of proportional guidance[67] is generally employed; whatever the technique, the orbit dynamics of the closure trajectory need no longer be central to the scheme. Small corrections are applied, based on differences between actual and predicted velocity vs range during

closure, to allow the nominal transfer trajectory to be maintained by the CV as the target is approached. As final closure occurs, braking maneuvers are performed to reduce any residual relative velocity to zero in the neighborhood of the target vehicle.

The rendezvous phase is complete when the chase and target vehicles are separated by a small distance, typically well inside 100 m, and have essentially zero relative velocity. This defines the stationkeeping phase, in which the available CV acceleration $\boldsymbol{a}$ is often assumed to dominate the differential accelerations in Eq. (4.162) due to the orbital dynamics effects. In such a case, for example,

$$\frac{d^2r}{dt^2} \ll a_r \tag{4.164}$$

and similarly for the other components. Thus, the motion in the near neighborhood of the TV is essentially rectilinear and dominated by the CV control maneuvers, provided transit times are kept small relative to the orbit period.

Note that stationkeeping is entirely possible even if the CV control authority is low; however, neglect of the orbital dynamics in local maneuvers is then not possible. Low-impulse stationkeeping control implies maneuvers having a duration significant with respect to the orbital period. Orbital dynamics effects will always be apparent in such cases. This is graphically demonstrated in the case of manned maneuvering unit activities in the vicinity of the shuttle.[68]

It is worth noting that the only completely passive stationkeeping positions possible are directly ahead of or behind the target vehicle in its orbit. Radial or out-of-plane offsets will, in the absence of control maneuvers, result in oscillations of the CV about the target during the orbital period.

Interest in automated rendezvous and docking, or "capture," has been of great interest recently, in the context of satellite servicing and retrieval and for the task of delivering cargo to the International Space Station via unmanned expendable launch vehicles.[69] As mentioned, this is a proven technology in the Russian space program (albeit not without numerous anomalies), but not so far implemented by the United States. The first such demonstration will occur with the NASA Demonstration of Automated Rendezvous Techniques (DART) program, which as this is written is progressing toward a planned 2004 launch.

References

[1]Farquhar, R. W., Muhonen, D. P., Newman, C. R., and Heuberger, H. S., "Trajectories and Orbital Maneuvers for the First Libration Point Satellite," *Journal of Guidance and Control*, Vol. 3, No. 6, 1980, pp. 549–554.

[2]O'Neill, G. K., *The High Frontier*, Morrow, New York, 1977.

[3]Heppenheimer, T. A., "Achromatic Trajectories and Lunar Material Transport for Space Colonization," *Journal of Spacecraft and Rockets*, Vol. 15, No. 3, 1978, pp. 236–239.

[4]Flandro, G. A., "Solar Electric Low-Thrust Missions to Jupiter with Swingby Continuation to the Outer Planets," *Journal of Spacecraft and Rockets*, Vol. 5, Sept. 1968, pp. 1029–1032.

[5]Hoyle, F., *Astronomy and Cosmology—A Modern Course*, Freeman, San Francisco, CA, 1975.

[6]Halliday, D., and Resnick, R., *Physics*, 3rd ed., Wiley, New York, 1977.

[7]Goldstein, H., *Classical Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1980.

[8]Dirac, P. A. M., *Directions in Physics*, Wiley-Interscience, New York, 1975.

[9]Kaplan, M. H., *Modern Spacecraft Dynamics and Control*, Wiley, New York, 1976.

[10]Danby, J. M. A., *Celestial Mechanics*, Macmillan, New York, 1962.

[11]Hamming, R. W., *Introduction to Applied Numerical Analysis*, McGraw-Hill, New York, 1971.

[12]Sheela, B. V., "An Empirical Initial Estimate for the Solution of Kepler's Equation," *Journal of the Astronautical Sciences*, Vol. 30, No. 4, 1982, pp. 415–419.

[13]Battin, R. H., *An Introduction to the Mathematics and Methods of Astrodynamics*, AIAA Education Series, AIAA, New York, 1987.

[14]Herrick, S., *Astrodynamics*, Vol. 2, Van Nostrand Reinhold, London, 1971.

[15]Bate, R. R., Mueller, D. D., and White, J. E., *Fundamentals of Astrodynamics*, Dover, New York, 1971.

[16]"GEODSS Photographs Orbiting Satellite," *Aviation Week and Space Technology*, Vol. 119, No. 26, 28 Nov. 1983, pp. 146–147.

[17]Gelb, A. (ed.), *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

[18]Nahi, N. T., *Estimation Theory and Applications*, Krieger, Huntingdon, NY, 1976.

[19]Wertz, J. R. (ed.), *Spacecraft Attitude Determination and Control*, Reidel, Boston, MA, 1978.

[20]Schmidt, S. F., "The Kalman Filter: Its Recognition and Development for Aerospace Applications," *Journal of Guidance and Control*, Vol. 4, Jan.–Feb. 1981, pp. 4–7.

[21]Mechtly, E. A., "The International System of Units," NASA SP-7012, 1973.

[22]Fliegel, H. F., and Van Flandern, T. C., "A Machine Algorithm for Processing Calendar Dates," *Communications of the ACM*, Vol. 11, Oct. 1968, p. 657.

[23]Sundman, K. F., "Memoire sur le Probleme des Trois Corps," *Acta Mathematica*, Vol. 36, 1913, pp. 105–179.

[24]Kaula, W. M., *An Introduction to Planetary Physics: The Terrestrial Planets*, Wiley, New York, 1968.

[25]Arfken, G. A., *Mathematical Methods for Physicists*, 2nd ed., Academic, New York, 1970.

[26]Kaula, W. M., *Theory of Satellite Geodesy*, Blaisdell, Waltham, MA, 1966.

[27]"A Satellite Freed of All but Gravitational Forces," *Journal of Spacecraft and Rockets*, Vol. 11, Sept. 1974, pp. 637–644.

[28]Merrigan, M. J., Swift, E. R., Wong, R. F., and Saffel, J. T., "A Refinement to the World Geodetic System 1984 Reference Frame," Inst. of Navigation, ION-GPS-2002, Portland, OR, Sept. 2002.

[29]Kershner, R. B., "Technical Innovations in the APL Space Department," *Johns Hopkins APL Technical Digest*, Vol. 1. No. 4, 1980, pp. 264–278.

[30]*STS User's Handbook*, NASA, Washington, DC, May 1982.

[31]Kaplan, M. H., Cwynar, D. J., and Alexander, S. G., "Simulation of Skylab Orbit Decay and Attitude Dynamics," *Journal of Guidance and Control*, Vol. 2, No. 6, 1979, pp. 511–516.

[32]U.S. Standard Atmosphere, National Oceanic and Atmospheric Administration, NOAA S/T 76-1562, U.S. Government Printing Office, Washington, DC, 1976.

[33]Jacchia, L. G., "Revised Static Models of the Thermosphere and Exosphere with Empirical Temperature Profiles," Smithsonian Astrophysical Observatory Special Rept. 332, 1971.

[34]Vinh, N. X., Busemann, A., and Culp, R. D., *Hypersonic and Planetary Entry Flight Mechanics*, Univ. of Michigan Press, Ann Arbor, MI, 1980.

[35]Shanklin, R. E., Lee, T., Samii, M., Mallick, M. K., and Capellari, J. O., "Comparative Studies of Atmospheric Density Models Used for Earth Orbit Estimation," *Journal of Guidance, Control, and Dynamics*, Vol. 7, March–April 1984, pp. 235–237.

[36]Jacchia, L. G., "Thermospheric Temperature, Density, and Composition: New Models," Smithsonian Astrophysical Observatory Special Rept. 375, March 1977.

[37]Liu, J. J. F., "Advances in Orbit Theory for an Artificial Satellite with Drag," *Journal of the Astronautical Sciences*, Vol. 31, No. 2, 1983, pp. 165–188.

[38]King-Hele, D., *Theory of Satellite Orbits in an Atmosphere*, Butterworths, London, 1964.

[39]Abramowitz, M., and Stegun, I., *Handbook of Mathematical Functions, Tables, and Graphs*, 10th Printing, National Bureau of Standards AMS-55, U.S. Government Printing Office, Washington, DC, Dec. 1972.

[40]Liou, J., Matney, M. J., Anz-Meador, P. D., Kessler, D. J., Jansen, M., and Theall, J. R., "The New NASA Orbital Debris Engineering Model ORDEM2000," NASA/TP-2002-210780, May 2002.

[41]Schaaf, S. A., and Chambre, P. L., "Flow of Rarefied Gases," *Fundamentals of Gas Dynamics*, edited by H. W. Emmons, Princeton Univ. Press, Princeton, MJ, 1958.

[42]Fredo, R. M., and Kaplan, M. H., "Procedure for Obtaining Aerodynamic Properties of Spacecraft," *Journal of Spacecraft and Rockets*, Vol. 18, July–Aug. 1981, pp. 367–373.

[43]Knechtel, E. D., and Pitts, W. C., "Normal and Tangential Momentum Accommodation Coefficients for Earth Satellite Conditions," *Astronautica Acta*, Vol. 18, No. 3, 1973, pp. 171–184.

[44]Smith, E., and Gottlieb, D. M., "Possible Relationships Between Solar Activity and Meteorological Phenomena," NASA TR X-901-74-156, 1974.

[45]Siegel, R., and Howell, J. R., *Thermal Radiation Heat Transfer*, 2nd ed., Hemisphere, Washington, DC, 1981.

[46]Jacobson, R. A., and Thornton, C. L., "Elements of Solar Sail Navigation with Applications to a Halley's Comet Rendezvous," *Journal of Guidance and Control*, Vol. 1, Sept.–Oct. 1978, pp. 365–371.

[47]Van der Ha, J. C., and Modi, V. J., "On the Maximization of Orbital Momentum and Energy Using Solar Radiation Pressure," *Journal of the Aeronautical Sciences*, Vol. 27, Jan. 1979, pp. 63–84.

[48]Lang, T. J., "Optimal Impulsive Maneuvers to Accomplish Small Plane Changes in an Elliptical Orbit," *Journal of Guidance and Control*, Vol. 2, No. 2, 1979, pp. 301–307.

[49]Ikawa, H., "Synergistic Plane Changes Maneuvers," *Journal of Spacecraft and Rockets*, Vol. 19, Nov. 1982, pp. 300–324.

[50]D'Amario, L. D., Byrnes, D. V., and Stanford, R. H., "Interplanetary Trajectory Optimization with Application to Galileo," *Journal of Guidance and Control*, Vol. 5, No. 5, 1982, pp. 465–468.

[51]Battin, R. H., "Lambert's Problem Revisited," *AIAA Journal*, Vol. 15, May 1977, pp. 703–713.

[52]Battin, R. H., Fill, T. J., and Shepperd, S. W., "A New Transformation Invariant in the Orbital Boundary-Value Problem," *Journal of Guidance and Control*, Vol. 1, Jan.–Feb. 1978, pp. 50–55.

[53]Battin, R. H., and Vaughan, R. M., "An Elegant Lambert Algorithm," *Journal of Guidance, Control, and Dynamics*, Vol. 7, Nov.–Dec. 1984, pp. 662–666.

[54]Small, H. W., "Globally Optimal Parking Orbit Transfer," *Journal of the Astronautical Sciences*, Vol. 31, No. 2, 1983, pp. 251–264.

[55]Hulkower, N. D., Lau, C. O., and Bender, D. F., "Optimum Two-Impulse Transfers for Preliminary Interplanetary Trajectory Design," *Journal of Guidance, Control, and Dynamics*, Vol. 7, July–Aug. 1984, pp. 458–462.

[56]Hamming, R. W., *Numerical Methods for Scientists and Engineers*, 2nd ed., McGraw-Hill, New York, 1973.

[57]Tang, C. C. H., "Co-Apsidal Autonomous Terminal Rendezvous in Mars Orbit," *Journal of Guidance and Control*, Vol. 3, Sept.–Oct. 1980, pp. 472–473.

[58]Clohessy, W. H., and Wiltshire, R. S., "Terminal Guidance System for Satellite Rendezvous," *Journal of Aerospace Sciences*, Vol. 27, 1960, pp. 653–658.

[59]Jones, J. B., "A Solution of the Variational Equations for Elliptic Orbits in Rotating Coordinates," AIAA Paper 80-1690, Aug. 1980.

[60]Stern, R. G., "Interplanetary Midcourse Guidance Analysis," MIT Experimental Astronomy Laboratory Rept. TE-5, Cambridge, MA, 1963.

[61]Dunning, R. S., "The Orbital Mechanics of Flight Mechanics," NASA SP-325, 1975.

[62]Chiarappa, D. J., "Analysis and Design of Space Vehicle Flight Control Systems: Volume VIII—Rendezvous and Docking," NASA Contractor Rept. CR-827, 1967.

[63]Schneider, A. M., Prussing, J. E., and Timin, M. E., "A Manual Method for Space Rendezvous Navigation and Guidance," *Journal of Spacecraft and Rockets*, Vol. 6, Sept. 1969, pp. 998–1006.

[64]Parten, R. P., and Mayer, J. P., "Development of the Gemini Operational Rendezvous Plan," *Journal of Spacecraft and Rockets*, Vol. 5, Sept. 1968, pp. 1023–1026.

[65]Young, K. A., and Alexander, J. D., "Apollo Lunar Rendezvous," *Journal of Spacecraft and Rockets*, Vol. 7, Sept. 1970, pp. 1083–1086.

[66]Jezewski, D. J., and Donaldson, J. D., "An Analytic Approach to Optimal Rendezvous Using the Clohessy-Wiltshire Equations," *Journal of the Astronautical Sciences*, Vol. 27, No. 3, 1979, pp. 293–310.

[67]Nesline, F. W., and Zarchan, P., "A New Look at Classical vs. Modern Homing Missile Guidance," *Journal of Guidance and Control*, Vol. 4, Jan.–Feb. 1981, pp. 78–85.

[68]Covault, C., "MMU," *Aviation Week and Space Technology*, Vol. 120, No. 4, 23 Jan. 1984, pp. 42–56.

[69]Polites, M. E., "An Assessment of the Technology of Automated Rendezvous and Capture in Space," NASA TP 1998-208528, July 1998.

Problems

4.1 A spacecraft intended to map the surface of Mars is desired to be placed into a sun synchronous orbit having a periapsis height of 500 km and a period of 1/7 of a Martian day (24.62 h). What should the orbital inclination be if a Martian year is 687 Earth days?

4.2 Because the spacecraft in problem 4.1 can only see a small portion of the planet during a low altitude pass, it is planned to map essentially all of the planet between the latitude limits imposed by the orbital inclination of problem 4.1, which you may assume to be 100° if you did not get the answer, by "walking" the periapsis around the planet. Each region of latitude will be mapped as the planet rotates beneath the orbit, and as the nodal line regresses. What initial value (or values) of the argument of perigee ω_0 should be selected to minimize the time required to conduct this mission? How long will it take to map the planet in detail?

4.3 Assume the approach velocity to Mars for the spacecraft in problem 4.1 to be $V_\infty = 4$ km/s. What should the B-plane miss distance be to achieve a 500-km altitude periapsis? What is the injection ΔV required at periapsis?

4.4 After the high-resolution mapping outlined in problem 4.2 is completed, it is desired to change the orbit plane to an inclination of 90° to map the polar regions. What is the minimum ΔV required to effect this plane change? Explain.

4.5 A target spacecraft occupies a circular orbit with a 100-minute period. It is desired to rendezvous with this spacecraft from a near-circular orbit having the same inclination. An initial $\Delta V = (-123, 81.4)$ m/s is used to initiate closure on the target from an initial position of $(x, y)_0 = (25, -75)$ km.

(a) Plot the approach to the target.

(b) What terminal maneuver is required to halt the approach in the neighborhood of the target?

4.6 For the initial parameters of problem 4.5, what is the proper initial closure maneuver for a desired rendezvous time of 25 min from the initial position?

4.7 A Landsat spacecraft is to be placed in a 600-km altitude circular sun-synchronous orbit. What inclination is required? If the spacecraft is launched from Vandenberg Air Force Base (latitude 34.5°N), what is the required launch azimuth?

4.8 A spacecraft intended to facilitate communications in high northerly latitude regions is placed in a 12-h Molniya orbit with $\omega = 270°$. What is the "hang time" above the northern hemisphere?

4.9 From the 1979 edition of the *American Ephemeris and Nautical Almanac*, p. 492, we find that the U.S. Naval Observatory (USNO) in Washington, D.C., is located at longitude and latitude $(\lambda, \phi) = (5$ hrs 08 min 15.75 s W, 38° 55′ 14.2″) and is at an altitude of 86 m. On p. 12 of the same almanac we note that on 20 Jan. 1979 at 0 h U.T., the Greenwich Sidereal Time (GST) was 7 hrs 55 min 6.975 s. What was the local sidereal time, in

degrees, at the USNO at 1200 hrs EST on 20 Jan. 1979? For reference, the rotation rate of the Earth can be expressed as $\omega_e = 2\pi$ radians/day $\times$ 1.0027379093 sidereal days/solar day.

4.10 A spacecraft is to be sent to Saturn. Assume the vehicle is initially in a circular Earth parking orbit with $r = 6600$ km. What is the required ΔV for a Hohmann transfer to Saturn? Assume a flyby only at the Saturn end, i.e., no orbit injection maneuvers at Saturn. You may find the following constants useful:

$$1 \text{ A.U.} = 1.496 \times 10^8 \text{ km} \qquad r_{\text{Saturn}} = 9.539 \text{ A.U.}$$

$$\mu_{\text{Earth}} = 398{,}600 \text{ km}^3/\text{s}^2 \qquad \mu_{\text{Saturn}} = 3.7934 \times 10^7 \text{ km}^3/\text{s}^2$$

$$\mu_{\text{Sun}} = 1.327 \times 10^{11} \text{ km}^3/\text{s}^2$$

4.11 Sketch the geometry for the hyperbolic Earth departure segment of problem 4.10. Be sure to show some reference direction, such as the Earth-sun vector, or the Earth orbital velocity vector, which ties the Earth-centered frame to heliocentric space. Given this reference direction, show the desired departure asymptote and the correct point of application for the departure ΔV. Compute for the departure hyperbola the values for e, β, θ_d, and swingby angle ψ and indicate them on your diagram. If you did not solve problem 4.10, assume $V_\infty = 10$ km/s for the departure hyperbola.

4.12 If you solved problem 4.10 correctly, you will note that the spacecraft arrives at Saturn at the apogee of its Hohmann trajectory with a heliocentric velocity of 4.2 km/s. Assume a sun-side swingby at Saturn with a periapsis of $r_p = 100{,}000$ km, and solve for the departure conditions at the completion of the gravity assist maneuver. Specifically, compute the heliocentric departure velocity and flight-path angle (or equivalently, the V_r and V_θ components of heliocentric velocity) upon leaving Saturn's vicinity. Sketch the swingby hyperbola at Saturn.

4.13 What is the ΔV required to go to the moon from
(a) a 185-km altitude circular Earth parking orbit,
(b) the perigee of a geostationary transfer orbit of dimension 6563 km $\times$ 42,164 km, and
(c) geostationary orbit?

4.14 At 1200 hrs EST on 20 Jan 1979 an orbiting upper stage injects a spacecraft in an orbit to the vicinity of the Moon. Following this maneuver, the Earth referenced orbital elements are found to be: $\pi = 13{,}625.24$ $\kappa\mu$, $\varepsilon = 1.0336$, $i = 28.5°$, $\Omega = 270°$, $\omega = 14.65°$, $\theta = 20°$. What are $\boldsymbol{r}$ and $\boldsymbol{V}$ in GEI coordinates?

4.15 How long does it take the spacecraft in problem 4.14 to cross lunar orbit at $r = 400{,}000$ km?

4.16 What is nominal orbit lifetime of an object in a circular, 28.5° inclination orbit at an altitude of 300 km, given a ballistic coefficient of 100 kg/m^2? Use the standard atmospheric model of Chapter 3.

4.17 An Earth monitoring satellite is to be placed into a 185-km altitude circular sun-synchronous orbit. What is the required orbit inclination? If launched from Vandenberg Air Force Base, latitude 34.5°N, what launch azimuths are possible? Which would you expect to use, and why?

5
Propulsion

Probably no single factor constrains the design of a space vehicle and the execution of its mission more than does the state of the art in propulsion technology. Ascent propulsion capability, together with the physical limitations imposed by celestial mechanics, sets the limits on payload mass, volume, and configuration that bound the overall design. The economics of space flight and our progress in exploiting space are driven inexorably by the cost per kilogram of mass delivered to orbit. Mankind's reach in exploring interplanetary space is limited by the energy available from current upper stages. Though the advent of the space shuttle has expanded many of the boundaries of the spacecraft design environment, it is still true that the scope of most space missions is ultimately set by propulsion system limitations.

Yet, despite its importance, ascent propulsion is probably the factor over which a spacecraft designer has the least control. Except for those involved directly in the areas of rocket engine, booster, or upper-stage design, most aerospace engineers will be in the position of customers with freight to be moved. A limited number of choices are available, and the final selection is seldom optimal for the given task, but is merely the least unsatisfactory. Rarely is a particular mission so important that a specific engine or launch vehicle will be designed to fit its needs. Indeed, there have been few boosters designed for space missions at all; most are converted Intermediate Range Ballistic Missiles (IRBM) and Intercontinental Ballistic Missiles (ICBM) or derivatives of those vehicles. The Saturn family, the space shuttle, the Proton, the Zenit, and Ariane are conspicuous exceptions, but even as this book is written, a significant fraction of payloads reach space on various derivatives of the Atlas, the Delta, the Titan, and the Soviet Semyorka ICBM.

We therefore take the view that ascent propulsion is essentially a "given" in the overall design. We do not explore in detail the multitude of considerations that go into the design of launch vehicles. The text by Sutton and Biblarz[1] is probably the best source for those seeking more detail in this area. Our purpose is to explore the factors that are involved in the selection of launch vehicles and upper stages for a given mission. We have tried to include a reasonably comprehensive discussion of the capabilities of the various vehicles, including those of both current and projected availability and which could be of interest.

However, injection into a specified trajectory does not end the consideration of propulsion systems required by the spacecraft systems engineer. Low-orbit satellites may need propulsion for drag compensation. Satellites in geosynchronous Earth orbit (GEO) require a similar system for stationkeeping purposes.

Many spacecraft will need substantial orbit adjustments or midcourse maneuvers. Attitude control systems will often employ small thrusters, either for direct control or for adjustment of spacecraft angular momentum. The future holds the prospect of the development of orbital transfer vehicles (OTV) for operations in Earth orbit. Planetary landers, such as the Surveyor, Apollo, and Viking missions to the moon and Mars, as well as many projected planetary landers, involve the development of specialized descent propulsion systems. For these and other reasons, the spacecraft designer must be familiar with propulsion system fundamentals.

5.1 Rocket Propulsion Fundamentals

5.1.1 Thrust Equation

The fundamental equation for rocket engine performance is the thrust equation,[2]

$$T = \dot{m}V_e + (p_e - p_a)A_e \tag{5.1}$$

where

T = thrust force
$\dot{m}$ = flow rate = $\rho_e V_e A_e$
ρ_e = fluid density at nozzle exit
V_e = exhaust velocity at nozzle exit
p_e = exhaust pressure at nozzle exit
p_a = ambient pressure
A_e = nozzle exit area

This equation is valid for reaction motors that generate thrust through the expulsion of a fluid stream without ingesting fuel or oxidizer from any source external to the vehicle. In aerospace applications, the working fluid is a gas, possibly nitrogen gas stored under pressure, combustion products from a variety of propellants, or hydrogen gas superheated by passage through a nuclear reactor.

If efficiency is at all important, the gas is made as hot as possible and expanded through a supersonic nozzle, as in Fig. 5.1, to increase V_e. The expansion ratio is usually made as high as possible, subject to considerations to be discussed later. This of course causes the term $p_e A_e$ to decrease, but the loss here is usually small compared to the gain in $\dot{m}V_e$. In all large operational engines to date, heating of the gas has been accomplished chemically. The working fluid is thus composed of the products of the combustion cycle producing the heat.

Because the dominant term in the thrust equation is $\dot{m}V_e$, it is customary to rewrite Eq. (5.1) as

$$T = \dot{m}[V_e + (p_e - p_a)A_e/\dot{m}] = \dot{m}V_{\text{eq}} \tag{5.2}$$

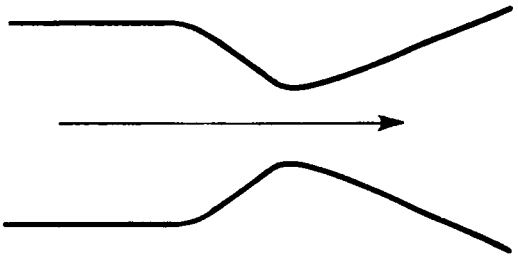


Fig. 5.1 Supersonic nozzle.

where V_{eq} is the equivalent exhaust velocity. The portion of V_{eq} due to the pressure term will nearly always be small relative to V_e and back-of-the-envelope performance calculations often make use of this fact by dropping V_{eq} for the more simply obtained V_e. However, the pressure term is by no means negligible when accurate results are desired, as later examples will show. Even rough calculations can sometimes require its inclusion. For example, the space shuttle main engine suffers about a 20% loss of thrust at sea level compared to vacuum conditions because of pressure effects.

5.1.2 Specific Impulse

The total change in momentum, or *total impulse*, of the expelled propellant (and hence, in the opposite direction, of the rocket) is

$$I = \int T\,dt = \int \dot{m} V_{eq}\,dt \tag{5.3}$$

If V_{eq} is constant over the burning time (never strictly true, even ignoring various transients for real motors, except for horizontal flight and flight in vacuum), Eq. (5.3) becomes

$$I = m_p V_{eq} \tag{5.4}$$

where m_p is the mass of propellant consumed. Note that it is not necessary in Eq. (5.3) to assume $\dot{m}$ is constant to obtain this result, but only to assume that any throttling used is done in such a way as to leave V_{eq} unaffected (a desirable but seldom achievable condition).

To focus on the efficiency of the engine rather than its size or duration of operation (which determine m_p), it is convenient to define the *specific impulse*,

$$I_{sp} = \frac{I}{m_p} = V_{eq} = \frac{T}{\dot{m}} \tag{5.5a}$$

which is seen to be the thrust per unit mass flow rate. This is a definition founded in basic physics. It is far more customary in engineering circles to use instead the weight flow to normalize the total impulse, yielding

$$I_{sp} = \frac{I}{m_p g} = \frac{V_{eq}}{g} = \frac{T}{\dot{m} g} \tag{5.5b}$$

With the definition of Eq. (5.5b), specific impulse is measured in seconds in any consistent system of units, a not inconsiderable advantage. Note, however, that Eq. (5.5a) reveals the fundamental physics of the situation; the specific impulse, or change in momentum per unit mass, is merely the equivalent exhaust velocity. Specific impulse is the most important single measure of rocket engine performance, because it relates in a fundamental way to the payload-carrying capability of the overall vehicle, as we shall see in later sections.

5.1.3 Nozzle Expansion

Returning to Eq. (5.1), it is of interest to explore some general considerations in the operation of rocket engines for maximum efficiency. It is clear from Eq. (5.1) that $p_e < p_a$ is undesirable, because the pressure term is then negative and reduces thrust. Also, this condition can be harmful when operating inside the atmosphere, because the exit flow will tend to separate from the walls of the nozzle, producing a region of recirculating flow that can under some conditions set up destructive vibrations due to unbalanced and shifting pressure distributions.

It is not immediately clear, but is true, that $p_e \gg p_a$ *is* also to be avoided. This situation basically indicates a failure to expand the exhaust nozzle as much as might be done, with a consequent loss of potentially available thrust. And again, within-the-atmosphere operation at very large exit-to-ambient pressure ratios can cause undesirable interactions of the exhaust plume with the external airflow.

It is thus ideal to have a close match between nozzle exit pressure and ambient pressure. There are practical limits to the degree to which this can be accomplished. An engine operating in vacuum would require an infinite exit area to obtain $p_e = 0$. Very large nozzles introduce a mass penalty that can obviate the additional thrust obtained. Large nozzles are more difficult to gimbal for thrust vector control and may be unacceptable if more than one engine is to be mounted on the same vehicle base. Furthermore, it is clear that engines intended for ascent propulsion cannot in any case provide an ideal pressure match at more than one altitude. An effective compromise is to operate the nozzle as much as possible in a slightly underexpanded condition, the effects of which are less harmful than overexpansion. Still, as a practical matter, overexpansion must often be tolerated. Shuttle main engines sized for sea level operation would be extremely inefficient for high-altitude operation. These engines operate with an exit pressure of about

0.08 atm and thus do not approach an underexpanded condition until an altitude of about 18 km is reached.

As noted earlier, choice of expansion ratio for conventional nozzle, fixed-area ratio engines is usually a compromise. For engines that operate from liftoff through the atmosphere and into space, as do most lower-stage engines, the compromise will be driven by performance and liftoff thrust requirements. Engines that provide most of the liftoff thrust and do not perform long in vacuum have lower optimal expansion ratios than those that operate solely in space. In this latter case, the highest practical expansion ratio is usually constrained by the available volume and increasing nozzle weight. Engines such as the space shuttle main engine (SSME) and the Atlas sustainer, which must operate at liftoff but perform much of their work in space, are the most difficult compromise, requiring a trade between vacuum performance, sea level performance, and other factors. The maximum expansion ratio is often limited by the need to prevent flow separation in the nozzle, with its resulting asymmetric side loads.

Even some space engines are limited by this problem. In the case of the J-2, a 250,000-lb thrust LO_2/LH_2 engine used in the second and third stages of the Saturn 5, it was desired to test the engines in a sea level environment even though all flight operations would be in vacuum. This avoided the expense of the very large vacuum test facilities that would have been needed for each test of a higher-expansion-ratio engine. The J-2 was marginal in its ability to maintain full flow at the nozzle exit under these conditions and was frequently plagued by flow separation and sideloads during testing. The problem was most annoying during engine startup and at off-design operation.

A variety of unconventional nozzle concepts have been suggested that have as their goal the achievement of optimum expansion at all altitudes. The concept that these nozzles have in common is a free expansion and deflection of the jet. The most prominent examples are plug or spike nozzles and the expansion-deflection (ED) nozzles.

A plug or spike nozzle is depicted in Fig. 5.2. The combustion chamber is a torus or, more probably, an annular ring of a number of individual combustors. The nozzle through which the gases exit the combustion chamber will converge to a sonic throat and may be followed in some cases by a diverging supersonic section as in conventional nozzles. This expansion, if used, will be small relative

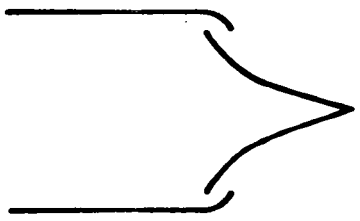


Fig. 5.2 Plug or spike nozzle.

to the engine design expansion ratio. The gas is directed inboard and slightly aft, the angle being defined by the overall characteristics of the engine. The gas impinges on the central plug, which is carefully contoured to turn the flow in the aft direction. The unconfined gas tends to expand, even as its momentum carries it inboard and along the plug. The boundary condition that must be satisfied by the outer sheath of gas is that it match the ambient pressure; the stream expands to achieve this condition. As the vehicle ascends and ambient pressure decreases, the stream expands accordingly, as shown in Fig. 5.3. Expansion ratio is thus always near-optimal, tending to infinity in vacuum.

In the initial concept, the central spike tapered to a point. It was quickly discerned that performance was equally good, and the mass much lower, if the point were truncated. The final variant of this concept is the Rocketdyne Aerospike. In this case, the spike is still more truncated, but a substantial secondary flow (provided by the turbine exhaust in a complete engine) is fed through the bottom of the plug to help maintain the core flow shape and provide an adequate base pressure.

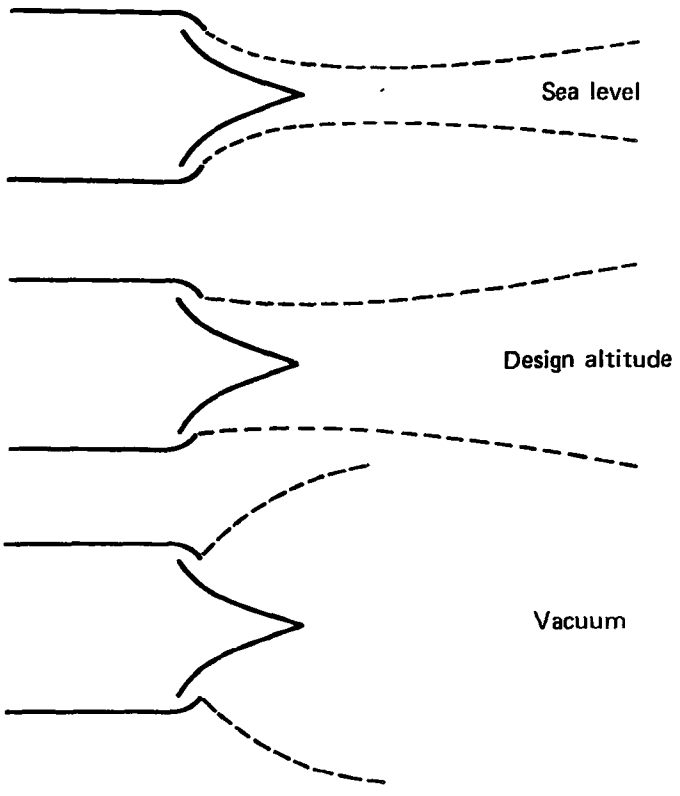


Fig. 5.3 Plug nozzle at various altitudes.

One disadvantage of the plug nozzle is that the heavy, high-pressure combustion section is the largest diameter of the engine rather than a small, compact cylinder, thus resulting in a heavy engine. The expansion-deflection concept was an attempt to obtain a pressure-compensating nozzle in a more nearly conventional overall shape. A central plug shaped like an inverted mushroom turns the flow from the combustion chamber outward and nearly horizontal, as in Fig. 5.4. The contoured outer skirt then turns the flow aft. The pressure compensation is supposed to come from the degree of expansion into the annular central space behind the plug. Without secondary flow this does not work well under all combinations of expansion ratio and ambient pressure, because the self-pumping action of the flow tends to close the flow behind the plug, creating a low-pressure area behind the plug base. This results in a conventional aerodynamic "pressure drag," which can seriously inhibit engine performance. A large secondary flow from the turbopump system or from an ambient air bleed may be required to obtain the desired performance.

A variation on these concepts is shown in Fig. 5.5. This is sometimes called the "linear plug," because the combustors and the deflection surface form a linear array. This concept is especially well suited to lifting body vehicles. This application may see the first flight use of a pressure-compensated nozzle for space vehicles. A linear aerospike engine was intended for the NASA/Lockheed Martin X-33 vehicle; however, the program was cancelled prior to flight for a combination of technical and fiscal reasons.

At first glance, the altitude compensating character of the plug or spike nozzle seems attractive. However, analysis of actual trajectories often shows the overall gains to be small, and possibly not ultimately beneficial when other factors, such

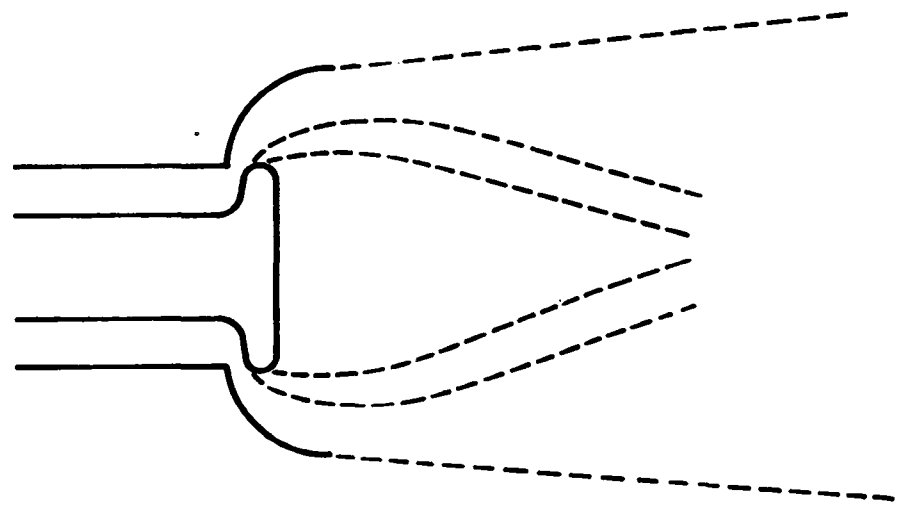


Fig. 5.4 Expansion-deflection nozzle.

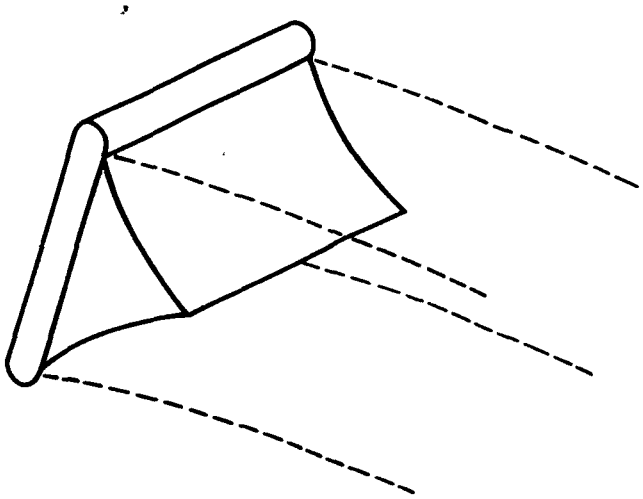


Fig. 5.5 Linear plug nozzle.

as the difficulty of thrust vector control, are considered. The choice of engine nozzle design for a launch vehicle should be carefully evaluated on a case-by-case basis. The various types of plug and spike nozzles do offer the possibility of more efficient integration into the structure of the vehicle in many cases, provided the engine is fixed. A fixed installation requires differential throttling or fluid injection for thrust vector control.

Extendable exit cones (EEC) have become common as a solution to the problem of launching upper stages containing motors that must operate efficiently in vacuum without being so large as to pose packaging problems for launch. In the EEC concept, the exit nozzle is designed in two sections, as shown in Fig. 5.6, where the second section is translated from its stored to its operational position by springs or pneumatic plungers. This technique is primarily applied to designs where a radiatively cooled, dump cooled, or ablative exit cone is used; it is not suitable where regenerative cooling of the extension is required. Advanced versions of the Pratt & Whitney RL-10 and numerous solid propellant upper stages use extendable exit cones of various types.

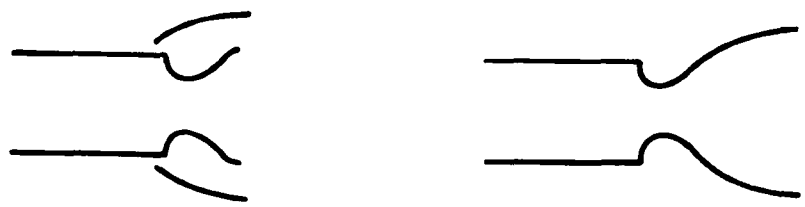


Fig. 5.6 Extendable nozzle.

5.1.4 Calculation of Specific Impulse

The exact calculation of rocket engine exhaust velocity and pressure, and hence specific impulse and thrust, is an exceedingly complex task requiring the numerical solution of a multidimensional coupled set of partial differential equations describing the fluid dynamic and chemical processes involved. However, surprisingly good results can be obtained by idealizing the rocket engine flowfield as a quasi-one-dimensional adiabatic, frictionless, shock-free flow of a calorically perfect gas having a fixed chemical composition determined by the combustion process. If this model is employed, the energy equation of gasdynamics,

$$c_p T_c = c_p T_e + \frac{V_e^2}{2} \tag{5.6}$$

may be combined with the isentropic pressure-temperature relation,

$$\frac{T_e}{T_c} = \left(\frac{p_e}{p_c}\right)^{(k-1)/k} \tag{5.7}$$

to yield for the exhaust velocity

$$V_e^2 = \frac{kR_{\text{gas}} T_c [1 - (p_e/p_c)^{(k-1)/k}]}{(k-1)} \tag{5.8}$$

where

k = ratio of specific heats, c_p/c_v.
p_e = nozzle exit pressure
p_c = combustion chamber pressure
T_c = combustion chamber temperature
R_{gas} = exhaust flow specific gas constant = $\mathcal{R}/\mathcal{M}$
$\mathcal{R}$ = universal gas constant
$\mathcal{M}$ = exhaust gas molecular weight

We use the less conventional k for the ratio of specific heats, as opposed to the more customary γ, in order to avoid confusion with the notation for flight-path angle used throughout this text.

The extent to which this result is useful to the designer is a matter for careful judgment. It establishes the parameters on which exhaust velocity, and hence specific impulse, depends. Thus, it is clear that high combustion temperature and low exhaust gas molecular weight are advantageous. High chamber pressure is also seen to be desirable, as is a low effective ratio of specific heats (k is always greater than unity).

However, additional complexities exist. Chamber temperature depends on the chemical reactions that take place during combustion; a highly energetic reaction

such as

$$H_2 + F_2 \rightarrow 2HF$$

will produce higher temperatures than a reaction such as

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

because of the inherent differences in bonding energy. However, the rate of energy release also affects the chamber temperature; the combustion process is essentially an equilibrium reaction whose rate depends on equilibrium "constants" that are functions of pressure. Reactions yielding products having a lower specific volume than the constituents, such as

$$2H_2 + O_2 \rightarrow 2H_2O$$

will proceed faster at higher pressure, releasing energy at a greater rate. The net result is a generally small but useful temperature increase due to chamber pressure, which is not seen in the above simplified model.

Also, most large, liquid-fueled rocket engines are regeneratively cooled, meaning that the thrust chamber and nozzle are cooled by the flow of propellant in a surrounding jacket prior to injection and combustion. This process, intended to protect the metal walls, removes little heat from the main flow (thus allowing the adiabatic flow assumption to be retained) but may raise the precombustion fuel temperature by several hundred degrees, thus raising the energy level of the propellant. This effect can add several seconds of I_{sp} compared to that predicted by Eq. (5.8).

The utility of k as a parameter describing the gas is somewhat questionable in a chemically reacting flow. For simple diatomic gases at temperatures below roughly 500 K, it is both theoretically and observably true that $k = 7/5$. For example, this is true for air (neglecting CO_2 and other minor constituents). At higher temperatures c_p and c_v are not constant (the gas is not "calorically perfect") and do not allow the static enthalpy to be written in the form

$$h = c_p T \tag{5.9}$$

as we did earlier in Eq. (5.6). The similarity parameter k thus has little meaning and does not appear in the basic equations that must be solved to obtain the exhaust velocity. However, it is commonly found that good results can be obtained using Eq. (5.8) and similar calorically perfect gas results, provided that an empirically determined "hot k" is used. This will be on the order of $k = 1.21–1.26$ for a wide range of fuels and oxidizers. Values of k for high temperature air may be found in Appendix B.

In computing specific impulse, as opposed to exhaust velocity, it is necessary to account for the pressure term. Again, it is typically not large, but is significant, because payload mass is highly sensitive to I_{sp}. Subject to the same

approximations as previously, the exit area is given by

$$\frac{A_e}{A^*} = \frac{1}{M_e}\left\{\frac{2[1+(k-1)M_e^2/2]}{k+1}\right\}^{\frac{1}{2}\frac{k+1}{k-1}} \tag{5.10a}$$

where

$M_e = V_e/a_e$ = exit Mach number

$a_e^2 = kRT_e$ = exit speed of sound

A^* = sonic throat area

It is also necessary to know $\dot{m}$ if the pressure effect on I_{sp} is to be assessed. Subject to the same calorically perfect gas approximations previously noted, it is found that

$$\dot{m} = p_e A^* \left\{\frac{k}{RT_c}\left(\frac{2}{k+1}\right)^{(k+1)/(k-1)}\right\}^{1/2} \tag{5.10b}$$

As an indication of available engine performance, Table 5.1 gives actual specific impulse for a variety of engines with varying fuel/oxidizer combinations.

5.1.5 Nozzle Contour

A characteristic of all supersonic flow devices, including rocket engine thrust chambers, is the use of a convergent-divergent nozzle to achieve the transition from subsonic to supersonic flow. Various shapes are possible for the subsonic converging section, because the flow is not particularly sensitive to the shape in this region. A simple cone, faired smoothly into cylindrical combustion zone and the rounded throat, is usually satisfactory. For the divergent section, a cone is the most straightforward and obvious shape and indeed was used in all early rocket designs. The chosen cone angle was a compromise between excessive length, mass, and friction loss for a small angle vs the loss due to nonaxial flow velocity for larger angles.

The classic optimum conical nozzle tends to have about a 15° half-angle. Such an angle was generally satisfactory for low-pressure engines operating at modest expansion ratios. As chamber pressures and expansion ratios increased in the search for higher performance, conic nozzles became unsatisfactory. The increased length needed in such cases results in greater weight and high moment of inertia, which causes difficulty in gimballing the motor for thrust vector control. This gave rise to the so-called "bell" or contoured nozzle. This concept involves expanding the flow at a large initial angle and then turning it so that it exits in a nearly axial direction (most nozzles will still have a small divergence angle, e.g., 2° at the exit). Design of the nozzle to achieve this turning of the flow

Table 5.1 Specific impulse for operational and prototype engines

Engine	Thrust lbf	Fuel	Oxidizer	I_{sp} sec	Expansion Ratio
Aerojet					
AJ110	9800 (vac)	UDMH/ N_2H_4	N_2O_4	320 (vac)	65:1
Atlantic Research 8096-39 (Agena)	17,000 (vac)	UDMH	H.P. nitric acid	300 (vac)	45:1
Daimler-Chrysler					
Aestus	6140 (vac)	MMH	N_2O_4	324 (vac)	83:1
Aestus II (w/Rocketdyne)	10,300 (vac)	MMH	N_2O_4	337.5	280:1
Morton Thiokol					
STAR 48	17,210	Solid		293 (vac)	55:1
STAR 37F	14,139	Solid		286 (vac)	41:1
Pratt & Whitney					
RL 10A3-3A	16,500	Liquid H_2	Liquid O_2	444 (vac)	61:1
RL 10A4-1	20,800	Liquid H_2	Liquid O_2	449 (vac)	84:1
RL 10A4-2	22,300	Liquid H_2	Liquid O_2	451 (vac)	84:1
RL 10B-2	24,750	Liquid H_2	Liquid O_2	464 (vac)	285:1
Rocket Research					
MR 104C	129 (vac)	N_2H_4	—	239 (vac)	53:1
MR 50L	5 (vac)	N_2H_4	—	225 (vac)	40:1
MR 103A	0.18 (vac)	N_2H_4	—	223 (vac)	100:1
Rocketdyne					
SSME	375,000 (sl)	Liquid H_2	Liquid O_2	361 (sl)	77.5:1
	470,000 (vac)			425.5 (vac)	
RS-27A (Delta 2)	200,000 (sl)	RP-1	Liquid O_2	255 (sl)	12:1
	237,000 (vac)			302 (vac)	
RS-68	650,000 (sl)	Liquid H_2	Liquid O_2	365 (sl)	21.5:1
	745,000 (vac)			410 (vac)	
RS-72	12,450 (vac)	MMH	N_2O_4	338.5	300:1
XLR-132	3570 (vac)	MMH	N_2O_4	$\geq$340	400:1
Russian					
NK-33	329,900 (sl)	Kerosene	Liquid O_2	297 (sl)	27:1
	368,000 (vac)			331 (vac)	
RD-120	175,000 (sl)	Kerosene	Liquid O_2	303 (sl)	106:1
	191,000 (vac)			350 (vac)	
RD-170	1,632,000 (sl)	RP-1	Liquid O_2	309 (sl)	36.4:1
	1,777,000 (vac)			331 (vac)	
RD-180	860,400 (sl)	RP-1	Liquid O_2	311 (sl)	36.4:1
	933,000 (vac)			337 (vac)	

(*continued*)

Table 5.1 Specific impulse for operational and prototype engines (continued)

Engine	Thrust lbf	Fuel	Oxidizer	I_{sp} sec	Expansion Ratio
SEP					
Viking 4B	177,000 (vac)	UH25	N_2O_4	293.5 (vac)	30.8:1
Snecma					
Vulcain	257,000 (vac)	Liquid H_2	Liquid O_2	431 (vac)	45:1
Vulcain-2	304,000 (vac)	Liquid H_2	Liquid O_2	433 (vac)	58.5:1
Vinci	40,500 (vac)	Liquid H_2	Liquid O_2	424 (vac)	?
TR1W					
TR1-201	9900 (vac)	UDMH/ N_2H_4	N_2O_4	303 (vac)	50:1
MMPS (Spacecraft)	88 lbf (vac)	MMH	N_2O_4	305 (vac)	180:1
MRE-5	4 lbf (vac)	N_2H_4	—	226 (vac)	?
United Technologies					
Orbus 6	23,800	Solid		290 (vac)	47:1
Orbus 21	58,560	Solid		296 (vac)	64:1

without producing undesired shock waves requires the application of the method of characteristics and is beyond the intended scope of this text. Bell or contour nozzles are often referred to by the percentage of length as compared to a 15° cone of the same expansion ratio. For example, an 80% bell has a length 80% of that of the equivalent conic nozzle.

The efficiency or thrust coefficient of bell and cone nozzles is essentially the same. Although the flow exiting the bell is more nearly axial, the losses involved in turning the flow tend to compensate for this advantage. Practical engine designs turn the flow quite rapidly after the sonic throat, a process that introduces various inefficiencies. Gradually contoured nozzles such as used in high-speed wind tunnels are possible but tend to be quite long and generally do not offer sufficient advantage to compensate for their weight, volume, and cost penalty.

5.1.6 Engine Cooling

A variety of cooling concepts have been proposed for use in rocket engines, many of which have seen operational use, often in combination. The most common approach for large engines with lengthy operating times is "regenerative cooling," mentioned earlier, where one of the propellants is passed through cooling passages in the thrust chamber and nozzle wall before being injected into the combustion chamber. This very effective and efficient approach is usually supplemented by film or boundary-layer cooling, where propellant is injected so as to form a cooler, fuel-rich zone near the walls. This is accomplished by the relatively simple means of altering the propellant distribution at the injector, which usually consists of a "shower head" arrangement of many small entrance ports for fuel and oxidizer. In

some cases injector orifices may be oriented to spray directly on the engine wall. Probably the most extreme example occurred in the pioneering V-2, which had a series of holes drilled just above the throat to bleed in raw fuel to protect the combustion chamber, which was fabricated from mild steel. In most regenerative cooling designs, the fuel is used as the coolant, although oxidizer has been used and is increasingly suggested for high-mixture-ratio, high-pressure LO_2/LH_2 engines.

Ablative thrust chambers are commonly employed in engines designed for a single use, in cases where neither propellant is an efficient coolant, or for operational reasons such as when deep throttling or pulse mode operation is required. When a wide range of throttling is available, propellant (and hence coolant) flow at the lower thrust settings may become so sluggish that fluid stagnation and overheating may occur. When pulsed operation is desired, as, for example, in thrusters used for on-orbit attitude or translation control, the volume of the coolant passages is incompatible with the requirement for short, sharp pulses. Ablative thrust chamber endurance of several thousand seconds has been demonstrated. In cases where throat erosion is critical, refractory inserts have been used. Ablative chambers are especially sensitive to mixture ratio distribution in the flow, with hot streaks causing severe local erosion, especially if oxidizer-rich.

Radiation-cooled thrust chambers have been extensively used in smaller engine assemblies. Refractory metals or graphite have been commonly used in the fabrication of such motors, which tend to be simple and of reasonably low mass and have nearly unlimited life. However, these desirable features are sometimes offset by the nature of radiation cooling, which causes difficulty in some applications. The outer surface of the thrust chamber, which rejects heat at temperatures approaching 1500 K, must have an unimpeded view of deep space. Furthermore, any object in view of the thrust chamber will be exposed to substantial radiative heating. A compact, vehicle-integrated engine installation such as might be used for a regeneratively or ablatively cooled engine is thus impossible. Fully radiatively cooled engines of more than a few thousand pounds thrust have not been demonstrated to date. This is due to the materials costs and systems integration difficulties of fabricating such engines. It may be noted, however, that some fairly large engines intended for upper-stage operation employ the extendable exit cones mentioned earlier, or fixed extensions, which are radiation-cooled.

Heat sink thrust chambers, where the chamber wall material simply accumulates the heat by bulk temperature increase during the burn, are fairly common as low-cost, short-duration ground test articles, which are required for injector performance characterization. They are rarely used in flight hardware, except as buried units subjected to brief, infrequent pulses. The ability of refractory metals such as niobium (columbium) to operate at very high temperatures allows use of a "hybrid" cooling scheme. The niobium thrust chamber/nozzle assembly acts as a refractory heat sink; however, the interior of the nozzle has a sufficiently good view of space that much of the heat energy can be radiated away, allowing long-duration operation. In most cases, boundary-layer cooling, i.e., excess fuel near the walls, helps minimize heat transfer. With

adequate external insulation, this assembly can be buried in structure. The space shuttle attitude control thrusters use this approach.

An interesting concept, worthwhile only with hydrogen, is dump cooling. Hydrogen, if heated to a few hundred degrees and exhausted through a nozzle, has a specific impulse equal to many bipropellant combinations. In such an engine, the bipropellants would be burned in the conventional manner and exhausted, while hydrogen would pass through the chamber walls, being heated in the process and exhausted through its own nozzle. Previous studies have not shown the performance to be worth the complexity of a three-propellant system, but future applications may be possible.

Such concepts as spray cooling, in which the liquid coolant is sprayed against the combustion chamber wall rather than caused to flow over it, can accommodate very high heat fluxes but have not been required by propulsion systems used to date. Similarly, transpiration cooling, where the coolant is uniformly "sweated" through a porous wall, has not shown enough performance advantage over less expensive and more conventional boundary-layer cooling to justify its use. However, transpiration cooling has found use in some LO_2/LH_2 injectors, such as those for the J-2 and RL-10.

The F-1 engine used an interesting variant in which turbopump exhaust gas was dumped into a double-walled nozzle extension and then, via a series of holes in the inner wall, into the boundary layer of the main stream. This cooled the extension while getting rid of the often troublesome turbine exhaust by entraining it in the main flow.

5.1.7 Combustion Cycles

Rocket engine combustion cycles have grown steadily more complex over the years as designers have sought to obtain the maximum possible specific impulse and thrust from hardware of minimum weight. The current state of the art in this field is probably exemplified by the SSME. However, basic designs remain in wide use, as exemplified by the fact that the simple pressure-feed system continues to be a method of choice where simplicity, reliability, and low cost are driving requirements.

As noted, a major driver in engine cycle development is the desire for higher performance. This translates to higher combustion chamber pressure, efficient use of propellant, and minimum structural mass. Because structural mass increases rapidly with tank size and pressure, the need for pump-fed, as opposed to pressure-fed, engines was recognized quite early. Dr. Robert H. Goddard began flying pump-fed engines in the 1920s to prove the concept, while the German rocket engineers at Peenemünde went immediately to pump-fed systems for the larger vehicles such as the V-2.

The early vehicles, of which the V-2 and the U.S. Army Redstone are classical examples, used a turbopump system in which the hot gas, which drove the turbine that in turn drove the pumps, was provided by a source

completely separate from the rocket propellants. The hot gas was obtained by decomposing hydrogen peroxide into water and oxygen, a process accomplished by spraying the peroxide and a solution of potassium permanganate into a reaction chamber. These substances were stored in pressurized tanks, an approach with the virtue of simplicity in that operation of the turbine drive was decoupled from that of the main propulsion system. Also, the low-temperature turbine exhaust made turbine design relatively simple. On the negative side, there was a considerable mass penalty because of the low energy of the hydrogen peroxide and because of the extra tanks required to hold the peroxide and the permanganate. The basic concept worked quite well, however, and was applied in a variety of systems well into the late 1950s. The various derivatives of the original Russian ICBM, such as the Soyuz and Molniya launchers, still use this approach.

Next to be developed was the bootstrap gas generator concept, in which a small fraction of the main propellant is tapped off at the pump outlet and burned in a gas generator to provide turbine drive gas. This approach has several advantages. The use of existing propellants saves weight because the increase in tank size to accommodate the turbine requirements imposes a smaller penalty than the use of separate tanks. Also, in order to provide gas temperatures tolerable for turbine materials, it is necessary to operate well away from the stoichiometric fuel-to-oxidizer ratio. This is usually done by running substantially on the fuel-rich side, which provides a nonoxidizing atmosphere for the turbine. Some vehicles, such as the recently-retired Ariane IV, run the gas generators near the stoichiometric ratio and cool the gas to an acceptable temperature by injecting water downstream of the combustion zone.

A number of systems developed in the 1950s and 1960s used the bootstrap gas generator approach. The first such vehicles were the Navaho booster and the Atlas, Titan, Thor, and Jupiter missiles. The F-1 and J-2 engines for the Saturn series used similar cycles, as in fact have most of the vehicles in the U.S. inventory of launchers, excluding the space shuttle.

Individual engine systems vary in detail regarding implementation, particularly in the starting cycle. Some use small ground start tanks filled with propellant to get the engines started and up to steady-state speed. In other cases, the start tanks are mounted on the vehicle and refilled from the main propellant tanks to allow later use with vernier engines providing velocity trim after main engine shutdown. Still others, use solid-propellant charges that burn for about a second to spin up the pumps and provide an ignition source for the gas generator. The J-2 used hydrogen gas from an engine-mounted pressure bottle, which was repressurized during the burn to allow orbital restart.

The F-1, the 1.5-million-lb-thrust first-stage engine for the Saturn 5, used no auxiliary starting system at all. By the time this engine was designed, it was recognized that a bootstrap gas generator system could be self-starting. Simply opening the valves at the tank pressure used to provide inlet pump head and igniting the propellants would start the pumps, which would increase combustion

chamber pressure, etc., in a positive-feedback process that continued until full thrust was obtained. On the F-1 this took some 8-9 seconds; however, because of the reduced structural loads associated with the slow start, this was an advantage. Figures 5.7–5.10 show some of the various engine cycles.

Of the workhorse engines of the past several decades, one in particular did not use the gas generator cycle. The Pratt & Whitney RL-10 LO_2/LH_2 series used in

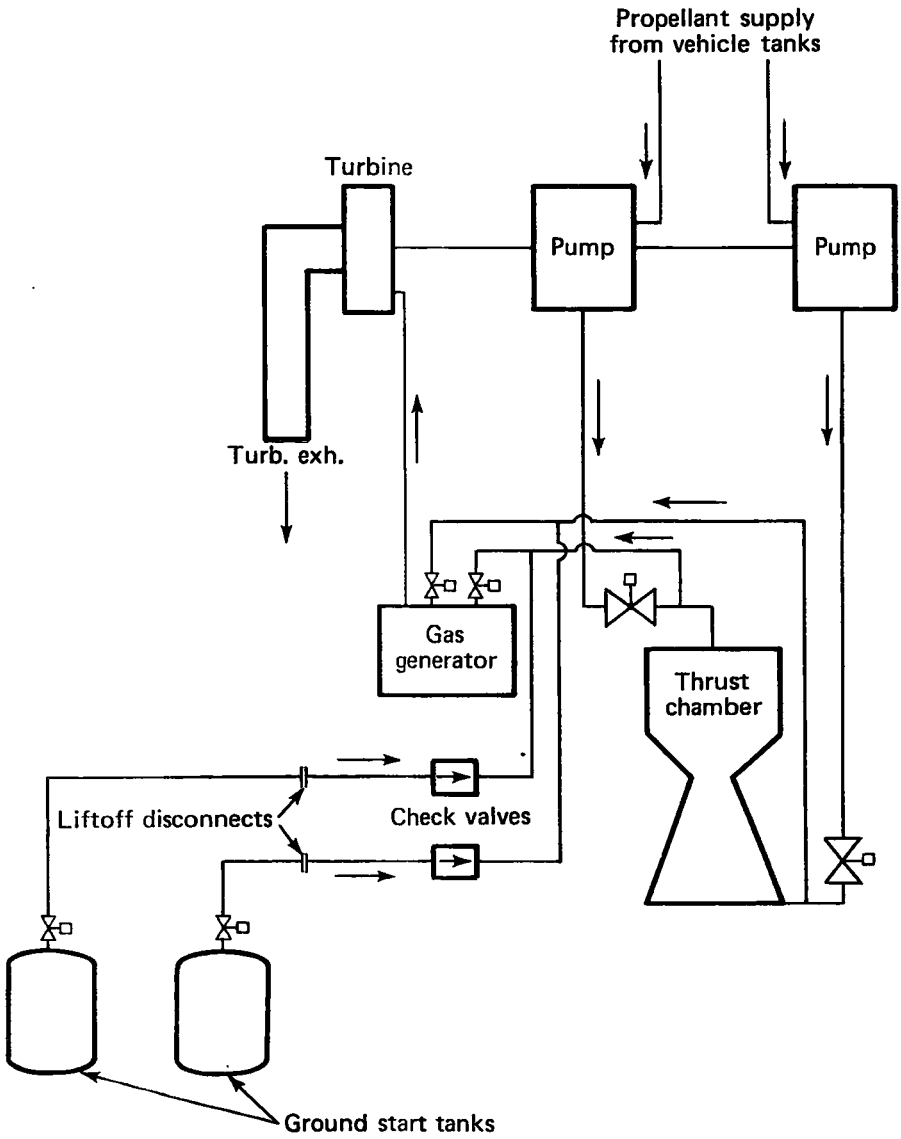


Fig. 5.7 Rocket engine diagram—gas generator cycle with ground start tanks.

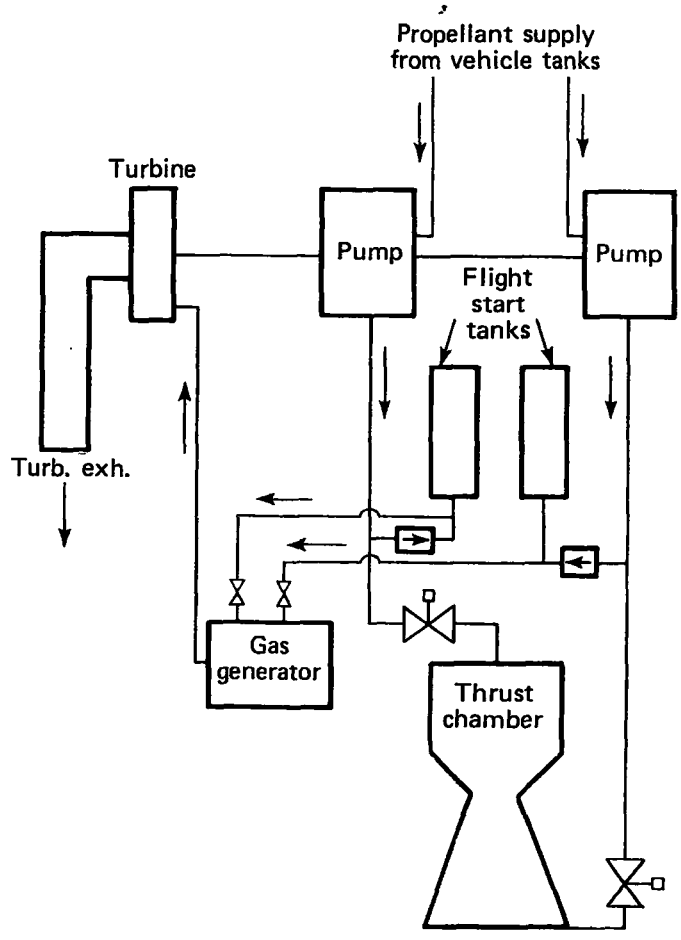


Fig. 5.8 Rocket engine diagram—gas generator cycle with flight start tanks.

the S-IV, Centaur, and Delta IV upper stages, as well as the DC-X test vehicle, takes advantage of the hydrogen fuel being heated in the thrust chamber cooling process and expands it through the turbine to provide energy to run the pumps. Essentially all the hydrogen is used for this purpose and then injected (as a gas) into the combustion chamber. In this cycle, none of the propellant is "wasted" as relatively low-energy turbine exhaust gas. Therefore, the overall performance tends to be better than that of the basic bootstrap cycle. The primary performance penalty for this system is the extra pumping energy required to counteract the pressure drop through the turbine. This engine allows a tank head start, as with the F-1, and seems more tolerant of throttling than most. The cycle is diagrammed in Fig. 5.11.

The SSME uses another more complex cycle. The liquid oxygen and liquid hydrogen are passed through completely separate turbopump packages driven by

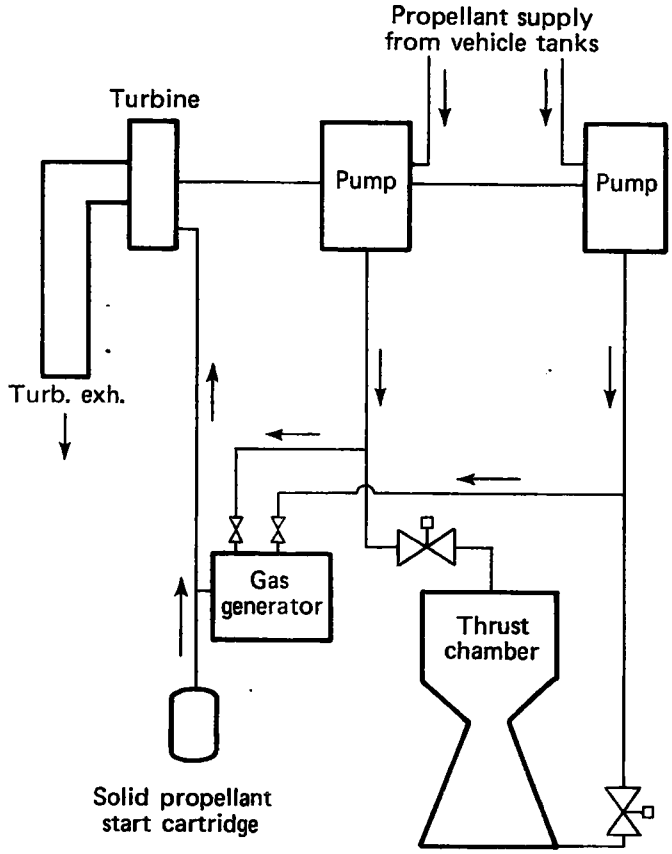


Fig. 5.9 Rocket engine diagram with solid propellant start cartridge.

separate pre-burners that maintain acceptable temperatures by off-stoichiometric combustion. The turbine gas is generated by burning a portion of the total main engine flow at off-mixture-ratio conditions. This gas then flows through a turbine to power the pumps. It is then dumped into the main combustion chamber along with the flow from the other pump loop. Additional propellant is added to form the main thrust chamber flow. Figure 5.12 shows the cycle. The SSME is most notable for operating at a much higher chamber pressure than its predecessors and therefore yields very high performance. Even higher performance would be possible, except for the requirement in the shuttle system for the engines to be operating at liftoff. The SSME is currently capable of throttling from 65 to 109% of rated thrust, which is 470,000 lbf in vacuum.

5.1.8 Combustion Chamber Pressure

Besides the obvious advantages indicated in Eq. (5.8) and discussed in Sec. 5.1.4 (Calculation of Specific Impulse), some additional benefits and problems

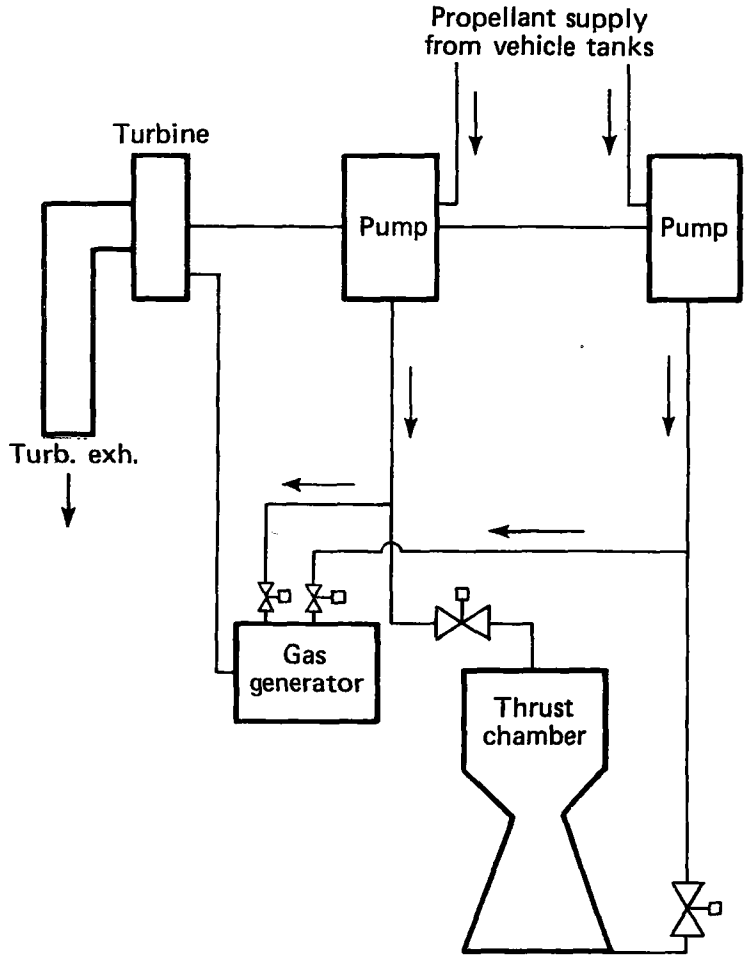


Fig. 5.10 Rocket engine diagram—gas generator with tank head start.

accrue to the use of high chamber pressure. At a given thrust level, a higher chamber pressure engine is more compact. This may lend itself to easier packaging and integration. A high-pressure engine may allow more flexibility in the choice of expansion ratio and may be more amenable to throttling down within the atmosphere. On the negative side, all of the internal plumbing becomes substantially heavier. Sealing of joints, welds, and valve seats becomes progressively more difficult as pressure increases. Leakage, especially of hot gas or propellants such as hydrogen, becomes more dangerous and destructive. The amount of energy required to operate the fuel pumps that inject the propellants into the combustion chamber increases with chamber pressure. Inasmuch as this energy is derived from propellant combustion as part of the overall engine cycle, it represents a "tax" on the available propellant energy that

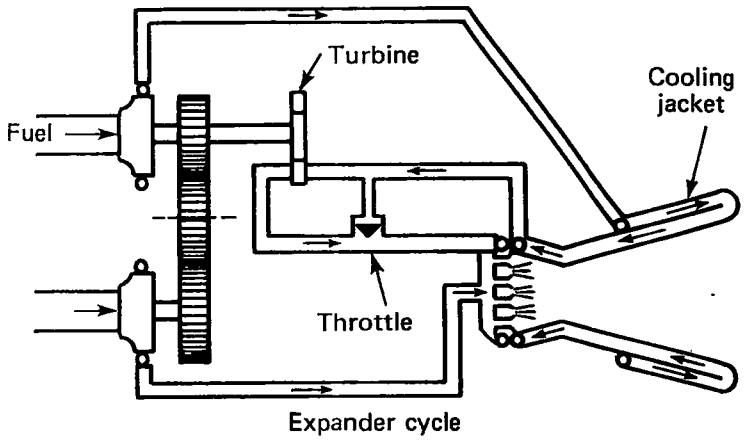


Fig. 5.11 Rocket engine diagram—expander cycle (RL-10).

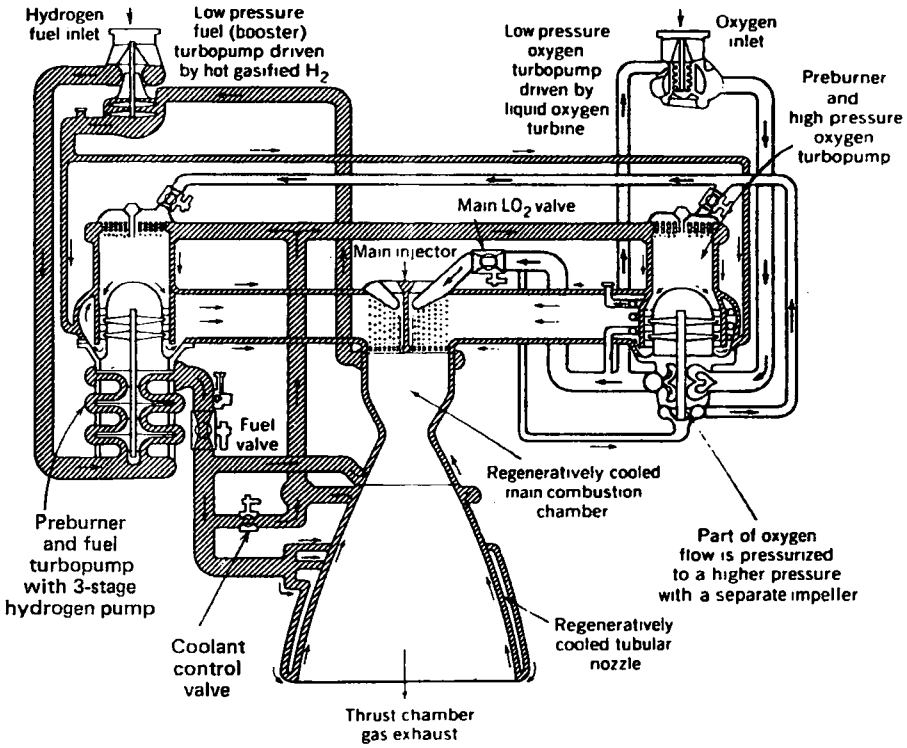


Fig. 5.12 Rocket engine diagram—preburner cycle (SSME).

mitigates the advantages of high-pressure operation. Finally, higher-pressure assemblies have historically demonstrated a greater tendency toward combustion instability than their lower-pressure counterparts.

5.2 Ascent Flight Mechanics

Rocket-powered ascent vehicles bridge the gap between flight in the atmosphere, governed both by gravitational and aerodynamic forces, and space flight, shaped principally by gravitational forces punctuated occasionally by impulsive corrections. Purely astrodynamic considerations were discussed in Chapter 4; in this section we discuss the mechanics of the powered ascent phase. We first consider the basic equations of motion in terms of the physical parameters involved, followed by a discussion of special solutions of the equations.

5.2.1 Equations of Motion

In keeping with our approach, we present the simplest analysis that treats the issues of salient interest to the vehicle designer. To this end, we consider the planar trajectory of a vehicle over a nonrotating spherical planet. The geometric situation is as shown in Fig. 5.13. The equations of motion are[3]

$$\frac{dV}{dt} = \frac{T\cos\alpha - D}{m} - g\sin\gamma \tag{5.11a}$$

$$V\frac{d\gamma}{dt} = \frac{T\sin\alpha + L}{m} - \left(g - \frac{V^2}{r}\right)\cos\gamma \tag{5.11b}$$

$$\frac{ds}{dt} = \frac{R}{r}V\cos\gamma \tag{5.11c}$$

$$\frac{dr}{dt} = \frac{dh}{dt} = V\sin\gamma \tag{5.11d}$$

$$\frac{dm}{dt} = -\dot{m}(t) \tag{5.11e}$$

$$L = \frac{1}{2}\rho V'^2 S C_L \tag{5.11f}$$

$$D = \frac{1}{2}\rho V'^2 S C_D \tag{5.11g}$$

$$g = g_s\left(\frac{R}{R+h}\right)^2 \tag{5.11h}$$

$$\alpha = \alpha(t) \tag{5.11i}$$

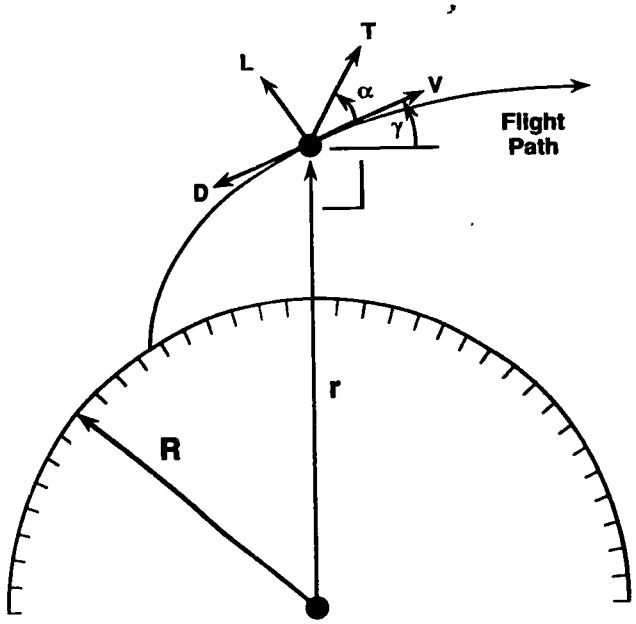


Fig. 5.13 Planar ascent from nonrotating planet.

where

V = inertial velocity magnitude
V' = speed relative to planetary atmosphere
R = planetary radius
h = height above surface
$r = R + h$ = radius from planetary center
s = down-range travel relative to nonrotating planet
γ = flight-path angle, positive above local horizon
T = thrust at time t
m = mass at time t
$\dot{m}$ = mass flow rate, a prescribed function
L = lift force, normal to flight path
D = drag force, parallel to flight path
C_L = lift coefficient
C_D = drag coefficient
ρ = atmospheric density
S = vehicle reference area for lift and drag
g = gravitational acceleration
g_s = surface gravitational acceleration
α = angle of thrust vector relative to flight path; i.e., vehicle pitch angle, a prescribed function

These equations are not solvable in closed form but may be integrated numerically, subject to appropriate initial conditions. Note that, in practice, this particular formulation would not necessarily be used for numerical calculations. If numerical integration is to be employed, it is often simplest to work in Cartesian coordinates directly with the vector equations,

$$\frac{d\boldsymbol{r}}{dt} = \boldsymbol{V} \tag{5.12a}$$

$$\frac{d\boldsymbol{V}}{dt} = f(\boldsymbol{r}, \boldsymbol{V}, t) \tag{5.12b}$$

where $f(\cdot)$ is the sum, per unit mass, of forces on the vehicle. When obtained, the results are easily transformed to a coordinate system that is more appropriate to ascent from a spherical planet. However, Eqs. (5.11) in the form given have the advantage of portraying the physical parameters of interest most directly and are used here for that reason.

The assumption of a nonrotating planet introduces three basic errors. First, Eqs. (5.11) are valid as written only in an inertial frame; neglect of planetary rotation involves neglect of the centrifugal and Coriolis forces generated by the transformation of the time derivatives to a rotating frame. Predictions of position and, to a smaller extent, velocity relative to the planetary surface will be in error if the rotational effect is omitted.

The atmosphere shares the planetary rotation, which tends to carry the vehicle along with it, thus altering the trajectory. Also, errors are introduced in the aerodynamic modeling of the flight vehicle if the atmosphere-relative velocity V' is not used.

Finally, planetary rotation aids the launch by providing an initial velocity in the direction of rotation. The extent to which this is helpful depends on the vehicle design and, as shown in Chapter 4, on the launch site latitude and launch azimuth.

None of these factors is important in the present discussion. Provided lift and drag are computed using planetary-relative velocity, rotating atmosphere effects are usually ignored except for reentry calculations, and often there as well. Coriolis and centrifugal terms are important at the preliminary design level when calculating ballistic missile trajectories, but usually not otherwise. Finally, the effect of planetary rotation on vehicle performance can be modeled simply by specifying the appropriate initial condition on the inertially referenced launch velocity.

The angle α, the vehicle pitch angle in conventional flight mechanics terminology, is a control variable. In general, it is desired that the vehicle adhere to some specific predetermined flight path (position and velocity history) as a function of time. This is accomplished by controlling the direction (and often the magnitude) of the thrust vector. It is the task of the ascent guidance system to

provide the required commands to follow the chosen trajectory. The guidance commands ultimately translate to specification of a prescribed vehicle attitude, represented here by the pitch angle α.

The assumption that the vehicle pitch angle defines the thrust axis alignment ignores small, transient variations about this mean condition that are commonly used for implementation of vehicle steering commands via thrust vector control. For example, some or all of the engines may be gimballed slightly (a 2 to 5° range is typical for lower stages, less than 1° for upper stages) to generate a force perpendicular to the thrust axis and hence a moment about the center of mass to allow control of vehicle attitude. It is the task of the vehicle autopilot to translate attitude requests from the guidance logic into engine gimbal angles for steering. Because the gimbal angles are typically small, preliminary calculations often omit this effect; i.e., the autopilot is not modeled, and it is assumed that the vehicle points as required to shape the trajectory. Once a suitable family of ascent trajectories is found, higher order models including guidance and autopilot functions are used to establish detailed performance.

As discussed previously, methods other than engine gimballing may be used to effect thrust vector control. These include nozzle injection, jet vanes or, when several engines are present, differential throttling. Finally, the vehicle may in some cases be steered, or at least stabilized, aerodynamically.

If the engines are throttled, as is the case with a space shuttle ascent, then T is also a control variable. Calculation of thrust from basic engine parameters was discussed in the previous section. If thrust is constant, a state variable may be eliminated, as Eq. (5.11e) integrates to yield

$$m(t) = m_p - \dot{m}(t - t_0) \tag{5.13}$$

The lift and drag coefficients contain the information on vehicle aerodynamic behavior. For a specified body shape, C_L and C_D are functions of Mach number, Reynolds number, and the angle of attack. Except at very low speeds, which constitute an insignificant portion of the ascent flight, the dependence on Reynolds number is unimportant. For the flight regimes of interest in typical ascent performance calculations, and for a given Mach number, C_L is proportional to α and C_D is proportional to α^2.

As before, the preceding statement contains the implicit assumption that the vehicle thrust axis is aligned with the geometric centerline, to which the aerodynamic angle of attack is referenced, and that the center of mass lies along the centerline. These assumptions are usually appropriate at the preliminary design level, but will rarely be strictly true. If the vehicle center of mass is offset from whatever aerodynamic symmetry axes exist, as is the case with the space shuttle, the thrust axis cannot be aligned with the centerline and the pitch angle will not equal the aerodynamic angle of attack. And, as mentioned, if the vehicle is steered via thrust vector control, transient offsets of the thrust axis from the center of mass are used to generate attitude control moments. These effects may often

be neglected for initial performance assessments; however, a complete six-degree-of-freedom simulation with guidance and autopilot models will include them.

For a particular vehicle shape, C_L and C_D are usually obtained as functions of Mach number and angle of attack from experimental data, taken either in wind tunnels or during flight tests. A wealth of such data[4,5] exists for various generic shapes of interest as well as for specific vehicles that have flown. For preliminary design purposes, data can usually be found that will be sufficiently representative of the actual vehicle. It recent times it has become possible to solve numerically the governing fluid dynamic equations appropriate to many vehicle configurations of interest. These computational fluid dynamic methods can often provide data outside the envelope of wind-tunnel test capabilities. As in so many areas we have discussed, the space shuttle program again provides an excellent example. Substantial effort was expended during the 1970s in learning to compute high-speed flowfields over space shuttle configurations. In some regimes, the information obtained represented the only aerodynamic performance data available prior to the first flight. Subsequent comparisons with flight data have shown generally excellent agreement. Theoretical methods and results obtained are surveyed by Chapman[6] and Kutler.[7]

In using computational methods, it is important to recognize that a variety of assumptions, including the method by which the computational grid is defined, can greatly affect results. It is crucial to verify CFD calculations by anchoring them with data from flight tests or wind tunnels.

The Mach number is given by

$$M = \frac{V}{a} \tag{5.14}$$

where a is the local speed of sound, which for perfect gases is a function of the temperature alone,

$$a^2 = kR_{\text{gas}}T \tag{5.15}$$

Temperature in turn is a prescribed function of the altitude h, usually according to the dictates of a standard atmosphere model. Very detailed models exist for the Earth[8] and Mars[9] and to a lesser extent for other planets. Actual atmospheric probe data are necessary to obtain a temperature profile; planets for which these data have not yet been obtained are often idealized by very simple models based on what can be observed at the planet's cloud tops. However obtained, the temperature information is used with the hydrostatic equation,

$$\mathrm{d}p = -\rho g \,\mathrm{d}h \tag{5.16}$$

and the perfect-gas equation of state,

$$p = \rho R_{\text{gas}} T \tag{5.17}$$

to allow $\rho = \rho(h)$ to be computed. If the temperature profile is piecewise linear (the usual fitting procedure), the resulting density function has one of two forms:

$$\rho = \rho_1 \exp\left[\frac{-g_s(h - h_1)}{R_{\text{gas}} T}\right] \tag{5.18a}$$

for isothermal layers, and

$$\rho = \rho_1 \left(\frac{T}{T_1}\right)^{-(1+g_s/aR_{\text{gas}})} \tag{5.18b}$$

for constant gradient layers where

$$T(h) = T_1 + a(h - h_1) \tag{5.19a}$$

and

$$a = \frac{T_2 - T_1}{h_2 - h_1} \tag{5.19b}$$

where $T_1 = T(h_1)$ and $T_2 = T(h_2)$ are constants from the measured temperature profile.

The preceding results are strictly true only for constant $g = g_s$ whereas, in fact, the gravitational acceleration varies according to Eq. (5.11h). Although the difference is rarely important, the preceding formulation applies exactly upon replacement of the altitude h in Eqs. (5.16–5.19) with the geopotential altitude,

$$h_G = \left[\frac{R}{R + h}\right]^2 h \tag{5.20}$$

5.2.2 Rocket Performance and Staging

Let us consider Eqs. (5.11) under the simplest possible circumstances; i.e., neglect lift, drag, and gravitational forces and assume no steering, so that α and γ are zero. These assumptions are a poor approximation to planetary ascent but may faithfully represent operation in space far away from planetary fields. More importantly, these conditions are also appropriate to the case of acceleration applied in a near-circular orbit, where the terms $(g\sin\gamma)$ and $(V^2/r - g)$ are nearly zero and lift and drag are absent. This is often the situation for orbital maneuvers or for injection into an interplanetary trajectory from a parking orbit.

In this case, Eqs. (5.11) reduce to

$$\frac{dV}{dt} = \frac{T}{m} = \frac{\dot{m}V_{eq}}{m} - \left(\frac{gI_{sp}}{m}\right)\frac{dm}{dt} \tag{5.21}$$

which integrates immediately to yield

$$\Delta V = gI_{sp}\ \ell n\left(\frac{m_i}{m_f}\right) = V_{eq}\ \ell n\left(\frac{m_i}{m_f}\right) \tag{5.22}$$

where m_i and m_f are the initial and final masses, respectively. Defining the mass ratio MR as

$$MR \equiv \frac{m_i}{m_f}$$

we have, for the change in velocity during the burn,

$$\Delta V = V_{eq}\ \ell n\ MR = gI_{sp}\ \ell n\ MR \tag{5.23}$$

If a burn to propellant exhaustion is assumed, this equation gives the maximum theoretically obtainable velocity increment from a single stage. Here we clearly see the desirability of high I_{sp} and a large mass ratio. This latter condition implies a vehicle consisting, as much as possible, of payload and propellant only.

It is often necessary to compute the propellant mass expended for a single ΔV maneuver; from Eq. (5.22) it is readily found that the propellant expenditure is

$$\delta m_p = \left\{1 - \exp\left[-\left(\frac{\Delta V}{gI_{sp}}\right)\right]\right\} m_i \simeq \left(\frac{\Delta V}{gI_{sp}}\right) m_i \tag{5.23a}$$

or

$$\delta m_p = \left\{\exp\left[-\left(\frac{\Delta V}{gI_{sp}}\right)\right] - 1\right\} m_f \simeq \left(\frac{\Delta V}{gI_{sp}}\right) m_f \tag{5.23b}$$

It is also useful to know the payload sensitivity to small changes in I_{sp} Again, from Eq. (5.22), it is found that

$$\frac{\delta m_f}{m_f} \simeq \left(\frac{\Delta V}{gI_{sp}}\right)\frac{\delta I_{sp}}{I_{sp}} \tag{5.23c}$$

A variety of dimensionless quantities are used to describe the allocation of mass to various portions of the rocket vehicle. Note

$$m_i = m_p + m_s + m_{pl} \tag{5.24}$$

where

m_p = total propellant mass
m_{pl} = payload mass
m_s = total structural mass (all other mass necessary to build and fly the vehicle, including tanks, engines, guidance, and other supporting structures)

If complete propellant depletion may be assumed, then

$$m_f = m_s + m_{pl} \tag{5.25}$$

Note that we do not require this assumption, and indeed it will never be satisfied exactly. Vehicles intended for multiple restarts will of course retain propellant for later use after each maneuver. Also, even if a given stage burns to depletion, there will remain some surplus fuel or oxidizer, because it is impossible to achieve exactly the required mixture ratio during the loading process. This excess fuel is termed ullage and is normally small. If significant propellant remains at engine cutoff, whether by accident or design, then the actual m_f must be used in performance calculations. When this is done intentionally, it will generally occur only with a single-stage vehicle or with the final stage of a multistage vehicle. If this is the case, the remaining propellant can be classed, for accounting purposes in what follows, with payload. Because we wish to consider the maximum attainable performance, we neglect any ullage in the analysis that follows.

In any case, we define the payload ratio λ as

$$\lambda = \frac{m_{pl}}{m_i - m_{pl}} \tag{5.26a}$$

or, with no ullage,

$$\lambda \simeq \frac{m_{pl}}{m_p + m_s} \tag{5.26b}$$

and the structural coefficient as

$$\varepsilon = \frac{m_s}{m_p + m_s} \tag{5.27a}$$

Again assuming no ullage at burnout,

$$\varepsilon \simeq \frac{m_f - m_{pl}}{m_p + m_s} \tag{5.27b}$$

The mass fraction η is also used frequently in place of the structural coefficient s:

$$\eta = \frac{m_p}{m_p + m_s} = 1 - \varepsilon \tag{5.28}$$

Assuming complete propellant depletion, the mass ratio becomes

$$MR = \frac{m_p + m_s + m_{pl}}{m_s + m_{pl}} \tag{5.29a}$$

or, in terms of the previously defined nondimensional quantities,

$$MR = \frac{1 + \lambda}{\varepsilon + \lambda} \tag{5.29b}$$

The advantage of a light structure (small ε) is clear. Because $(m_s + m_{pl})$ appears as a unit, structural mass trades directly for payload. The launch vehicle designer works to keep the structural coefficient as small (or propellant fraction as large) as possible.

There are limits on the minimum structural and control hardware required to contain and burn a given mass of propellant. We shall examine these limits in more detail later, but consider as an example the shuttle external tank (ET), which carries no engines and very little other equipment. On STS-1 (the first shuttle mission) the ET had an empty mass of approximately 35,100 kg and carried about 700,000 kg of propellant, yielding a structural coefficient of 0.0478 from Eq. (5.27). Subsequent modifications to the tank design produced a lightweight ET with a mass of approximately 30,200 kg and a structural coefficient of $\varepsilon = 0.0414$.

The most recent version of the ET, the superlight weight tank, makes extensive use of 2195 aluminum-lithium alloy and the lessons learned from earlier models. This tank weighs about 27,000 kg and carries 721,000 kg of propellant for a structural coefficient of $\varepsilon = 0.0361$. Results such as these represent the currently practical limits for a vehicle that must ascend from Earth. Thus, improvement in overall performance must be sought in other areas, chiefly (at least for chemical propulsion systems) by means of vehicle staging.

Staging is useful in two ways. First and most obviously, expended booster elements are discarded when empty, so that their mass does not have to be accelerated further. A second consideration is that the engines needed for initial liftoff and acceleration of the fully loaded vehicle are usually too powerful to be used after considerable fuel has burned and the remaining mass is lower. Even in unmanned vehicles where crew stress limits are not a factor, the use of very high

acceleration can cause much additional mass to be used to provide structural strength.

The analysis for a multistage vehicle is similar to that for a single stage. Assuming sequential operation of an N-stage vehicle, the convention is to define

m_{i_n} = nth stage initial mass, with upper stages and payload
m_{f_n} = nth stage final mass, with upper stages and payload
m_{s_n} = structural mass of nth stage alone
m_{p_n} = propellant mass for nth stage

The initial mass of the nth stage is then

$$m_{i_n} = m_{p_n} + m_{s_n} m_{i_{n+1}} \tag{5.30}$$

It is thus clear that the effective payload for the nth stage is the true payload plus any stages above the nth. The payload for stage N is the original m_{pl} from the single-stage analysis. By analogy with this earlier case, we define for the nth stage the ratios

$$\lambda_n = \frac{m_{i_{n+1}}}{m_{i_n} - m_{i_{n+1}}} \tag{5.31}$$

$$\varepsilon_n = \frac{m_{s_n}}{m_{i_n} - m_{i_{n+1}}} \simeq \frac{m_{f_n} - m_{i_{n+1}}}{m_{i_n} - m_{i_{n+1}}} \tag{5.32}$$

$$MR_n = \frac{m_{i_n}}{m_{f_n}} \simeq \frac{1 + \lambda_n}{\varepsilon_n + \lambda_n} \tag{5.33}$$

With these definitions, the basic result of Eqs. (5.23) still applies to each stage:

$$\Delta V_n = g I_{\text{sp}_n} \ell n MR_n \tag{5.34}$$

The total ΔV is the sum of the stage ΔV_n:

$$\Delta V = \sum_{n=1}^{N} \Delta V_n \tag{5.35}$$

and the mass ratio is the product of the stage mass ratios:

$$MR = \prod_{n=1}^{N} MR_n \tag{5.36}$$

The approach outlined in the preceding equations must be applied with care to parallel-burn configurations such as the Atlas, Delta, Ariane, Soyuz, space shuttle, Titan 3, etc. This is because fuel from more than one stage at a time may be used prior to a staging event, thus complicating the allocation of mass among the various stages. For example, m_{i_1} for the shuttle consists of the shuttle orbiter, external tank, and two solid rocket boosters (SRB). Following SRB separation,

m_{i_2} consists of the Orbiter and the external tank, less the fuel burned by the shuttle main engines prior to SRB separation. This complicates the definition of m_{s_n} and m_{p_n}; however, no fundamental difficulties are involved. Of greater concern is the fact that the SRBs and the shuttle main engines have substantially different I_{sp}. In such cases, staging analysis as presented here may be of little utility.

Hill and Peterson[2] examine the optimization of preliminary multistage design configurations, subject to different assumptions regarding the nature of the various stages. In the simplest case, where ε and I_{sp} are constant throughout the stages ("similar stages"), it is shown that maximum final velocity for a given m_{pl} and initial mass m_{i_1} occurs when $\lambda_n = \lambda$, a constant for all stages, where

$$\lambda = \frac{(m_{pl}/m_{i_1})^{1/N}}{1 - (m_{pl}/m_{i_1})^{1/N}} \tag{5.37}$$

The similar-stage approximation is unrealizable in practice; very often the last stage carries a variety of equipment used by the whole vehicle. Even if this is not the case, there are economies of scale that tend to allow large stages to be built with structural coefficients smaller than those for small stages. If we assume fixed I_{sp} for all stages but allow ε to vary, then for fixed m_{pl}, and m_{i_1}, maximum final velocity occurs for

$$\lambda_n = \frac{\alpha\varepsilon_n}{1 - \varepsilon_n - \alpha} \tag{5.38}$$

where α is a Lagrange multiplier obtained from the constraint on the ratio of payload to initial mass given by

$$\frac{m_{pl}}{m_{i_1}} = \prod_{n=1}^{N} \frac{\lambda_n}{1 + \lambda_n} = \prod_{n=1}^{N} \frac{\alpha\varepsilon_n}{(1 - \varepsilon_n - \alpha + \alpha\varepsilon_n)} \tag{5.39}$$

If all stages have both varying ε_n and I_{sp}, then again for MR and number of stages N, it is found that the maximum velocity is obtained with stage payload ratios:

$$\lambda_n = \frac{\alpha\varepsilon_n}{gI_{sp_n}(1 - \varepsilon_n) - \alpha} \tag{5.40}$$

where α is again found from the constraint

$$\frac{m_{i_1}}{m_{pl}} = \prod_{n=1}^{N} \frac{1 + \lambda_n}{\lambda_n} = \prod_{n=1}^{N} \frac{(1 - \varepsilon_n)(gI_{sp_n} - \alpha)}{\alpha\varepsilon_n} \tag{5.41}$$

With α known, λ_n may be found, and the mass ratio

$$MR_n = \frac{1 + \varepsilon_n}{\varepsilon_n + \lambda_n} \tag{5.42}$$

computed for each stage.

Finally, if it is desired to find the minimum gross mass for m_{pl}, final velocity V, and N with both ε_n and I_{sp_n} known variables,

$$\lambda_n = \frac{1 - \varepsilon_n MR^n}{MR_n - 1} \tag{5.43}$$

where

$$MR_n = \frac{1 + \alpha g I_{sp_n}}{\alpha \varepsilon_n g I_{sp_n}} \tag{5.44}$$

and α is found from the constraint on final velocity,

$$V = \sum_{n=1}^{N} g I_{sp_n} \ell n \left(\frac{\alpha g I_{sp_n} + 1}{\alpha \varepsilon_n g I_{sp_n}} \right) \tag{5.45}$$

As stated previously, α in Eqs. (5.38–5.45) is a Lagrange multiplier resulting from the inclusion of a constraint equation. In general, it will be found necessary to obtain the roots of Eqs. (5.39), (5.41), and (5.45) numerically.

Again, we point out that the preceding results ignore the effects of drag and gravity. Essentially, these are free-space analyses and are thus of questionable validity for ascent through a gravity field with steering maneuvers and atmospheric drag. Furthermore, the results of this section are inapplicable to the case of parallel burn or other consequential staging configurations. Though for detailed performance analysis it will be necessary to resort to the direct numerical integration of Eqs. (5.11), the results given earlier are useful for preliminary assessment.

5.2.3 Ascent Trajectories

The objective of the powered ascent phase of a space mission is to put the payload, often desired to be as large as possible, into a specified orbit. The manner in which this is done is important because small changes in the overall ascent profile can have significant effects on the final payload that can be delivered, as well as on the design of the ascent vehicle itself. The usual desire in astronautics is to maximize payload subject to constraints imposed by structural stress limits, bending moments, aerodynamic heating, crew comfort, range safety requirements, mission-abort procedures, launch site location, etc.

During the ascent phase the rocket vehicle must in most cases satisfy two essentially incompatible requirements. It must climb vertically away from the Earth at least as far as necessary to escape the atmosphere and must execute a turn so that, at burnout, the flight-path angle has some desired value, usually near zero. Few if any missions are launched directly into an escape orbit; thus, a satisfactory closed orbit is practically a universally required burnout condition for the ascent phase of a mission, unless it is a sounding rocket or ICBM flight. Except on an airless planet, high altitude at orbit injection is needed to prevent immediate reentry, and near-horizontal injection is usually necessary to prevent the orbit from intersecting the planet's surface.

A typical powered ascent into orbit will begin with an initially vertical liftoff for a few hundred feet, which is done to clear the launch pad prior to initiating further maneuvers. In general, the launch vehicle guidance system will be unable to execute pitch maneuvers about an arbitrary axis but will require such maneuvers to be done in a particular vehicle plane. This may also be a requirement due to the vehicle aerodynamic or structural configuration, as with the Titan 3 or space shuttle. In any case, if this plane does not lie along the desired launch azimuth, the rocket must roll to that azimuth prior to executing any further maneuvers. Following this roll, a pitch program is initiated to turn the vehicle from its initially vertical ascent to the generally required near-horizontal flight-path angle at burnout. The pitch program is often specified in terms of an initial pitch angle at some epoch (not necessarily liftoff) and a desired angular rate, $d\alpha/dt$, as a function of time. This closes Eqs. (5.11) and allows integration of the trajectory from the launch pad to burnout conditions.

Detailed examination of Eqs. (5.11) reveals a number of energy loss mechanisms that degrade ascent performance. These can be classified as thrust losses, drag losses, gravity losses, and steering losses. The selection of an ascent trajectory is governed by the desire to minimize these losses subject to the operational constraints mentioned earlier. Problems such as this are classically suited to the application of mathematical and computational optimization techniques, areas in which much theoretical work has been done. Examples with application to ascent trajectory optimization include the work of Bauer et al.,[10] Well and Tandon,[11] Brusch,[12] and Gottlieb and Fowler.[13]

Detailed discussion of these techniques is beyond the scope of this book and to some extent is also beyond the scope of the current state of the art in actual launch operations. In practice, ascent profiles are often optimized for given vehicles and orbital injection conditions through considerable reliance on trial and error and the experience of the trajectory designer. We will examine in this section some of the basic considerations in trajectory design and the tradeoffs involved in the selection of an ascent profile.

Thrust loss has already been discussed; here we are speaking of the degradation in specific impulse or thrust due to the pressure term in Eq. (5.1) when exit pressure is less than ambient pressure. If the engine is sized for sea level operation, it is then much less efficient than a fully expanded engine for the

high-altitude portions of its flight. However, any rocket engine delivers better performance at higher altitudes. It is thus advantageous, from the point of efficient utilization of the propulsion system, to operate the vehicle at high altitudes as much as possible.

The dependence of vehicle drag on atmospheric density, flight velocity, angle of attack, and body shape was discussed in Sec. 5.2.1 (Equations of Motion). From Eq. (5.14b), it is clearly desirable to operate at high altitudes, again as early as possible in the flight, because drag is proportional to atmospheric density. On the other hand, it is advantageous to ascend slowly to minimize the effect of the squared velocity in regions of higher density.

Gravity losses are those due to the effect of the term ($g\sin\gamma$) in Eq. (5.11a). To minimize this term, it is desirable to attain horizontal flight as soon as possible. Also, careful consideration will show that, to minimize gravity losses, the ascent phase should be completed as quickly as possible, so that energy is not expended lifting fuel through a gravity field only to burn it later. Other factors being equal, a high thrust-to-weight ratio is a desirable factor; an impulsive launch, as with a cannon, is the limiting case here but is impractical on a planet with an atmosphere. However, electromagnetic mass-drivers, which are essentially electric cannons, have been proposed for launching payloads from lunar or asteroid bases to the vicinity of Earth for use in orbital operations.[14] The opposite limiting case occurs when a vehicle has just sufficient thrust to balance its weight; it then hangs in the air, expending its fuel without benefit.

Finally, steering loss is that associated with modulating the thrust vector by the $\cos\alpha$ term in Eq. (5.11a). Clearly, any force applied normal to the instantaneous direction of travel is thrust that fails to add to the total vehicle velocity. Thus, any turning of the vehicle at all is undesirable. If done, it should be done early, at low speeds. This is seen in Eq. (5.11b), where, if we specify for example a constant flight-path angular rate (i.e., constant $d\gamma/dt$), the required angle of attack varies as

$$\alpha = \sin^{-1}\left\{\frac{V\,d\gamma/dt + (g - V^2/r)\cos\gamma - L/m}{T/m}\right\} \tag{5.46}$$

It is seen that larger flight velocities imply larger angles of attack to achieve a fixed turning rate.

The preceding discussion shows the essential incompatibility of the operational techniques that individually reduce the various ascent losses. Early pitchover to near-horizontal flight, followed by a long, shallow climb to altitude, minimizes steering and gravity losses but dramatically increases drag and aggravates the problem by reducing the operating efficiency of the power plant. Similarly, a steep vertical climb can minimize drag losses while obtaining maximum engine performance, at the price of expending considerable fuel to go in a direction that is ultimately not desired. Experience reported by Fleming and Kemp[15] indicates that the various energy losses result in typical first-stage

burnout velocities about 70% of the theoretical optimum as given in Eqs. (5.23) for the 0-g drag-free case.

There are a number of special cases in pitch rate specification that are worthy of more detailed discussion. The first of these is the gravity turn, which is defined by the specification of an initial flight-path angle γ and the requirement that the angle of attack be maintained at zero throughout the boost. In this way, no thrust is wasted in the sense of being applied normal to the flight path. All thrust is used to increase the magnitude of the current velocity, and α is controlled to align the vehicle (and hence the thrust vector) along the current velocity vector. Because the term $g \cos\gamma$ in Eq. (5.11b) produces a component normal to the current flight path, a gradual turn toward the horizontal will be executed for any case other than an initially vertical ascent. Setting the angle of attack to zero and solving Eq. (5.11b) for pitch rate, we find

$$\frac{d\gamma}{dt} = \frac{L/m - (g - V^2/r)\cos\gamma}{V} \tag{5.47}$$

where we note that the lift L is generally small and is zero for rotationally symmetric vehicles at zero angle of attack. It is seen that low velocity or small flight-path angle increases the turn rate.

This approach would seem to be most efficient, as with zero angle of attack the acceleration V is maximized. However, this is strictly true only for launch from an airless planet. In the case of an Earth ascent, a rocket using a gravity turn would spend too much time at lower levels in the atmosphere, where other factors act to offset the lack of steering loss. Selection of higher initial values of γ, for more nearly vertical flight, does not generally allow the turn to horizontal to be completed within the burn time of the rocket. In general, gravity turns may comprise portions of an ascent profile but are unsuitable for a complete mission. An exception is powered ascent from an airless planet such as the moon; the trajectories used for Lunar Module ascent flight closely approximated gravity turns.

The case of constant flight-path angle is also of interest. Particularly with the final stage, the launch vehicle spends much of its time essentially above the atmosphere and accelerating horizontally to orbital velocity, with no need to turn the vehicle. In this case, Eq. (5.11b) is solved to yield for the pitch angle:

$$\alpha = \sin^{-1}\left\{\frac{m(g - V^2/r)\cos\gamma - L}{T}\right\} \tag{5.48}$$

Most trajectories can be approximated by combinations of these two segments, plus the constant pitch rate turn noted earlier. In practice, once a desired trajectory is identified, implementation is often in the form of a series of piecewise constant steps in pitch rate, $d\alpha/dt$, chosen to approximate a more complicated curve. Such profiles tend in general to follow a decaying exponential

of the form[15]

$$\frac{d\alpha}{dt} = Ae^{-K(t-t_0)} \tag{5.49}$$

where

A = amplitude factor
K = shape factor
t_0 = time bias

For such a case, Fleming and Kemp develop a convenient trajectory optimization method that allows substantial reductions in the time required to design representative two-stage ascent profiles. However, realistic ascent profiles can also be considerably more complex, as illustrated by the launch sequence[16] for STS-1, summarized in Table 5.2. Space shuttle ascent guidance strategy and algorithms are reported by McHenry et al.,[17] Schleich,[18,19] Pearson,[20] and Olson and Sunkel.[21]

5.2.4 Rocket Vehicle Structures

As has been discussed, it is the sophistication of the electrical, mechanical, and structural design that produces low values of structural coefficients for each stage. Some general rules may be observed. Large stages tend to have lower values of ε than smaller stages. As mentioned, this is because some equipment required to construct a complete vehicle tends to be relatively independent of vehicle size. Also, the mass of propellant carried increases with the volume enclosed, but the mass of the tankage required to enclose it does not. Denser fuels allow more structurally efficient designs for given specific impulse, because a smaller structure can enclose a larger mass of propellant. This factor tends to remove some of the theoretical performance advantage of liquid hydrogen, particularly for first-stage operation, and was a reason it was not selected for use on the first stage of the Saturn 5.

This area is the province of the structural design specialist, and its details are beyond the scope of this book. For preliminary design calculations and assessments, as well as to provide a "feel" for what is reasonable and possible, we include in Table 5.3 data on a wide range of vehicle stages and the structural coefficient for each.

5.3 Launch Vehicle Selection

5.3.1 Solid vs Liquid Propellant

The late 1950s and early 1960s were a time of strong debate between the proponents of solid propulsion and those of liquid propulsion. At stake was the

Table 5.2 STS-1 ascent timeline

Time	Altitude	Comments
8 s	400 ft	120° combined roll/pitch maneuver for head-down ascent at 90° launch azimuth.
32 s	8000 ft	Throttle back to 65% thrust for maximum pressure at 429 kTorr.
52 s	24,000 ft	Mach 1, throttle up to 100%.
1 min 53 s	120,000 ft	Mach 3, upper limit for ejection seats.
2 min 12 s	27 n mile	Mach 4, 5, SRB jettison.
4 min 30 s	63 n mile	Mach 6.5, limit of return to launch site (RTLS) abort. Pitch from +19 to −4°. Initially lofted trajectory for altitude in case of single engine failure.
6 min 30 s	70 n mile	Mach 15, peak of lofted trajectory.
7 min	68 n mile	Mach 17, 3-*g* acceleration limit reached. Throttle back to maintain 3 *g* maximum.
8 min 32 s	63 n mile	Main engine cutoff; 81 × 13 n.mi. orbit.
8 min 51 s	63 n mile	External tank jettison.
10 min 32 s	57 n mile	OMS[a] 1 burn, attain 130 × 57 n.mi. orbit.
44 min	130 n mile	OMS 2 burn, attain 130 × 130 n.mi. orbit.

[a]Orbital Maneuvering System.

direction of development of launch vehicles for the Apollo lunar mission and possibly even larger vehicles beyond that.

Solid propellants offer generally high reliability and high mass fraction resulting, respectively, from a relative lack of moving parts and high propellant density. Liquid-propellant engines generally achieve higher specific impulse and better thrust control, including throttling, restart capability, and accurate thrust termination. Development of liquid oxygen/liquid hydrogen stages with high

Table 5.3 Structural coefficient vs mass

Stage	Mass	Structural coefficient
	LO_2/LH_2 stages	
Ariane-4 3rd stage	9400 kg	0.127
Centaur III	22,960 kg	0.09
Delta IV-H	30,710 kg	0.11
Saturn SIV-B	105,000 kg	0.093
Saturn SII	437,727 kg	0.078
	Earth-storable or LO_2/hydrocarbon	
Delta II 2nd stage	6499 kg	0.077
Titan III 2nd stage	33,152 kg	0.081
Titan III 3rd stage	124,399 kg	0.051
Ariane IV 2nd stage	36,600 kg	0.098
Ariane IV 3rd stage	160,000 kg	0.0875
Atlas	110,909 kg	0.036

specific impulse and good mass fraction has led to extensive use of this propellant combination for upper stages.

Considerable effort has gone into the development of solid rockets having some of the desirable liquid motor characteristics such as controlled thrust termination, multiple burns, and throttling. Various thrust termination schemes such as quenching and explosively activated vent ports have been successfully developed. Multiple burn and throttling concepts have been less productive, due in part to the fact that such features greatly increase the complexity of the motor and vitiate one of its main advantages, that of simplicity. (It should be noted that the simplicity of solid rocket motors refers to their operational characteristics. The design and fabrication of high performance solid boosters is a complex and demanding exercise.)

In some cases preprogrammed thrust variations can be used to accomplish for solid rocket motors what is done by throttling liquids. As solid propellant technology has matured, it has become possible to tailor the thrust vs time profile in a fairly complex manner. Except for the fact that the profile is set when the motor is cast and cannot be varied in response to commands, this thrust profile tailoring is almost as good as throttling for purposes such as moderating inertial or aerodynamic loads during ascent.

Thrust vector control has progressed substantially as well. Most early solid motor systems were spin stabilized. Although this is still common practice, it is not satisfactory for large vehicles or those requiring precise guidance. In such cases three-axis control is required. Early attempts to attain it included the use of jet vanes as in the liquid propellant V-2 or Redstone, or jetavator rings. These devices are swivel-mounted rings that surround the nozzle exit and are activated to dip into the flow, thus deflecting it. Multiple nozzles, each with a ring, can

provide full three-axis control. Such an approach was used on the early Polaris missiles.

Vanes and rings are prone to failures through erosion and thermal shock effects and also reduce performance by introducing drag into the exhaust stream. Approaches circumventing these difficulties include the use of gimballed nozzles and nozzle fluid injection. The most notable use of fluid injection was found in the Titan 3 series of solid rocket boosters. Each booster carried a tank of N_2O_4 (nitrogen tetroxide) under pressure. Four banks of valves located orthogonally on the nozzle exit cone were used to control the injection of the N_2O_4 through the nozzle wall into the supersonic flow. In this concept, the intruding fluid produces a local shock wave that generates a downstream high-pressure region and a consequent flow deflection. Multiple valves provide various levels of sideloading and enhance reliability.

The fluid injection approach is simple and reliable, but again introduces performance losses due to the mass of the injection system and to the generation of shock waves used to turn the supersonic flow.

The minimum performance degradation for a thrust vector control system is obtained through the use of a gimballed nozzle. Providing a nozzle for a solid-propellant motor that can move freely under substantial thrust and aerodynamic loads while preventing leakage of very hot, high-pressure, and frequently corrosive gas is a major design challenge. Advances in mechanism and structural design as well as materials engineering have been required. Successful designs have evolved to meet requirements ranging from the small, multiple nozzle configurations of Minuteman and Polaris to the much larger single nozzle solid rocket boosters used on the space shuttle. A reminder of the difficulty involved in designing a gimballed nozzle for a solid rocket motor is provided by the in-flight failure of the Inertial Upper Stage (IUS) during the STS-6 mission in April 1983. The difficulty was traced to overheating and erosion of a fluid bearing seal in the nozzle gimbal mechanism.[22]

Solid-propellant rockets are most useful for applications requiring high thrust from a compact package in a single burn. First stage propulsion as on the Ariane 5, Titan 3 and 4, and shuttle and first stage auxiliary propulsion as on Ariane 4, Delta, and Atlas are prime examples of solid motor applications. Another major application is for apogee and perigee "kick motors" for Earth-orbiting spacecraft, typically communications satellites deployed in geostationary orbit. High thrust is not necessarily a virtue for these applications, but reliability, ease of integration, and simplified ground operational requirements are often crucial.

Liquid-propulsion systems are, in an operational sense, more complex than their solid counterparts. There must be some type of propellant flow control and some means of feeding propellant to the combustion chamber. In even the simplest systems, this requires the use of active components and introduces additional possibilities of failure. However, liquids offer flexibility in thrust levels, burn time, and number of burns, coupled with generally higher specific impulse.

Hybrid propulsion systems, normally consisting of a solid fuel and a liquid oxidizer, offer some of the advantages (and disadvantages) of both systems. Performance is better than solids but, as a rule, is inferior to liquid systems. Hybrid motors throttle readily, are easily restarted and lend themselves to low-cost production. These systems received considerable attention during the 1960s, but further development languished afterward. Interest in hybrids is again on the rise, in part because of their almost total freedom from the risk of detonation in even the most violent impact.[23]

Of particular interest for liquid-propellant engines is their ability to be throttled. Although not trivial to develop, this capability is far easier to include in a liquid-propellant engine than in a solid rocket motor and is mandatory in many applications such as planetary landers (Surveyor, Apollo, Viking). Throttling capability is also highly desirable for some ascent propulsion systems, such as the space shuttle, to reduce structural loads and to obtain greater control over the ascent profile.

The primary difficulty in throttling is in maintaining an adequate injector pressure drop. For most types of combustion chamber injectors, the pressure drop across the injector is crucial to ensuring good atomization and mixing of the propellants. Adequate atomization and mixing are crucial to high performance and smooth operation. With a fixed injector area, pressure drop varies as the square of the flow rate according to

$$\Delta p = \dot{m}/2\rho C_d^2 A^2 \tag{5.50}$$

where

$\dot{m}$ = mass flow rate
A = total injector orifice cross-sectional area
C_d = injector orifice discharge coefficient
ρ = propellant density
Δp = pressure drop across injector

For throttling ratios up to about 3:1 this can be tolerated by designing for an excessively high pressure drop at full thrust. This is probably acceptable only for relatively small systems, because the higher supply pressure will result in a substantially heavier overall system.

The lunar module descent engine required approximately a 10:1 ratio between full and minimum thrust using the liquid-propellant combination of nitrogen tetroxide and a 50/50 mixture of hydrazine and unsymmetric dimethylhydrazine. The propellant flow rate was controlled by variable-area cavitating venturi tubes (see Fig. 5.14). The contoured movable pintles were connected mechanically to a movable sleeve on the single element coaxial injector. As the venturi pintles moved to change the flow rate, the injector sleeve moved to adjust injector area to

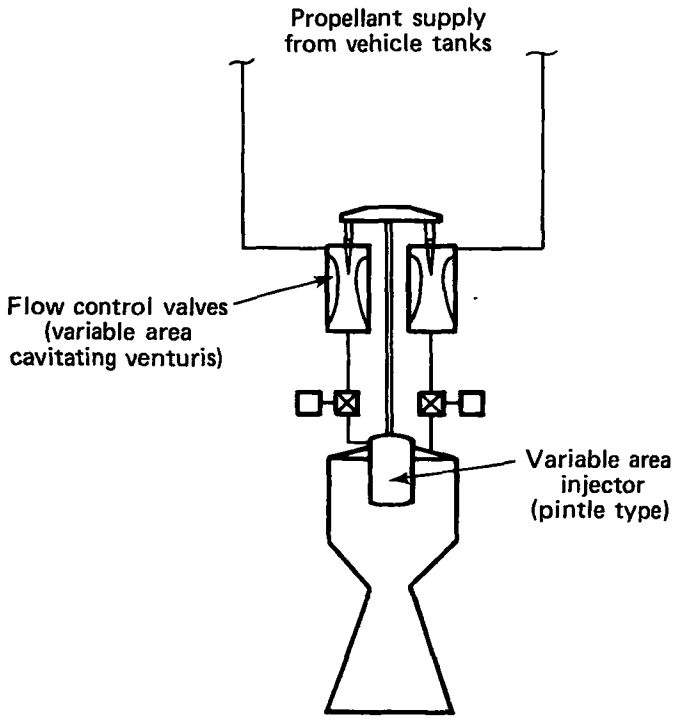


Fig. 5.14 Rocket engine diagram—throttling engine (LMDE type).

maintain a suitable pressure drop. This system could throttle over a range of nearly 20:1 while maintaining satisfactory performance.

Alternative approaches such as injecting an inert gas into the injector to entrain the liquid and maintain the flow momentum during throttling have been tried with somewhat less success. Multielement concentric tube injectors (Fig. 5.15), which were used in the RL-10 and J-2, seem to offer the best overall performance with liquid hydrogen and are more tolerant of throttling than other fixed orifice injectors. This is probably due to the high velocity of the hydrogen and its tendency to flash quickly to a gas as the pressure drops. This tendency is enhanced by the temperature increase incurred in thrust chamber cooling. Micro-orifice injectors (Fig. 5.16) are also less dependent upon pressure for atomization and mixing and thus are able to tolerate a wider range of injector pressure drops.

Design of engine throttling systems is complex, and the details are beyond the scope of this text and are only peripherally relevant to the overall system design problem. The spacecraft systems designer must be aware of the problems and potential solutions in order to allow evaluation of competing concepts. From a systems design point of view, the best approach to engine throttling may be to minimize the extent to which it is needed. The Surveyor spacecraft demonstrated a design approach that avoided the need for a difficult-to-attain deep-throttling

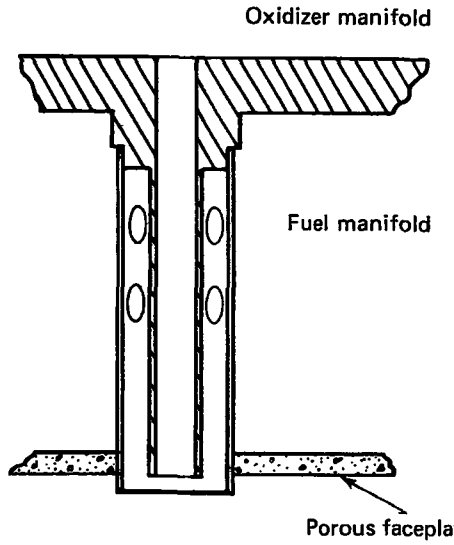


Fig. 5.15 Cross section—single element of concentric tube injector.

capability. A solid-propellant motor was used to remove most of the lunar approach velocity. Small liquid-propellant engines that provided thrust vector control during the solid motor burn then took over control of the descent following termination of the solid rocket burn. A relatively small amount of velocity was removed during the final descent, with emphasis placed on controlling the maneuver along a predetermined velocity vs altitude profile until essentially zero velocity was reached just above the surface. For this descent maneuver sequence, a throttle ratio of 3:1 was adequate.

Selection of a liquid propellant, solid propellant, or some combination of these for ascent, upper stage, or spacecraft propulsion is another example of a design issue with no single "correct" resolution. In general, if the problem can be solved at all, it can be solved in more than one way, selection of the final system will depend on such factors as cost, component availability, environmental considerations, etc., as well as traditional engineering criteria such as reliability and performance.

5.3.2 Launch Vehicles and Upper Stages

The number and variety of launch vehicles and upper stages, at least those built in the U.S., decreased substantially following the development of the space shuttle. However, the 1986 *Challenger* accident and the growing reluctance of the military to depend solely on the shuttle led to a resurrection of several

formerly available vehicles. Since the mid-1980s launch vehicle development has proceeded with evolution of existing designs as well as development of new vehicles. Some of the stages intended for use in launching commercial payloads from shuttle, an activity terminated after *Challenger*, are no longer produced. Further changes in both commercial launch options and U.S. government space transportation policies can be expected as a result of the 2003 *Columbia* accident; however, as this is written, the nature and details of such changes have not been defined.

We include a discussion of currently available launch vehicles and upper stages, primarily to convey a sense of typical vehicle capability, requirements, and constraints. Questions concerning the details of launch vehicle interface requirements, orbital performance, etc., should in all cases be referred to a current edition of the user's guide available from the manufacturer of the particular vehicle.

With the changing world political climate, U.S. payloads are now flying on Russian and Chinese launch vehicles as well as U.S. vehicles and the European Ariane series. Some commercially developed vehicles have also appeared.

Launch vehicles at present fall into two categories: the U.S. Space Transportation System (STS) and expendable vehicles. The expendable family comprises a variety of vehicles, many of them derived from military IRBMs and ICBMs, which can accommodate payloads ranging from a few kilograms in LEO to several thousand kilograms in interplanetary trajectories.

5.3.2.1 Space shuttle payload accommodations. Even though basic aspects of the space shuttle system will be familiar to most readers, we will describe it briefly here for the sake of completeness. The major components of the system are shown in Fig. 5.17. The central component of the shuttle "stack" is the Orbiter, a delta-winged aerospacecraft that contains the crew accommodations and cargo bay as well as main, auxiliary, and attitude propulsion systems and propellant for orbital maneuvers, along with power generation and control functions. The SRBs are used during the first 2 min of flight, after which they are jettisoned and recovered for refurbishment and reuse.

The external tank (ET) carries the entire supply of LO_2 and LH_2 for the main propulsion system. It is normally jettisoned just prior to orbital insertion and is destroyed during reentry over the Indian Ocean (for a due-East launch from Cape Canaveral). It is worth noting that little extra energy is required to retain the ET through injection into orbit, where the tank material and the residual propellant it carries could in some circumstances be quite useful. Many scenarios have been advanced for the use of surplus ETs during heavy construction work in LEO, though so far none have come to pass.

Generally speaking, shuttle payloads are carried in the cargo bay, which provides a clear space 15 ft in diameter and 60 ft long (4.6 × 18.3 m). Limited capacity for experiments also exists in the main cabin, which of course are

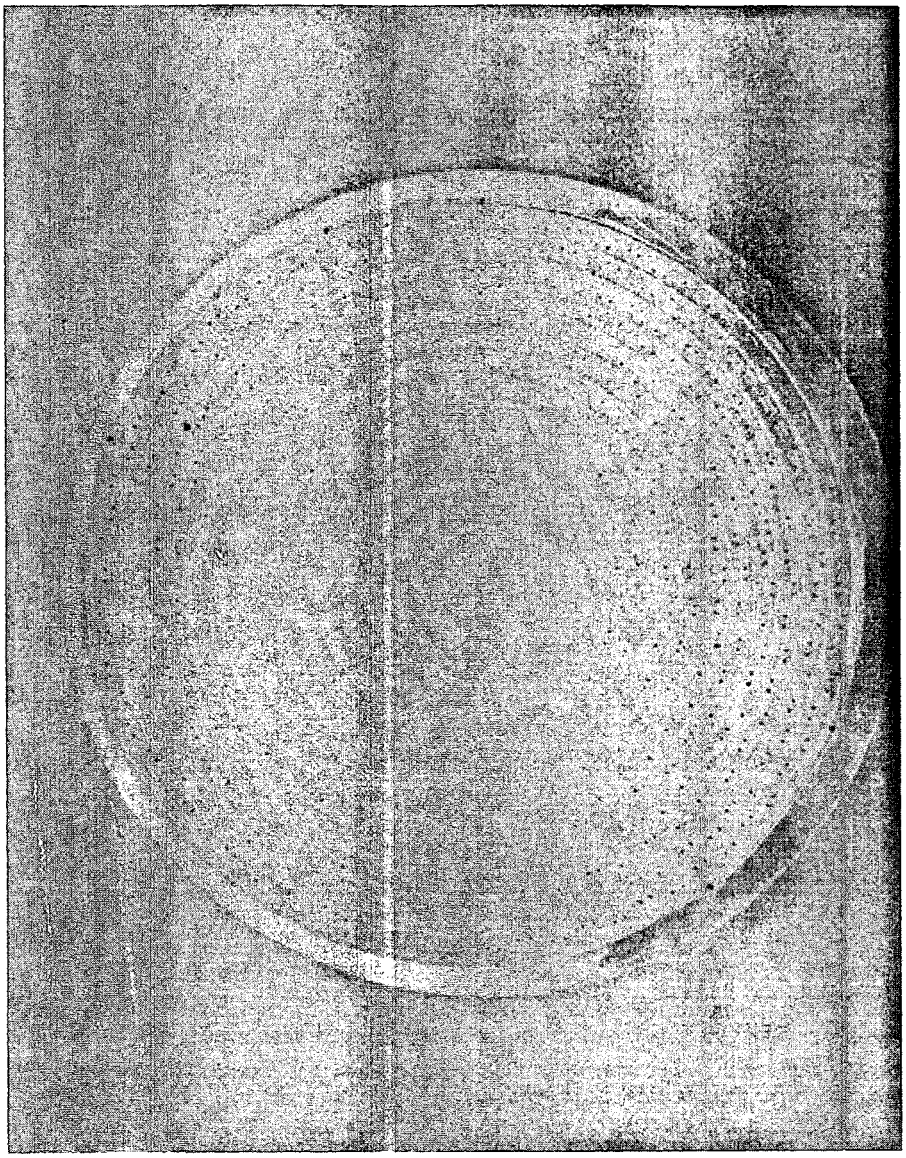


Fig. 5.16 Shuttle OMS engine injector. (Courtesy of Aerojet.)

restricted to those that do not require access to space and that pose no hazard to the flight crew and Orbiter systems. One of the first examples in this regard was the continuous-flow electrophoresis experiment by McDonnell Douglas (now Boeing) and Ortho Pharmaceuticals, which flew aboard the shuttle even during its initial test flights.

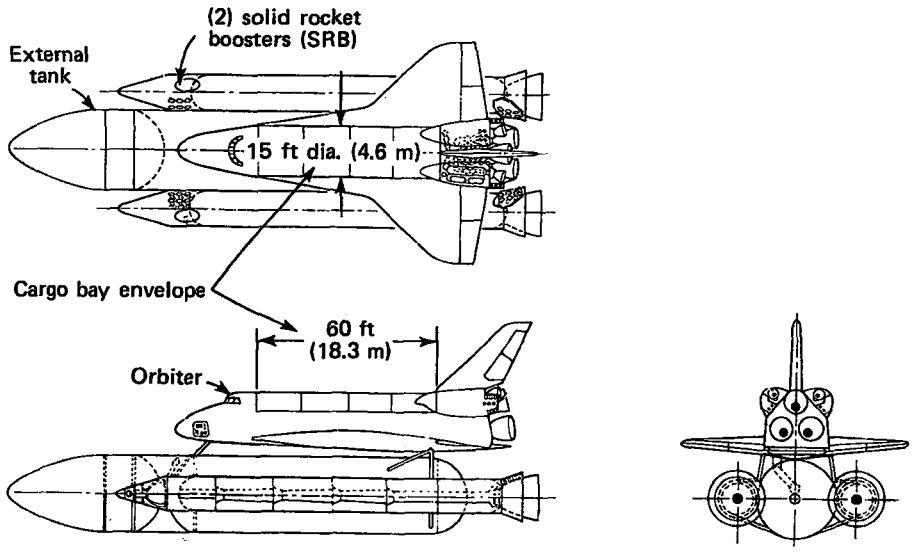


Fig. 5.17 Space shuttle flight system.

As an aside, various external means of carrying cargo have been suggested (but not flown), most prominent among them being the so-called aft cargo container, which would fit behind the ET and would be especially useful for payloads for which the 15-ft diameter constraint of the payload bay poses a problem. A disadvantage is that the ET must be carried into orbit also and, as discussed, this does involve some performance penalty.

The payload bay is not pressurized and thus will see essentially ambient pressure during ascent, orbital flight, and descent. Access to space is obtained by opening double doors that expose the full length of the payload bay to space. Because the doors also hold radiators needed for thermal control of the Orbiter, they must be opened soon after orbit insertion and must remain so until shortly before reentry. Payloads must therefore be designed to withstand the resulting environment, which may involve both extended cold soaking and lengthy periods of insolation. Space environmental effects are discussed in more detail in Chapter 3.

Provisions for mounting payloads in the bay are unique in launch vehicle practice and reflect the dual rocket/airplane nature of the vehicle, as well as the desire to accommodate a variety of payloads in a single launch. The support system shown in Fig. 5.18 consists of a series of support points along the two longerons that form the "doorsills" of the bay and along the keel located along the bottom (referenced to landing attitude) of the bay. Proper use of four of these attach points to apply restraint in selected directions as shown in Fig. 5.18a can provide a statically determinant attachment for even a large item such as an upper stage or a spacelab module.

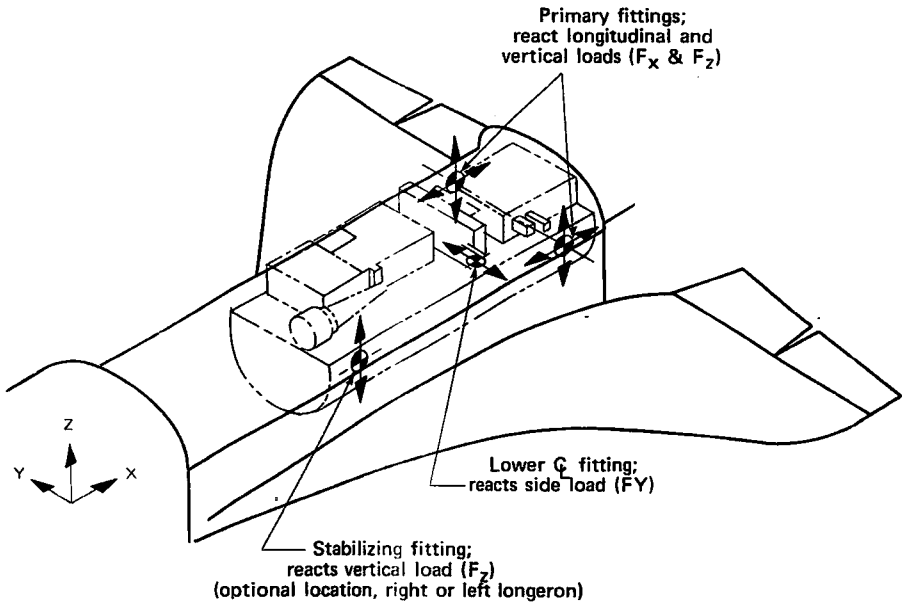


Fig. 5.18a Statically determinant shuttle payload attach points.

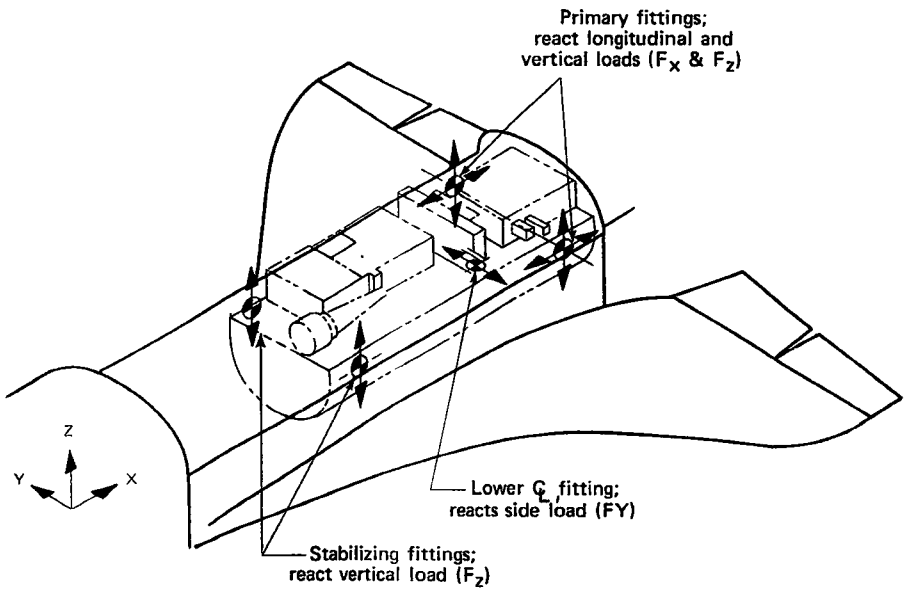


Fig. 5.18b Five-point payload retention system (indeterminate).

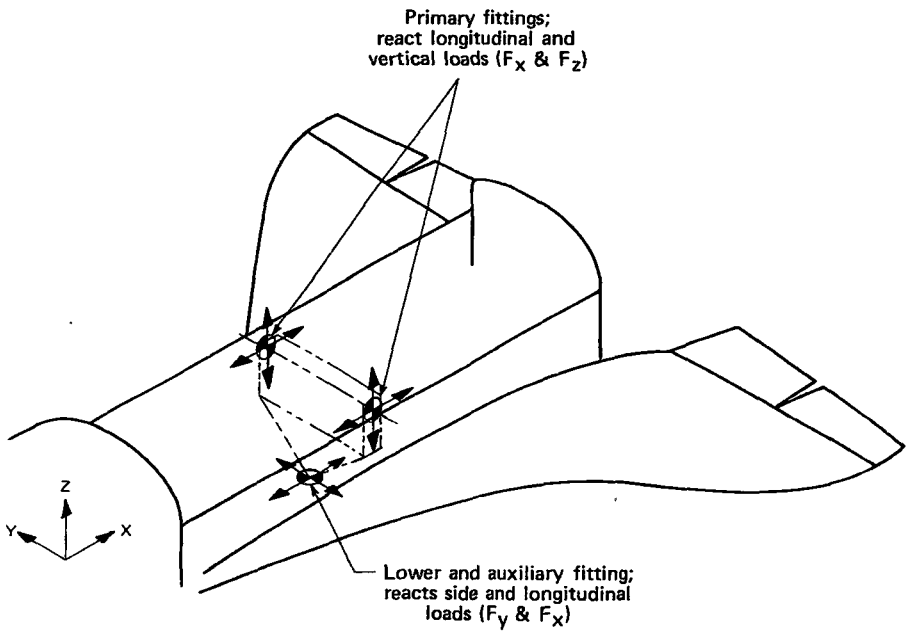


Fig. 5.18c Three-point payload retention system (determinate).

Other payload attachment accommodations are also available. Spacecraft and upper stages are usually mounted in a cradle or adapter that provides an interface to the shuttle attach points. The usual purpose of this procedure is to accommodate deployment mechanisms and to provide mounting locations for various auxiliary equipment. These structures impose some penalty on total shuttle payload mass and volume and in theory at least could be eliminated by having the upper stage or payload interface directly to the shuttle attachment points. However, this may also impose a mass penalty or design constraint on the payload, because the interface structure often performs a load leveling or isolation function that, in its absence, would be required of the payload itself.

A variety of methods are used to deploy payloads from the Orbiter bay, depending on such factors as the payload stabilization mode and instrument requirements. The manipulator arm may also be used to deploy payloads or to grapple with and return them to the bay.

Payloads can be installed in the bay while the shuttle is in the horizontal attitude in the Orbiter Processing Facility (OPF). This is typically done several weeks prior to launch, and the payload will thus remain with the shuttle throughout erection, mating to the ET, and rollout to the pad. Environmental control may not be maintained throughout all these operations, and for some payloads this may be unacceptable. In such cases, the payload may be installed vertically on the pad using the Rotating Service Structure (RSS). This is probably most desirable for spacecraft on large upper stages such as IUS. Advantages may

also exist in allowing later commitment to a particular spacecraft and in maintenance of environmental control.

The shuttle is basically a low-orbit transportation system. Generally, the orbit achieved is nearly circular in the range of 300–400 km. The original Shuttle design was intended to carry a 29,500-kg (65,000-lbm) payload into a 300-km circular orbit at the 28.5° inclination which results from a due-East launch from Cape Canaveral. The earliest shuttle vehicles, *Columbia* and *Challenger*, had a payload capability in the 24,000- to 26,000-kg range and therefore did not fully meet the design requirement, whereas the later vehicles, *Discovery*, *Atlantis*, and *Endeavour* met or exceeded it. *Endeavour* was the replacement for the ill-fated *Challenger* and incorporated all the improvements developed over the intervening years.

If a higher orbit or an inclination differing significantly from 28.5° is required, the payload must become lighter. It is notable that, for missions supporting assembly of the International Space Station in its 51.6° inclination orbit at 400 km, the payload capacity is about 16,000 kg.

Obviously, for missions beyond the maximum shuttle altitude (about 1000 km), auxiliary propulsion on the payload spacecraft is mandatory. Most traffic beyond LEO is destined for geosynchronous orbit, although particular scientific or military missions such as the Global Positioning System will require other orbits, as discussed in Chapter 2. A small but significant number of missions will be intended for lunar, planetary, or other deep space targets. A variety of upper stages have been developed or are under development to satisfy these requirements and will be discussed in some detail in later sections. The spacecraft designer is not restricted to the use of these stages, however, and may elect to design his own propulsion system as part of the spacecraft. As an example, the former Hughes Aircraft Corporation (now Boeing) has elected to design its own propulsion stage for many of its geosynchronous orbit communications spacecraft.

Even where full orbital transfer capability is not included with the basic spacecraft design, some auxiliary propulsion is often required. Again using the GEO example, most upper stages provide only the capability for inserting the spacecraft into the so-called geosynchronous transfer orbit (GTO), i.e., the apogee raising maneuver from the initially circular orbit. This maneuver will require (see Chapter 4) on the order of 2.5 km/s. A second maneuver of about 1.8 km/s combining a plane change and a perigee raising burn must be done at the GTO apogee. The motor for this "apogee kick" is often designed as an integral part of the spacecraft.

The specialized stage approach, using mostly existing components, may become more popular in the future, because available stages are not often optimal for a given task. This choice will be influenced not only by the nature of the payload but by the number of missions to be flown, because a custom stage may be economically justifiable for use with a series of spacecraft, but impractical for a single application.

The sole shuttle launch capability exists at the Kennedy Space Center (KSC) in Florida. Launches from this facility can achieve orbital inclinations between approximately 28.5° and 57° without difficulty. The lower limit is determined by the latitude of the launch site, as discussed in Chapter 4, and the upper limit results from safety constraints on SRB and ET impact zones. Inclinations outside these launch azimuth bounds can be (and have been) attained from KSC by using a "dogleg" maneuver on ascent, or by executing a plane change maneuver once in orbit. Both of these procedures result in a reduction in net payload delivered to the desired orbit.

Orbital inclination for many missions is not a critical parameter, and when this is so, KSC is the launch site of choice. However, as discussed in Chapter 2, many missions (e.g., communications satellites) require equatorial orbits or (as with military reconnaissance spacecraft) near-polar orbits. The requirement for a high-altitude equatorial orbit is met rather easily from KSC by executing any required plane change at the apogee of the geosynchronous transfer orbit, which can be chosen to occur at the equator. As discussed in Chapter 4, such a plane change imposes a ΔV requirement of approximately 0.8 km/s when performed separately and less when combined with the usually desired circularization maneuver.

Because, due to safety constraints, polar orbits cannot be achieved from KSC, such requirements have been met by expendable launches from Vandenberg AFB (VAFB) near Lompoc, California. This site can accommodate orbits with inclinations from about 55° to slightly retrograde. Prior to the *Challenger* accident it had been planned to conduct shuttle launches from this facility as well, even though performance penalties are substantial, as seen by comparing Fig. 5.19 and Fig. 5.20. However, the increased concern for safety in the wake of the *Challenger* accident, as well as cost, schedule, and facility concerns, resulted in cancellation of these plans. We have included Figs. 5.19 and 5.20 for historical interest, and to provide some insight into the performance impact of launching north or south, without the benefit of the Earth's rotation.

5.3.2.2 Expendable launch vehicles. During the period of space shuttle conceptual design and early development, it was frequently stated that when the STS became fully operational, expendable launch vehicles would become extinct. This proved incorrect, and expendable vehicles have continued to meet the majority of U.S., not to mention international, space launch requirements. Many expendable vehicles in use today have their origins in the 1960s; others, such as the European Space Agency's Ariane launcher family, were designed more recently. These vehicles are considered in the following sections.

Ariane. This family of launch vehicles was developed by a French-German consortium and marketed by the semiprivate organization Arianespace. The original Ariane vehicle was developed specifically as a competitor for the shuttle, and in particular for the lucrative geosynchronous orbit market. Ariane's design

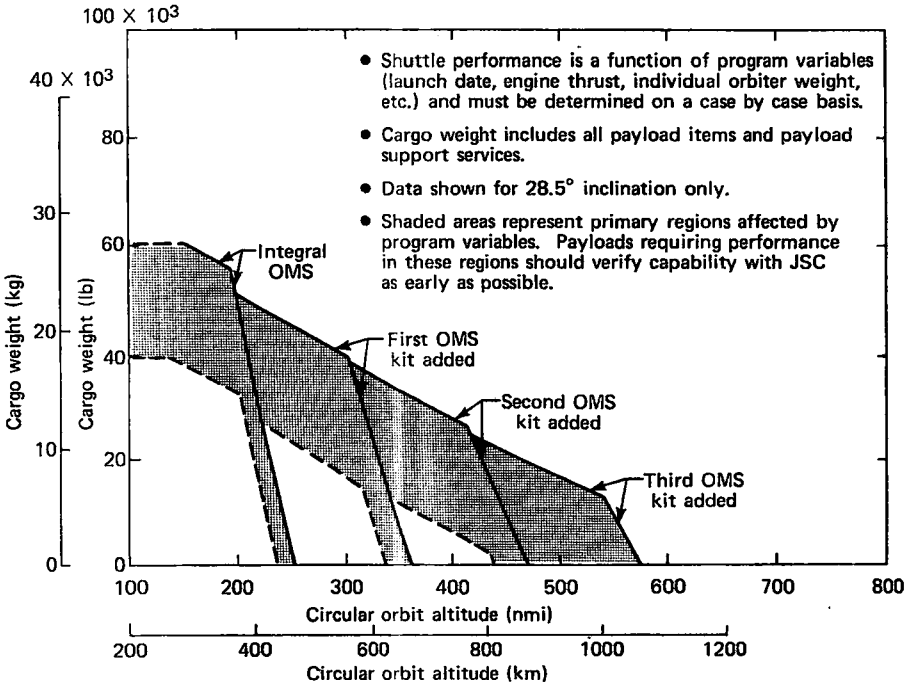


Fig. 5.19 Cargo weight vs circular orbital altitude for KSC launch.

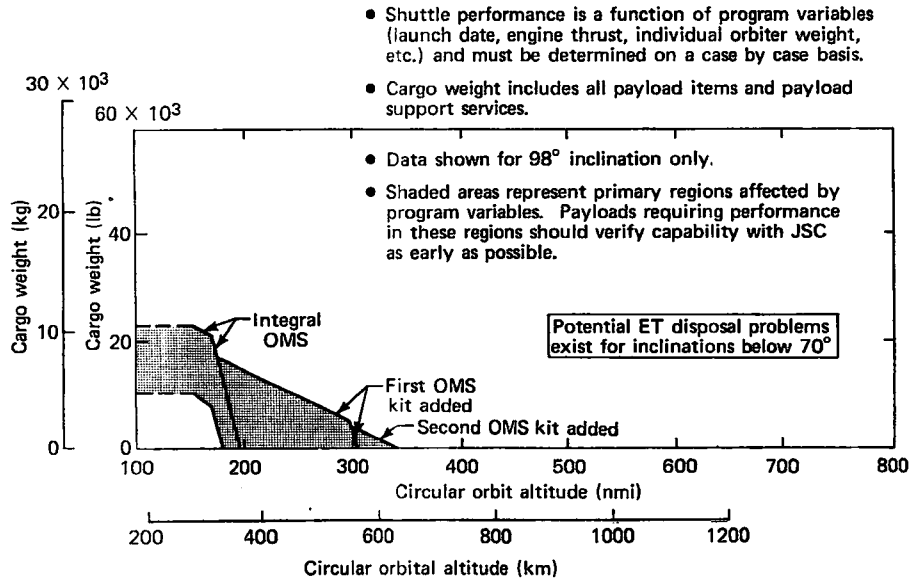


Fig. 5.20 Cargo weight vs circular orbital altitude—VAFB launch.

was antithetical to that of the STS, consisting as it does of a very conventional three-stage expendable launcher. Some early consideration was given to recovery and reuse of the first stage, but such plans were never implemented.

The basic Ariane 1 was optimized for delivery of payloads to GTO, an orbit with a perigee of 200 or 300 km and an apogee at the geosynchronous altitude of 36,000 km. Its third stage did not have an orbital restart capability, and there was thus no ability to coast in LEO prior to transfer orbit initiation. This inhibited application of Ariane to planetary missions as well as to others requiring multiple maneuvers. Despite this restriction, however, clever mission design allowed Ariane to launch several deep space missions. The payload spacecraft was placed in a highly elliptical parking orbit, similar to GTO, until the proper alignment with the departure asymptote occurred, at which point a propulsion stage was ignited to initiate the remainder of the mission.

As noted, the early Ariane was a conventional three-stage vehicle with all stages using pump-fed liquid propellants. The lower two stages burned Earth-storable propellants, and the third used cryogenic liquid oxygen and hydrogen. All Ariane launches are from Kourou, French Guiana, on the northeast coast of the South American continent. This site is only about 5° north of the equator, and thus has significant performance advantages for low inclination orbits and ample open sea areas to the east for down-range stage impact. Numerous upgrades were implemented, culminating in the very reliable Ariane 4 vehicle.

In tribute to the foresight of the designers and strategists who conceived and implemented Ariane, it should be noted that it has become a dominant GTO launch vehicle. In recent years, Ariane has carried over 50% of all GTO payloads, far more than any other single vehicle system. The last launch of Ariane 4 was conducted in early 2003; this vehicle has subsequently been phased out in favor of the larger Ariane 5.

The Ariane 5 family of vehicles is an essentially new design, and offers a major upgrade from earlier models. To the casual observer, Ariane 5 resembles the shuttle stack without the Orbiter. In the basic Ariane 5, two parallel-staged, 6.4 MN thrust solid rocket motors lift a central LOX/hydrogen tank, which feeds a 1.1 MN thrust Vulcain rocket engine. In this version, a storable-propellant upper stage carries the payload to GTO. Figures 5.21a–e present the launch capability of the Ariane 5, while Fig. 5.22 shows the configuration options.

Developed in response to the ever-increasing size of geostationary satellites, the basic Ariane 5 can deliver payloads of approximately 18,000 kg to LEO and 6800 kg to a 7° inclination geostationary transfer orbit. The original concept was that the vehicle could carry two geostationary communications satellites on a single launch. The actual satellite mass is limited to about 6000 kg in such a case, with the extra mass being accorded to the carrier frame. However the weight growth of such satellites has been so rapid that this capability is already of only marginal utility. Various performance upgrades are in progress as this is written.

The initial Ariane 5 upgrade replaces the core engine with the improved 1.3 MN thrust Vulcain-2, which offers a vacuum I_{sp} of up to 440 seconds, in

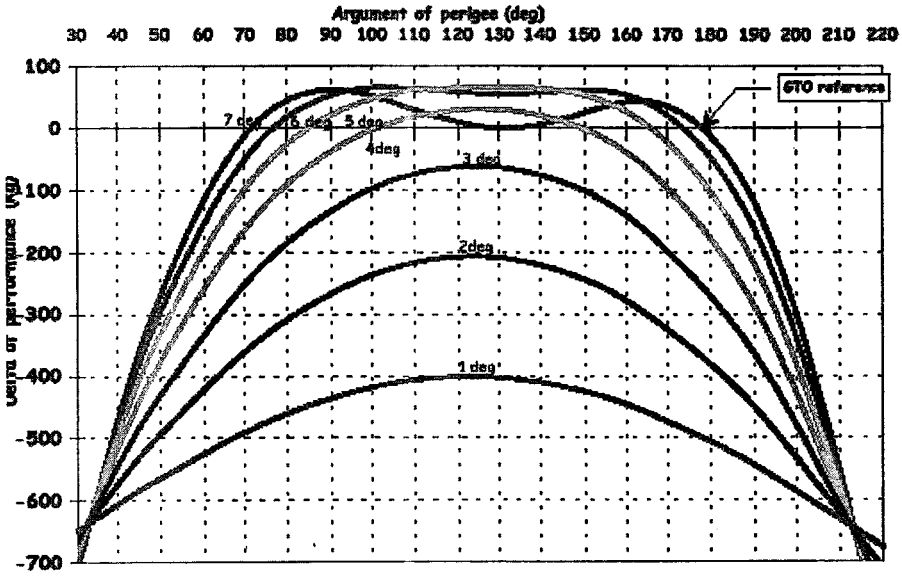


Fig. 5.21a Ariane 5G performance vs GTO inclination and argument of perigee. (GTO reference: 6640 kg to 560 km × 35890 km orbit.) (Courtesy Arianespace.)

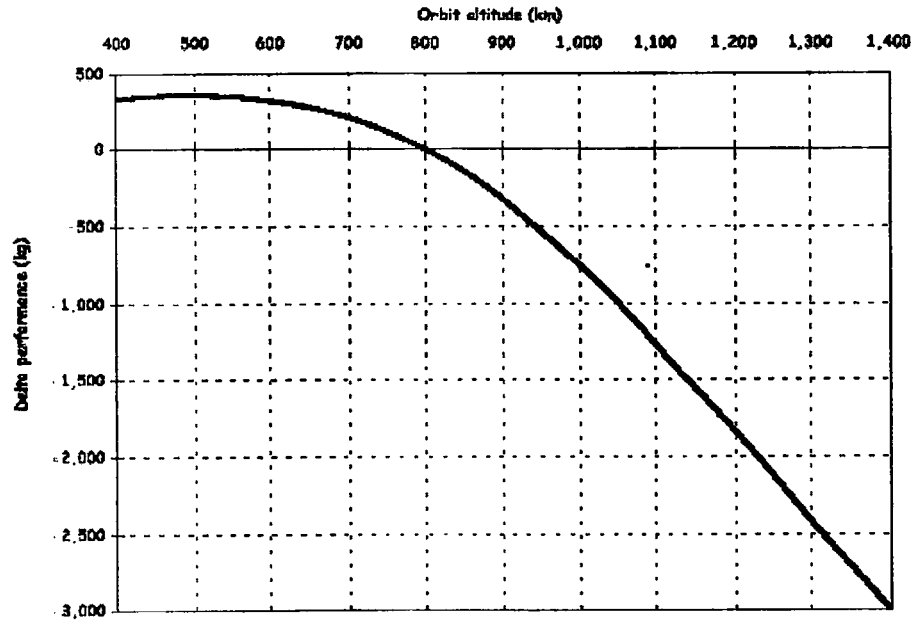


Fig. 5.21b Ariane 5G performance to sun-synchronous orbit. (Reference: 9500 kg to 800 km circuit orbit.) (Courtesy Arianespace.)

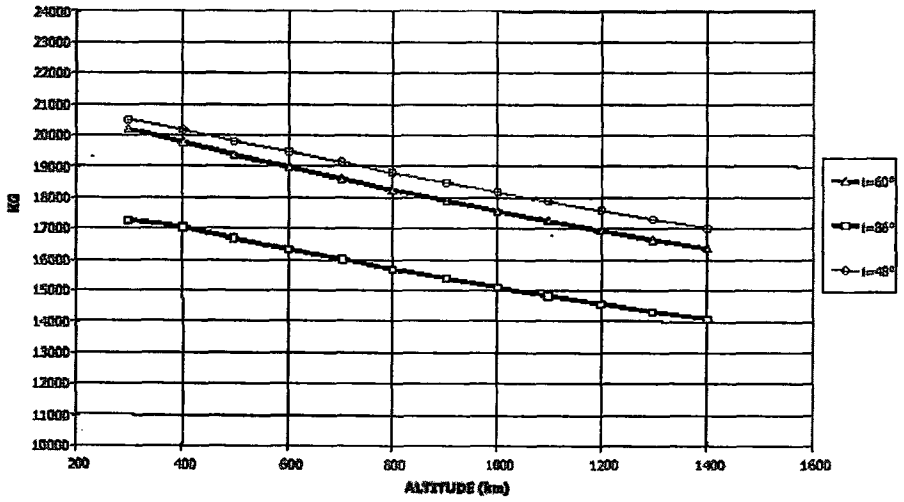


Fig. 5.21c Ariane 5-ES performance to low Earth orbit. (Courtesy Arianespace.)

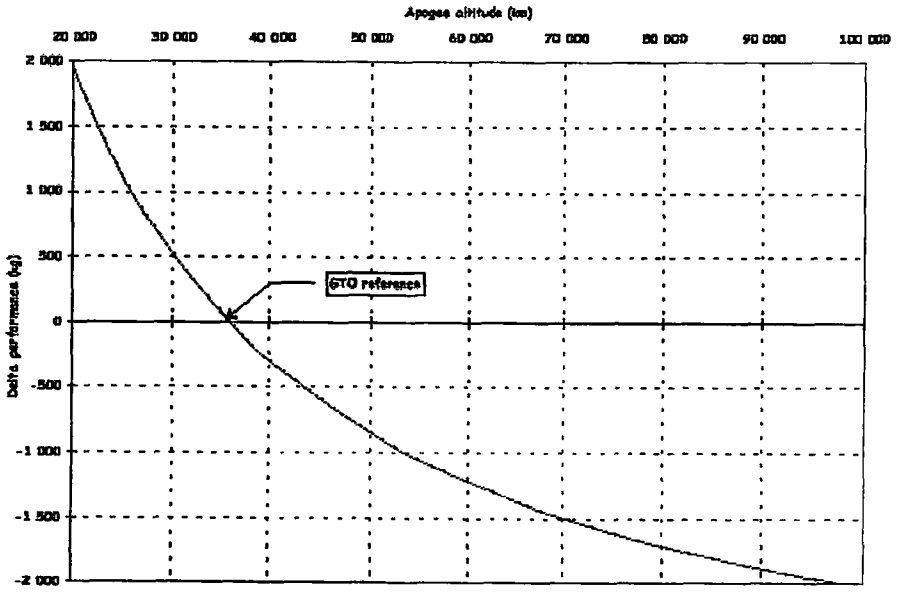


Fig. 5.21d Ariane 5-ECA performance to GTO as a function of transfer orbit apogee altitude. (GTO Reference: 10050 kg to 250 km × 35950 km orbit.) (Courtesy Arianespace.)

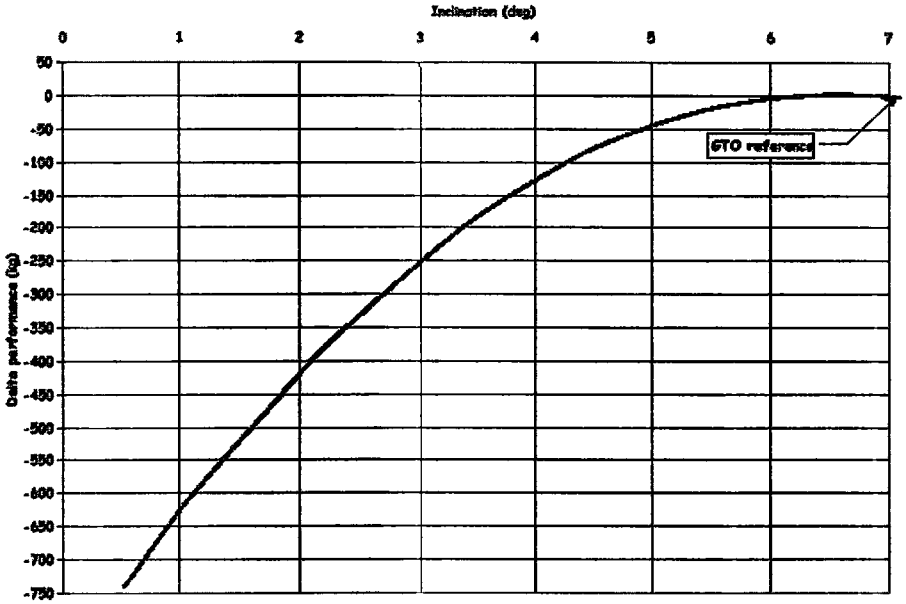


Fig. 5.21e Ariane 5-ECA performance to GTO as a function of transfer orbit inclination. (GTO Reference: 10050 kg to 250 km × 35950 km orbit.) (Courtesy Arianespace.)

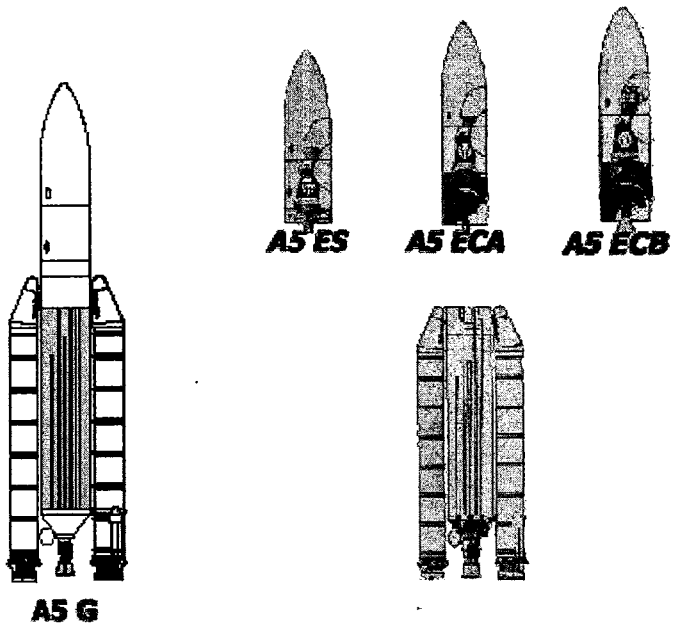


Fig. 5.22 Ariane 5 launch vehicle family. (Courtesy Arianespace.)

comparison to the 431 second vacuum I_{sp} of the basic Vulcain. This results in a GTO payload increase to approximately 7400 kg. Subsequent upgrades involve the replacement of the storable-propellant upper stage with higher-performance cryogenic upper stages. The Ariane 5 ESC-A uses the cryogenic Ariane 4 third stage, resulting in a GTO payload increase to approximately 10,000 kg. A later version, the ESC-B, will be powered by the new 180 kN thrust LOX/hydrogen Vinci engine, and will offer a 12 tonne GTO payload capability. Table 5.4 provides configuration data for the basic Ariane 5 vehicle and Table 5.5 shows the orbital injection accuracy to be expected.

The operational reliability of Ariane 5 has been similar to that of other new launch vehicles. Over the course of its first sixteen flights, involving two different models, two catastrophic failures occurred, and two missions delivered payloads to the wrong orbit. However, it is to be expected that Ariane 5 will ultimately settle into the workhorse role previously demonstrated by its predecessors.

Atlas. The Atlas in various forms and combinations has been a major element of the U.S. space program since the late 1950s. Originally designed as an ICBM, the basic Atlas provided significant payload capability to LEO. This was first demonstrated in 1958 when an entire "bare Atlas" (i.e., no upper stages) was put in orbit as part of Project Score. Ostensibly a communications experiment, the mission probably had more significance as a counter to Soviet propaganda and as a national morale booster. In any case, it was a portent of future developments.

A modified version of the operational Atlas D was used to launch the four manned orbital Mercury missions, beginning with John Glenn's three-orbit flight of 20 February 1962. The Mercury flights employed no upper stages; most other Atlas applications have exploited the efficiency of additional staging to augment the basic vehicle. The most common early Atlas upper stage was the Lockheed Agena, a storable liquid-propellant vehicle designed for use with both the Thor and Atlas boosters and later adapted for use with the Titan 3B. The Atlas-Agena

Table 5.4 Ariane 5 configuration data

	Solid Rocket Booster	Stage 1 (Core)	Stage 2
Length	31.2 m	29 m	4.5 m
Diameter	3.0 m	5.4 m	5.4 m
Mass	230,000 kg	170,000 kg	10,900 kg
Propellant	HTPB	LOX/LH2	N2O4/MMH
Mass	230,000 kg	155,000 kg	9700 kg
Engine		Vulcain	Aestus
Thrust	6360 kN (vac)	1120 kN (vac)	27.5 kN (vac)
Isp	273 sec (vac)	430 sec (vac)	324 sec (vac)
Number	2	1	1
Burn time	123 sec	590 sec	800 sec

launched a considerable variety of payloads, including most early planetary missions, and in modified form served as a docking target and orbital maneuvering stage for four two-man Gemini missions in 1966. However, it is no longer used.

The most capable of the early Atlas derivatives was the Atlas-Centaur. This vehicle used a modified Atlas first stage and the LH_2/LO_2 Centaur as the second stage. This system has evolved into a highly reliable, adaptable vehicle that has launched many scientific and commercial spacecraft.

The original Atlas is unique among launch vehicles in its use of a "balloon" tank structure. The propellant tanks themselves were used as the primary airframe structure, and were constructed from welded, thin-gage stainless steel. Without substantial internal pressure, or tension supplied by external means, the tank structure could not support itself or the payload. Although this unique design feature complicated ground-handling procedures, it allowed a structural coefficient to be achieved that is still unmatched among liquid-propellant vehicles. The later Centaur structure followed the same design approach.

Although a great variety of Atlas configurations has been used, most are no longer in production. The Atlas-Centaur very nearly became a casualty of the early-1980s purge of expendable launch vehicles. However, General Dynamics (now Lockheed Martin) entered the commercial market with derivatives of the Atlas-Centaur. The initial version, designated Atlas I, was a strengthened and slightly upgraded version of the late-model Atlas-Centaur. The primary change was an increase in nose-fairing diameter from 3.05 m (10 ft), the same as the tank diameter, to a choice of a 4.19 m diameter × 12.2 m (13.75 ft × 40.1 ft) or a 3.3 m diameter × 10.36 m (10.8 ft × 34 ft) fairing, depending on payload size.

General Dynamics then introduced the Atlas II, with increased tank length and thrust in the first stage and a propellant mixture ratio change and length increase in Centaur. Also, the vernier engines used for roll control because the earliest versions were replaced with a hydrazine monopropellant system. The Atlas IIA incorporated these changes, plus extendable nozzles on the Centaur engines and upgraded avionics. The Atlas IIAS added two Castor II solid-propellant strap-ons to the Atlas IIA. First launched in 1993, the Atlas IIAS offers a GTO payload of approximately 3700 kg, and as of this writing is the sole member of this vehicle class remaining in production.

The Centaur is a three-axis stabilized upper stage, initially developed in the mid-1960s, and since evolved into a true workhorse of the U.S. space program. Spinning payloads are mounted upon a spin table that is locked in place for launch. After final orbit insertion, the spin table is unlocked and spun up to the desired rate using small rockets. When proper spin rate and vehicle attitude are achieved, the payload is released. Following separation, the Centaur can maneuver to a new attitude and apply a ΔV to allow a safe distance from the payload to be maintained. Typical separation rates are 0.5–1 m/s.

The Centaur has the capability for a large number of restarts and thus can fly complex mission profiles, including multiple payload deployment. Early

Table 5.5 Ariane 5 injection accuracy (1σ)

Semi-major axis	a	40 km
Eccentricity	e	0.00041
Inclination	i	0.02°
Argument of perigee	ω	0.15°
Longitude of ascending node	Ω	0.15°

capability allowed two engine starts, with minor modifications required for additional burns. As many as seven restarts have been demonstrated on a single mission. Injection accuracy for the Centaur is summarized in Table 5.6.

The Atlas IIAS is launched from Cape Canaveral Air Force Station and Vandenberg AFB, California. As always, because of the 28.5° latitude of the KSC launch site, performance suffers for delivery to an equatorial orbit.

The Atlas family relied for decades exclusively on the Rocketdyne MA-5 three-engine cluster (two booster engines fall away partway through the flight while a central sustainer engine fires throughout the launch phase). But beginning in 2000, Lockheed-Martin (which absorbed the General Dynamics launch vehicle program) introduced two new versions of the Atlas which use a two-barrel Russian-made RD-180 rocket engine. This engine, a derivative of a highly successful four-barrel engine used in Russian launchers, is used in the first stage of the Atlas III and Atlas V launch vehicles. The high performance of these engines substantially increased payload capacity. The family of Atlas III configurations is summarized in Table 5.7.

The newest member of the Atlas family is the Atlas V design, which no longer uses the "stainless steel balloon" structural concept. The lower stage of Atlas V is a self-supporting aluminum structure, which reduces some of the operational problems inherent in the pressure-stabilized design, and allows easier use of strap-on solid rocket motors. (It is worth noting that "self-supporting" refers to ground loads. Most, if not all, liquid propellant launch vehicles require pressure in the tanks to provide resistance to compression loads and required structural rigidity during flight.) The Atlas V was developed as part of the USAF Evolved

Table 5.6 Atlas/Centaur injection accuracy (3σ)

LEO insertion (1111 km × 63.4°)		
Semi-major axis	a	± 19.4 km
Inclination	i	± 0.15°
Cutoff velocity	V	80 m/s
GTO insertion (167 km × 35941 km × 27°)		
Perigee	r_p	± 2.4 km
Apogee	r_a	± 117 km
Inclination	i	± 0.02°

Table 5.7 Atlas III configuration data

	Stage 1	Stage 2 Atlas IIIA (Centaur)	Stage 2 Atlas IIIB (Stretched Centaur)
Length	28.5 m	10.2 m	11.7 m
Diameter	3.0 m	3.0 m	3.0 m
Gross mass	195,630 kg	18,960 kg	22,960 kg
Propellant	LOX/RP	LOX/LH2	LOX/LH2
Propellant mass	181,903 kg	16,780 kg	20,830 kg
Engine	RD-180	RL-10A-4-2	RL-10A-4-2
Thrust	3820 kN (sl)	99,000 N	99,200 N
I_{sp}	311 sec (sl)	451 sec (vac)	451 sec (vac)
Number	1	1	2
Nominal burn Time	180 sec	770 sec	455 sec

Expendable Launch Vehicle (EELV) program, and flew successfully for the first time in 2002. Configuration data for Atlas V are summarized in Table 5.8.

Performance of some of the various Atlas configurations is presented in Figs. 5.23a–f. Figures 5.24a and b depict the current configurations.

Delta. The Delta launch vehicle system, developed by McDonnell Douglas Astronautics, began as a derivative of the Thor IRBM. As was the case with its contemporary, the Atlas, a variety of upper stage systems have been used with the Thor, including Agena, Able, Able-Star, and various solid motors. The Thor-Able was a mating of the Vanguard second and third stages (a storable liquid and a spinning solid, respectively) to the Thor in order to achieve better performance

Table 5.8 Atlas V configuration data

	Solid Motors	Stage 1 (Common Core)	Stage 2 (Centaur)	Stage 2 (Stretched Centaur)
Length	17.7 m	32.4 m	10.2 m	11.7 m
Diameter	1.55 m	3.8 m	3.0 m	3.0 m
Mass	40,824 kg	310,045 kg	18,960 kg	23,220 kg
Propellant	HTPB	LOX/RP	LOX/LH2	LOX/LH2
Mass	38,770 kg	284,350 kg	16,780 kg	20,410 kg
Engine	Atlas 5 SRB	RD-180	RL-10A-4-1	RL-10A-4-2
Thrust	1134 kN (sl)	3820 kN (sl)	99,000 N (vac)	99,000 N (vac)
I_{sp}	275 sec (sl)	311 sec (sl)	451 sec (vac)	451 sec (vac)
Number	0–5	1	1	2
Burn Time	94 sec	236 sec	894 sec	429 sec

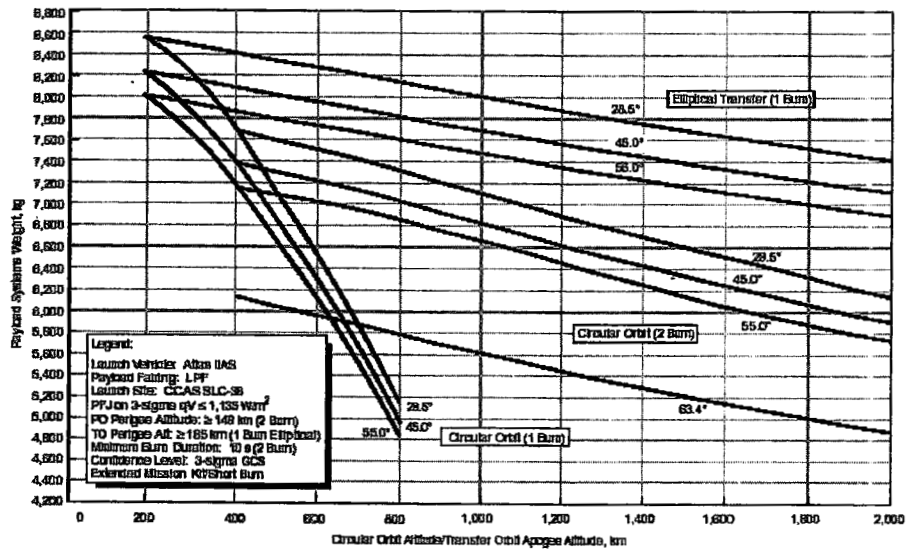


Fig. 5.23a Atlas IIAS performance to low Earth orbit from Cape Canaveral. (Courtesy Lockheed Martin Corporation and International Launch Services.)

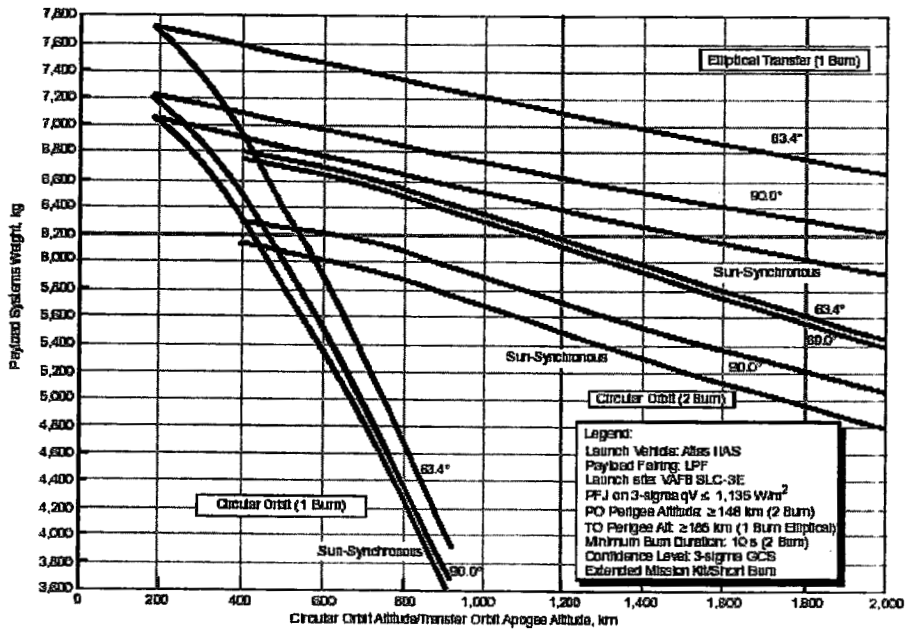


Fig. 5.23b Atlas IIAS performance to low Earth orbit from Vandenberg AFB. (Courtesy Lockheed Martin Corporation and International Launch Services.)

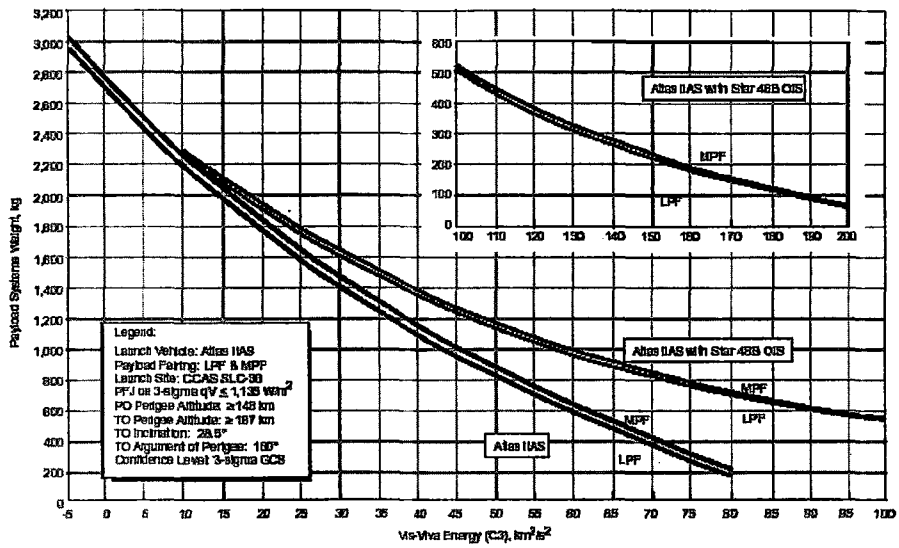


Fig. 5.23c Atlas IIAS Earth escape performance from Cape Canaveral. (Courtesy Lockheed Martin Corporation and International Launch Services.)

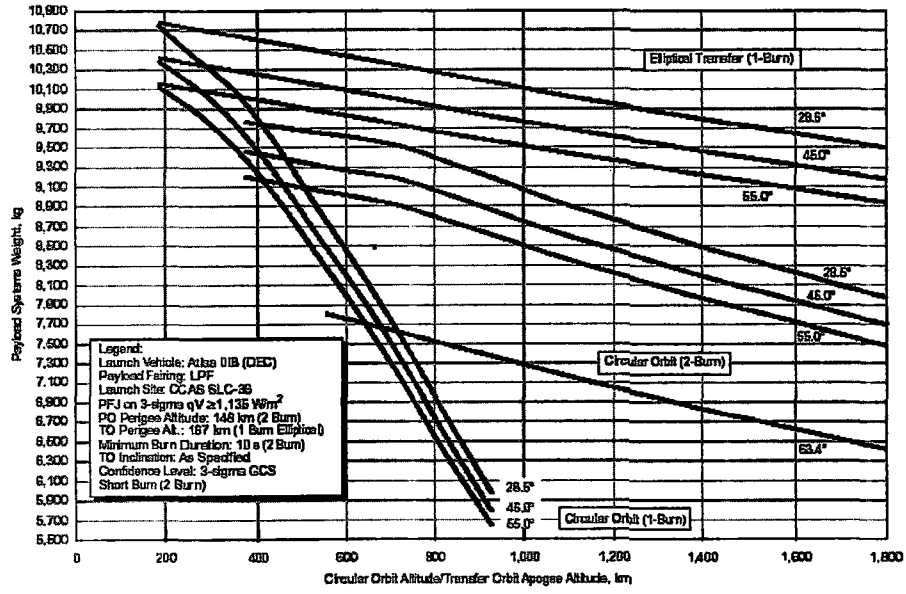


Fig. 5.23d Atlas IIB DEC performance to low Earth orbit from Cape Canaveral. (Courtesy Lockheed Martin Corporation and International Launch Services.)

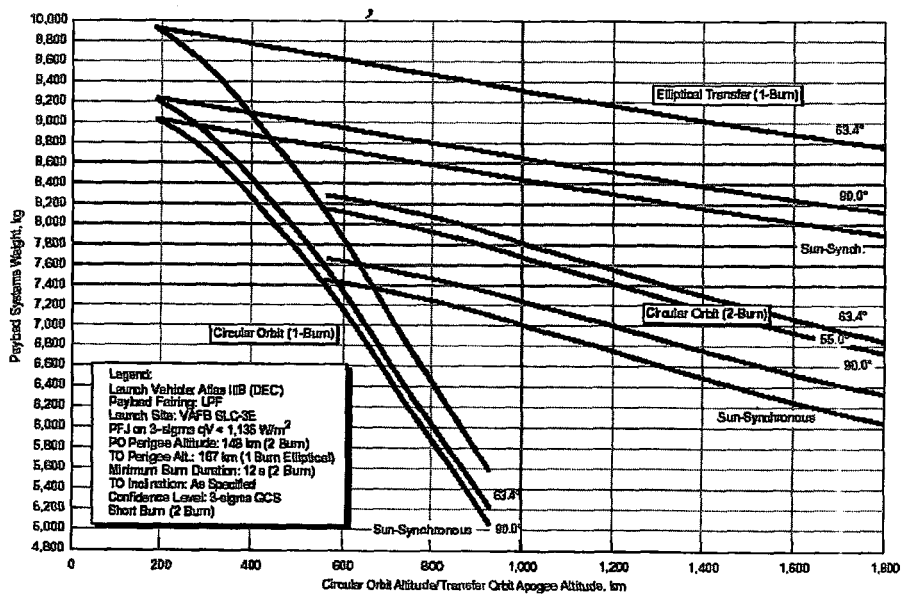


Fig. 5.23e Atlas IIIB DEC performance to low Earth orbit from Vandenberg AFB. (Courtesy Lockheed Martin Corporation and International Launch Services.)

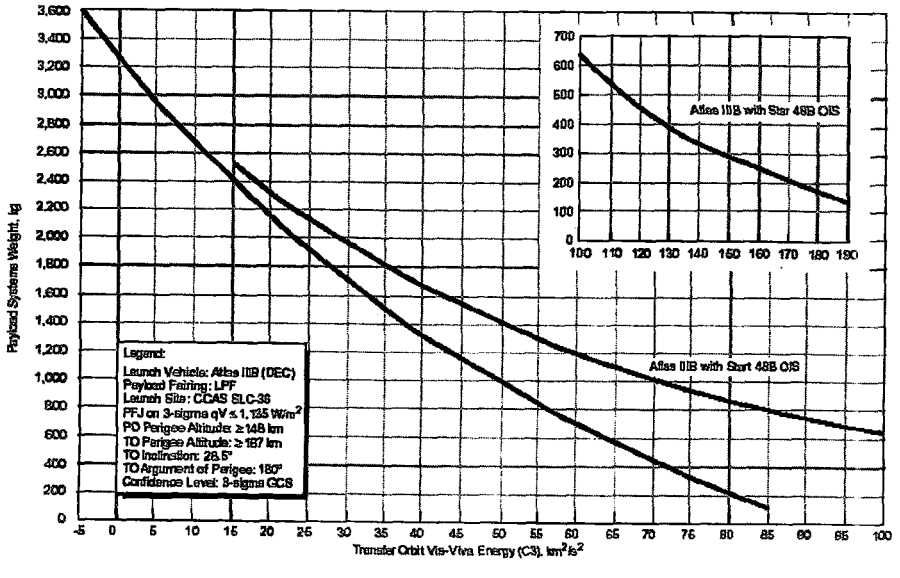


Fig. 5.23f Atlas IIIB DEC Earth escape performance from Cape Canaveral. (Courtesy Lockheed Martin Corporation and International Launch Services.)

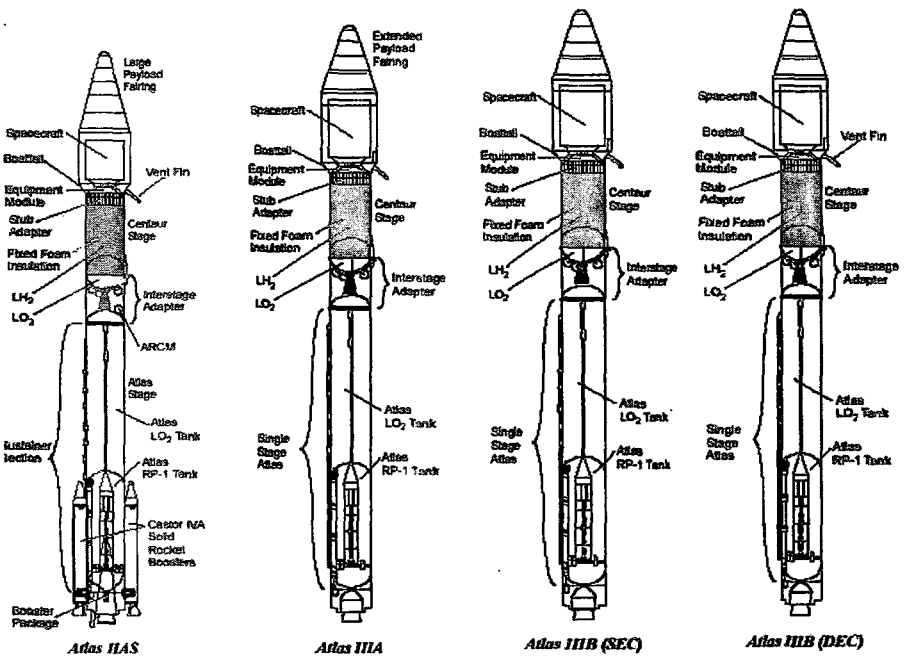
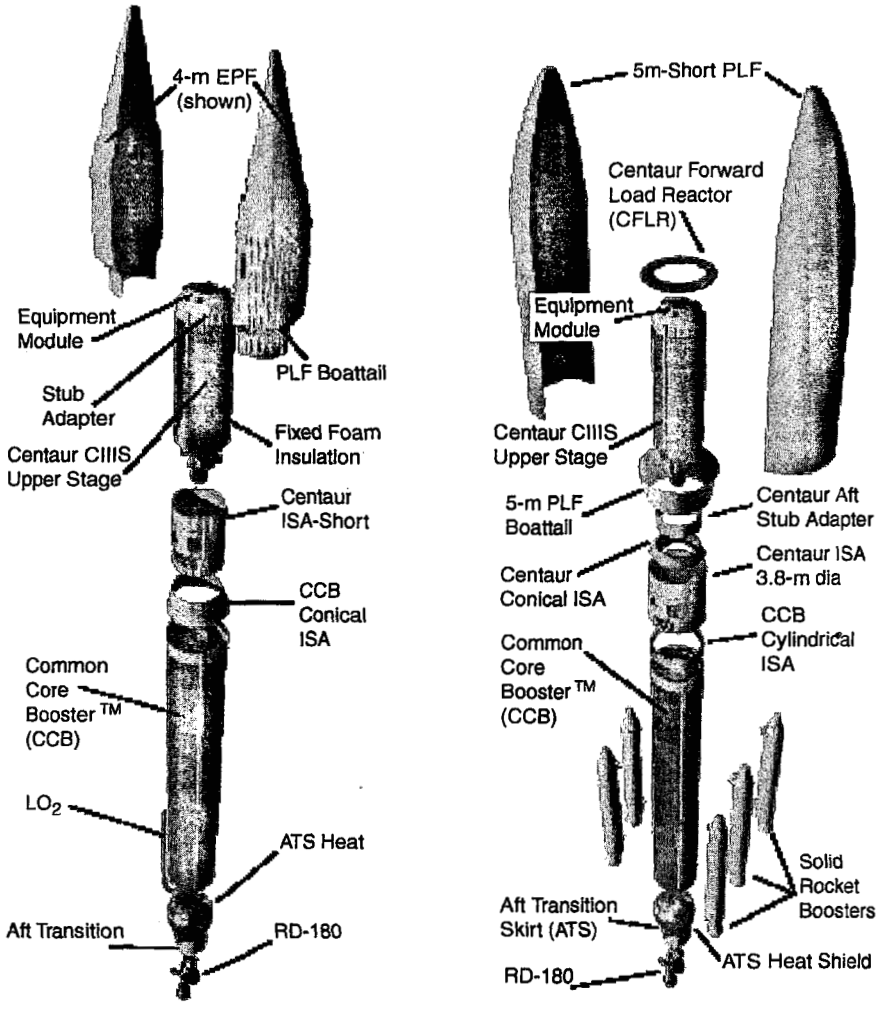


Fig. 5.24a Atlas IIAS, IIIA, and IIIB configurations. (Courtesy Lockheed Martin Corporation and International Launch Services.)

than offered by the small Vanguard first stage. The Thor-Able could deliver a few hundred pounds into low Earth orbit and a few tens of pounds into an escape trajectory. This vehicle, developed in the late 1950s, rapidly gave way to the Thor-Delta, a vehicle of similar appearance but improved performance and reliability.

Over the ensuing two decades the vehicle evolved considerably, with the first stage gaining length and assuming a cylindrical rather than a combined conical/cylindrical form. The second stage grew to equal the eight-foot diameter of the first, and has evolved through several engines. The first-stage engine remained much the same in its design, while the core engine was uprated from the original 150,000 lbf to over 200,000 lbf. Relatively early in the evolution of the vehicle, the concept of increasing liftoff thrust with strap-on solid-propellant motors was developed. The maximum number of such motors has now grown to nine fired in a 6/3 sequence to avoid excessive peak loads, and the motors have increased substantially in size and thrust. Figure 5.25 shows the evolution of the Delta launch vehicle family, and Fig. 5.26 displays profiles of the most common current configurations.

A solid-propellant third stage is an option depending on the mission and payload. If the third stage is used, it is mounted on a spin table. Prior to separation from the three-axis stabilized second stage, small rockets bring the spin rate up to



Atlas V 400 *Atlas V 500*

Fig. 5.24b Atlas V configurations. (Courtesy Lockheed Martin Corporation and International Launch Services.)

the desired level. A typical value would be 50–60 rpm. The third stage is normally used to obtain geosynchronous transfer or interplanetary injection velocities.

It is clear that a substantial family of Delta launcher variants exists. In order to simplify identification, a four-digit code is used to identify a particular model (e.g., Delta 3914 or Delta 3920). Table 5.9 explains the code. Only the 6000 and 7000 or Delta II series vehicles are available as this is written, but this still provides a substantial range of launch options. Configuration summary data are provided in Table 5.10.

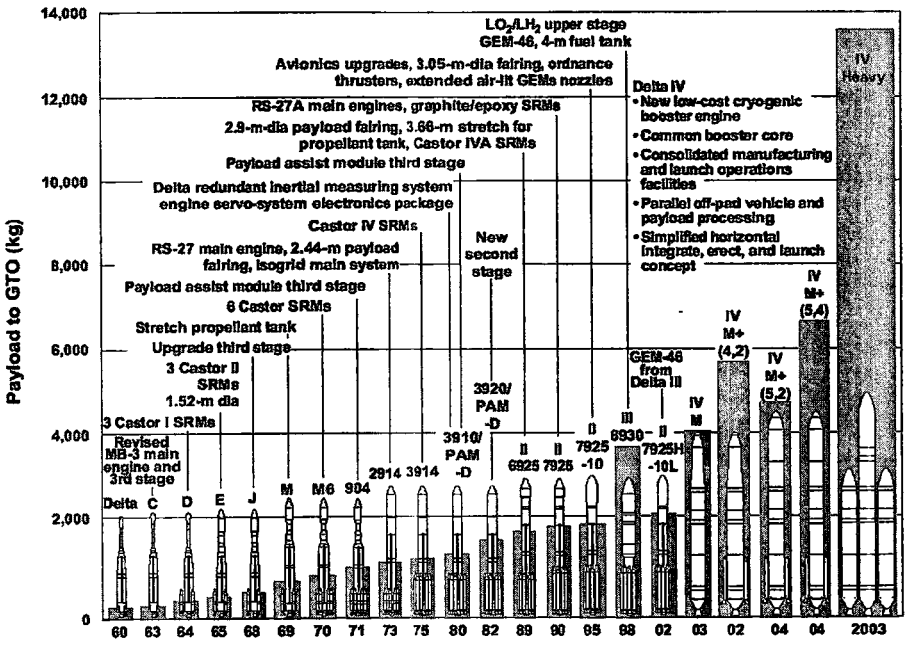


Fig. 5.25 Delta launch vehicle evolution. (Courtesy Boeing Company/Delta Launch Services.)

Delta II launches are available from the Eastern Space and Missile Center (Cape Canaveral) and the Western Space and Missile Center (Vandenberg Air Force Base). Figures 5.27–5.30 show the performance capabilities of the Delta II 7920 and 7925 configurations from both sites.

Accuracy of Delta II low-orbit injection is quite good, with typical $\pm 3\sigma$ accuracy margins of 18.5 km in altitude and 0.05° in inclination. Addition of the spin-stabilized but unguided third stage naturally degrades these values. The 1–2.5° pointing errors typical of these stages result in $\pm 3\sigma$ inclination errors of 0.2°.

Recently, McDonnell Douglas (now Boeing) developed a higher performance commercial derivative of the Delta family called Delta III. An upgraded Rocketdyne RS-27A remains as the core engine. The lower portion of the core remained at an eight-foot diameter, but the upper portion was expanded in order to increase propellant capacity within the same length. Nine GEM solid propellant rocket motors surround the first stage. All the motors have gimbaled nozzles. The second stage features a high performance, extendable nozzle version of the Pratt & Whitney RL-10 engine burning liquid hydrogen and liquid oxygen. The configuration is shown in Fig. 5.26.

The first two launches of Delta III were failures and, as this is written, the vehicle is being phased out.

An entirely new vehicle, designated Delta IV, has been developed under the auspices of the USAF Evolved Expendable Launch Vehicle (EELV) program.

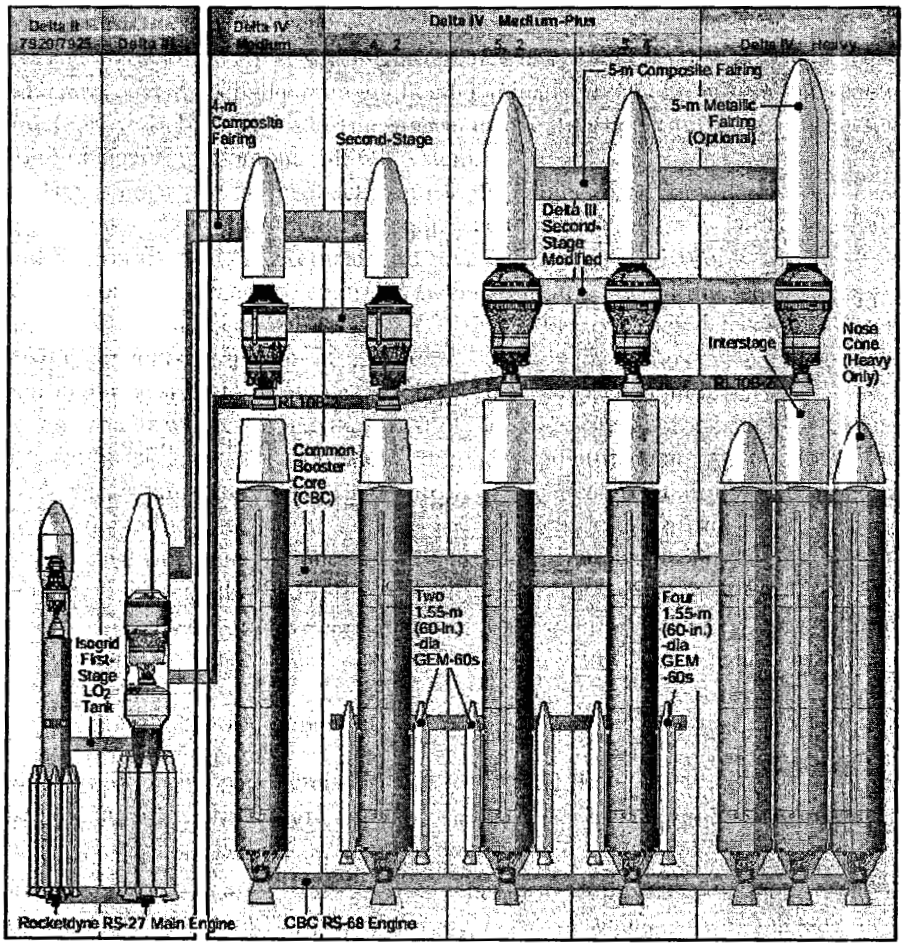


Fig. 5.26 Delta launch vehicle family. (Courtesy Boeing Company/Delta Launch Services.)

This vehicle uses liquid oxygen and liquid hydrogen in the lower stages, which are powered with the newly-developed Rocketdyne RS-68 engine.

The Delta IV has a variety of configurations as shown in Fig. 5.26. The Common Booster Core (CBC) is the basic building block, used alone in the smallest version and with a variety of solid propellant strap-on boosters to form the Delta IV M(edium) class. The largest variant uses a cluster of three CBCs to form the Delta IV H(eavy) class.

Upper stages for Delta IV use liquid hydrogen and oxygen as propellants, with a single RL-10 engine. The H-model uses a heavier upper stage than does the M-model. Performance data for representative Delta IV configurations are provided in Figs. 5.31–5.34. Table 5.11 presents the configuration data for Delta IV, while injection accuracy data are given in Table 5.12.

Table 5.9 , Delta II identification

Delta *WXYZ-Q*
W: First-stage identification
0: Long tank Thor, Rocketdyne MB-3 engine. (The "0" is usually omitted, in which case the configuration code is 3 digits.)
1: Extended long tank Thor, Rocketdyne MB-3 engine, Castor II solids.
2: Extended long tank Thor, Rocketdyne RS-27 engine, Castor II solids.
3: Extended long tank Thor, Rocketdyne RS-3 engine, Castor IV solids.
4: Extended long tank Thor, Rocketdyne MB-3 engine, Castor IVA solids.
5: Extended long tank Thor, Rocketdyne RS-27 engine, Castor IVA solids.
6: Extra extended long tank Thor, Rocketdyne RS-27 engine, Castor IVA solids.
7: Extra extended long tank Thor, Rocketdyne RS-27A engine, GEM - 40 solids.
8: Delta III shortened first stage, Rocketdyne RS-27A engine, GEM - 46 solids.
X: Number of solids
3–9: Number of first-stage strap-on solid motors, as in first-stage identification.
Y: Second stage
0: Aerojet AJ10-118F engine.
1: TRW TR-201 engine.
2: Aerojet AJ10-118K engine.
3. Pratt & Whitney RL10B-2 engine.
Z: Third stage
0: No third stage.
3: Thiokol TE364-3 engine.
4: Thiokol TE364-4 engine.
5: PAM-D derivative STAR-48B.
6: STAR 37-FM.
Q: Fairing Type
None: Standard Fairing (9.5 ft. for Delta II).
– 8: 8 ft. fairing.
– 10: 10 ft. composite fairing.
– 10L: 10 ft. stretched composite fairing.

Titan. The Titan family of launch vehicles was derived from the Titan ICBM family, which is no longer a part of the U.S. strategic arsenal. The Titan 1 was a two-stage ICBM using cryogenic liquid propellants; it first flew in 1959. This vehicle saw little use and was rapidly replaced by the Titan 2, a substantially different system using storable hypergolic propellants (N_2O_4 oxidizer and 50/50 N_2H_4-UDMH fuel) in both stages. First flown in 1962, this launcher was subsequently man-rated for use in the Gemini program in the mid-1960s. Twelve missions, including 10 manned flights in 20 months, were conducted between 1964 and 1966.

The Titan 3 family was developed in an effort to provide a flexible, high-capability launch system using existing technology in a "building block" approach. The Titan 3A added two 120-in diameter solid rocket boosters as a "zeroth stage" configuration to an uprated Titan 2 (also 120 in diameter). The

Table 5.10 Delta II configuration data

	Solid Strap-Ons	Stage 1	Stage 2	Stage 3 (optional)
Length	13.0 m	26.1 m	6.0 m	2.0 m
Diameter	1.0 m	2.4 m	2.4 m	1.25 m
Mass	13,080 kg	101,900 kg	6953 kg	2141 kg
Propellant	Solid	LOX/RP	N2O4/A-50	Solid
Mass	11,766 kg	95,800 kg	6004 kg	2,009 kg
Engine	GEM	RS-27A	AJ-10-118K	Star-48B
Thrust	499.1 kN (vac)	889.6 kN (vac)	43.6 kN (vac)	66.4 kN (vac)
I_{sp}	273.8 sec (vac)	301.7 sec (vac)	319.2 sec (vac)	292.2 sec (vac)
Number	3–9	1	1	1
Burn time	63 sec	265 sec	440 sec	55 sec

solids may be ignited in parallel with the basic Titan first stage, or fired first with first-stage ignition occurring shortly before burnout of the solids. The choice depends upon specific trajectory requirements. A restartable upper stage, the so-called Transtage, was added to provide capability for complex orbital missions.

Subsequent versions flew without solid motors (Titan 3B), with five-segment solids (Titan 3C), without a Transtage (Titan 3D), and with a Centaur upper stage (Titan 3E), and, in its final version (Titan 34D) with the IUS. At this point the

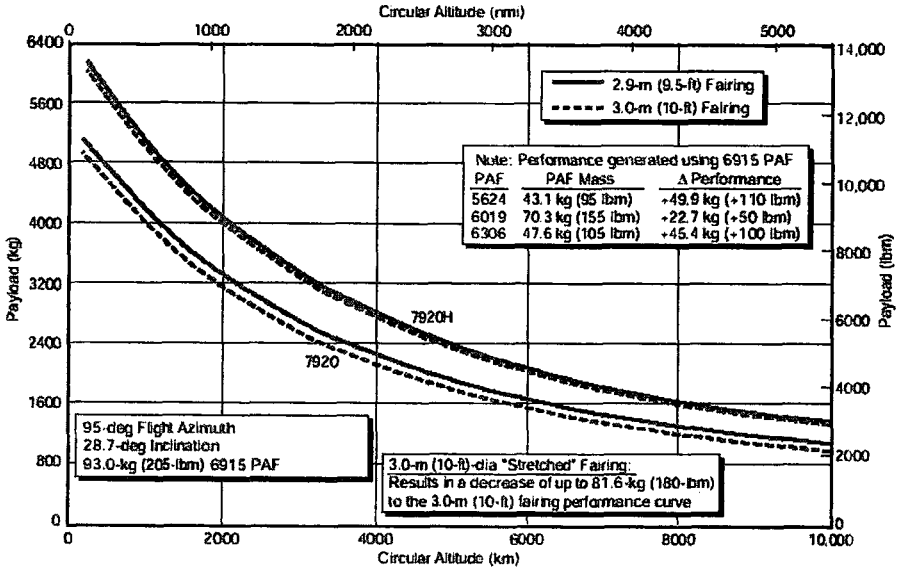


Fig. 5.27 Delta II 7920/7920H performance to low Earth orbit from Cape Canaveral. (Courtesy Boeing Company/Delta Launch Services.)

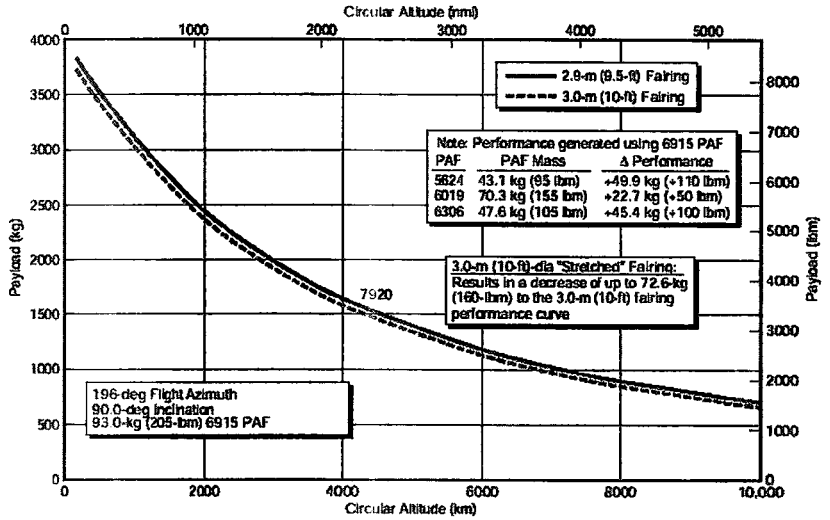


Fig. 5.28 Delta II 7920 performance to low Earth orbit from Vandenberg AFB. (Courtesy Boeing Company/Delta Launch Services.)

design was upgraded to the Titan 4, which used seven-segment solid boosters, had an uprated Centaur upper stage, and could boost shuttle-class payloads to geostationary orbit. With the advent of the EELV vehicles, the Titan IV was retired in 2003.

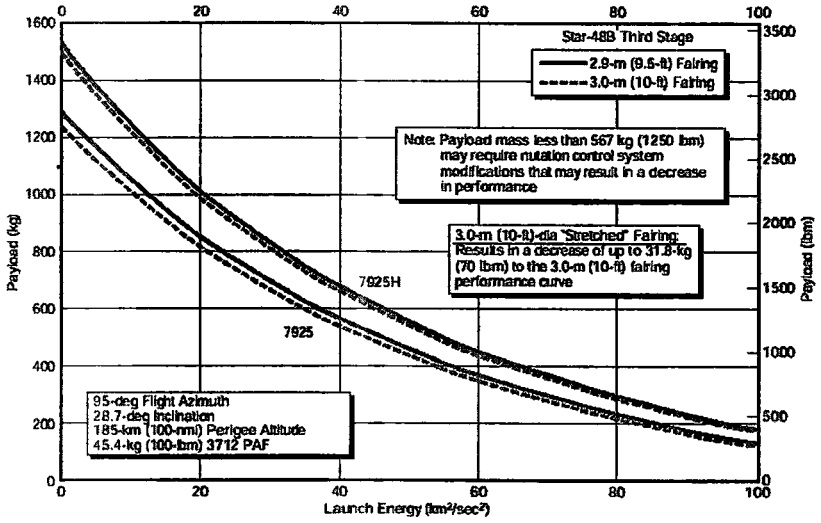


Fig. 5.29 Delta II 792X/792XH Earth escape performance from Cape Canaveral. (Courtesy Boeing Company/Delta Launch Services.)

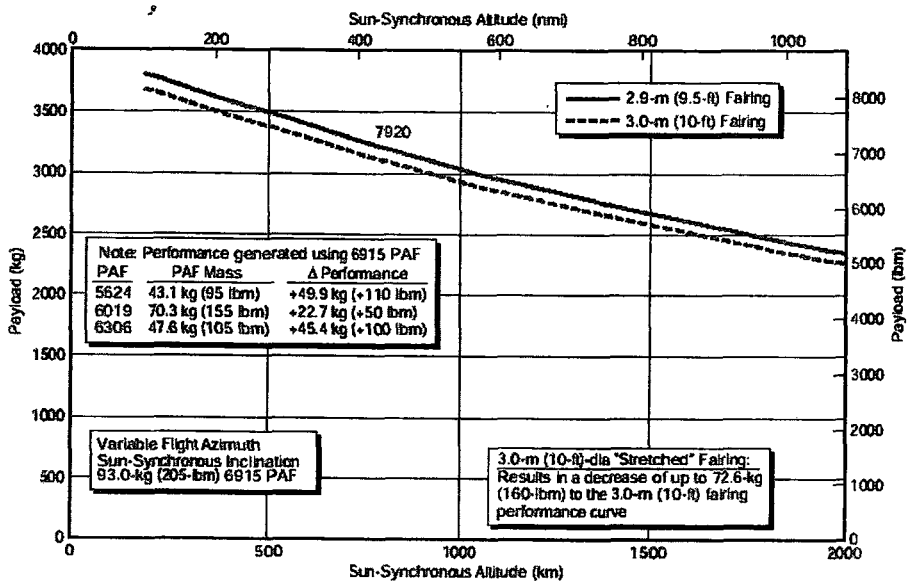


Fig. 5.30 Delta II 7920 performance to sun-synchronous orbit from Vandenberg AFB. (Courtesy Boeing Company/Delta Launch Services.)

Soyuz launcher. Like its American counterparts, the original Soviet ICBM (Semyorka, or "Number 7") and its derivatives has a long space-launch history. Early in its test program, this vehicle launched the Soviet's (and the world's) first artificial satellites and, in slightly upgraded form launched the first human into space. The basic concept still soldiers on with various upgrades and a variety of

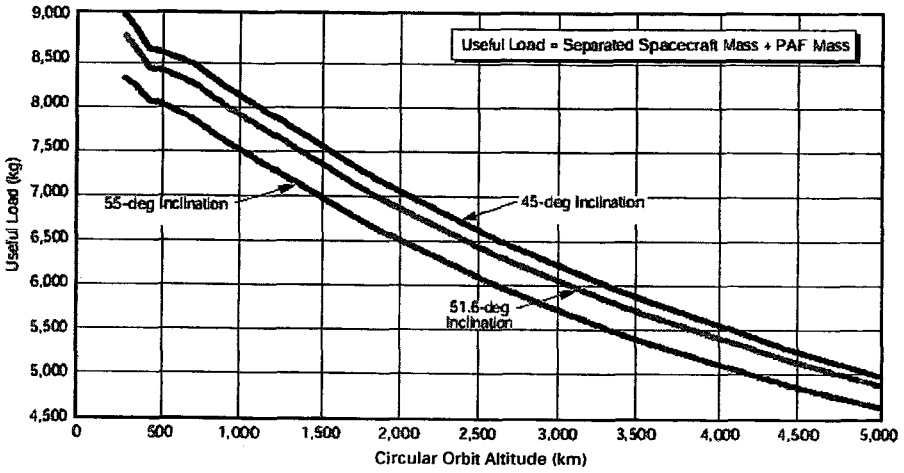


Fig. 5.31 Delta IV-M performance to low Earth orbit from Cape Canaveral. (Courtesy Boeing Company/Delta Launch Services.)

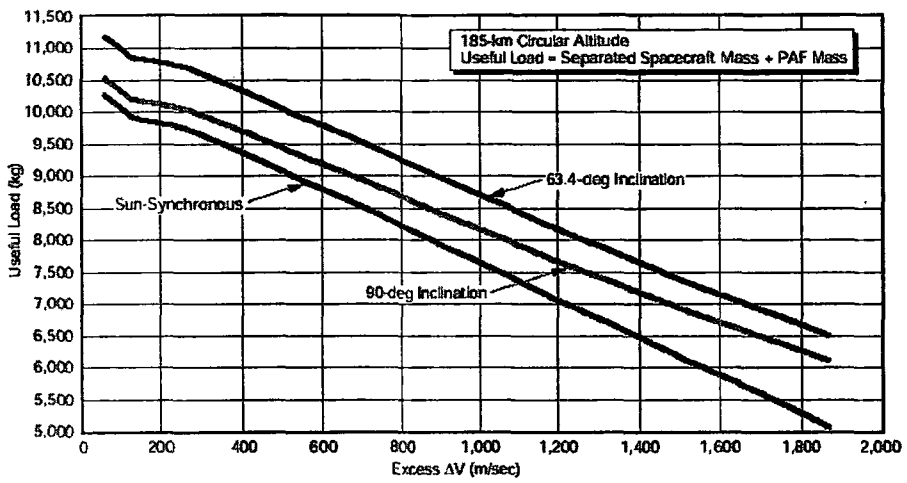


Fig. 5.32 Delta IV-M capability to Molniya polar and sun-synchronous orbits from Vandenberg AFB. (Courtesy Boeing Company/Delta Launch Services.)

upper stages. The most common variant is the version used to launch the Soyuz and Progress spacecraft which previously serviced the various Soviet/Russian space stations and which now supply the International Space Station (ISS). The Molniya version also takes its name from the spacecraft that was its first payload. As this is written, a new version called Aurora is in development, and arrangements have been made to launch the vehicle commercially from the European Space Agency facility in Kourou.

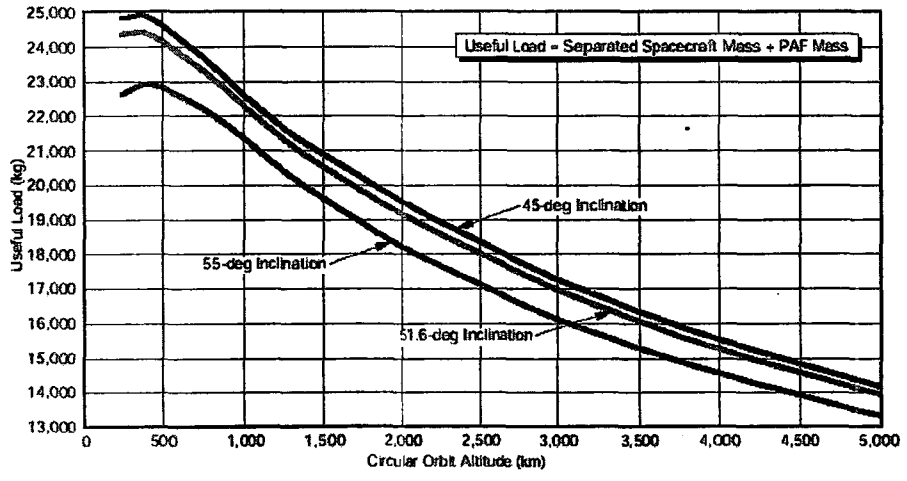


Fig. 5.33 Delta IV-H performance to low Earth orbit from Cape Canaveral. (Courtesy Boeing Company/Delta Launch Services.)

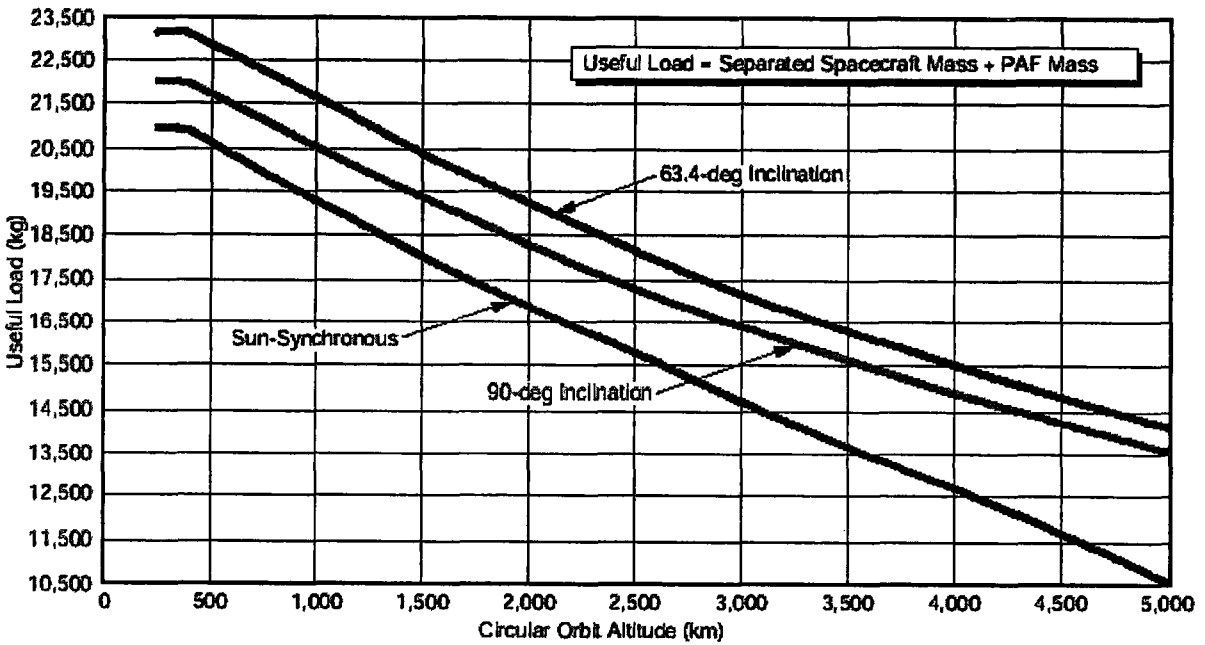


Fig. 5.34 Delta IV-H performance to Molniya, polar, and sun-synchronous orbits from Vandenberg AFB. (Courtesy Boeing Company/ Delta Launch Services.)

Table 5.11 Delta IV configuration data

	Solid Strap-Ons	Stage 1 (Common Core)	Stage 2 (4 m Fairing)	Stage 2 (5 m Fairing)
Length	15.2 m	36.6 m	12.2 m	13.7 m
Diameter	1.52 m	5.13 m	4.0 m	5.1 m
Mass	(Non-TVC/ TVC) 19,082/ 19,327 kg	218,030 kg	23,130 kg	30,840 kg
Propellant	HTPB	LOX/LH2	LOX/LH2	LOX/LH2
Mass	17,045 kg	200,000 kg	24,410 kg	27,200 kg
Engine	GEM-60 (ground/air start)	RS-68	RL-10B-2	RL-10B-2
Thrust	606.1/626.5 kN	2886.0 kN (sl)	110.1 kN (vac)	110.1 kN (vac)
I_{sp}	273.8 sec	365 sec (sl)	462.4 sec (vac)	462.4 sec (vac)
Number	0–4	1	1	1
Burn time	78 sec	249 sec	850 sec	1125 sec

Zenit. The first stage of the Zenit launcher is derived from the liquid propellant strap-on boosters of the short-lived Soviet heavy-lift launcher Energia. It is launched from the Plesetsk Cosmodrome, and suffers a substantial payload penalty on missions to geostationary orbit. To overcome this problem, companies in Russia, Sweden, and the U.S. have banded together to use a modified Zenit in the innovative Sea Launch concept, in which a modified Zenit is launched from a floating platform positioned near the equator.

Long March. The People's Republic of China has produced a family of launch vehicles called Long March. The original derivation was from strategic missile technology, and uses the operationally simple but environmentally undesirable nitrogen tetroxide-hydrazine blend propellant combination. The Long March has achieved some success in the commercial launch market.

Table 5.12 Delta IV injection accuracy (3σ)

LEO insertion (500 km × 90°)		
Circular orbit altitude	h	± 7.4 km
Inclination	i	± 0.04°
GTO insertion (185 km × 35786 km × 27°)		
Perigee	r_p	± 5.6 km
Apogee	r_a	± 93 km
Inclination	i	± 0.2°

Pegasus. The small payload market (e.g., less than 400 kg to LEO) in the U.S. is dominated by Orbital Science Corporation's Pegasus vehicle. This vehicle consists of a winged solid propellant first stage, solid propellant second and third stages, and an optional monopropellant orbit adjustment stage utilized on some flights. The Pegasus is launched from beneath a modified L-1011 carrier aircraft. The wing is primarily used to assist in the pull-up from horizontal flight to the proper climb angle, and is only lightly loaded after that. Performance data for the current model, Pegasus XL, is shown in Fig. 5.35.

After the usual record of mixed results following its introduction in 1990, Pegasus has accumulated an enviable flight history. Through mid-2003, 31 of 34 missions (involving two different models) have been successful, with no failures since late 1996.

Taurus. The Taurus launch vehicle is basically a ground-launched version of Pegasus, wherein a "Stage 0" — the base stage — is substituted for the L-1011 carrier aircraft. This provides a payload upgrade to the 1400 kg range. Two versions of the base stage have been used, a Peacekeeper ICBM first stage and a Castor 120 solid rocket motor. Through mid-2003, six Taurus launches have been conducted, with one failure.

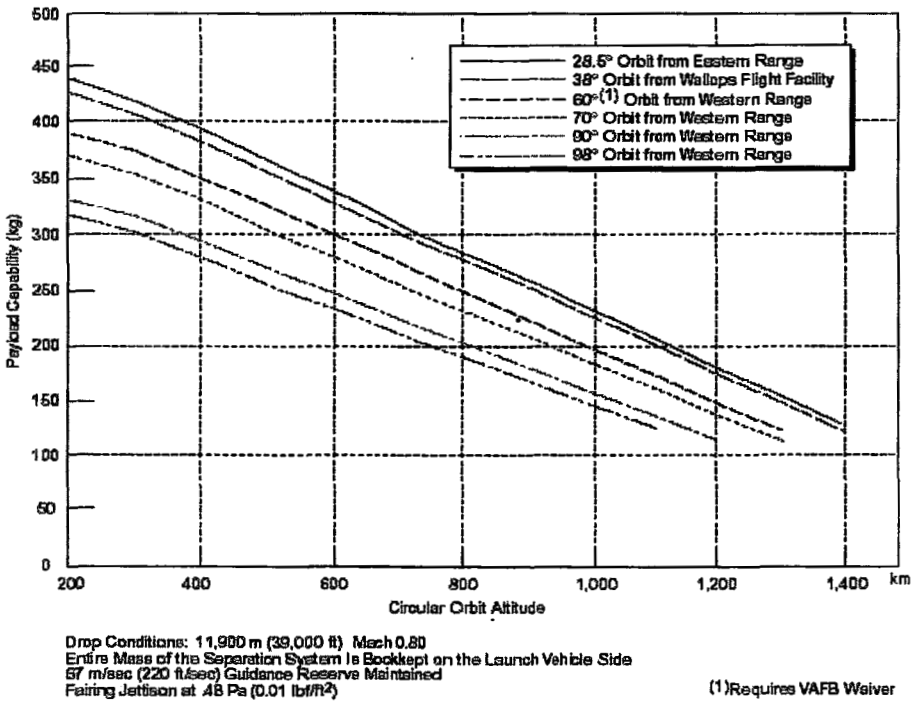


Fig. 5.35 Pegasus XL performance. (Courtesy Orbital Sciences Corporation.)

Other launch vehicles. A significant variety of other launch vehicles, both U.S. and foreign, exist. However, our purpose is to provide some acquaintance with the major options, rather than an exhaustive discussion of all alternatives. Thus, many of these are beyond the intended scope of this text. Readers desiring information on other vehicles or more detail on those discussed herein are referred to the excellent compendium by Isakowitz et al.[24]

5.3.2.3 Upper stages. An extensive variety of upper-stage vehicles has evolved over the last four decades. Some of this development was driven by the fact that the space shuttle is strictly a low-orbit vehicle and requires an upper stage of some type for delivery of payloads beyond LEO. Many of the upper stages derived to meet this need have been adapted for use on other launchers as well. In some cases, e.g., the Centaur, the stages are derivatives of previously existing stage designs.

Centaur. The Centaur has been discussed earlier in connection with its current role as an integral part of the Atlas family of vehicles. In fact, the Centaur is, in one form or another, the oldest operational upper stage in the U.S. fleet, and retains an enviable record of operational reliability and high performance. As indicated earlier, there are presently two configurations, the smaller of which comprises the upper stage of the Atlas IIIA, with the larger version utilized on the Atlas IIIB and Atlas V. Vehicle properties are given in Table 5.7; both versions may be flown with considerable propellant offload to allow optimization for a given mission.

The Centaur was also developed as a high-performance upper stage for the Titan and space shuttle vehicles (Centaur *G* and *G'*). However, primarily because of safety concerns arising in the wake of the *Challenger* accident, the shuttle/Centaur program was canceled in 1986. However, the Centaur *G* and *G'* stages were used as upper stages for the Titan 4 vehicle.

IUS. As noted earlier, the IUS was originally conceived as a low-cost interim upper stage for the shuttle pending development of a high-performance Space Tug. A variety of IUS concepts, based mostly on existing liquid-propellant stages, was examined. The final selection, however, was a concept employing combinations of two basic solid-propellant motors, the 6000-lb motor and the 20,000-lb motor, producing approximately 25,000 and 62,000 lb of thrust, respectively. The basic two-stage vehicle would consist of a large motor first stage and small motor second stage, a combination optimized for delivery of payloads to geostationary orbit. Heavier payloads into other orbits would be handled by a twin-stage vehicle whose two stages both consisted of large motors. Both these combinations had substantial planetary capability for lower-energy missions such as those to the moon, Mars, or Venus. High-energy planetary missions were to be handled by a three-stage vehicle consisting of two large and one small motors.

Complexity and rising cost forced cancellation of the two larger versions, leaving only the basic two-stage vehicle.

Payload Assist Modules (PAM-A and PAM-D). The Centaur and IUS are excessively large for small- and medium-size payloads of the class historically launched by the Atlas-Centaur or Delta vehicles. To meet the requirements for shuttle deployment of these vehicles, McDonnell Douglas developed the PAM. Two vehicles, the -A and -D models, were developed, the former denoting Atlas-Centaur class performance and the latter Delta class capability. Both were spin-stabilized vehicles using solid rocket motors. Avionics were provided for vehicle control for the period from shuttle ejection (via springs) through postburn separation. The PAM-D stages saw considerable use during the early 1980s, launching commercial satellites from shuttle. With the cessation of commercial launches from shuttle following the *Challenger* accident, the use of these stages has largely ceased. However, the PAM-D stage can be used as the third stage of the Delta expendable launch vehicle.

References

[1]Sutton, G. P., and Biblarz, O., *Rocket Propulsion Elements*, 7th ed., Wiley-Interscience, New York, 2000.

[2]Hill, P. G., and Peterson C. R., *Mechanics and Thermodynamics of Propulsion*, Addison-Wesley, Reading, MA, 1965.

[3]Vinh, N. X., Busemann, A., and Culp, R. D., *Hypersonic and Planetary Entry Flight Mechanics*, Univ. of Michigan Press, Ann Arbor, MI, 1980.

[4]Horner, S. F., *Fluid Dynamic Lift*, Horner Fluid Dynamics, Bricktown, NJ, 1965.

[5]Horner, S. F., *Fluid Dynamic Lift*, Horner Fluid Dynamics, Bricktown, NJ, 1975.

[6]Chapman, D. R., "Computational Aerodynamics Development and Outlook," *AIAA Journal*, Vol. 17, Dec. 1979, pp. 1293–1313.

[7]Kutler, P., "Computation of Three-Dimensional Inviscid Supersonic Flows," *Progress in Numerical Fluid Dynamics*, Springer-Verlag Lecture Notes in Physics, Vol. 41, 1975.

[8]U.S. Standard Atmosphere, National Oceanic and Atmospheric Administration, NOAA S/T 76-1562, U.S. Government Printing Office, Washington, DC, 1976.

[9]Kliore, A. (ed.), "The Mars Reference Atmosphere," *Advances in Space Research*, Vol. 2, No. 2, Committee on Space Research (COSPAR), Pergamon, Elmsford, NY, 1982.

[10]Bauer, G. L., Cornick, D. E., Habeger, A. R., Peterson, F. M., and Stevenson, R., "Program to Optimize Simulated Trajectories (POST)," NASA CR-132689, 1975.

[11]Well, K. H., and Tandon, S. R., "Rocket Ascent Trajectory Optimization via Recursive Quadratic Programming," *Journal of the Astronautical Sciences*, Vol. 30, No. 2, April–June 1982, pp. 101–116.

[12]Brusch, R. G., "Trajectory Optimization for the Atlas-Centaur Launch Vehicle," *Journal of Spacecraft and Rockets*, Vol. 14, Sept. 1977, pp. 541–545.

[13]Gottlieb, R. G., and Fowler, W. T., "Improved Secant Method Applied to Boost Trajectory Optimization," *Journal of Spacecraft and Rockets*, Vol. 14, Feb. 1977, pp. 201–205.

[14]Chilton, F., Hibbs, B., Kolm, H., O'Neill, G. K., and Phillips, J.,,"Mass-Driver Applications," *Space Manufacturing from Nonterrestrial Materials*, Vol. 57, edited by G. K. O'Neill, Progress in Astronautics and Aeronautics, AIAA, New York, 1977.

[15]Fleming, F. W., and Kemp, V. E., "Computer Efficient Determination of Optimum Performance Ascent Trajectories," *Journal of the Astronautical Sciences*, Vol. 30, No. 1, Jan.–March 1982, pp. 85–92.

[16]Covault, C., "Launch Activity Intensifies as Liftoff Nears," *Aviation Week & Space Technology*, Vol. 114, April 1981, pp. 40–48.

[17]McHenry, R. L., Brand, T. J., Long, A. D., Cockrell, B. F., and Thibodeau, J. R. III, "Space Shuttle Ascent Guidance, Navigation, and Control," *Journal of the Astronautical Sciences*, Vol. 27, No. 1, Jan.–March 1979, pp. 1–38.

[18]Schleich, W. T., "The Space Shuttle Ascent Guidance and Control," AIAA Paper 82-1497, Aug. 1982.

[19]Schleich, W. T., "Shuttle Vehicle Configuration Impact on Ascent Guidance and Control," AIAA Paper 82-1552, Aug. 1982.

[20]Pearson, D. W., "Space Shuttle Vehicle Lift-Off Dynamics Occurring in a Transition from a Cantilever to a Free-Free Flight Phase," AIAA Paper 82-1553, Aug. 1982.

[21]Olson, L., and Sunkel, J. W., "Evaluation of the Shuttle GN&C During Powered Ascent Flight Phase," AIAA Paper 82-1554, Aug. 1982.

[22]Goodfellow, A. K., Anderson, T. R., and Oshima, M. T., "Inertial Upper Stage/Tracking Data Relay Satellite (IUS/TDRS) Mission Post-Flight Analysis," *Proceedings of the AAS Rocky Mountain Guidance and Control Conference*, AAS Paper 84-050, Feb. 1984.

[23]Morring, Frank, "Test Puts Hybrid Rockets Back on the Table," *Aviation Week & Space Technology*, Vol. 158, 3 Feb 2003, p. 50.

[24]Isakowitz, S. J., Hopkins, J. P., Jr, and Hopkins, J. B., *International Reference Guide to Space Launch Systems*, 3rd ed., American Institute of Aeronautics and Astronautics, 1999.

Problems

5.1 Assume that a spacecraft headed for Saturn has a mass of 500 kg, and that the upper stage supplying the required ΔV of 7 km/s uses lox/hydrogen with $I_{sp} = 444$ s. Assuming the stage has a structural ratio $\varepsilon = 0.1$, what total mass in low Earth orbit is required to send the payload to Saturn?

5.2 A lunar transfer vehicle masses 50 tonnes (metric tons, or 1000 kg) fully fueled, has a dry mass of 15 tonnes, which includes 5 tonnes of cargo. Assume lox/hydrogen engines with $I_{sp} = 445$ seconds are used. How much ΔV does this vehicle provide?

5.3 A typical geosynchronous spacecraft requires 15 m/s/yr for orbital stationkeeping and momentum dumping. The satellite must last 10 years, and has a dry mass of 4000 kg. What is the fuel budget for monopropellant hydrazine with $I_{sp} = 220$ s?

5.4 With bipropellant, we obtain $I_{sp} = 300$ s for the case in Problem 3, and only 210 kg of fuel are required for the total mission. Whether or not you solved Problem 5.3, assume 300 kg were required for monopropellant in that case. Do you think mono- or bipropellant would be the better choice? Why or why not?

5.5 A design team is considering a shuttle-compatible upper stage for use an unmanned reusable orbital transport vehicle (OTV). The OTV will use a proven lox/hydrogen engine with a specific impulse of 446 seconds. The nominal mission requirement is to deploy the OTV in a 185 km altitude circular shuttle orbit at 28.5° inclination, after which it must ferry a satellite to geostationary orbit for release. The OTV must carry sufficient propellant to enable it to return to the shuttle orbit for later retrieval. To be shuttle compatible, it cannot mass more than 24,000 kg, including fuel and payload. You may assume that the fuel-dump problem required in the case of a shuttle abort has been addressed.

(a) What is the sequence of orbital maneuvers required?
(b) What is the total ΔV capability required of the OTV?
(c) Assuming a structural coefficient of $\varepsilon = 0.10$ for this vehicle, what is the payload capability for the mission?
(d) What mass of propellant (LH_2 and LO_2) will be required?
(e) What is the payload sensitivity to I_{sp}?
(f) Could this vehicle be useful as a manned OTV? Why or why not?

5.6 The SSME weighs 6600 lbs, uses lox/hydrogen propellants, and has a chamber pressure of approximately 2750 psi, propellant combustion temperature of 3517 K, an expansion ratio of 77.5:1, a vacuum I_{sp} of 452.5 s, and sea-level and vacuum thrust levels of 375,000 lbf and 470,000 lbf, respectively. The ratio of specific heats is approximately $k = 1.22$, and the molecular weight of the combustion products is 16.

(a) What is the mass flow rate?
(b) What is the exit Mach number?
(c) At what altitude is the nozzle properly expanded?
(d) What is the nozzle exit area?

5.7 A new two-stage expendable liquid-fuel rocket is being designed, as a reference mission, to put 3000 kg into a due-East 200 km altitude circular orbit from Cape Canaveral. The "ideal velocity" required for this mission is found by POST analysis to be 9.5 km/s. Preliminary design estimates for structural and performance parameters are

Parameter	*Stage 1*	*Stage 2*
η	0.9	0.8
I_{sp}	290 s	320 s

Additional constraints exist. It is desired that the second stage burnout acceleration be limited to 5 g, to avoid over-stressing potential payloads, and

that the first stage thrust-to-weight ratio be at least 1.4. Because, in general, it costs more to manufacture large, heavy objects than it does to manufacture smaller, lighter objects, it is desired that the vehicle have the minimum gross mass to accomplish its task.

(a) What is the gross mass for the optimal vehicle?
(b) What are the stage masses?
(c) What ΔV is contributed by each stage?
(d) What is the burn time for each stage?
Hint: It may be found that an iterative spreadsheet optimization is as efficient as the analytical approach presented in this text.

5.8 The required burnout velocity V_{bo} of a ballistic missile on a maximum range trajectory is given by (see e.g. *Fundamentals of Astrodynamics* by Bate, Mueller, and White, p. 293)

$$V_{bo} = \left(\frac{2(\mu/r_{bo})\sin\Psi/2}{1+\sin\Psi/2}\right)^{1/2}$$

where

$r_{bo} = R + h_{bo}$ = burnout radius
h_{bo} = burnout altitude $\cong$ 22 km
R = radius of Earth = 6378 km
μ = Earth gravitational constant = 398,600 km^3/s^2
$\Psi = s/R$ = range angle
s = surface range

For the purposes of this problem, "s" is a conservative approximation to the total range, which includes contributions (on the order of tens of kilometers) due to atmospheric ascent and reentry that are ignored here.

Military intelligence analysts are concerned about the possibility that a rogue nation may acquire an existing low-accuracy intermediate-range ballistic missile (IRBM) and, by putting a sophisticated upper stage on the missile, enable it to carry a nuclear weapon over an intercontinental range (e.g., 10,000 km). For analysis purposes, assume that any payload capability over 250 kg is considered threatening. Assume also that an upper stage with $\eta = 0.8$ and $I_{sp} = 300$ s is within the rogue nation's capability. Is there reason to be concerned?

5.9 Because the moon has no atmosphere, the Apollo lunar module could make effective use of a pure gravity-turn trajectory. It is most efficient for such a vehicle to burn out at a relatively low altitude, then coast in what is essentially the outbound leg of a Hohmann transfer (see Chapter 4) to the desired peak altitude, at which point a circularization burn was executed. For the Apollo missions, the CSM was parked in a nominally circular

orbit, 70 n.mi. altitude above the lunar surface. The later-design (J-series) lunar module ascent stage had a dry mass of 2130 kg, and a fully-fueled mass of 4760 kg. The ascent engine had a thrust of 15,260 N, $I_{sp} = 300$ s, and an ascent burn time of 7 m, 30 s. Assume a vertical liftoff for 10 seconds, and design a gravity-turn trajectory to allow the lunar module to coast to the desired altitude.

5.10 From *Isakowitz*, we note that each space shuttle solid rocket booster (SRB) has the following parameters:

$M_i = 590{,}000$ kg	$I_{sp} = 267$ s (vacuum)
$M_p = 502{,}000$ kg	$A_e/A^* = 7.5$
$M_i = 88{,}000$ kg	$t_{burn} = 123$ s
$p_c = 6.33 \times 10^6$ N/m^2	$T = 11.79 \times 10^6$ N (sea-level)

The new super-lightweight external tank, as cited earlier, has the following parameters:

$$M_i = 748{,}000 \text{ kg}$$
$$M_s = 27{,}000 \text{ kg}$$
$$M_p = 721{,}000 \text{ kg}$$

Shuttle main engine parameters were given above in Problem 5.6. Assume these components were used to develop a new, expendable launch vehicle with four SRBs attached to the existing external tank and SSMEs. What payload capability would exist for an ideal ΔV of 9.5 km/s?

6
Atmospheric Entry

6.1 Introduction

In the early days of space exploration, the problem of controlled atmospheric entry was as difficult and constraining as that of rocket propulsion itself. Although the technology is relatively mature and well understood today, it remains true that any Earth orbital mission for which the payload must be recovered, or any interplanetary mission targeted for a planet with an atmosphere, must address the issue of how to get down as well as how to get up. This obviously includes manned missions, and indeed some of the most challenging areas of manned spacecraft design are associated with systems and procedures for effective atmospheric entry.

It is worth noting that in several hundred missions over more than four decades of manned spaceflight, there has been only one fatal launch accident, the *Challenger* 51-L mission. Numerous launch abort procedures have evolved, and a variety of these (on-pad, in-flight, and abort-to-orbit) have been exercised in particular cases. In contrast, there have been three missions involving fatal reentry system failures of one kind or another (Soyuz 1, Soyuz 11, and *Columbia* STS-107), and several other "close calls." Though the technology of atmospheric entry is relatively mature, it remains very exacting in its demands, a characteristic deriving in part from the general lack of plausible abort scenarios following a primary system failure.

Atmospheric entry technology is a highly interdisciplinary area of space vehicle design. This is due to the many different functions that must be satisfied by the atmospheric entry system, and to the wide range of flight regimes and conditions encountered during a typical entry.

Basically, the atmospheric entry system must provide controlled dissipation of the combined kinetic and potential energy associated with the vehicle's speed and altitude at the entry interface. By controlled dissipation, we imply that both dynamic and thermal loads are maintained within acceptable limits during entry. This requires a carefully designed flight trajectory and often a precision guidance system to achieve the desired results. Control of the vehicle in response to guidance commands implies control of lift and drag throughout the flight. This is a nontrivial task for entry from Earth orbit, because it spans an aerodynamic flight range from subsonic speeds to Mach 25, and even higher speeds are encountered for hyperbolic entry. Finally, the entry system must provide suitable provisions

for surface contact, usually with constraints on the landing location and vehicle attitude. These and other issues are addressed in this chapter.

6.2 Fundamentals of Entry Flight Mechanics

6.2.1 Equations for Planar Flight

Figure 6.1 shows the geometry of atmospheric entry for planar flight over a spherical, nonrotating planet. Aside from the fact that thrust is normally zero for entry flight, the flight mechanics are the same as those discussed in Chapter 5 in connection with ascent vehicle performance. As in Sec. 5.2, we take this model to be the simplest one that allows presentation of the important phenomena, and Eqs. (5.11) will again apply. Assuming a nonthrusting entry, we have

$$\frac{dV}{dt} = -\frac{D}{m} - g\sin\gamma \tag{6.1a}$$

$$\frac{V d\gamma}{dt} = \frac{L}{m} - \left(g - \frac{V^2}{r}\right)\cos\gamma \tag{6.1b}$$

$$\frac{ds}{dt} = \left(\frac{R}{r}\right)V\cos\gamma \tag{6.1c}$$

$$\frac{dr}{dt} = \frac{dh}{dt} = V\sin\gamma \tag{6.1d}$$

$$L = \tfrac{1}{2}\rho V'^2 S C_L \tag{6.1e}$$

$$D = \tfrac{1}{2}\rho V'^2 S C_D \tag{6.1f}$$

$$g = g_s\left[\frac{R}{R+h}\right]^2 \tag{6.1g}$$

where

V = inertial velocity magnitude
V' = speed relative to planetary atmosphere
R = planetary radius
h = height above surface
$r = R + h$ = radius from planetary center
s = down-range travel relative to nonrotating planet
γ = flight-path angle, positive above local horizon
m = vehicle mass
L = lift force, normal to flight path
D = drag force, parallel to flight path
C_L = lift coefficient
C_D = drag coefficient
ρ = atmosphere density
S = vehicle reference area for lift and drag
g = gravitational acceleration
g_s = surface gravitational acceleration

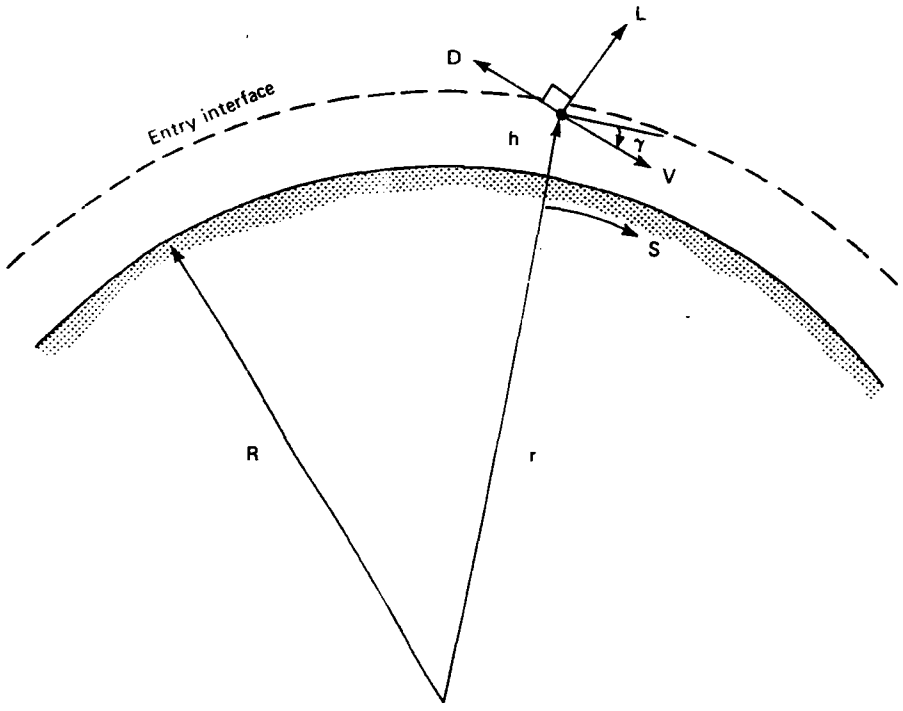


Fig. 6.1 Atmospheric entry geometry.

These equations may be integrated forward in time subject to prescribed entry interface conditions (r_e, V_e, γ_e), a model for the atmosphere density $\rho(h)$, and specified values of the vehicle control parameters C_L and C_D. Indeed, this would be essential prior to specification of a flight vehicle configuration and entry trajectory. However, the comments in Sec. 5.2 in regard to direct numerical integration apply here as well. Such a procedure will produce more accurate results (which incidentally justifies the use of a more sophisticated mathematical and physical model), but at the cost of considerable loss of insight.

To obtain the broader perspective that is possible with an analytical solution, three simplifying assumptions are employed:

1) The atmosphere density is approximated by

$$\rho(h) = \rho_s e^{-\beta h} \tag{6.2}$$

where

$h = r - R$

β^{-1} = scale height, assumed constant

ρ_s = surface density

2) The gravitational acceleration g is assumed constant.
3) In Eq. (6.1b) we employ the approximation

$$\frac{V^2}{r} \simeq \frac{V^2}{R} \simeq \frac{V^2}{r_e} \tag{6.3}$$

Some comments on these assumptions are in order.

The use of atmosphere models has been discussed in Chapters 4 and 5 with respect to orbit decay and ascent vehicle flight. Although for numerical calculations a more detailed model such as the U.S. Standard Atmosphere[1] or the Committee on Space Research (COSPAR) 1986 International Reference Atmosphere might be employed,[2] such models are inappropriate for analytical work where closed-form results are desired.

Using the ideal gas law

$$P = \rho R_{\text{gas}} T \tag{6.4}$$

and the hydrostatic equation

$$\mathrm{d}p = -\rho g \, \mathrm{d}r \tag{6.5}$$

we obtain the differential relation

$$\frac{\mathrm{d}\rho}{\rho} = -\left(\frac{g}{R_{\text{gas}}} + \frac{\mathrm{d}T}{\mathrm{d}r}\right)\frac{\mathrm{d}r}{T} = -\beta \, \mathrm{d}r \tag{6.6}$$

where

R_{gas} = specific gas constant
T = absolute temperature
P = pressure

Equation (6.2) is the integrated result of Eq. (6.6) subject to the assumption of constant scale height β^{-1}. Since by definition

$$\beta = -\frac{(g/R_{\text{gas}} + \mathrm{d}T/\mathrm{d}r)}{T} \tag{6.7}$$

it is seen that the assumption of constant scale height requires a locally isothermal atmosphere and fixed gravitational acceleration.

The Earth's atmosphere contains regions of strong temperature gradient, with resulting substantial variations in scale height. For entry analysis as given here, it is customary to select β^{-1} for the best fit according to some criteria. Chapman[3] recommends a weighted mean for β^{-1} of 7.165 km, and Regan[4] suggests 6.7 km as a better high-altitude approximation. Both Vinh et al.[5] and Regan[4] give extensive discussions of atmosphere models.

If flight in a particular altitude region is of primary interest, as may sometimes be the case, then ρ_s may be equated to the density near the altitude of interest. Careful selection of β^{-1} then allows a better fit of the density model to local conditions, at the expense of greater deviation elsewhere.

Gravitational acceleration varies according to Eq. (6.1g). For the Earth, with the entry interface altitude commonly taken by convention as 122 km (400,000 ft), variations in g amount to no more than 4%, an acceptable error at this level of study. Even less error results if the reference altitude is chosen to lie between the surface and the entry interface.

The remaining assumption that variations in $(1/r)$ are negligible in Eq. (6.1b) contributes an error of about 2% over the entry altitude range of interest for the Earth. This is insignificant in comparison with other approximations thus far employed. In this chapter we will consistently use $(1/R)$ to replace $(1/r)$ in Eq. (6.1b) and derivations that follow from it. Other choices are possible, with $(1/r_e)$ being the most common alternative.

Two independent variable transformations are normally employed in conjunction with the assumptions discussed earlier. It is customary to eliminate time and altitude in favor of density through the kinematic relation

$$\frac{\mathrm{d}}{\mathrm{d}t} = \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)\frac{\mathrm{d}}{\mathrm{d}r} = V\sin\gamma\frac{\mathrm{d}}{\mathrm{d}r} \tag{6.8}$$

and the density model

$$\mathrm{d}\rho = -\rho_s e^{-\beta h}\beta\,\mathrm{d}h \tag{6.9}$$

or

$$\frac{\mathrm{d}}{\mathrm{d}r} = -\beta\rho\frac{\mathrm{d}}{\mathrm{d}\rho} \tag{6.10}$$

With some additional manipulations,[6] Eqs. (6.1a) and (6.1b) are transformed to yield

$$\frac{\mathrm{d}(V^2/gR)}{\mathrm{d}\rho} = \left(\frac{SC_D}{m}\right)\left(\frac{1}{\beta}\sin\gamma\right)\left(\frac{V^2}{gR}\right) + \frac{2}{\rho\beta R} \tag{6.11a}$$

$$\frac{\mathrm{d}(\cos\gamma)}{\mathrm{d}\rho} = \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) - \left(\frac{gR}{V^2} - 1\right)\frac{\cos\gamma}{\rho\beta R} \tag{6.11b}$$

These are the reduced planar equations for flight over a nonrotating spherical plane. The reduced equations still cannot be integrated directly and thus require further approximations to obtain closed-form results. Nonetheless, they are worthy of some examination at this point.

In reduced form, the dependent variables are the non-dimensional energy (V^2/gR) and γ, with ρ the independent variable. Because at the surface of the planet

$$gR = \frac{\mu}{R} = \frac{GM}{R} = V_{\text{circ}}^2 \tag{6.12}$$

it is seen that the entry energy is normalized to the circular velocity at the planetary radius. If r_e is used instead of R, as discussed earlier, then the entry energy is referenced to the circular velocity at entry altitude.

Specification of entry interface conditions (V_e, γ_e, ρ_e) is sufficient to determine a particular trajectory subject to the fixed parameters in Eqs. (6.11). Trajectory solutions take the form of velocity and flight-path angle as a function of density, the independent variable. The location is obtained, if required, from Eqs. (6.1c) and (6.1d), with Eq. (6.2) relating altitude to density.

Four parameters control the solution of Eqs. (6.11): two define the vehicle and two define the relevant planetary characteristics. The vehicle parameters are the lift-drag ratio L/D and the ballistic coefficient m/SC_D. The planetary entry environment is determined by the radius R and the atmosphere scale height β^{-1}.

In the following sections we will examine various solutions obtained by first simplifying and then integrating Eqs. (6.11). Such solutions implicitly assume the controlling parameters given earlier to be constant. Obviously they are not; we have devoted considerable discussion to this point both here and in Sec. 5.2. The sources of parameter variation can be both natural and artificial. That is, L/D and the ballistic coefficient will vary considerably over the Mach 25 to 0 entry flight regime due to the differing flowfield dynamics. Additionally, however, the vehicle L/D is the primary control parameter available to the trajectory designer for tailoring the entry profile. Substantial mission flexibility can be gained with judicious L/D control. This cannot be modeled in the closed-form results derived from Eqs. (6.11) and offers another reason why detailed trajectory design requires a numerical approach.

The reader should note that it is common practice, particularly when English units are employed, to define the ballistic coefficient based on weight, i.e., mg/SC_D. When this is done, the ballistic coefficient C_B will have units of pounds per square feet or Newtons per square meters. We avoid this practice and will consistently express C_B in units of kilograms per square meter.

Some discussion of the physical significance of the terms in the reduced equations is in order, because in later sections we will obtain approximate solutions based on assuming a flat Earth, no gravity, small flight-path angle, etc. Equations (6.1) clearly show the influence of various terms such as lift, drag, and centrifugal force. The physical identity of the various terms is not as clear in the reduced form of Eqs. (6.11).

The first term on the right-hand side of Eq. (6.11a) is the reduced drag, and the second term ($2/\rho\beta R$) is the reduced form of the tangential gravitational

component, $g \sin\gamma$, in Eq. (6.1a). Depending on the vehicle configuration and flight conditions, one of these terms may be dominant.

The first term on the right-hand side of Eq. (6.11b) is the reduced lift force. The term $(g - V^2/r)$ in Eq. (6.1b) gives the net normal force contribution of gravity and centrifugal acceleration. The corresponding term on the right-hand side of Eq. (6.11b) is obvious; note, however, that gR/V^2 is the gravitational term, whereas "1" is the reduced centrifugal term.

The surface density ρ_s appears only through Eq. (6.2), which relates density to altitude. Care must be exercised in some cases to avoid the introduction of unrealistic density values (i.e., those corresponding to negative altitudes) when using integrated results from Eqs. (6.11). This will be seen in later sections where ballistic and skip entry are discussed.

As pointed out, Eqs. (6.11) are not directly integrable in the general case, and, if numerical integration is to be employed, there is little reason to use the reduced equations. If closed-form results are to be obtained, it is thus necessary to simplify the analysis even further. Such simplification yields several possible first-order entry trajectory solutions, classically denoted as ballistic, equilibrium glide, and skip entry. These solutions may be adequate within a restricted range of conditions, but most results are approximate only and are primarily suited to initial conceptual design. However, they are very useful in demonstrating the types of trajectories that can exist and the parameters that are important in determining them.

6.2.2 Ballistic Entry

First-order ballistic entry analysis involves two assumptions in addition to those thus far employed. By definition of ballistic entry, zero lift is assumed. We also employ the approximation

$$\frac{1}{\beta R} \simeq 0 \tag{6.13}$$

which results in dropping terms in Eqs. (6.11) where βR is in the denominator. Some examination of these assumptions is in order.

The zero-lift approximation is often quite accurate and can be made more so when desired. Entry bodies possessing axial symmetry and flown at zero angle of attack will fly nominally ballistic trajectories. In a practical vehicle, small asymmetries will always produce an offset between the center of mass and the center of pressure. This causes the vehicle, unless it is spherical, to fly aerodynamically trimmed at some angle of attack, inducing a lift force. However, this may be dealt with by slowly rolling the vehicle during the entry to cancel out any forces normal to the velocity vector. For example, the Mercury spacecraft was rolled at a nominal 15°/s rate during reentry.

Substantially higher roll rates are employed for ballistic entry of intercontinental ballistic missile (ICBM) warheads. The aerodynamic forces

generated in this case can give rise to a significantly more complex entry trajectory. Such topics are considerably beyond the scope of this text. Platus[7] gives an excellent summary of the state of the art in ballistic entry analysis.

The second approximation is somewhat less valid. Although it is true that βR is typically large (approximately 900 for the Earth), this does not justify dropping all terms in Eqs. (6.11) where it appears in the denominator. In particular, $(2/\rho\beta R)$ represents the reduced gravitational force along the trajectory. By omitting it, we assume that the drag force dominates, which is not always true. The drag is always small and usually comparable to the tangential gravitational force at the entry interface. Toward the end of the entry phase, when the velocity becomes small, $(2/\rho\beta R)$ will again dominate, and neglecting it will lead to inaccuracy.

From the preceding comments it is seen that our second ballistic entry assumption corresponds to neglecting gravity with respect to drag in Eq. (6.1a) and to neglecting the difference between gravitational and centrifugal force in Eq. (6.1b). Thus, first-order ballistic entry may be viewed as a zero-g, flat-Earth solution.

In any case, if terms containing $(1/\beta R)$ are dropped and zero lift is assumed, Eq. (6.11b) integrates immediately to yield

$$\cos\gamma = \cos\gamma_e \tag{6.14}$$

i.e., the flight-path angle remains constant at the entry value.

The validity of this result is obviously somewhat questionable. Intuition and experience suggest that for shallow entry angles, such as those that are required for manned flight, the vehicle will undergo a lengthy high-altitude deceleration and then, its energy depleted, nose over into a nearly vertical trajectory. Also, the shallow entry angle produces a lengthier entry, with consequently more time for gravity to curve the flight path. As discussed earlier, this is the problem with neglecting $(1/\beta R)$ in Eqs. (6.11). Nonetheless, when entry occurs at a reasonably steep angle, as is typical for an ICBM, the flight path is indeed nearly straight. This is graphically illustrated in numerous time exposure photographs of entry body flight tests. Ashley[6] suggests that the surprisingly shallow value of $-\gamma_e > 5°$ is sufficient to yield a good match to the ballistic entry assumptions.

Equation (6.11a) may be integrated subject to our assumptions to yield

$$V = V_e \exp\left[\left(\frac{1}{2\beta}\right)\left(\frac{\rho_s}{\sin\gamma_e}\right)\left(\frac{SC_D}{m}\right)\exp(-\beta h)\right] \tag{6.15}$$

for the velocity as a function of altitude and entry flight-path angle. Of possibly greater interest is the derivative of velocity, the acceleration, which can be shown to have a peak value of

$$a_{\max} = -\left(\frac{\beta V_e^2}{2e}\right)\sin\gamma_e \tag{6.16}$$

which occurs at altitude

$$h_{\text{crit}} = \left(\frac{1}{\beta}\right) \ln\left[\left(\frac{-1}{\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{\rho_s}{\sin \gamma_e}\right)\right] \tag{6.17}$$

and velocity

$$V_{\text{crit}} = \frac{V_e}{e} \tag{6.18}$$

If the altitude of peak deceleration is to be positive, the argument of the logarithm in Eq. (6.17) must be greater than unity. Assuming hypervelocity impact with the ground is to be avoided, the useful range of entry angles is defined by

$$0 < -\sin \gamma_e < \left(\frac{SC_D}{m}\right)\left(\frac{\rho_s}{\beta}\right) \tag{6.19}$$

The ballistic entry results just given should be used with caution. For example, Eq. (6.16) predicts zero peak deceleration for $\gamma_e = 0$, a grazing entry. The first-order analysis provides a very poor model of the entry in such a case. The second-order analysis by Chapman,[3] summarized in Fig. 6.2, shows that entry from low circular Earth orbit has an irreducible deceleration load of about $8g$, which occurs for flight-path angles between 0 and $-1°$. For flight-path angles steeper than about $-5°$, the theories are in reasonable agreement as to trend, though the first-order theory underpredicts the deceleration by about $1g$.

Figure 6.3 shows a typical shallow angle ballistic entry solution, obtained by numerically integrating Eqs. (6.1). The sharp peak in entry deceleration and the rather high value of that peak are characteristic of ballistic entry. As indicated earlier, the peak is seriously underestimated in this case by Eq. (6.16), which predicts a $3.6g$ maximum deceleration.

Note also in Fig. 6.3 the difference between the inertial and Earth-relative entry speed. The Earth-relative (hence atmosphere-relative) speed should be employed for V_e; however, the approximations inherent in a first-order solution are such that correcting for atmosphere-relative velocity may not be important. If used, the appropriate correction is

$$V_e' = \left[1 - \left(\frac{\omega r_e}{V_e}\right)\cos i\right] V_e \tag{6.20}$$

where

ω = angular velocity of Earth, 7.292×10^{-5} rad/s
V_e = inertial entry velocity
V_e' = Earth-relative entry velocity
i = orbital inclination
r_e = entry interface radius

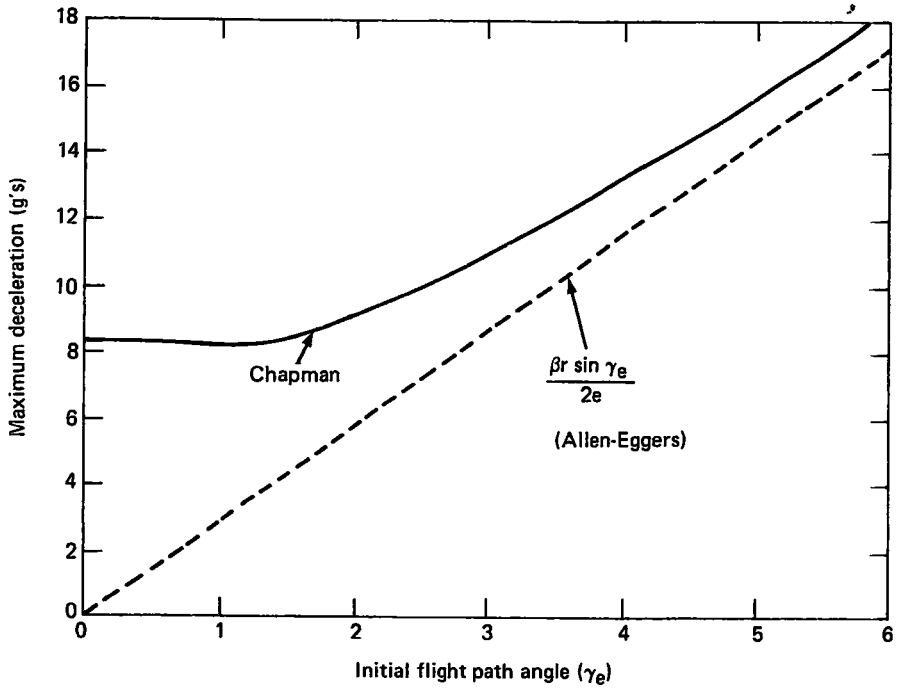


Fig. 6.2 Deceleration loads for ballistic entry Earth orbit.

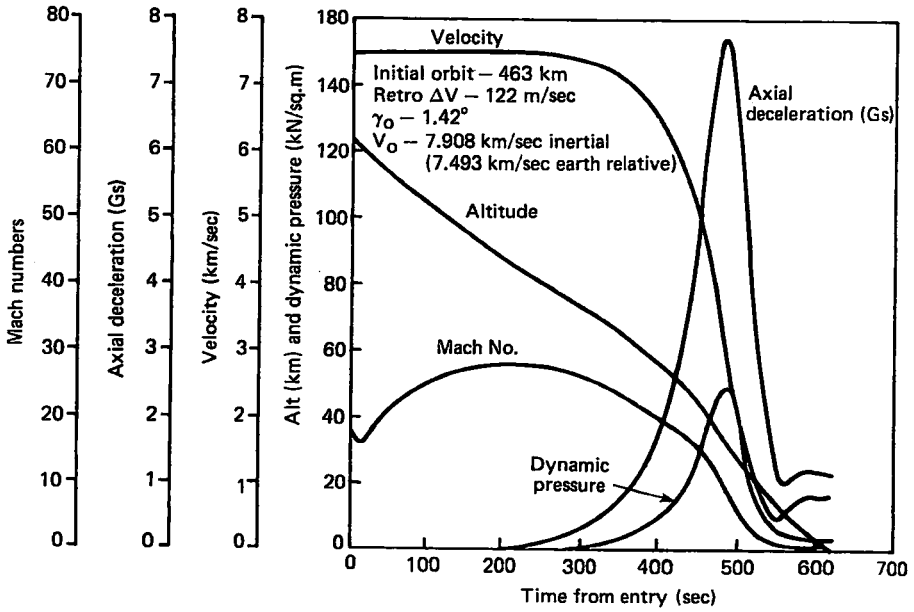


Fig. 6.3 Typical ballistic entry from Earth orbit.

As mentioned, ballistic entry from Earth orbit will require loads of $8g$ or more. Entry at hyperbolic speed is practical only at fairly steep angles and consequently very high deceleration if skip-out is to be avoided. It will be seen in later sections that a moderate value of L/D greatly eases the entry dynamic load.

Purely ballistic entry has a somewhat limited range of application; however, within this range it is quite useful and is widely employed. It is simple to mechanize, requiring little or no guidance beyond stabilization for the deorbit burn, if any. Entry and landing accuracy is determined primarily by the precision to which V_e and γ_e are controlled, knowledge of the ballistic coefficient, and variations in atmosphere properties. However, relatively large dispersions in the controlling parameters can usually be tolerated without disaster; the technique is quite robust when, again, one is within its range of applicability. Table 6.1 gives the ballistic entry vehicle dispersions used in the Mercury program.[8] Actual flight performance was somewhat better than these estimates.

Ballistic entry has seen application to numerous vehicles, including the manned U.S. Mercury spacecraft and Russian Vostok/Voshkod series, as well as unmanned spacecraft, including Discoverer, Pioneer Venus, and Galileo. The Mercury procedure is typical of those used for entry from low Earth orbit, 200–500 km.

The Mercury entry sequence was initiated by performing a deorbit burn of approximately 150 m/s at a pitch attitude of 34° above the horizontal. This resulted in a nominal flight-path angle of $-1.6°$ at the 122-km entry interface. At this point, indicated by a 0.05-g switch, a 15°/s roll rate was initiated, with pitch and yaw attitude rates controlled to zero. Attitude control was terminated at 12–

Table 6.1 Mercury spacecraft entry dispersions

		Dispersions, n mile		
Error source	Tolerance	Overshoot	Undershoot	Cross range
Perigee altitude	± 0.05 n mile	11.0	11.0	
Eccentricity	± 0.0001	15.6	15.6	
Inclination	± 0.10°			6.0
Pitch angle	± 6.9°	65.0	10.0	
Yaw angle	± 8.1°			15.0
Retrofire velocity	± 2.4%	85.0	85.0	
Down-range position	± 5 n mile	5.0	5.0	
Cross-range position	± 5 n mile			5.0
Drag coefficient	± 10%	5.8	5.8	
Atmosphere density	± 50%	15.4	15.4	
Winds		2.5	2.5	4.0
Root-sum-squared (RSS) Total		110.1	89.4	17.4

15-km altitude upon deployment of a drogue stabilization parachute, followed by main parachute deployment at 3-km altitude. The flight time from retrofire to landing was about 20 min, and the range was approximately 5500 km.

The overwhelming historical base of ballistic entry flight experience lies with unmanned satellite reconnaissance vehicles, which in some cases utilized small entry vehicles for return of film canisters from orbit.[9] These vehicles performed shallow-angle, Mercury-type reentries and routinely landed within 20 km of their intended targets. Over 300 such flights took place successfully.

A steep ballistic entry can be much more accurate, and for this reason it has been favored for use with ICBMs. High dynamic loads, on the order of several hundred g, are possible, but unmanned vehicles can be designed to withstand this. Purely ballistic entry can yield a targeting accuracy on the order of a few hundred meters under such conditions.

The first-order ballistic entry solution given here follows the classical treatment of Allen and Eggers[10] and as previously emphasized is valid only at relatively steep flight-path angles. Higher order theories that are more suited to shallow-angle entry are available.[3,5] However, none of these treatments produces closed-form results suited to rapid analysis of vehicle loads in preliminary design. It is our view that, if more accuracy is required, direct numerical integration of the basic equations is preferred to a numerical solution from a second-order theory.

6.2.3 Gliding Entry

In contrast to ballistic entry, first-order gliding entry analysis assumes that the vehicle generates sufficient lift to maintain a lengthy hypersonic glide at a small flight-path angle. Clearly this is an idealization. Substantial lift is readily obtained at hypersonic speeds, and it is possible to achieve shallow-angle gliding flight over major portions of the entry trajectory. However, a practical vehicle configuration for an extended hypersonic glide will be poorly suited to flight at low supersonic and subsonic speeds. Toward the end of its flight, such a vehicle must fly at a steeper angle to maintain adequate airspeed for approach and landing control.

This is readily illustrated by the space shuttle entry profile.[11–13] The entry guidance phase is initiated at the entry interface altitude of 122 km with the flight-path angle typically about $-1.2°$. It terminates when the shuttle reaches an Earth-relative speed of about 760 m/s (Mach 2.5), at an altitude of approximately 24 km and a distance to the landing site of about 110 km. This phase of flight covers a total range of roughly 8500 km, with the average flight-path angle on the order of $-1°$. The shuttle's hypersonic glide phase is considerably longer than was the case for preceding manned vehicles.

Upon completion of entry guidance, the terminal area energy management (TAEM) phase is initiated. The goal of this procedure is to deliver the orbiter to the runway threshold at the desired altitude and speed for approach and landing.

This phase of flight covers a range of 110 km while descending through 24 km of altitude, at an average flight-path angle of about $-12°$, an order of magnitude steeper than that for the hypersonic phase.

The results of this section, although inadequate in the terminal flight regime, may well be appropriate for the major portion of a gliding entry. In keeping with the small angle assumption noted earlier, we assume $\sin \simeq \gamma$, $\cos \gamma \simeq 1$, and hence $\mathrm{d}(\cos\gamma)/\mathrm{d}\rho \simeq 0$. With these approximations Eq. (6.11b) is reduced to an algebraic equation for energy as a function of density:

$$\frac{V^2}{gR} = \frac{1}{[1 + (R/2)(SC_D/m)(L/D)\rho]} \tag{6.21}$$

or, from Eq. (6.2),

$$\frac{V^2}{gR} = \frac{1}{[1 + (R/2)(SC_D/m)(L/D)\rho_s e^{-\beta h}]} \tag{6.22}$$

Equation (6.21) may be differentiated with respect to ρ and substituted into Eq. (6.11a) to solve for the flight-path angle. Consistent with the assumption of small γ, we neglect the tangential component of gravitational acceleration and obtain

$$\sin\gamma \simeq \gamma \simeq \frac{-2}{[\beta R(L/D)(V^2/gR)]} \tag{6.23}$$

Note that, although the flight-path angle γ is assumed to be small and its cosine constant, γ is not itself assumed constant.

Equation (6.22) is an equilibrium glide result, where the gravitational force cancels the sum of the centrifugal and lift forces. This is readily seen by noting the physical identity of the various terms in Eq. (6.11b). Of course, the equilibrium is not exact because the derivative term in Eq. (6.11b) is not identically zero. For this reason the trajectory given by Eq. (6.22) is sometimes, and more correctly, referred to as a pseudoequilibrium glide.

To obtain the acceleration along the trajectory, note from Eq. (6.1a),

$$a = \frac{\mathrm{d}V}{\mathrm{d}t} \simeq -\frac{D}{m} = -\left(\frac{V^2}{2}\right)\left(\frac{SC_D}{m}\right)\rho \tag{6.24}$$

where again we neglect the gravitational acceleration along the flight path. Solving Eq. (6.21) for ρ and substituting above gives the tangential acceleration,

$$\frac{a}{g} = \frac{V^2/gR - 1}{L/D} \tag{6.25}$$

experienced by the vehicle during the equilibrium glide. Note that the maximum deceleration is encountered as the vehicle slows to minimum speed. Here we see the advantage of even small values of L/D in moderating entry deceleration

loads. For a Gemini-class vehicle, with a hypersonic L/D of approximately 0.2,[8] the peak load is $5g$, substantially lower than for a Mercury-style ballistic entry at the same flight-path angle. For the shuttle, which flies a major portion of its entry profile with a hypersonic L/D of about 1.1, Eq. (6.25) predicts essentially a $1g$ reentry. These results are consistent with flight experience.

By integrating the velocity along the entry trajectory, the total range of the equilibrium glide is found to be

$$s = \tfrac{1}{2}R\left(\frac{L}{D}\right)\ln\left[\frac{1}{1 - V_e^2/gR}\right] \tag{6.26}$$

Clearly, the greatest range is obtained when entry is performed at the maximum vehicle L/D.

As an example, consider a space shuttle entry at an atmosphere-relative speed of 7.5 km/s with a hypersonic L/D of 1.1 assumed. With these representative values, Eq. (6.26) yields a predicted range of about 8000 km, in good agreement with flight experience. This may be somewhat fortuitous, because the shuttle in fact uses substantial lift modulation during entry to achieve landing point control. This is shown in Fig. 6.4 for the STS-2 reentry.[14] Nonetheless, an L/D of 1.1 closely approximates the high-altitude, high-speed portion of the entry, and it is this portion that obviously has the most effect on total range.

Equation (6.26) is of limited utility when the entry speed approaches the reference circular velocity. In this case the argument of the logarithm is almost singular and hence extremely sensitive to the value of the entry velocity. This reflects a limitation of the first-order theory rather than any real physical effect. Equation (6.26) is also invalid for the supercircular entry, because the logarithm becomes imaginary in this case.

Figure 6.5 shows an entry trajectory simulation result for the British horizontal takeoff and landing (HOTOL) vehicle concept. HOTOL was a mid-1980s design for an unmanned, reusable, single-stage-to-orbit vehicle intended for runway launch and landing. Because it was to be quite light, it needed to fly a high, shallow-angle gliding entry to minimize peak dynamic and thermal loads. As shown, the result is an entry profile that closely approximates the pseudoequilibrium glide trajectory during the high-speed portion of the flight.

The first-order equilibrium glide solution given earlier is due to Eggers et al.[15] and is discussed by later authors, including Ashley[6] and Vinh et al.[5] As with ballistic entry, higher order solutions are available but in our view are as annoying to implement as a complete numerical solution and yield little additional insight compared to the first-order theory.

6.2.4 Skip Entry

Gliding entry flight is not restricted to the equilibrium glide condition discussed in the previous section. Of particular interest is the case of supercircular entry with sufficient lift to dominate the gravitational and

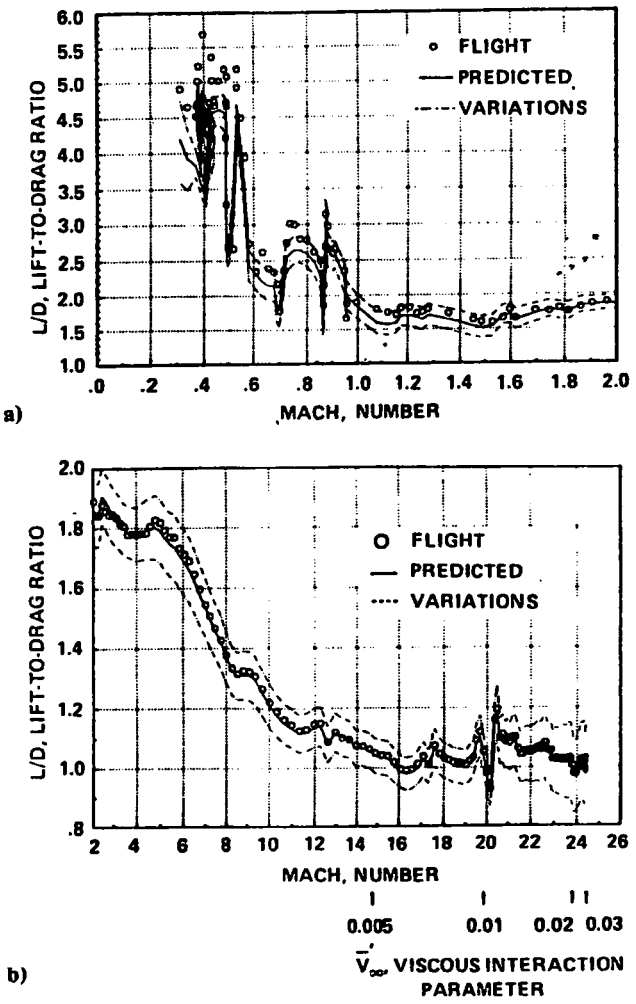


Fig. 6.4 Lift modulation for STS-2 reentry.

centrifugal forces. This is essentially the first-order ballistic entry model with lift added. With proper selection of parameters, the so-called skip or skip-glide entry may be obtained.

Consider the high-speed entry of a lifting vehicle at an initially negative flight-path angle. As always, the vehicle and atmosphere parameters are considered constant. With lift dominant over gravity, the flight path will be turned upward ($d\gamma/dt > 0$) so that the vehicle enters the atmosphere, reaches a certain minimum altitude, pulls up, and eventually exits the atmosphere at reduced speed. Provided the exit velocity and flight-path angle are properly controlled, a brief Keplerian phase ensues, followed by a second entry that occurs somewhat downrange from the first, as shown schematically in Fig. 6.6.

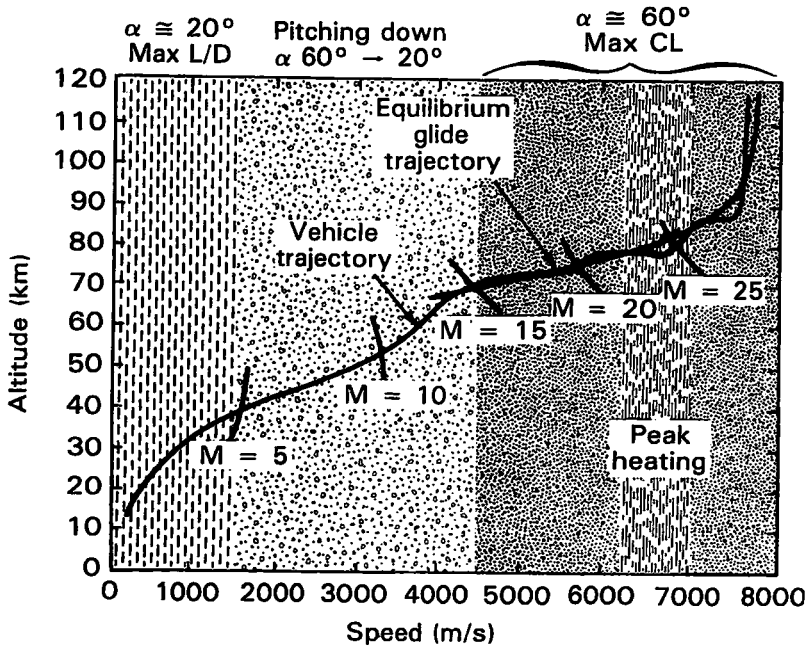


Fig. 6.5 HOTOL entry trajectory.

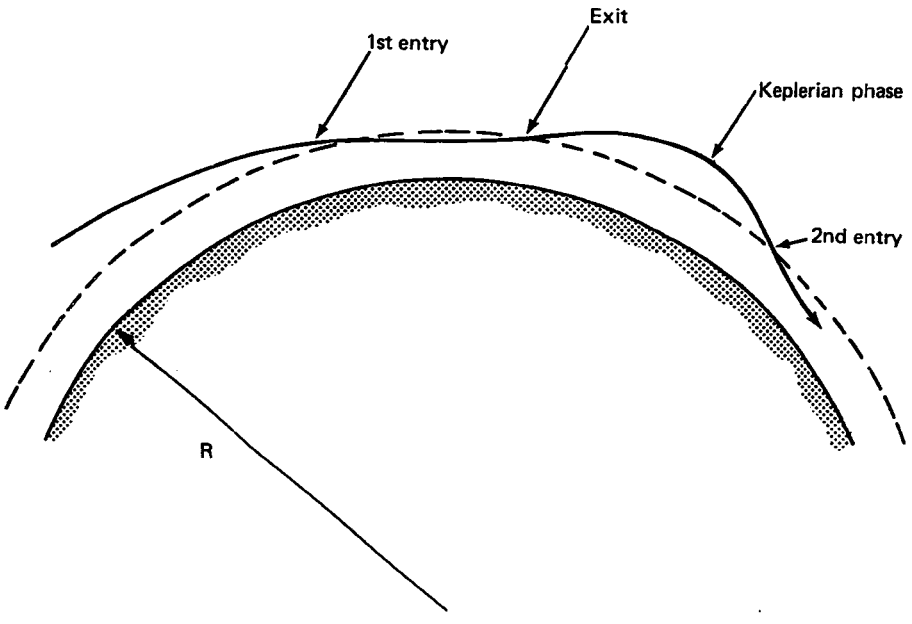


Fig. 6.6 Skip entry trajectory.

This type of trajectory offers considerable flexibility in the control of high-speed entry. For example, at lunar return speeds of about 11 km/s, the kinetic energy to be dissipated is about twice that of a typical low-Earth-orbit entry. This results in a very challenging thermal control problem, especially if the entry is required to occur in a single pass. With the skip entry, however, the vehicle can reduce its velocity sufficiently on its first pass to guarantee Earth capture. The brief suborbital lob allows radiative cooling and is followed by a second entry phase at lower speed.

Skip entry is also useful for range control and allows the vehicle to land in places that could not be reached via a single entry phase. Again, lunar return provides an example, this time in connection with the Soviet Zond series of unmanned lunar probes. Astrodynamic constraints forced the initial entry point for the return vehicles to occur at latitudes well to the south of the USSR. The entry vehicles were brought in over the Indian Ocean and, following a pronounced skip, were targeted to land in the USSR.

With the previously mentioned assumptions (0g, flat Earth) for a first-order model, Eqs. (6.11) are reduced to

$$\frac{\mathrm{d}(V^2/gR)}{\mathrm{d}\rho} = \left(\frac{SC_D}{m}\right)\left(\frac{1}{\beta \sin\gamma}\right)\left(\frac{V^2}{gR}\right) \tag{6.27a}$$

$$\frac{\mathrm{d}(\cos\gamma)}{\mathrm{d}\rho} = \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) \tag{6.27b}$$

Assuming, as always, constant ballistic coefficient and L/D, Eq. (6.27b) integrates immediately to yield the flight-path angle as a function of density (hence altitude):

$$\cos\gamma = \cos\gamma_e + \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right)\rho \tag{6.28}$$

with the approximation $\rho_e \simeq 0$. Since

$$\frac{\mathrm{d}V}{\mathrm{d}\gamma} = \left(\frac{\mathrm{d}V}{\mathrm{d}\rho}\right)\left[\frac{\mathrm{d}\rho}{\mathrm{d}(\cos\gamma)}\right]\left[\frac{\mathrm{d}(\cos\gamma)}{\mathrm{d}\gamma}\right] = -\frac{V^2/gR}{L/D} \tag{6.29}$$

the velocity as a function of flight-path angle is found to be

$$V = V_e \exp\left[-\frac{(\gamma - \gamma_e)}{(L/D)}\right] \tag{6.30}$$

Equations (6.28) and (6.30) constitute the first-order solution for gliding flight with lift large in comparison to other forces and range small with respect to the planetary radius. Though we are discussing skip entry, Vinh et al.[5] point out that these approximations are also appropriate to gliding entry at medium or large flight-path angle.

For the skip entry, however, we note that $\gamma = 0$ defines the pull-up condition. Equation (6.28) then yields the density at pull-up,

$$\rho_{\text{pullup}} = \rho_{\max} = \frac{2\beta(1 - \cos \gamma_e)}{(SC_D/m)(L/D)} \tag{6.31}$$

and Eq. (6.30) gives the corresponding velocity,

$$V_{\text{pullup}} = V_e \exp\left[\frac{\gamma_e}{L/D}\right] \tag{6.32}$$

Care must obviously be taken to ensure that the pull-up density corresponds to a positive altitude. Though this is not typically a problem for the Earth, it can be a constraint when considering skip entry at a planet, such as Mars, with a tenuous atmosphere.

Observing that both entry and exit occur at the same defined altitude (and hence the same density, often assumed zero), the exit flight-path angle is simply

$$\gamma_{\text{exit}} = -\gamma_e \tag{6.33}$$

From Eq. (6.30), the exit velocity is then

$$V_{\text{exit}} = V_e \exp\left[\frac{2\gamma_e}{L/D}\right] \tag{6.34}$$

The acceleration along the trajectory is found to be[5]

$$a = \tfrac{1}{2}\rho V^2 \left[1 + \left(\frac{L}{D}\right)^2\right]^{1/2} \left(\frac{SC_D}{m}\right) \tag{6.35}$$

Maximum deceleration occurs at a small negative flight-path angle, i.e., just prior to pull-up. However, the value at pull-up ($\gamma = 0$) is nearly the same and is much more easily obtained; hence,

$$a_{\max} \simeq a_{\text{pullup}} = \left[1 + \frac{1}{(L/D)^2}\right]^{1/2} (1 - \cos \gamma_e)\beta V_e^2 \exp\left[\frac{2\gamma_e}{(L/D)}\right] \tag{6.36}$$

Taken together, Eqs. (6.34) and (6.36) imply the existence of an entry corridor, an acceptable range of flight-path angles, for supercircular skip entry. The lower (steep entry) bound on γ_e is determined for a given L/D by the acceptable deceleration load. For a manned vehicle, a reasonable maximum might be $12g$, the design limit for the Apollo missions. The upper (shallow entry) bound on γ_e for supercircular entry is determined by the requirement that the exit velocity be reduced to a sufficiently low level to allow the second phase of entry to occur within a reasonable time. The Apollo command module, for example, had battery power for only a few hours after the service module was jettisoned and could not tolerate a lengthy suborbital lob.

As an example, consider an Apollo-type entry with a vehicle L/D of 0.30, a lunar return speed of 11 km/s, and an atmospheric scale height of 7.1 km. Using the $12g$ maximum acceleration design limit selected for Apollo, the steepest allowed entry angle is found from Eq. (6.36) to be $-4.8°$. Assuming circular exit velocity to be the maximum acceptable (the vehicle will not go into orbit because the flight-path angle is nonzero at the exit interface, which is itself too low for a stable orbit), we find from Eq. (6.34) that the shallowest possible entry is $-2.9°$.

An indication of the accuracy and limitations of the first-order skip entry analysis presented here is obtained by noting that the Apollo 11 entry was initiated at a velocity and flight-path angle of 11 km/s and $-6.5°$, respectively. The $12g$ undershoot (steep entry) boundary was $-7°$ and the overshoot (shallow-angle) boundary was $-5°$. Figure 6.7 shows the predicted and actual L/D for the Apollo vehicle.[16]

Table 6.2 provides a summary of the landing accuracy demonstrated by the manned Apollo 7–17 missions.[17] Apollo 7 and 9 were Earth orbital missions only. Note that the actual entry guidance accuracy was better than indicated in Table 6.2, because of the effect of wind drift following parachute deployment. In a modern implementation, using steerable parachutes and global positioning system (GPS) guidance for entry and landing control, substantially better performance would be achieved.

The angular bounds on the entry corridor can be extrapolated backward along the entry hyperbola to yield the required B-plane targeting accuracy. This is done via the results of Section 4.2.4 (Motion in Hyperbolic Orbits).

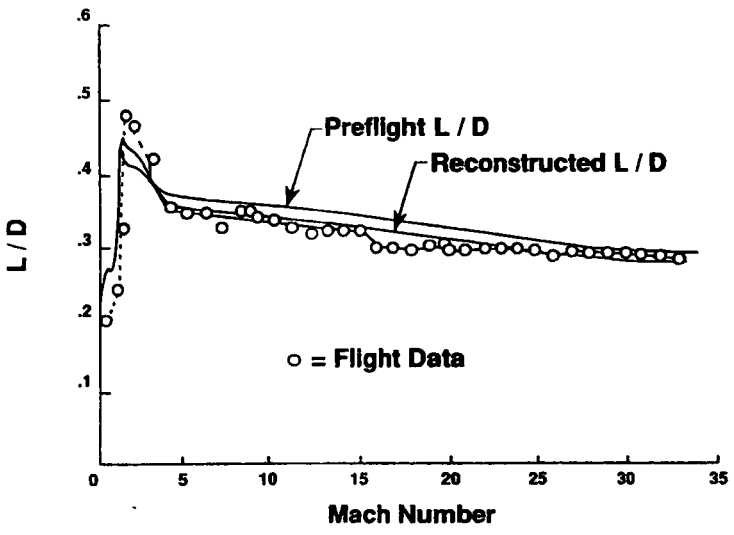


Fig. 6.7 Apollo spacecraft lift/drag ratio.

Table 6.2 Apollo landing accuracy

Mission	Distance from target, miles[a]
Apollo 7	1.9
Apollo 8	1.4
Apollo 9	2.7
Apollo 10	1.3
Apollo 11	1.7
Apollo 12	2.0
Apollo 13	1.0
Apollo 14	0.6
Apollo 15	1.0
Apollo 16	3.0
Apollo 17	1.0

[a]Best estimate based upon recovery ship positioning accuracy, command module computer data, and trajectory reconstruction.

The preceding discussion of entry corridor limits, though relevant, is oversimplified. In addition to errors introduced by the first-order model, other limitations must be considered. The total entry heating load is aggravated by an excessively shallow entry, whereas the heating rate (but not usually the total heat load) is increased by steepening the flight path. Either case may be prohibitive for a particular vehicle and may modify the entry corridor width determined solely from acceleration and exit velocity requirements.

The constant L/D assumption is an unnecessarily restrictive artifact of the analytical integration of the equations of motion. A more benign, and thus safer, entry can be obtained at an initially shallow angle with the lift vector negative, i.e., with the vehicle rolled on its back. Once the vehicle has been pulled into the atmosphere in this manner, it may be rolled over and flown with positive lift to effect the skip. This strategy was employed for the Apollo lunar return.[16]

As implied earlier, a skip entry sequence was possible with the Apollo command module and was initially selected as the nominal entry mode. Refinement of the entry guidance and targeting philosophy ultimately led to the use of a modulated-lift entry in which a full skip-out was avoided in favor of a trajectory that retained aerodynamic control throughout a nominal entry. However, the full skip phase was still available for trajectory control in the event of an off-nominal entry.[16] Because of the conservative aerothermodynamic design of the Apollo vehicle, heating loads were not a factor in entry corridor definition.

6.2.5 Cross-Range Maneuvers

Thus far we have assumed that the entry trajectory lies in the plane of the initial orbit. However, if a lifting entry vehicle is banked (lift vector rotated out of

the vertical plane defined by r, V), then a force normal to the original orbit plane is generated and the vehicle flies a three-dimensional trajectory. This may be done with both gliding and skip entry profiles as discussed earlier.

The dynamics of three-dimensional flight within the atmosphere are beyond the intended scope of this text, and we will not engage in a detailed study of lateral maneuvers. As with the planar trajectories discussed previously, however, first-order results are available[18] that yield considerable insight into the effect of banking maneuvers. Because cross-range control is often of interest even in the preliminary stages of entry vehicle and trajectory design, we will consider here some results of first-order three-dimensional entry analysis.

It is usually of interest to examine the maximum "footprint," or envelope of possible landing points, to which an entry vehicle can be steered. In-plane or down-range control for a lifting vehicle is attained through modulation of the lift-to-drag ratio. As seen in Eq. (6.26), maximum range is attained with flight at the highest available L/D. A landing at lesser range can be achieved by flying energy-dissipating maneuvers up and down in the entry plane, or back and forth across the initial plane. Cross-range maneuvers that do not cancel result in a lateral offset of the landing point, at some expense in downtrack range.

To effect a lateral maneuver, a lifting entry vehicle must bank to obtain a turning force normal to the initial plane and, upon attaining the desired heading change, reduce the bank angle again to zero. For maximum lateral range, the bank angle modulation must be performed in such a way that the downtrack range is not unduly reduced, or else the cross-range maneuver will not have time to achieve its full effect. There will thus be an optimum bank angle history that allows the maximum possible cross-range maneuver for a given down-range landing point.

For analytical purposes, the optimum, time-varying, bank angle history must be replaced by an equivalent constant value that provides similar results while allowing integration of the equations of motion. Although justified by the mean value theorem of integral calculus, this procedure renders invalid any consideration of the trajectory history, preserving only the maximum capability information. If in addition we assume an equilibrium glide with small changes in heading angle, first-order (Eggers[18]) and second-order (Vinh et al.[5]) results for maximum lateral range may be obtained. To first order, the angular cross range is

$$\phi = \left(\frac{\pi^2}{48}\right)\left(\frac{L}{D}\right)^2 \sin 2\sigma \tag{6.37}$$

where

σ = optimum constant bank angle
ϕ = "latitude" angle attained relative to great-circle "equatorial" plane of initial entry trajectory

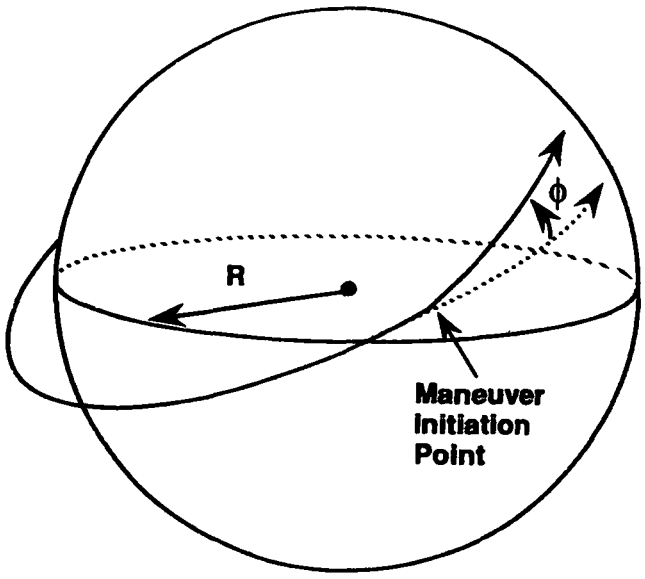


Fig. 6.8 Cross-range reentry geometry.

and, in terms of distance over the planetary surface, we have

$$s_\perp = R\phi \tag{6.38}$$

Figure 6.8 shows the geometry for a cross-range maneuver.

The use of $(L/D)_{max}$ in Eq. (6.37) implies, as with the planar equilibrium glide, that maximum cross range is achieved with flight at maximum L/D. Note that the Eggers solution yields $\sigma = 45°$ for the optimum constant bank angle. This is intuitively reasonable, because it implies that use of the vehicle lift vector is evenly divided between turning ($\sigma = 90°$) and staying in the air ($\sigma = 0°$) long enough to realize the result of the turn.

Vinh et al.[5] find that Eq. (6.37) overpredicts the cross-range travel that can be achieved with a given vehicle L/D. This is shown in Fig. 6.9, which compares the Eggers solution,[18] the second-order result of Vinh et al.,[5] and the cross range achieved with the true optimal bank angle history. It is seen that, for a vehicle L/D of 1.5 or less (small enough that even the maximum possible cross-range angle remains relatively small), the theories are in reasonable agreement.

As an example, consider the Gemini spacecraft with, as stated previously, an L/D of 0.2. Equations (6.37) and (6.38) yield a maximum cross-range travel of 52.4 km, or about 28.3 n mile. This is in excellent agreement with the actual Gemini vehicle footprint data, shown in Fig. 6.10.[8]

The shuttle, with its much higher L/D, exhibits correspondingly greater cross range capability. This allows the shuttle to land routinely at Edwards Air Force

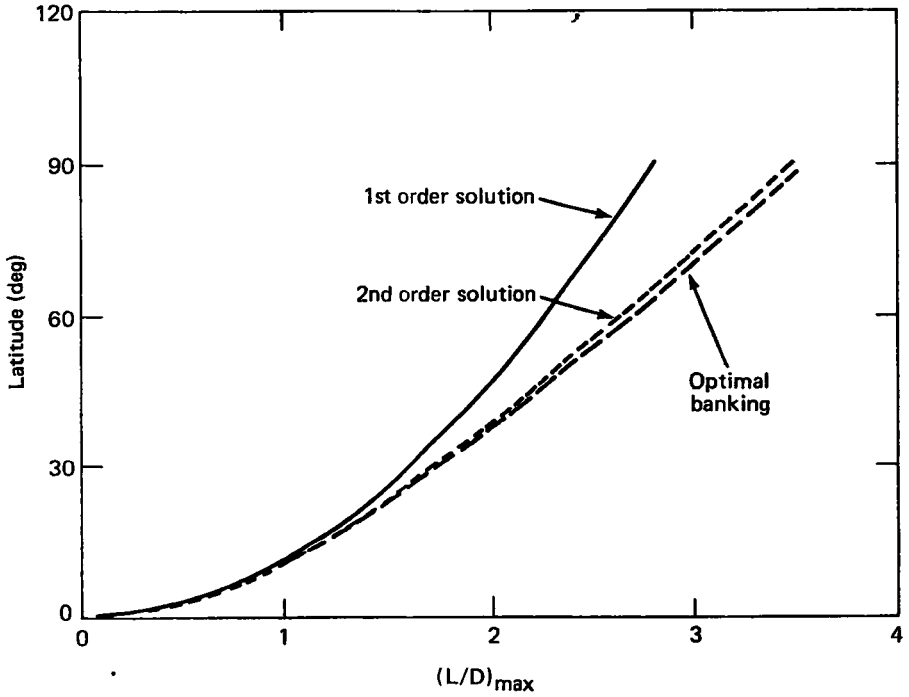


Fig. 6.9 Cross-range capability for varying lift-drag ratio.

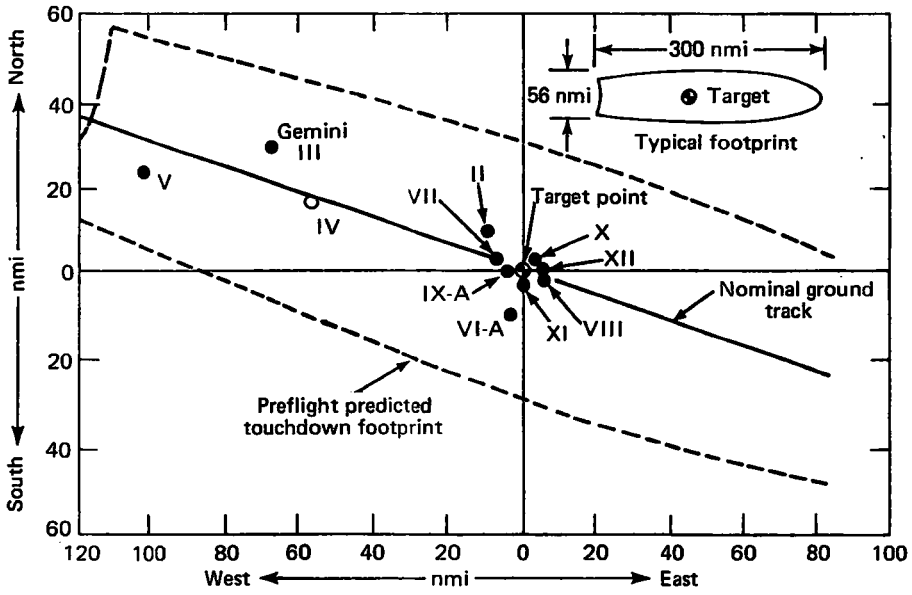


Fig. 6.10 Gemini landing footprint.

Base (34°N latitude) out of a 28.5° inclination orbit. This would be impossible for a low L/D vehicle. The maximum shuttle cross-range capability demonstrated to date has been 1465 km,[13] very close to the maximum value predicted by Eq. (6.38) for a vehicle with a hypersonic L/D of 1.1.

6.2.6 Loh's Second-Order Solution

The first-order atmospheric entry results presented earlier adequately demonstrate the various entry profiles that can be obtained for particular initial conditions and vehicle parameters. However, these solutions have two important limitations that preclude their use for anything more sophisticated than initial conceptual design.

The obvious restriction of the first-order solutions is their accuracy. All results derived thus far incorporate the assumptions of constant gravitational force, strictly exponential atmosphere and fixed reference altitude. These assumptions were discussed in Sec. 6.1 and contribute a total error on the order of 5%. An additional modeling error in the integrated results is the assumption of constant vehicle ballistic coefficient and L/D.

However, the more troublesome aspect of the first-order theory is the fact that one must know in advance what type of entry is under consideration. For example, the vehicle is assumed to be in either a steep ballistic or a shallow gliding trajectory. Shallow-angle ballistic entry and semiballistic entry at low but nonzero L/D do not readily fit into either category. A first-order theory that can address a wide range of entry interface conditions and vehicle parameters is not available.

As mentioned, second-order theories exist. We have also stated that in our view the implementation of higher order solutions to approximate equations is more troublesome and has less benefit than direct numerical integration of the basic governing equations. There is one possible exception to this policy that is worthy of consideration here, and that is Loh's empirically derived second-order theory.[19] We include a brief discussion of the theory more for its pedagogical value than for its practical utility.

Recall in Eq. (6.11b) that the term

$$\Delta \equiv \left(\frac{1}{\rho\beta R}\right)\left(\frac{gR}{V^2} - 1\right)\cos\gamma \tag{6.39}$$

represents the difference between the gravitational and centrifugal force normal to the entry flight path. The various first-order solutions assume either that this term is zero (ballistic, skip) or that the rate of change of the flight-path angle is small (equilibrium glide). Another possible assumption, intuitively reasonable but discovered by Loh through a comprehensive numerical investigation, is that Δ is a nonzero constant throughout entry. Modeling errors are then due to

departures from this value rather than to complete omission of the term, and are of second order.

With Loh's assumption of constant Δ, Eq. (6.11b) becomes

$$\frac{\mathrm{d}(\cos\gamma)}{\mathrm{d}\rho} \simeq \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) - \Delta \tag{6.40}$$

and integrates immediately to yield

$$\cos\gamma = \cos\gamma_e + \left[\left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) - \Delta\right](\rho - \rho_e) \tag{6.41}$$

To integrate the tangential equation, we introduce the additional assumption, used previously in the first-order analysis, that the tangential component of gravitational force is negligible in comparison with drag. This allows us to drop $(2/\rho\beta R)$ in Eq. (6.11a) and write

$$\sin\gamma \frac{\mathrm{d}(V^2/gR)}{\mathrm{d}\rho} \simeq \left(\frac{1}{\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{V^2}{gR}\right) \tag{6.42}$$

As discussed previously, this assumption is quite reasonable everywhere except near the entry interface, where drag is small and comparable to the tangential gravitational force. Indeed, it is found that the major area of difficulty with Loh's analysis lies in obtaining a match between the endo- and exoatmospheric trajectories at the entry interface.

From Eq. (6.40), note that

$$\Delta = -\frac{\mathrm{d}(\cos\gamma)}{\mathrm{d}\rho} + \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) = \sin\gamma\frac{\mathrm{d}\gamma}{\mathrm{d}\rho} + \left(\tfrac{1}{2}\beta\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) \tag{6.43}$$

or

$$\sin\gamma\frac{\mathrm{d}\gamma}{\mathrm{d}\rho} = \Delta - \left(\frac{1}{2\beta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{L}{D}\right) = \Delta = \text{const} \tag{6.44}$$

Dividing this result into Eq. (6.42) gives

$$\frac{\mathrm{d}(V^2/gR)}{\mathrm{d}\gamma} = \left(\frac{1}{\beta\Delta}\right)\left(\frac{SC_D}{m}\right)\left(\frac{V^2}{gR}\right) \tag{6.45}$$

which immediately integrates to

$$\ln\left[\frac{(V^2/gR)}{(V_e^2/gR)}\right] = \left(\frac{1}{\beta\Delta}\right)\left(\frac{SC_D}{m}\right)(\gamma - \gamma_e) \tag{6.46}$$

As written, Eqs. (6.41) and (6.46) are not materially different from Eqs. (6.28) and (6.30). Taking $\Delta \equiv 0$ reproduces the first-order solution for skip or non-equilibrium gliding entry. However, substituting the definition of Δ into Eqs.

(6.41), (6.43), and (6.46) allows the flight-path angle

$$\cos\gamma = \frac{\cos\gamma_e + (1/2\beta)(L/D)(SC_D/m)(\rho - \rho_e)}{1 + (1/\beta R)(gR/V^2 - 1)(1 - \rho_e/\rho)} \tag{6.47}$$

to be determined. The density can be found as a function of flight-path angle and velocity as

$$\rho - \rho_e = \left[\left(\frac{gR}{V^2} - 1\right)\left(\frac{m}{SC_D}\right)\left(\frac{\cos\gamma}{R}\right)\ln\left(\frac{V^2}{V_e^2}\right)\right] \div \left[(\gamma - \gamma_e) + \tfrac{1}{2}\left(\frac{L}{D}\right)\ln\left(\frac{V^2}{V_e^2}\right)\right] \tag{6.48}$$

Equations (6.47) and (6.48) constitute Loh's second-order solution for atmospheric entry. They are a coupled set of transcendental equations connecting the three variables V, γ, and ρ. Typically, we wish to specify the density (or altitude) and obtain velocity and flight-path angle as a result. This requires a numerical root-finding scheme that may easily be as complex as directly integrating the differential equations of motion. The solution is slightly simpler, computationally, if we are able to make the usual assumption that $\rho_e \simeq 0$ in Eq. (6.47).

Loh[19] shows excellent results with this theory over a wide range of entry flight conditions and vehicle parameters. Speyer and Womble[20] have verified this conclusion numerically during their investigation of three-dimensional trajectories. The authors perform an interesting variation of Loh's analysis in which Δ is explicitly included as a constant in the differential equations of motion, which are then integrated numerically. Speyer and Womble show that, by periodically updating Δ with recent trajectory values, even better results than claimed by Loh can be obtained.

6.3 Fundamentals of Entry Heating

Up to this point we have considered only the particle dynamics of atmospheric entry, wherein the vehicle is completely characterized by its L/D and ballistic coefficient. This determines the flight trajectory and allows assessment of the vehicle acceleration and dynamic pressure loads, the down-range and cross-range travel, and the sensitivity of these quantities to the entry conditions and vehicle parameters.

Of equal importance are the thermal loads imposed on the vehicle during entry. These are of two types: the total heat load and the instantaneous heating rate. The total heat load is obviously a concern in that the average vehicle temperature will increase with the energy input. The allowed heating rate, either local or body-averaged, is a concern because of the thermal gradient induced

from a heat flux according to Fourier's law:

$$q = -\kappa \nabla T \tag{6.49}$$

where

q = power per unit area, W/m^2
κ = thermal conductivity, W/mK
∇T = gradient of temperature, K/m

In materials with a nonzero coefficient of thermal expansion, a temperature gradient causes differential expansion and mechanical stress in the vehicle wall material.

Tradeoffs between allowed heating rate and total heat load are often necessary. Sustained high-energy flight at high altitude (e.g., gliding entry) reduces the instantaneous heating rate but, by extending the duration of the flight, may unacceptably increase the total heat absorbed. A more rapid, high-drag entry usually reduces the total energy input at the expense of incurring a very high local heating rate and may in addition result in unacceptable dynamic loads.

Entry vehicle heating results from the dissipation of the initial total (kinetic plus potential) energy through two heat transfer mechanisms, convection and radiation. Convective heating occurs when the air, heated by passage through a strong bow shock in front of the vehicle, bathes the wall in a hot fluid stream. If the air is hot enough, significant thermal radiation will occur as well. Radiative heat transfer is important when the entry velocity is greater than about 10 km/s and may be significant at considerably lower speeds.

Peak aerodynamic heating will usually occur in stagnation point regions, such as on a blunt nose or wing leading edge. However, turbulent flow along the vehicle afterbody can under some conditions produce a comparable or greater heat flux. Conversely, delayed onset of turbulence (i.e., turbulent transition at a higher than expected Reynolds number) can produce a substantially cooler aft body flow than expected.

Thermal control is a major entry vehicle design challenge. As Regan[4] notes, the specific kinetic energy that is dissipated during entry from low Earth orbit is about 3×10^7 J/kg. This is sufficient to vaporize a heat shield composed of pure carbon ($h_v = 6 \times 10^7$ J/kg) and equal to half the initial vehicle mass. If this is to be prevented, then the major portion of the entry kinetic and potential energy must be deflected to the atmosphere rather than the vehicle. A good aerothermodynamic design will allow only a few percent of this energy to reach the vehicle.

6.3.1 Thermal Protection Techniques

The design and analysis of an entry vehicle and flight profile to meet the thermal protection requirement is a multidiscipline task involving aerodynamics,

chemistry, flight mechanics, structural analysis, and materials science. Three basic approaches to entry vehicle thermal control have evolved: heat sinking, radiative cooling, and ablative shielding.

The heat sink technique, as the name implies, uses a large mass of material with a high melting point and high heat capacity to absorb the entry heat load. The initial Mercury spacecraft design utilized this approach, employing a beryllium blunt body heat shield. This design was used on the unmanned tests and on the first manned Mercury-Redstone suborbital flights. However, the increased system weight for protection against the order-of-magnitude higher orbital entry heat load forced the use of an ablative shield on the subsequent orbital missions. The second manned suborbital mission tested the ablative heat shield. This weight penalty is a typical and important limitation of the heat sink approach to entry thermal control.

The principle of radiative cooling is to allow the outer skin of the vehicle to become, literally, red hot due to the convectively transferred heat from the flowfield around the vehicle. Blackbody radiation, primarily in the infrared portion of the spectrum, then transports energy from the vehicle to the surrounding atmosphere. Convective heating to the vehicle is proportional to the temperature difference between the fluid and the wall, whereas the energy radiated away is in proportion to the difference in the fourth powers of the fluid and wall temperatures. The net result is that thermal equilibrium can be reached at a relatively modest skin temperature provided that the rate of heating is kept low enough to maintain near-equilibrium conditions.

Radiative cooling obviously requires excellent insulation between the intensely hot outer shell and the internal vehicle payload and structure. This is exactly the purpose of the shuttle tiles, the main element of the shuttle thermal protection system. Essentially a porous matrix of silica (quartz) fibers, these tiles have such low thermal conductivity that they can literally be held in the hand on one side and heated with a blowtorch on the other.

As stated, radiative cooling relies on equilibrium, or near-equilibrium, between the entry vehicle and its surroundings to shed the absorbed heat load. This is most easily achieved in a lengthy, high-altitude gliding entry where the instantaneous heating rate is minimized as the speed is slowly reduced. The vehicle aerodynamic design (ballistic coefficient and L/D), the entry flight trajectory, and the heat shield material selection are intimately related when radiative cooling is used. This complicates the design problem; however, significant mass savings are possible when a system-level approach is taken.[21]

A potential problem with an insulated, radiatively cooled vehicle having a lengthy flight time is that ultimately some heat will soak through to the underlying structure. Coolant fluid may thus need to be circulated through the vehicle so that this energy can be radiated away to a portion of the surroundings, such as the aft region, which is cooler. This can occur even if the flight time is sufficiently short that in-flight cooling is not required. Such is the case with the

shuttle, which must be connected to cooling lines from ground support equipment if postflight damage to the aluminum structure is to be prevented.

Although heat sinking is best suited to a brief, high-drag entry and radiative cooling is more appropriate for a gliding trajectory, ablative cooling offers considerably more flexibility in the flight profile definition. Ablative cooling is also typically the least massive approach to entry heat protection. These advantages accrue at the expense of vehicle (or at least heat shield) reusability, which is a pronounced benefit of the other techniques.

Ablative cooling occurs when the heat shield material, commonly a fiberglass-resin matrix, sublimes under the entry heat load. When the sublimed material is swept away in the flowfield, the vehicle is cooled. This process can produce well over 10^7 J/kg of effective energy removal. Ablative cooling has been the method of choice for most entry vehicles, including the manned Mercury, Gemini, and Apollo vehicles.

6.3.2 Entry Heating Analysis

From the theory of viscous fluid flow[22] it is known that the flowfield about an atmospheric entry vehicle develops a thin boundary layer close to the body to which viscous effects, including skin friction and heat transfer, are confined. The heat flux to the wall is proportional to the local temperature gradient,

$$q_w = \kappa\left(\frac{\partial T}{\partial y}\right)_w = \varepsilon\sigma T_w^4 \tag{6.50}$$

where y is the coordinate normal to the wall. The temperature gradient is obtained from the boundary-layer flowfield solution, determined from the boundary-layer edge properties and wall conditions. The edge conditions in turn follow from the inviscid solution for the flow over the entry vehicle. The vehicle heat transfer analysis is thus dependent on knowledge of the flowfield.

The right-hand equality in Eq. (6.50) implies that, in the steady state, iteration between the convective and radiative heat flux equations will be necessary to fix the equilibrium wall temperature.

The difficulty of obtaining an accurate solution for the high-speed flowfield around an entry vehicle can hardly be overstated. The fluid is a chemically reacting gas, possibly not in equilibrium, probably ionized, and with potentially significant radiative energy transfer. Vehicle surface properties such as roughness and wall catalycity influence the flowfield and heat transfer analysis.

We note in passing that this partially ionized flowfield is the "plasma sheath" that interferes with air-to-ground communication during major portions of entry flight. This was the cause of the "communications blackout" familiar to readers who recall the early Mercury, Gemini, and Apollo manned flights. The advent of the Tracking Data Relay Satellite System (TDRSS) constellation (see Chapter

11) has at least somewhat alleviated this problem; antennas on the leeward side of the space shuttle can generally complete the link to a TDRSS satellite. However, the roll maneuvers necessary for shuttle landing point control can, and periodically do, disrupt communications during the entry phase.

The entry flight regime is equally demanding of an experimental approach. It is at present impossible to conduct a wind-tunnel experiment that simultaneously provides both Mach and Reynolds numbers appropriate to entry flight. Thus, the space shuttle received the first true test of its performance during its first flight, a potentially hazardous situation, because, unlike its predecessors, the shuttle was not flight tested in an unmanned configuration.

A recurrent theme in this text is that recourse to all available analytical sophistication is desirable, even essential, prior to critical design and development. However, preliminary design and mission feasibility assessment would be virtually impossible without the use of simpler, less accurate, techniques. Accordingly, we rely on an approach to entry heating analysis first given by Allen and Eggers.[10] This approach assumes the primary source of energy input to be convective heating from the laminar boundary-layer flow over the entry vehicle. In this case, the local heating rate as given by Eq. (6.50) may be correlated with the total enthalpy difference across the boundary layer:[22]

$$q_w = \kappa\left(\frac{\partial T}{\partial y}\right)_w = \left(\frac{\kappa}{C_p}\right)\left(\frac{Nu_L}{L}\right)(H_{oe} - H_w) = \left(\frac{Nu_L}{Pr}\right)\left(\frac{\mu}{L}\right)H_{oe}\left(1 - \frac{H_w}{H_{oe}}\right) \quad (6.51)$$

where

Nu_L = Nusselt number
$Pr = \mu C_p/\kappa =$ Prandtl number
κ = thermal conductivity
μ = fluid viscosity
C_p = fluid heat capacity at constant pressure
$H = V^2/2 + C_pT$ = total enthalpy
V = freestream velocity

The Nusselt number Nu_L, is a parameter based on both the fluid properties and on the particular flow situation. The reader should consult Chapter 9 for a further discussion of the role of this parameter in heat transfer analysis. The subscript L implies that the Nusselt number is based on an appropriate length scale L for the particular type of boundary-layer flow in question. The choice of length scale obviously varies with the nature of the flow geometry; as we shall see, it will often be a characteristic parameter such as the nose or wing leading edge radius.

The Prandtl number is a fluid property, ranging from 0.71 to 0.73 for air below 9000 K, but is often taken as unity for approximate calculations. The notation for fluid viscosity μ is almost universal; we trust that, in the context of the present analysis, it will not be confused with the gravitational parameter $\mu = GM$ used

elsewhere. The subscripts 'oe' and 'w' denote local flow conditions at the outer edge of the boundary layer and at the wall, respectively.

Total enthalpy is conserved for the inviscid flow across the normal shock portion of the bow wave. Because it is this stream that wets the body, we have

$$H_{oe} = H = h + \frac{V^2}{2} = C_pT + \frac{V^2}{2} \cong \frac{V^2}{2} \tag{6.52}$$

The unsubscripted parameters denote, as usual, freestream or approach conditions. The right-hand approximate equality follows from the high-speed, low-temperature nature of the upstream flow. For example, assume an entry vehicle at 80-km altitude with $T = 200$ K, $V = 6000$ m/s, and $C_p = 1005$ J/kg·K. Then $V^2/2C_pT = 90$, and the thermal energy content of the air provides a negligible contribution to the total enthalpy.

Multiplying and dividing Eq. (6.51) by $(\rho V)_{oe}$ yields an equivalent result,

$$q_w = (\rho V)_{oe}\left(\frac{Nu_L}{PrRe_L}\right)H_{oe}\left(1 - \frac{H_w}{H_{oe}}\right) = (\rho V)_{oe}StH_{oe}\left(1 - \frac{H_w}{H_{oe}}\right) \tag{6.53}$$

where

$St = Nu/PrRe$ = Stanton number
$Re = \rho VL/\mu$ = Reynolds number

The derivation of Eq. (6.53) makes it clear that the Reynolds number Re is referenced to the same (as yet unspecified) length scale as the Nusselt number and to boundary-layer edge values of density and velocity. This result offers no apparent simplification; however, exploiting the Reynolds analogy for laminar boundary-layer flow,[22] we note that

$$St \cong \frac{C_f}{2} \tag{6.54}$$

where C_f is the local skin friction coefficient. This approximation is typically valid to within about 20%. For example, White[22] shows that the Reynolds analogy factor $2St/C_f$ varies between 1.24 and 1.27 over the subsonic to Mach 16 range for laminar flow over a flat plate.

With the Reynolds analogy and Eq. (6.52), Eq. (6.53) becomes

$$q_w = \tfrac{1}{4}(\rho V)_{oe}V^2\left(1 - \frac{H_w}{H_{oe}}\right)C_f \tag{6.55}$$

Equation (6.55) shows that the heating rate to the body depends on the local wall temperature through the term $(1 - H_w/H_{oe})$. Because the flow is stagnant at the wall, $H_w \cong C_pT_w$, with the equality exact if T_w is low enough (below about 600 K) that the gas may be assumed calorically perfect. It is a conservative assumption, consistent with other approximations adopted here, to assume the

wall to be sufficiently cool that H_w/H_{oe} is small. The heating rate is then

$$q_w \cong (\rho V)_{oe} V^2 C_f/4 \tag{6.56}$$

and we see that for a reasonably cool wall, the gross heat-transfer rate is independent of the body temperature.

Integration over the body wall area S_w gives the total heating rate (power input) to the body,

$$Q = \rho V^3 S_w C_F/4 \tag{6.57}$$

where C_F is the body-averaged skin friction coefficient defined as

$$C_F = \left(\frac{1}{S_w}\right)\int C_f\left[\frac{(\rho V)_{oe}}{\rho V}\right] ds \tag{6.58}$$

Again, the subscript "oe" denotes local boundary-layer outer edge values. As usual, the upstream or approach velocity V is found as a function of density ρ from trajectory solutions such as those obtained in Sec. 6.2. Skin friction coefficient calculations are discussed in more detail in a subsequent section.

6.3.3 Total Entry Heat Load

The total heat load (energy) into the vehicle can be obtained from Eq. (6.57),

$$\frac{dE}{dV} = \left(\frac{dE}{dt}\right)\left(\frac{dt}{dV}\right) = Q\frac{dt}{dV} = 2Q\left(\frac{m}{SC_D}\right)\left(\frac{1}{\rho V^2}\right) = \tfrac{1}{2}\left(\frac{m}{SC_D}\right) S_w C_F V \tag{6.59}$$

where as usual we have dropped the tangential gravitational force in Eq. (6.1a). Upon integrating from the entry velocity to the final velocity,

$$E = \tfrac{1}{4} m(V_e^2 - V_f^2)\left(\frac{S_w C_F}{SC_D}\right) \tag{6.60}$$

If, as is usually the case, the final velocity is effectively zero, the total heat load has the particularly simple form

$$\frac{E}{(\frac{1}{2}mV_e^2)} = \tfrac{1}{2}\frac{S_w C_F}{SC_D} \tag{6.61}$$

Equation (6.61) is valid with any entry profile (ballistic, glide, or skip) and for any vehicle sufficiently "light" that it slows before hitting the ground. This is the same requirement as for a deceleration peak with ballistic entry, and Eq. (6.19) may therefore be used to define a "light" vehicle. A dense vehicle on a steep ballistic trajectory may fail to meet this criterion. If this is the case, the first-order ballistic entry velocity profile given by Eq. (6.15) will be quite accurate and

may be substituted in Eq. (6.60) to yield

$$\frac{E}{(\frac{1}{2}mV_e^2)} = \frac{1}{2}\left(\frac{S_w C_F}{SC_D}\right)\left\{1 - \exp\left[\left(\frac{\rho_s}{\beta \sin \gamma_e}\right)\left(\frac{SC_D}{m}\right)\right]\right\} \tag{6.62}$$

By failing the light vehicle criterion of Eq. (6.19), the exponent has magnitude less than unity, and the expansion $e^{\alpha} \simeq 1 + \alpha$ may be employed inside the brackets to yield the "heavy body" result,

$$\frac{E}{\frac{1}{2}mV_e^2} \simeq -\frac{1}{2m}\frac{S_w C_F \rho_s}{\beta \sin \gamma_e} \tag{6.63}$$

Equation (6.61) provides the rationale for the classical blunt body entry vehicle design. The total heat load is minimized when the skin friction drag C_F is small compared to the total drag C_D, and the wetted area S_w is as small as possible in comparison with the reference projected area S. Both of these conditions are met with an entry vehicle having a rounded or blunt shape.

Equation (6.63) shows that a dense ballistic entry vehicle should have a slender profile to minimize the total skin friction and hence the heat load.

6.3.4 Entry Heating Rate

The body-averaged heating rate is also of interest and is found from Eq. (6.57):

$$q_{\text{avg}} \equiv \frac{Q}{S_w} = \frac{1}{4}\rho V^3 C_F \tag{6.64}$$

The average heating rate can be found once a trajectory profile giving velocity as a function of atmospheric density (hence altitude) is specified. This can be done as an adjunct to a numerical solution, or by substituting the previously obtained first-order results for ballistic, equilibrium glide, and skip trajectories. Using this latter approach, we obtain

$$q_{\text{avg}} = \frac{1}{4}C_F \rho V_e^3 \exp\left[\left(\frac{3}{2}\right)\left(\frac{SC_D}{m}\right)\left(\frac{\rho}{\beta \sin \gamma_e}\right)\right] \tag{6.65}$$

for the ballistic entry heating rate as a function of density. Similarly,

$$q_{\text{avg}} = \frac{1}{2}\left(\frac{m}{SC_D}\right)\left(1 - \frac{V^2}{gR}\right) g C_F \frac{V}{L/D} \tag{6.66}$$

gives the heating rate for gliding entry vs velocity. Finally,

$$q_{\text{avg}} = \frac{1}{2}\left(\frac{m}{SC_D}\right)\beta V_e^3 \exp\left[-\frac{3(\gamma - \gamma_e)}{L/D}\right]\frac{(\cos \gamma - \cos \gamma_e)}{L/D} \tag{6.67}$$

is the heating rate for skip entry as a function of flight-path angle.

It is usually of greatest interest to find the value of the maximum body-averaged heating rate, as well as the altitude (or density) and velocity at which this rate occurs. This maximum heating rate will often constrain the entry trajectory. For ballistic entry, the maximum heating rate and critical trajectory conditions are

$$q_{\text{avg}_{\max}} = -\left(\frac{C_F}{6e}\right)\left(\frac{m}{SC_D}\right)\beta V_e^3 \sin \gamma_e \tag{6.68}$$

$$\rho_{\text{crit}} = -\left(\frac{2}{3}\right)\left(\frac{m}{SC_D}\right)\beta \sin \gamma_e \tag{6.69}$$

$$V_{\text{crit}} = \frac{V_e}{e^{1/3}} \tag{6.70}$$

For equilibrium gliding entry, we find

$$q_{\text{avg}_{\max}} = \left(\frac{g^3 R}{27}\right)^{\frac{1}{2}} \frac{m/SC_D}{L/D} \tag{6.71}$$

$$\rho_{\text{crit}} = \left(\frac{4}{R}\right)\frac{m/SC_D}{L/D} \tag{6.72}$$

$$V_{\text{crit}} = \left(\frac{gR}{3}\right)^{\frac{1}{2}} \tag{6.73}$$

The corresponding parameters are slightly more difficult to obtain for skip entry. To obtain explicit algebraic results, it is necessary to assume small γ_e^5. This assumption is nearly always satisfied, and the results are

$$q_{\text{avg}_{\max}} \cong \left(\frac{\beta}{4}\right)\left(\frac{1}{L/D}\right)\left(\frac{m}{SC_D}\right)\gamma_e^2 V_e^3 \exp\left[\frac{3\gamma_e}{L/D}\right] \tag{6.74}$$

$$\rho_{\text{crit}} \simeq 2\beta\gamma_e \frac{(m/SC_D)}{L/D} \tag{6.75}$$

$$V_{\text{crit}} \simeq V_e \exp\left[\frac{\gamma_e}{L/D}\right] \tag{6.76}$$

$$\gamma_{\text{crit}} \simeq -\frac{3\gamma_e^2}{L/D} \tag{6.77}$$

We urge caution in the application of the results given here. The heat load calculations of this section implicitly incorporate the approximations in the trajectory solutions for ballistic, gliding, or skip entry. For example, we have seen that the first-order ballistic entry analysis underpredicts the acceleration load for shallow entry angles. Because this result is incorporated in Eqs. (6.60) and (6.65), it is expected that at shallow entry angles the ballistic entry heating rate would be underpredicted and the total heat load overpredicted. This is the case, as shown in Fig. 6.11, which compares the first-order theory with that of Chapman.[3]

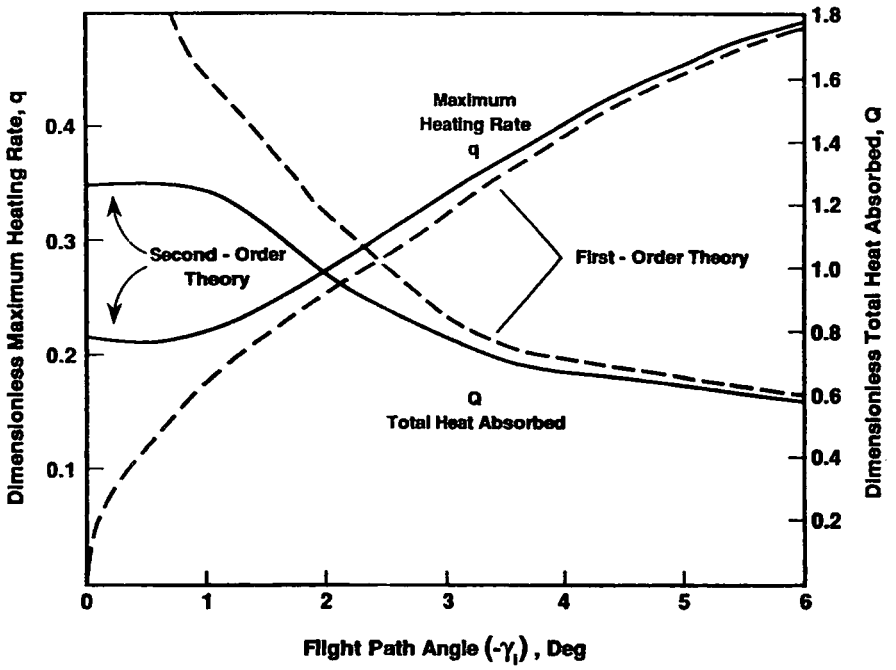


Fig. 6.11 Comparison of first- and second-order entry heating analysis.

Moreover, the entry heating analysis is itself approximate, because it assumes laminar boundary-layer heating, invokes the Reynolds analogy to eliminate the Stanton number, ignores radiant energy input, and neglects vibrational and chemical excitation ("real gas effects") in the gas. These assumptions are quite valid at low speeds, below about 2 km/s, but become progressively less so as typical atmospheric entry speeds are approached. There are some mitigating effects; for example, neglect of radiant heating partially offsets the calorically perfect gas assumption. High-altitude entry flight, with its attendant low Reynolds number, tends to favor laminar flow, particularly for short, blunt vehicles.

For gliding entry vehicles, the situation may be different. While at high altitude, it is likely that the vehicle will encounter laminar flow. As atmospheric density increases, however, it is to be expected that at some point the boundary layer will transition to turbulent flow, with attendant higher drag and, by Reynolds analogy, a higher heating rate. This is obviously a situation to be avoided for as long as possible, leading to the requirement that the wetted surface of a gliding entry vehicle be as smooth and regular as possible, thus avoiding any premature "tripping" of the boundary layer into turbulent flow.

We must point out that all too frequently even the most sophisticated calculations yield poor accuracy. Prabhu and Tannehill[23] compared shuttle flight

data with theoretical heat-transfer results using a state-of-the-art flowfield code together with both equilibrium air and calorically perfect gas models. It was found that, provided a proper value of k (the ratio of specific heats) is chosen (the authors recommend $k = 1.2$), the calorically perfect gas model does as well as the equilibrium air model. In some cases substantially better agreement with flight data was obtained with the simpler model!

Other space shuttle flight experience further illustrates the points discussed earlier. For STS missions 1–5, Williams and Curry[24] show generally excellent agreement between preflight analysis and flight data, particularly in the higher-temperature regions. Heating in the cooler, leeside areas (where the flowfield is typically quite complex) was significantly lower than preflight predictions, even those based on wind-tunnel data. Throckmorton and Zoby[25] attributed this to delayed onset of turbulent flow as compared with subscale wind-tunnel test results.

Subsequent data obtained over the course of over a hundred flights of the space shuttle has revealed some quite complex behavior. The space shuttle typically experiences transition to turbulent boundary-layer flow on the wetted underside at about Mach 8, but has encountered it prior to Mach 11 on approximately 20% of flights, and in one case (STS-73) as early as Mach 19. Early onset of turbulent transition has been attributed to excessive surface roughness, and in particular to partially dislodged "gap filler" material, placed between the shuttle thermal protection tiles to impede hot gas flow in these interstices. The gap fillers can apparently loosen in flight and intrude into the boundary-layer flow, causing early transition and, in some cases, unexpected damage to the thermal protection system.

Of those flights on which early transition to turbulent flow has been observed, some 60% have demonstrated asymmetric transition, i.e., one wing goes turbulent while the other remains laminar, resulting in a significant differential drag and imposing a lateral moment on the vehicle.[13]

This has significant flight-control implications for shuttle as well as for future hypersonic entry vehicle designs. It is necessary to ensure that the combination of reaction control thrusters and aerodynamic surfaces is capable of exerting sufficient control authority, for a sufficient length of time, to overcome the lateral disturbing moment until the other wing also transitions to turbulent flow. It is, of course, also necessary to ensure that the overall heating rate under such adverse conditions remains within the thermal protection system design limits.

Finally, shuttle heat-transfer flight experience has varied with time. Scott[26] discusses the effects of wall catalysis on orbiter heat transfer and notes that the heat flux has increased from flight to flight as the shuttle tile properties change with age and use.

Cumulative uncertainties as to model validity argue for due caution in interpreting the results of all heat-transfer analysis. We regard entry heating analysis as presented here to be an order-of-magnitude theory, useful in

preliminary design but unsuited for detailed work. Even detailed calculations are not generally regarded as accurate to better than 10%.

6.3.5 Skin Friction Coefficient

The body-averaged skin friction coefficient C_F is seen to be a key parameter in determining both the heating rate and the total heat load for an entry vehicle. As shown by Eq. (6.58), C_F is determined by integration of the local skin friction coefficient C_f over the body. C_f is defined by

$$C_f = \frac{2\tau_w}{(\rho V)_{oe}} \tag{6.78}$$

where

$$\tau_w = \mu\left(\frac{\partial V}{\partial \gamma}\right)_w \tag{6.79}$$

is the boundary-layer shear stress at the wall.

Clearly, the boundary-layer flowfield solution must be known to evaluate the wall shear stress and skin friction coefficient. Because the skin friction coefficient was introduced to avoid precisely this difficulty, further approximation is required. To this end, we include some results from laminar boundary-layer theory, which, when used with judgment, allow estimation of C_F for preliminary vehicle design.

From low-speed boundary-layer theory we have the classical result for incompressible laminar flow over a flat plate[27] that

$$C_f = 0.664/Re_x^{1/2} \tag{6.80}$$

where Re_x is the Reynolds number referenced to boundary-layer edge conditions and to the x or streamwise coordinate as measured from the leading edge of the plate:

$$Re_x = \frac{(\rho V)_{oe} x}{\mu} \tag{6.81}$$

The streamlines that wet the outer edge of the boundary layer obey the steady flow continuity result

$$(\rho V)_{oe} = \rho V \tag{6.82}$$

Combining Eqs. (6.82) and (6.58) and integrating over a plate of unit width and length L yields the low-speed result

$$C_F = 1.328/Re_L^{1/2} \tag{6.83}$$

Flat-plate theory is useful in aerodynamics because most portions of a flight vehicle are of a scale such that the local body radius of curvature dwarfs the boundary-layer thickness. Thus, most of the body appears locally as a flat plate, and good approximate results for skin friction can be obtained by ignoring those portions, small by definition, which do not. This assumption can be invalid for flight at very high altitude, where the reduced density lowers the Reynolds number and produces a thicker boundary layer.

Equation (6.83) can be extended to high-speed, hence compressible, flow through the reference-temperature approach.[22] It is found that, in the worst case (adiabatic wall), $C_F/\sqrt{(Re_L)}$ varies from 1.328 at low speed to approximately 0.65 at Mach 20. Compressibility thus has an important but not overwhelming effect on skin friction coefficient and, for entry heating calculations such as presented here, may with some justification be ignored or included in an ad hoc fashion. In any case, the use of the low-speed value is conservative from an entry heating viewpoint.

6.3.6 Stagnation Point Heating

Both the total heat load and the body-averaged heating rate are important in entry analysis, because either may constrain the trajectory. Their relative importance will depend on the entry profile and vehicle parameters, and, as we have mentioned, relief from one is usually obtained by aggravating the other.

Of equal importance is the maximum local heating rate imposed on any part of the entry vehicle, which determines the most severe local thermal protection requirement. With the possible exception of local afterbody hot spots due to turbulent effects and shock–boundary-layer interactions, the body heating rate is maximized at the stagnation point. Any realistic vehicle design will have a blunt nose or wing leading edge, and this will be a region of stagnation flow, shown schematically in Fig. 6.12.

The stagnation region behind a strong normal shock is one of particularly intense heating. For example, at an entry speed of Mach 25 the perfect gas shock tables[28] yield $T_{t2}/T_1 = 126$, where T_1 is the freestream static temperature and T_{t2} is the stagnation temperature behind the shock. Assuming $T_1 = 166$ K for the standard atmosphere at 80 km, the total temperature behind the shock is 20,900 K! For comparison, the surface temperature of the sun is approximately 5780 K.

Such extreme temperatures are of course not attained. The previous calculation assumes the atmosphere to be a calorically perfect gas for which the enthalpy and temperature are related by

$$h = C_p T \tag{6.84}$$

where the heat capacity C_p is a constant, 1005 J/kg·K for air. In fact, a major fraction of the available thermal energy is used to dissociate and ionize the air

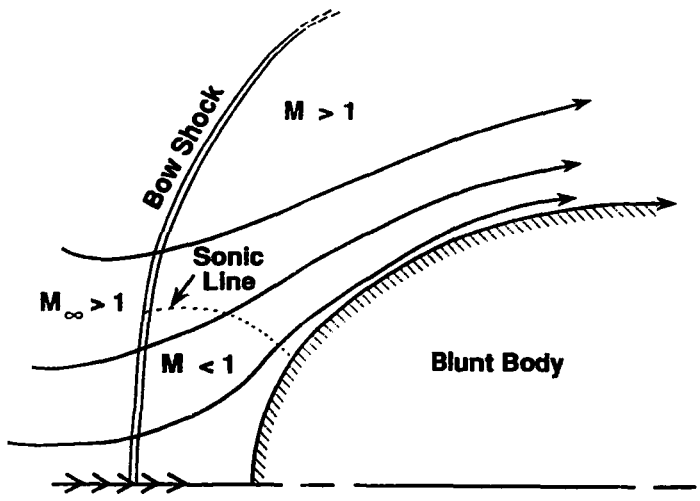


Fig. 6.12 Stagnation point flow.

molecules, effectively increasing the heat capacity of the gas and lowering the stagnation temperature. For a shuttle entry, the nose cap stagnation region reaches a peak temperature of approximately 1650 K (2500°F).[24]

The preceding example is interesting and informative regarding attempts to predict individual flowfield properties for high-speed and therefore high-energy flows. A cautionary note should be added, however. The wall heat flux q_w is the parameter of importance in entry vehicle design and is driven by the total enthalpy difference $(H_{oe}-H_w)$ between the wall and the outer edge of the boundary layer. The temperature difference is not the relevant parameter, despite what Eq. (6.50) would imply. For a calorically perfect gas, where Eq. (6.84) applies, no distinction between temperature and enthalpy need be made. In a chemically reacting gas, dissociation and ionization will alter the balance between effective heat capacity and temperature and thus significantly affect the flowfield. However, the net effect on the boundary-layer flowfield total enthalpy difference $(H_{oe}-H_w)$ and hence the wall heat flux may be small.

The implication is that the neglect of real gas effects, although horrifying to a physical gas dynamicist, may be fairly reasonable for our purposes. This is especially true when chemical equilibrium exists in the boundary-layer flowfield, an approximation that is reasonable in the stagnation region. Also, the assumption of a non-equilibrium boundary-layer flow with a fully catalytic wall, so that surface equilibrium exists by definition, yields similar results.

Our approximate analysis of stagnation heating relies again on Eq. (6.51), which we restate here:

$$q_w = (Nu_L/Pr)(\mu/L)H_{oe}(1 - H_w/H_{oe}) \tag{6.85}$$

Previously we rearranged this equation to employ the Stanton number instead

of the Nusselt number, then used Reynolds' analogy to cast the results in terms of the skin friction coefficient. This was done because skin friction data are more easily obtained and generalized, if only empirically, than are heat transfer data when the complete flowfield solution is not available. However, the boundary-layer flow in the stagnation region shown in Fig. 6.12 is sufficiently well understood that a more direct approach is possible.

In the low-speed stagnation region behind a strong bow shock, incompressible flow theory applies. For such a flow over a rounded nose or wing leading edge, the Nusselt number is found to be[29]

$$Nu_L = \eta Pr^{\frac{2}{5}} \left(\frac{K\rho}{\mu}\right)^{\frac{1}{2}} L \tag{6.86}$$

where K is the stagnation point velocity gradient in the x, or streamwise, direction at the edge of the boundary layer:

$$K = \left(\frac{dV_{oe}}{dx}\right)_{sp} \tag{6.87}$$

and the subscript "sp" denotes stagnation point conditions. For axisymmetric flow, $\eta = 0.763$,[29] whereas for two-dimensional flow, such as over a wing leading edge, $\eta = 0.570$.[22] Employing Eq. (6.52), Eq. (6.85) now becomes

$$q_w = \tfrac{1}{2}\eta Pr^{-0.6}(\rho_{oe}\,\mu_{oe})_{sp}^{\frac{1}{2}}(1 - H_w/H_{oe})V^2(dV_{oe}/dx)_{sp}^{\frac{1}{2}} \tag{6.88}$$

The stagnation point velocity gradient $(dV_{oe}/dx)_{sp}$ is evaluated for high-speed flow by combining the Newtonian wall pressure distribution with the boundary-layer momentum equation and the inviscid flow solution at the stagnation point. This yields[22]

$$K = \left(\frac{dV_{oe}}{dx}\right)_{sp} = \left(\frac{V}{R_n}\right)\left(\frac{2\rho}{\rho_{oe}}\right)^{\frac{1}{2}} \tag{6.89}$$

where R_n is the nose radius of curvature. The term (ρ/ρ_{oe}) is the density ratio for the inviscid flow across a normal shock at upstream Mach number M:

$$\rho/\rho_{oe} = [(k-1)M^2 + 2]/(k+1)M^2 \tag{6.90}$$

This ratio varies from unity at Mach 1 to $(k-1)/(k+1)$ at infinite Mach number.

Equation (6.89) provides the well-known result that stagnation point heating varies inversely with the square root of the nose radius. This does not imply that a flat nose eliminates stagnation point heating; the various approximations employed invalidate the model in this limiting case. It remains true, however, that stagnation point heating scales with leading edge radius of curvature as given earlier.

Equation (6.88) is a perfect gas result and omits the effects of vibrational and chemical excitation. The landmark analysis of stagnation point heating including these effects was given by Fay and Riddell[30] and later extended by Hoshizaki[31] and by Fay and Kemp[32] to include the effects of ionization. Experimental work in support of these theories includes that of Rose and Stark[33] and Kemp et al.[34] We summarize here the important conclusions from this work.

Fay and Riddell[30] found the stagnation point heat flux for a nonradiating "binary gas" consisting of atoms (either O or N) and molecules (N_2 or O_2) to be

$$q_w = \tfrac{1}{2}\eta Pr^{-0.6}(\rho_{oe}\,\mu_{oe})_{sp}^{0.4}(\rho_w\mu_w)_{sp}^{0.1}(1 - H_w/H_{oe})V^2(dV_{oe}/dx)_{sp}^{\frac{1}{2}} \times [1 + (Le^{\varepsilon} - 1)h_d/H_{oe}] \tag{6.91}$$

where

$\varepsilon = 0.52$ for equilibrium boundary-layer flow
$= 0.63$ for frozen flow with fully catalytic wall
$= -\infty$ for frozen flow with noncatalytic wall
Le = Lewis number = 1.4 for air below 9000 K
$h_d = \sum c_i(\Delta h_f^\circ)_i$ = average dissociation energy
$c_i = i$th species concentration
$(\Delta h_f^\circ)_i = i$th species heat of formation[35]

The Fay and Riddell analysis, which agrees quite well with experimental data for typical Earth orbital speeds, modifies Eq. (6.88) by the factor

$$D = [1 + (Le^{\varepsilon} - 1)h_d/H_{oe}](\rho_w\mu_w/\rho_{oe}\mu_{oe})_{sp}^{0.1} \tag{6.92}$$

which is due to dissociation. Kemp and Riddell[36] show this factor to increase the stagnation heat flux by about 20% over the calorically perfect gas result for entry from low Earth orbit. [This quite reasonably tempts the engineer seeking a preliminary result simply to use Eq. (6.88), and then to increase the result by 20% to obtain a conservative answer; see also Eq. (6.93).]

A few comments on the use of Eq. (6.91) are in order. The equation is evaluated in the forward direction, i.e., the wall temperature is specified and the heat flux computed. If q_w rather than T_w is known, the wall temperature must be found by iteration.

The freestream density ρ and velocity V are known from the trajectory solution. Specification of density fixes, through the standard atmosphere model, freestream pressure and temperature. Given the wall temperature (which may well be specified as an upper bound for design), and state relations for the gas comprising the chemically reacting boundary layer, the quantities P_w, μ_w, h_d, and H_w may be computed. The gas properties may be determined from first principles[35] or, with somewhat less effort, found in tables.[37,38] Reasonably

accurate empirical relationships such as Sutherland's viscosity law[22] are also useful.

Although hand-calculator evaluation of Eq. (6.91) is feasible, it is somewhat tedious and therefore to be avoided when possible. Kemp and Riddell[36] used the Fay and Riddell result to correlate stagnation point heating for entry from Earth orbit as a function of freestream density and velocity, obtaining

$$q_w = 20{,}800\,\text{Btu/ft}^2/s\,(\rho/\rho_s)^{\frac{1}{2}}(V/V_{\text{circ}})^{3.25}(1 - H_w/H_{\text{oe}})(1\,\text{ft}/R_n)^{\frac{1}{2}} \quad (6.93a)$$

or

$$q_w = 1.304 \times 10^8\,\text{W/m}^2(\rho/\rho_s)^{\frac{1}{2}}(V/V_{\text{circ}})^{3.25}(1 - H_w/H_{\text{oe}})(1\,\text{m}/R_n)^{\frac{1}{2}} \quad (6.93b)$$

in SI units. Orbital velocity $V_{\text{circ}} = 26{,}000\,\text{ft/s} = 7.924\,\text{km/s}$, and surface density $\rho_s = 0.002378\,\text{slug/ft}^3 = 1.225\,\text{kg/m}^3$ were assumed; the correlation is claimed accurate to within 5%. Note that the cooled wall assumption, $H_w = 0$, gives a conservative result.

6.3.7 Free Molecular Heating

Thus far we have discussed only continuum flow results; stagnation heating in rarefied flow may be important when considering satellites that orbit, or at least have periapsis, at very low altitudes. Free molecular heating is also relevant during launch vehicle ascent flight; indeed, it is usually this constraint that determines the lowest altitude, typically around 100 km, at which the payload shroud can be jettisoned.

The free molecular heating rate will be of the form

$$q_w = \alpha\sigma\rho V^3 \quad (6.94)$$

where α is an unknown constant, ρ is the atmospheric density and σ is an accommodation coefficient, upper bounded by unity but more commonly in the range 0.6–0.8, and which accounts for the energy transfer efficiency of the impacting atmosphere particles into the vehicle.

Kemp and Riddell[36] correlated numerous experimental results for stagnation point heating in the free molecular flow regime, yielding

$$q_w = 2.69 \times 10^7\,\text{Btu/ft}^2/\text{s}\,\sigma(\rho/\rho_s)(V/V_{\text{circ}})^3 \quad (6.95)$$

With the constants combined, the free molecular stagnation point heating rate at low Earth orbital speeds becomes

$$q_w = \tfrac{1}{2}\sigma\rho V^3 \quad (6.96)$$

6.4 Entry Vehicle Designs

In previous sections we have seen that the key entry vehicle parameters are the ballistic coefficient C_B, the lift-drag ratio L/D, and the body radius of curvature at the nose or wing leading edge. The topic of entry vehicle aeroshell design to achieve suitable combinations of these parameters is, in detail, somewhat beyond the scope of this book. Consequently, our discussion in this area will be of a qualitative nature only.

Figure 6.13 shows vehicle L/D vs C_B for a range of typical entry vehicle aerodynamic designs.[39]

The greatest amount of flight experience has been accumulated with the simplest entry vehicle designs and flight profiles, i.e., ballistic or semiballistic capsules, blunted cones, etc. The flight characteristics of such vehicles are relatively well understood, a consequence resulting in part from their somewhat limited flexibility in mission design. The subject of ballistic and semiballistic (i.e., low L/D) entry vehicle design and flight experience has already been discussed and needs no further treatment here.

Experience with hypersonic winged or lifting-body vehicles has been more restricted. Over an 11-year period ending in 1968, the X-15 manned research rocket plane carried out 199 flights, reaching a maximum speed of Mach 6.7 and a

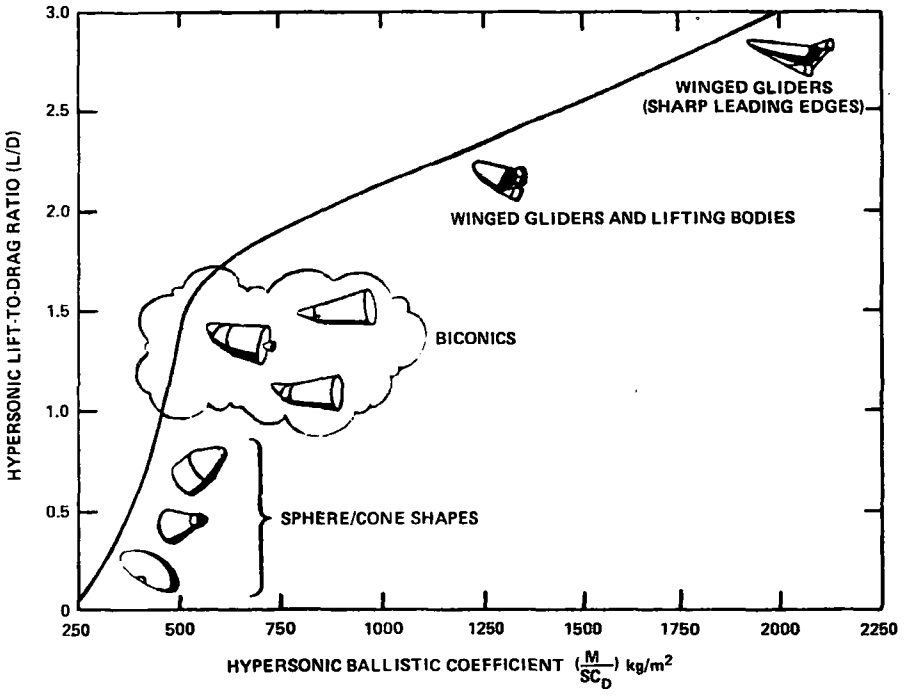


Fig. 6.13 Entry vehicle shapes.

maximum altitude (not on the same flight) of 108 km. The program explored many now well-understood, but then unknown, aspects of hypersonic flight. Among the many key X-15 contributions are the discovery that hypersonic boundary layers tend to be turbulent rather than laminar, and the first demonstration of lifting reentry techniques.[40] However, because entry heating rates are proportional to the cube of the vehicle velocity, it is clear that the X-15 was able to explore only a small fraction of the overall atmospheric entry flight envelope.

Many lifting-body designs have been flown subsonically, in crucial demonstrations of low-speed handling characteristics essential for approach and landing. Still, the highest speed achieved by any piloted lifting-body vehicle to date is Mach 1.86, and the highest altitude 27.5 km, both by the HL-10 at NASA's Dryden Flight Research Center[40] in February 1970. Numerous high-speed, high-altitude subscale lifting-body tests have been conducted by both the United States and Russia; however, many of the results of these tests are of restricted availability.

As of this writing, with well over 100 missions having been flown, the space shuttle has accumulated by far the greatest wealth of modern, openly available hypersonic performance data, albeit over a limited range of vehicle and flight profile parameters. As has been noted earlier, flight performance has been close to theoretical predictions, with the exceptions generally associated with the difficult-to-predict transition from laminar to turbulent flow along the aft body,

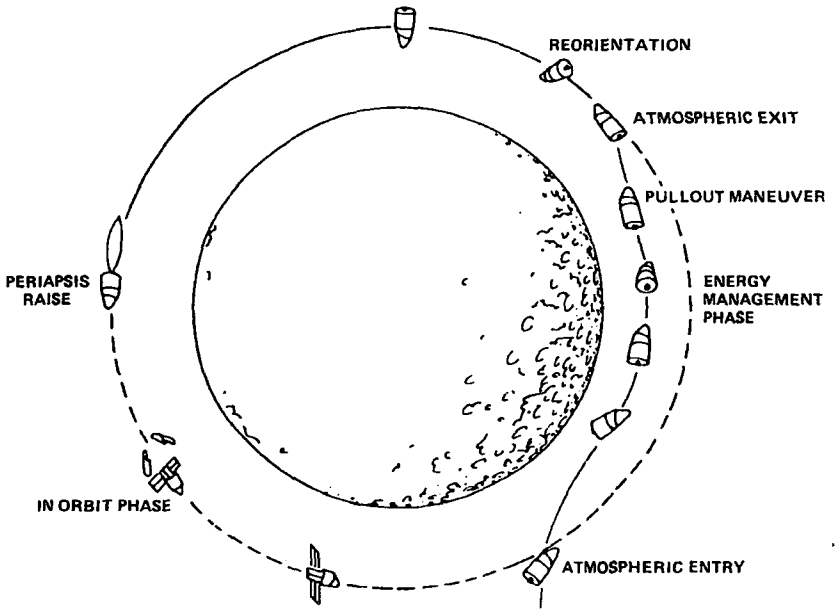


Fig. 6.14 Aerocapture flight plan.

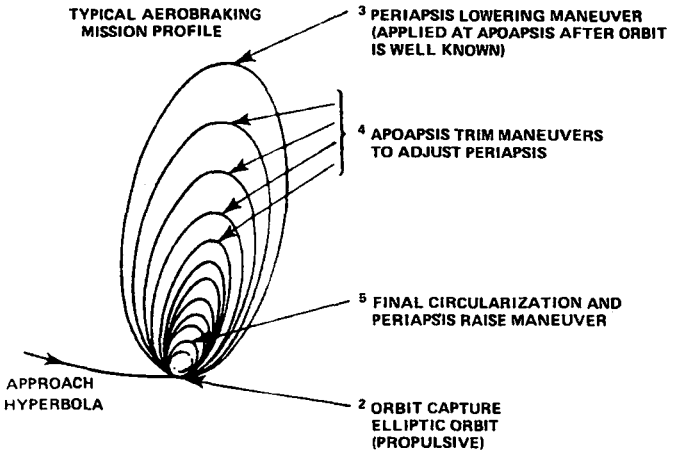
and sometimes even on the fully-wetted underside. As discussed earlier, this latter phenomenon has been linked in general to the surface roughness of the surface tiles, and in particular to the unique nature of the shuttle's tile-and-gap-filler thermal protection system.

6.5 Aeroassisted Orbit Transfer

A technique of great promise and extensive current interest for advanced space operations is that of aeroassisted orbit transfer. Many analyses[41] have demonstrated that propulsive requirements for both interplanetary and orbital operations can be significantly reduced with maneuvers that utilize the atmosphere of a nearby planet for braking or plane change ΔV.

Figure 6.14 shows this concept as applied to aerocapture of an interplanetary spacecraft into a low orbit at a target planet. The concept is also applicable for transfer from high orbit to low orbit around a given planet.

As discussed earlier, aerocapture is a technique requiring a fairly sophisticated, high L/D aeroshell design that imposes significant configuration



TYPICAL AEROBRAKING PARAMETERS

PLANET	VENUS	EARTH	MARS
VEHICLE MASS, kg	1630	1630	1630
DRAG BRAKE DIAMETER, m	9	9	9
INITIAL ORBIT PERIOD, hr	12	12	12
PERIAPSIS ALTITUDE, km	132	86	123
INITIAL APOAPSIS ALTITUDE, km	37,218	18,247	40,033
PERIAPSIS VELOCITY, m/s	9594	4604	10,363
SHIELD TEMP, k	664	571	592
FINAL PERIAPSIS ALTITUDE, km	691	556	692
NUMBER OF ORBITS	244	53	344
ELAPSED TIME, DAYS	44	10	61
ΔV TO CIRCULARIZE, m/s (at 300 km)	205	142	225

Fig. 6.15 Aerobraking scenario.

constraints on the internal payload. In return, it offers the maximum flexibility in the entry flight trajectory design and control. Where the entry requirements are less severe, lower L/D or even ballistic designs may be suitable and usually lead to more advantageous packaging arrangements. Lower L/D generally demands a higher level of approach guidance accuracy than the more capable high L/D designs, which can compensate with atmospheric maneuvers for relatively coarse entry accuracy. Aeroassisted orbit transfer with low or zero L/D is commonly denoted "aerobraking." An application to low-orbit planetary capture is shown in Fig. 6.15. Again, the same scenario can be employed for transfer from high orbit to low orbit about a planet. The technique has been demonstrated at Venus by the Pioneer Venus Orbiter and the Venus Orbiter Imaging Radar spacecraft, and at Mars by several Mars orbiting spacecraft.

References

[1]U.S. Standard Atmosphere, National Oceanic and Atmospheric Administration, NOAA S/T 76- 1562, 1976.

[2]Fleming, E. L., Chandra, S., Schoeberl, M. R., and Barnett, J. J., "Monthly Mean Global Climatology of Temperature, Wind, Geopotential Height, and Pressure for 0–120 km," NASA TM-100697, 1988.

[3]Chapman, D. R., "An Approximate Analytical Method for Studying Entry into Planetary Atmospheres," NACA TN-4276, 1958.

[4]Regan, F. J., *Re-Entry Vehicle Dynamics*, AIAA Education Series, AIAA, New York, 1984.

[5]Vinh, N. X., Busemann, A., and Culp, R. D., *Hypersonic and Planetary Entry Flight Mechanics*, Univ. of Michigan Press, Ann Arbor, MI, 1980.

[6]Ashley, H., *Engineering Analysis of Flight Vehicles*, Addison-Wesley, Reading, MA, 1974.

[7]Platus, D. H., "Ballistic Re-Entry Vehicle Flight Dynamics," *Journal of Guidance, Control, and Dynamics*, Vol. 5, Jan.–Feb. 1982, pp. 4–16.

[8]"Guidance and Navigation for Entry Vehicles," NASA SP-8015, Nov. 1968.

[9]Mayer, R. T., "MOSES (Manned Orbital Space Escape System)," *Journal of Spacecraft and Rockets*, Vol. 20, March–April 1983, pp. 158–163.

[10]Allen, H. J., and Eggers, A. J., "A Study of the Motion and Aerodynamic Heating of Missiles Entering the Earth's Atmosphere at High Supersonic Speeds," NACA TR-1381, 1958.

[11]Harpold, J. C., and Graves, C. A., Jr., "Shuttle Entry Guidance," *Journal of the Astronautical Sciences*, Vol. 27, July–Sept. 1979, pp. 239–268.

[12]Harpold, J. C., and Gavert, D. E., "Space Shuttle Entry Guidance Performance Results," *Journal of Guidance, Control, and Dynamics*, Vol. 6, Nov.–Dec. 1983, pp. 442–447.

[13]Hale, N. W., Lamotte, N. O., and Garner, T. W., "Operational Experience with Hypersonic Flight of the Space Shuttle," AIAA Paper AIAA-2002-5259, Oct. 2002.

[14]Romere, P. O., and Young, J. C., "Space Shuttle Entry Longitudinal Aerodynamic Comparisons of Flight 2 with Preflight Predictions," *Journal of Spacecraft and Rockets*, Vol. 20, Nov.–Dec. 1983, pp. 518–523.

[15]Eggers, A. J., Allen, H. J., and Neice, S. E., "A Comparative Analysis of the Performance of Long-Range Hypervelocity Vehicles," NACA TN-4046, 1957.

[16]Graves, C. A., and Harpold, J. C., "Re-Entry Targeting Philosophy and Flight Results from Apollo 10 and 11," AIAA Paper 70-28, Jan. 1970.

[17]"Apollo Program Summary Report," NASA TM-X-68725, April 1975.

[18]Eggers, A. J., "The Possibility of a Safe Landing," *Space Technology*, edited by H. S. Seifert, Wiley, New York, 1959, Chap. 13.

[19]Loh, W. H. T., "Entry Mechanics," *Re-Entry and Planetary Entry Physics and Technology*, edited by W. H. T. Loh, Springer-Verlag, Berlin, 1968.

[20]Speyer, J. L., and Womble, M. E., "Approximate Optimal Atmospheric Entry Trajectories," *Journal of Spacecraft and Rockets*, Vol. 8, Nov. 1971, pp. 1120–1125.

[21]Wurster, K. E., "Lifting Entry Vehicle Mass Reduction Through Integrated Thermostructural/Trajectory Design," *Journal of Spacecraft and Rockets*, Vol. 20, Nov.–Dec. 1983, pp. 589–596.

[22]White, F. M., *Viscous Fluid Flow*, McGraw-Hill, New York, 1974.

[23]Prabhu, D. K., and Tannehill, J. C., "Numerical Solution of Space Shuttle Orbiter Flowfields Including Real-Gas Effects," *Journal of Spacecraft and Rockets*, Vol. 23, May-June 1986, pp. 264-272.

[24]Williams, S. D., and Curry, D. M., "Assessing the Orbiter Thermal Environment Using Flight Data," *Journal of Spacecraft and Rockets*, Vol. 21, Nov.–Dec. 1984, pp. 534–541.

[25]Throckmorton, D. A., and Zoby, E. V., "Orbiter Entry Leeside Heat Transfer Data Analysis," *Journal of Spacecraft and Rockets*, Vol. 20, Nov.–Dec. 1983, pp. 524–530.

[26]Scott, C. D., "Effects of Nonequilibrium and Wall Catalysis on Shuttle Heat Transfer," *Journal of Spacecraft and Rockets*, Vol. 22, Sept.–Oct. 1985, pp. 489–499.

[27]Liepmann, H. W., and Roshko, A., *Elements of Gasdynamics*, Wiley, New York, 1957.

[28]Ames Research Staff, "Equations, Tables, and Charts for Compressible Flow," NACA Rept. 1135, 1953.

[29]Sibulkin, M., "Heat Transfer Near the Forward Stagnation Point of a Body of Revolution," *Journal of the Aeronautical Sciences*, Vol. 19, Aug. 1952, pp. 570–571.

[30]Fay, J. A., and Riddell, F. R., "Theory of Stagnation Point Heat Transfer in Dissociated Air," *Journal of the Aeronautical Sciences*, Vol. 25, Feb. 1958, pp. 73–85.

[31]Hoshizaki, H., "Heat Transfer in Planetary Atmospheres at Super-Satellite Speeds," *ARS Journal*, Oct. 1962, pp. 1544–1552.

[32]Fay, J. A., and Kemp, N. H., "Theory of Stagnation Point Heat Transfer in a Partially Ionized Diatomic Gas," *AIAA Journal*, Vol. 1, Dec. 1963, pp. 2741–2751.

[33]Rose, P. H., and Stark, W. I., "Stagnation Point Heat Transfer Measurements in Dissociated Air," *Journal of the Aeronautical Sciences*, Vol. 25, Feb. 1958, pp. 86–97.

[34]Kemp, N. H., Rose, P. H., and Detra, R. W., "Laminar Heat Transfer Around Blunt Bodies in Dissociated Air," *Journal of the Aerospace Sciences*, July 1959, pp. 421–430.

[35]Anderson, J. D., Jr., *Modern Compressible Flow*, McGraw-Hill, New York, 1982.

[36]Kemp, N. H., and Riddell, F. R., "Heat Transfer to Satellite Vehicles Reentering the Atmosphere," *Jet Propulsion*, 1957, pp. 132–147.

[37]Hilsenrath, J., and Klein, M., "Tables of Thermodynamic Properties of Air in Chemical Equilibrium Including Second Virial Corrections from 1500 K to 15,000 K," Arnold Engineering Development Center, Rept. AEDC-TR-65-68, 1965.

[38] Stull, D. R., *JANAF Thermochemical Tables*, National Bureau of Standards, NSRDS-NBS 37, 1971.

[39] Cruz, M. I., "The Aerocapture Vehicle Mission Design Concept—Aerodynamically Controlled Capture of Payload into Mars Orbit," AIAA Paper 79-0893, May 1979.

[40] Hallion, R. P., *On the Frontier: Flight Research at Dryden, 1946–1981*, NASA History Series, NASA SP-4303, 1984.

[41] Walberg, G. D., "A Survey of Aeroassisted Orbit Transfer," *Journal of Spacecraft and Rockets*, Vol. 22, Jan.–Feb. 1985, pp. 3–18.

Bibliography

Cohen, C. B., and Reshotko, E., "Similar Solutions for the Compressible Laminar Boundary Layer with Heat Transfer and Pressure Gradient," NACA TN-3325, 1955.

Florence, D. E., "Aerothermodynamic Design Feasibility of a Mars Aerocapture Vehicle," *Journal of Spacecraft and Rockets*, Vol. 22, Jan.–Feb. 1985, pp. 74–79.

Hansen, C. F., "Approximations for the Thermodynamic and Transport Properties of High Temperature Air," NASA TR-R-50, 1959.

Miller, C. G., Gnoffo, P. A., and Wilder, S. E., "Measured and Predicted Heating Distributions for Biconics at Mach 10," *Journal of Spacecraft and Rockets*, Vol. 23, May–June 1986, pp. 251–258.

Vinh, N. X., Johannesen, J. R., Mease, K. D., and Hanson, J. M., "Explicit Guidance of Drag-Modulated Aeroassisted Transfer Between Elliptical Orbits," *Journal of Guidance, Control, and Dynamics*, Vol. 9, May–June 1986, pp. 274–280.

Problems

All Earth-referenced problems in this section may be solved assuming:

$R_E = 6378$ km
$h_e = 122$ km (entry interface altitude)
$r_e = R_E + h_e = 6500$ km
$g = 9.8\ \text{m/s}^2$
$\beta = 0.1354\ \text{km}^{-1}$
$\rho_s = 1.225\ \text{kg/m}^3$

6.1 The Mercury spacecraft was designed to perform a ballistic reentry with initial conditions

$m = 1350$ kg $\qquad V_e = 7.5$ km/s (air relative)
$C_D = 1.5 \qquad S = 2.8\ \text{m}^2$

(a) If the maximum design reentry deceleration was to be $8g$, what was the desired entry flight-path angle for the Mercury spacecraft?
(b) At what height did maximum deceleration occur?

6.2 Vehicle parameters for Apollo, a lifting entry vehicle, were

$$V_e = 11.2\,\text{km/s (air relative)} \qquad m = 5600\,\text{kg}$$
$$\frac{L}{D} = 0.30\,\text{(hypersonic)} \qquad S = 12.0\,\text{m}^2$$

(a) For lunar return missions, Apollo performed a mild version of the skip entry. A most important factor in this type of entry is ensuring that, upon completion of the skip, spacecraft velocity is low enough, and at a low enough angle, to prevent returning the vehicle to a long, high orbit prior to the second entry. Choose a reasonable velocity constraint, and find the entry angle constraint that results.

(b) It is equally important not to dig too deeply into the atmosphere on the first pass, as unacceptable g loads will result. Assuming maximum loads at the pull-up point of the skip entry (not exactly true, but close), find the flight-path angle constraint that guarantees an acceptable g load.

6.3 The shuttle orbiter performs a reentry in which the angle of attack, and consequently the L/D, can vary considerably depending on the particular entry requirements for a given mission. However, let us assume that a basic entry profile can be approximated by the use of an average or typical value of L/D of 1.05, which corresponds to $\alpha = 40°$. This turns out to be reasonable for flight, above Mach 12, which comprises most of the entry flight regime (AIAA JSR, Jan.–Feb. 1983). Let us define completion of the high-speed phase of entry as occurring at 25-km altitude and 750 m/s velocity (roughly Mach 2). Assuming that $\gamma_e = -1.2°$,

(a) What is the approximate range from entry interface until terminal phase initiation?

(b) What is the approximate time of flight for (a)?

(c) What is the approximate shuttle orbiter cross-range capability using the average L/D?

6.4 In the mid-1960s some tentative consideration was given to a lunar flyby mission using an uprated (i.e., more heat shielding) Gemini spacecraft. Of particular concern in this mission concept was the Earth entry phase. Assume a lunar return air-relative entry speed of $V_e = 11.2$ km/s, with Gemini vehicle parameters

$$S = 4.1\,\text{m}^2 \qquad \frac{L}{D} = 0.19$$
$$m = 2200\,\text{kg} \qquad C_D = 1.5$$

(a) What would have been the entry-angle bounds for this mission, assuming an acceptable peak g load of $12g$ and a desired maximum skip-out velocity of 7 km/s?
(b) What would have been the pull-up altitude in the steep entry angle case?

6.5 Assume that the Gemini spacecraft reentry from Earth orbit could be reasonably modeled as a shallow angle equilibrium glide. With $V_e = 7.5$ km/s and $\gamma_e = -2°$,
(a) What was the spacecraft velocity at the point where forward progress had ceased (i.e., flight path angle $\gamma = -90°$, or vertical descent)?
(b) What was the total range from entry interface to landing?
(c) What was the maximum cross-range capability?

6.6 Gemini was able to perform a ballistic entry by executing a continuous roll throughout the entry phase; this was in fact done on Gemini 5 when spacecraft guidance failed. Assuming a ballistic entry,
(a) What was the maximum entry acceleration?
(b) At what altitude did maximum acceleration occur?

6.7 A spacecraft is injected into a Hohmann transfer trajectory to Mars (see Chapter 4), at which point its mass is $m = 2000$ kg. Neglecting midcourse corrections and other concerns relating to guidance and navigation, it is determined that periapsis radius on the approach hyperbola to Mars will be 3970 km (590 km altitude). You may assume the periapsis velocity of the spacecraft on hyperbolic approach to be $V_p = 5.348$ km/s.
(a) Assuming it is desired to inject propulsively into a circular Mars orbit at the 590-km altitude, what ΔV is required?
(b) Assuming an I_{sp} of 300 s for the injection rocket, what mass of fuel is required?
(c) Assume now that the orbital injection of part (a) is to be accomplished via a combined maneuver, with an atmospheric braking phase (initially at the hyperbolic approach velocity) followed by a propulsive maneuver to reach the desired 590-km altitude circular orbit. The vehicle must exit the Martian atmosphere at or below escape velocity, and the allowable lower bound on altitude during the atmospheric entry phase is 30 km (to avoid the possibility of hitting mountains). What is the entry corridor at Mars in terms of allowable entry flight-path angles? The relevant vehicle and planetary parameters are

$$\rho_s = 0.011\,\text{kg/m}^3 \qquad \beta^{-1} = 10\,\text{km}$$
$$h_e = 100\,\text{km} \qquad R_M = 3380\,\text{km}$$

$$\frac{L}{D} = 1.0 \qquad C_D = 2.0$$
$$S = 10\,\text{m}^2$$

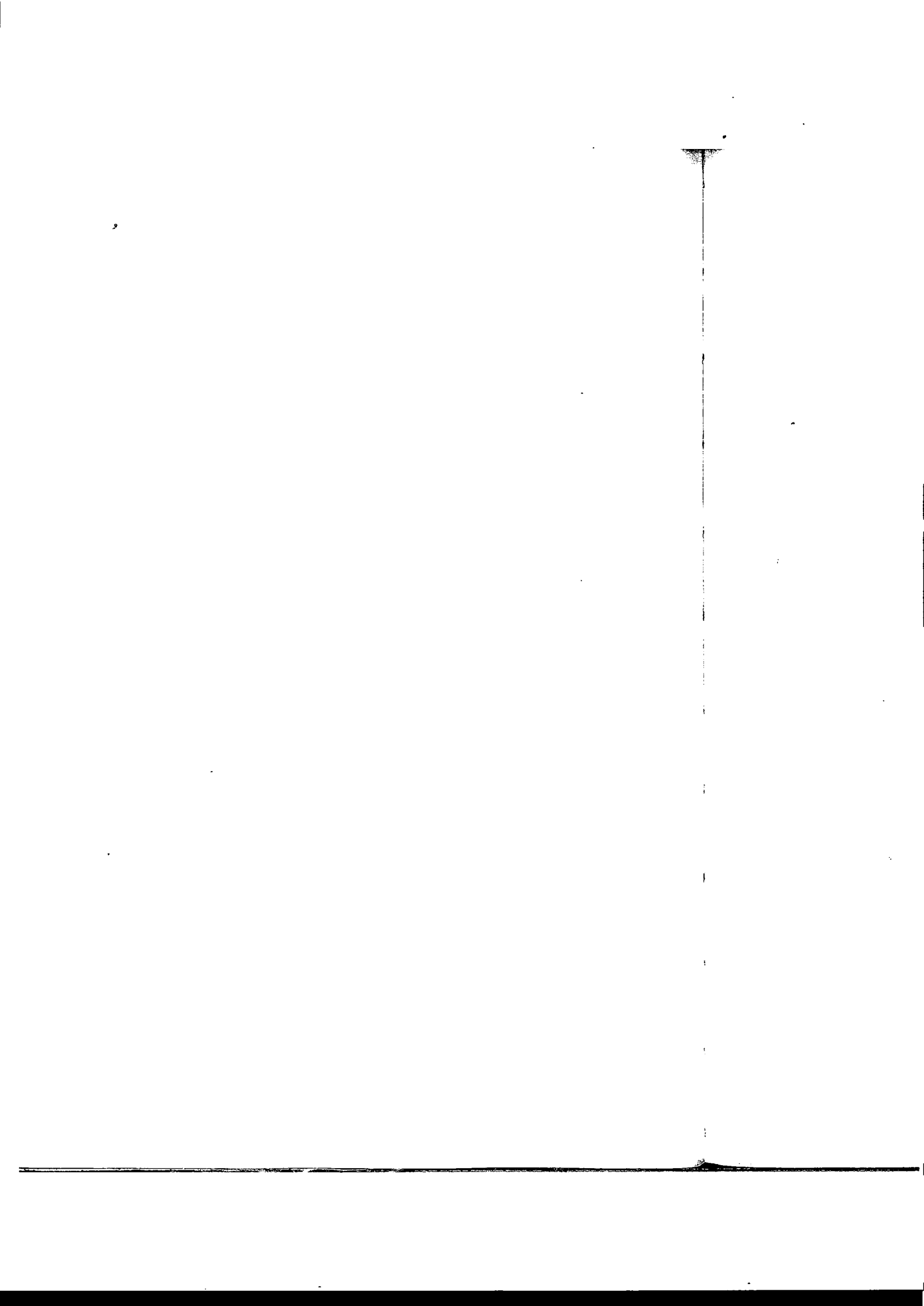
(d) If we choose an entry flight-path angle for this skip maneuver of $\gamma = -10°$, what mass of fuel is now required for the orbit injection maneuver, assuming the same I_{sp} as in part (b)?

6.8 It is desired to have a simple, fail-safe crew emergency return vehicle for use on a space station. The program manager edicts that, to minimize cost and complications, a basic ballistic entry vehicle will be used. However, the entry g load is edicted by medical authorities to be $8g$ or less. Can you do it? What is the maximum absolute value of entry flight-path angle?

6.9 A proposed design for Apollo-like lunar return vehicle featuring a reusable aerobrake with $L/D = 0.5$ reenters the atmosphere at 11 km/s, reducing speed and exiting the atmosphere at lower velocity. It coasts to apogee, where its trajectory is then circularized into a low parking orbit. Assume the desired exit velocity is 7.9 km/s and that the entry/exit interface altitude is 122 km. What should the entry flight-path angle be? Assume $\beta^{-1} = 7$ km and $SC_D/m = 0.0175\ \text{m}^2/\text{kg}$.

6.10 In problem 6.9 what is the minimum altitude the vehicle reaches, assuming constant L/D? Whether or not you solved Problem 6.9, assume for this problem $\gamma_0 = -5°$, $\beta^{-1} = 7$ km, and $SC_D/m = 0.0175\ \text{m}^2/\text{kg}$. Use the simple exponential atmosphere model, with surface density $1.225\ \text{kg/m}^3$.

6.11 What is the maximum acceleration experienced along the trajectory of problem 6.9? Again, whether or not you solved problem 6.9, assume $\gamma_0 = -5°$, $\beta^{-1} = 7$ km, and $SC_D/m = 0.0175\ \text{m}^2/\text{kg}$.



7
Attitude Determination and Control

7.1 Introduction

In this chapter we discuss what is often considered to be the most complex and least intuitive of the space vehicle design disciplines, that of attitude determination and control. The authors agree with this assessment, but would add that the more complex aspects of the subject are of primarily theoretical interest, having limited connection with practical spacecraft design and performance analysis. Exceptions exist, of course, and will be discussed here because of their instructional value. However, we believe that the most significant features of attitude determination and control system (ADCS) design can be understood in terms of rigid body rotational mechanics modified by the effects of flexibility and internal energy dissipation. At this level, the subject is quite accessible at the advanced undergraduate or beginning graduate level.

Even so, we recognize that the required mathematical sophistication will be considered excessive by many readers. Attitude dynamics analysis is necessarily complex due to three factors. Attitude information is inherently vectorial, requiring three coordinates for its complete specification. Attitude analysis deals inherently with rotating, hence noninertial, frames. Finally, rotations are inherently order dependent in their description; the mathematics that describes them therefore lacks the multiplicative commutativity found in basic algebra.

In the following discussion, we attempt to alleviate this by appealing to the many analogies between rotational and translational dynamics and, as always, by stressing applications rather than derivations of results. Those requiring more detail are urged to consult one of the many excellent references in the field. Hughes[1] provides an especially good analytical development of attitude dynamics analysis and includes extensive applications to practical spacecraft design. Wertz[2] offers a definitive text on operational practices in attitude determination, as well as including brief but cogent summaries of many other topics of interest in space vehicle design.

Attitude determination and control is typically a major vehicle subsystem, with requirements that quite often drive the overall spacecraft design. Components tend to be relatively massive, power consuming, and demanding of specific orientation, alignment tolerance, field of view, structural frequency response, and structural damping. As we will see, effective attitude control system design is unusually demanding of a true systems orientation.

7.2 Basic Concepts and Terminology

7.2.1 Definition of Attitude

Spacecraft *attitude* refers to the angular orientation of a defined body-fixed coordinate frame with respect to a separately defined external frame. The spacecraft body frame may be arbitrarily chosen; however, some ways of defining it offer more utility than others, as we will see. The external frame may be one of the "inertial" systems discussed in Chapter 4 (GCI or HCI), or it may be a non-inertial system such as the local vertical, local horizontal (LVLH) frame, which is used to define the flight path angle (Fig. 4.9).

Astute readers will note that we have mentioned only the angular orientation between a spacecraft and an external frame, whereas in general some translational offset will also exist between the two. This is illustrated in Fig. 7.1, and leads to the question of the influence of parallax in performing spacecraft attitude measurements with respect to the "fixed" stars, which serve as the basis for inertial frames.

The concept of parallax is shown in Fig. 7.2. As seen, measurements of angles with respect to a given star will differ for frames whose origins are located apart. However, in almost all cases of practical interest parallax effects are insignificant

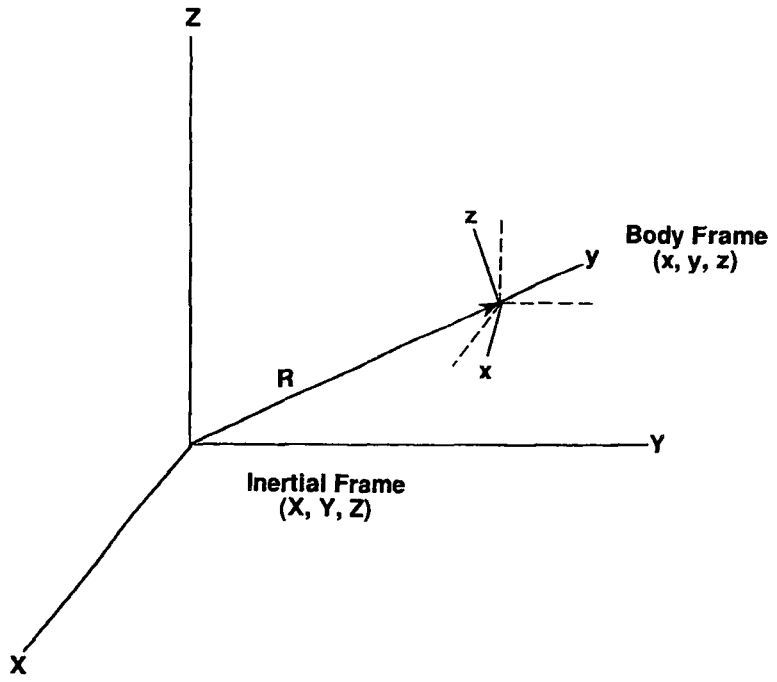


Fig. 7.1 Spacecraft body frame referred to inertial frame.

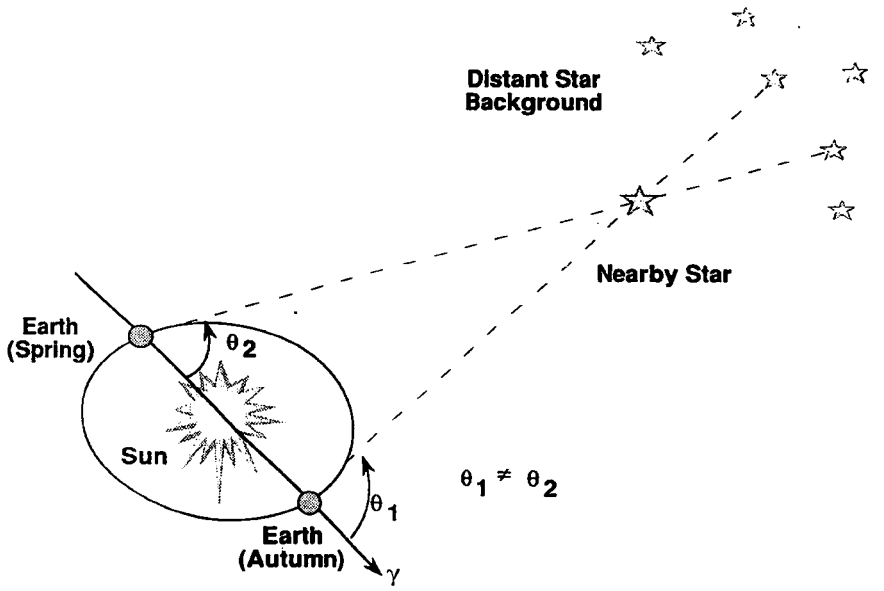


Fig. 7.2 Example of stellar parallax.

for spacecraft. The nearest star system, α-Centauri, is approximately 4.3 lightyears (LY) from Earth. Using the Earth's orbital diameter as a baseline, and making measurements six months apart, an object will show a parallax of 1 arcsecond at a distance of 3.26 LY, a quantity defined for obvious reasons as a *parsec*. Thus, even α-Centauri has a parallax of only about 0.75 arcsecond; all other stars have less. For most practical purposes, then, the location of a spacecraft will not influence measurements made to determine its attitude.

As always, exceptions exist. The European Space Agency's Hipparcos spacecraft was placed in orbit in 1989 for the purpose of making astrometric measurements of the parallax of some 120,000 relatively nearby stars, down to about 10^{th} magnitude, so that their distances could be more accurately determined. Angular measurement errors of order 0.002 arcsecond (about 10 nrad) were sought. (The Hubble Space Telescope is also designed to make such measurements, though not to this level of precision.) Obviously, the "error" due to stellar parallax is precisely the measurement sought by these missions.

As another example, HST is required to track and observe moving objects within the solar system to within 0.01 arcsec. At this level, parallax errors induced by HST movement across its Earth-orbital baseline diameter of 13,500 km are significant. Mars, for example, periodically approaches to within approximately 75 million km of Earth. During a half-orbit of HST it would then appear to shift its position by about 180 arcseconds rad, or roughly 36 arcseconds, against the background of fixed guide stars. The tracking accuracy requirement

would be grossly violated if this apparent motion were not compensated for in the HST pointing algorithm.

Attitude determination refers to the process (to which we have already alluded) of measuring spacecraft orientation. *Attitude control* implies a process, usually occurring more or less continuously, of returning the spacecraft to a desired orientation, given that the measurement reveals a discrepancy. In practice, errors of both measurement and actuation will always exist, and so both these processes take place within some tolerance.

Errors will result from inexact execution of reorientation maneuvers that are themselves based on inexact measurements, and will in addition arise from disturbances both internally and externally generated. The spacecraft is not capable of responding instantly to these disturbances; some time is always consumed in the process of measuring an error and computing and applying a correction. This leads to a typical pointing history such as shown in Fig. 7.3. Close examination of this figure reveals several features of interest.

The low-frequency, cyclic departure from and restoration to an average value is the result of the error detection and correction process implemented by the ADCS. It is roughly periodic, an artifact of the finite interval required to sense an error and implement a correction. This fundamental period, τ, implies a limit to the frequency response of the spacecraft, called the *bandwidth* or *passband*, of about $1/\tau$ Hz. A disturbance (such as an internal vibration or external impulsive torque), which has a frequency content higher than this, is simply not sensed by the spacecraft ADCS. Only the longer-term integrated effect, if such exists, is correctable.

This inability to sense and respond to high-frequency disturbances produces the *jitter* on the signal shown in Fig. 7.3. Jitter then refers to the high-frequency (meaning above the spacecraft passband) discrepancy between the actual and desired attitude. Attitude *error*, as we will use it henceforth, implies the low-frequency (within the passband) misalignment that is capable of being sensed and acted upon.

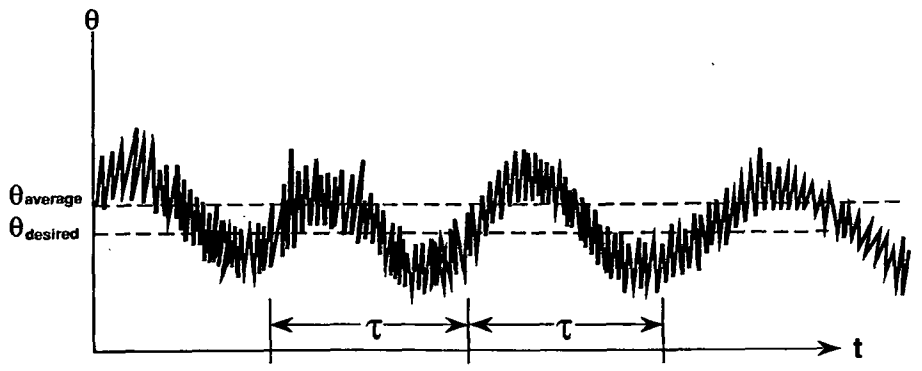


Fig. 7.3 Typical spacecraft pointing history.

We will return shortly to the discussion of attitude jitter. For the moment, note that a long-term integration (several τ periods) of the data in Fig. 7.3 would clearly yield an average value θ_a displaced from the desired value θ_d. This *bias* in the attitude could be due to sensor or actuator misalignment, to the effects of certain types of disturbances, or to more subtle properties of the control algorithm. Note further that θ_d is not always (and maybe not even very often) a constant. If not, we are said to have a *tracking* problem, as opposed to the much simpler constant-angle *pointing* problem. Tracking at higher rates or nonconstant rates generally yields poorer average performances than does pointing or low-rate tracking, or requires more complex engineering to achieve comparable performance.

7.2.2 Attitude Jitter

Spacecraft attitude jitter is almost universally discussed in statistical terms, a view consistent with the fact that the jitter is, by definition, not subject to ADCS influence, and is therefore "random" in that sense. Continuing in this vein, we note that by subtracting the average value θ_a from the data, we produce by definition a zero-mean history such as shown in Fig. 7.4. The smooth central curve results from filtering the data to remove the jitter, i.e., the components above the spacecraft passband. This curve is what we have earlier denoted the attitude error.

If we subtract the smooth central curve as well, we retain only the jitter, as shown in Fig. 7.5. This jitter can have sources both deterministic and random. An example of the former could be the vibration of an attitude sensor due to an internal source at a structural frequency above the control system passband. Random jitter may be due to many causes, including electronic and mechanical noise in the sensors and actuators. Our use of the term "noise" in this sense somewhat begs the question; perception of noise often depends on who is using the data. The spacecraft structural engineer will regard only the electronic effects as noise; the structural vibrations are, if included in the data, "signal" to him. The

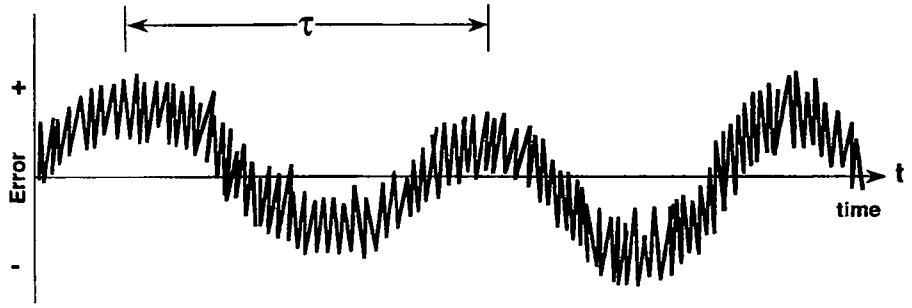


Fig. 7.4 Zero-mean attitude history.

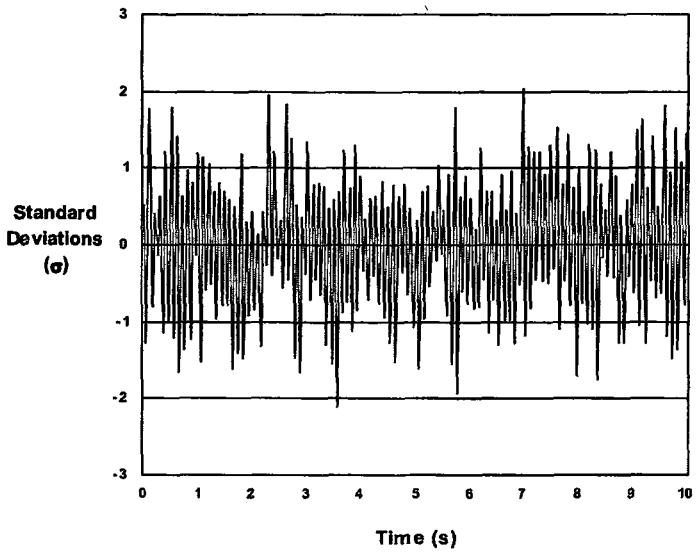


Fig. 7.5 Attitude jitter.

sensor designer may, given the data with mechanical effects removed, find much value in isolating the electronic disturbances. To the ADCS engineer, it is all noise, but understanding its source characteristics may be instrumental in removing or coping with it.

Because of the crucial importance to ADCS design of controlling attitude jitter, some discussion of the approach to its modeling is in order here. Readers conversant with the terminology of probability and statistics will have no difficulty with the subsequent discussion; others may wish to review the slightly broader discussion in Appendix A, or one of the many available references in the field.

Because of the way we have constructed Fig. 7.5, the time history of the attitude jitter has a mean value of zero, as noted earlier. Further, it is usually profitable in ADCS analysis to assume that the jitter is random, and that at any instant in time its amplitude has a *Gaussian* or *normal* probability distribution. In the language of probability theory, then, we view the jitter as a zero-mean Gaussian *random process*. A Gaussian distribution is fully characterized by only two parameters, its mean (zero in this example) and its variance, always denoted as σ^2.

The seemingly restrictive (but enormously convenient) assumption of Gaussian process statistics is usually quite well satisfied in practice. This results from application of the central limit theorem of statistics,[3] which loosely states that the sum of many independent zero-mean probability distributions converges in the limit to a Gaussian distribution. In practice (and this is a forever surprising result) "many" may be as few as four or five, and rarely more than 10, unless we

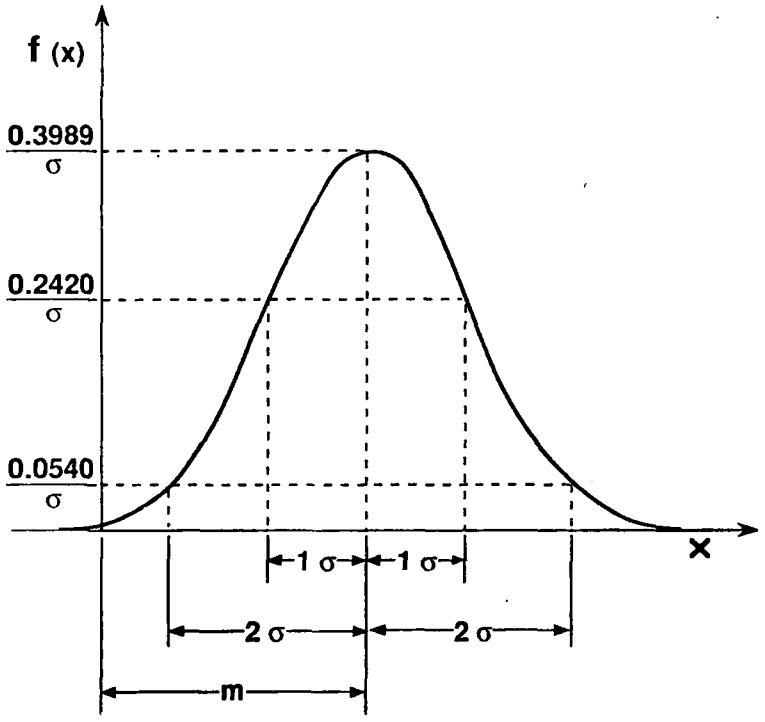


Fig. 7.6 Gaussian probability density function.

are at the extremes of the normal curve shown in Fig 7.6. Because we usually have very many independent noise sources in a system, in most cases we can rely quite comfortably on the assumption of zero-mean Gaussian noise.

If the jitter amplitude data are squared, we obtain the instantaneous *power* in the signal, usually called the *noise power*, $N(t)$. If the amplitude data are Gaussian distributed, $N(t)$ is Gaussian as well, though now with mean $N_0 > 0$. If $N(t)$ is averaged over time and found to yield the same process statistics (mean and variance only, for a Gaussian) at any epoch, the noise process is said to be *stationary*. Note carefully that a stationary process does not necessarily produce the same values of noise power, $N(t)$, at two different times, t_1 and t_2. Rather, stationarity implies that $N(t_1)$ and $N(t_2)$ are sample values drawn from an underlying distribution having the same process statistics (N_o and σ^2 for a Gaussian) at any time.

Each jitter time history can be viewed as being only one example of an ensemble of possible sequences. If an average across the ensemble of sequences would yield the same constant process statistics as the average over a given time sequence, the process is further said to be *ergodic*; the time average and the ensemble average are the same. If a process is ergodic, it is of course stationary,

but not conversely. In nearly all practical applications, ergodicity is assumed, even though such an assumption can be difficult to verify.

If the power spectral density (see Chapter 12) of the jitter is constant across all frequencies, the noise is said to be *white*, while if not constant, it is of course *colored*. These terms derive from the fact that the noise power, if constant at all frequencies (colors), is white by analogy to white light in optics. White noise cannot truly exist, as it possesses infinite total signal power; however, in usual applications the assumption of white Gaussian noise (WGN) is nearly universal. It is also reasonable, in that the system passband is often quite narrow with respect to the variations in the noise spectrum. Thus, in any such narrow segment, the noise power may indeed be approximately constant. Moreover, even highly colored noise can often be represented by the process of filtering an initially white noise input. Thus, the assumption of WGN processes is often both realistic and analytically convenient.

Under the zero-mean WGN jitter model, we note that the maximum amplitude excursion seen in Fig. 7.5 can be loosely said to fall at about the 3σ point. (Strictly, 99.73% of the data from a Gaussian distribution fall within $\pm 3\sigma$ of the mean.) This defines the corresponding 2σ and 1σ levels, at approximately 95.4% and 68.3%, respectively. Attitude jitter specifications are most commonly quoted in terms of either 1σ or 3σ performance levels. To discuss the average value requires more care; as mentioned repeatedly, the average jitter amplitude is zero. This is not a useful concept in characterizing the system performance. However, if we square the data, average over time interval τ, and then square-root the result (the so-called root-mean-square, or rms, operation), we obtain a useful average system jitter. In essence, we can more usefully describe the jitter in terms of its power rather than its amplitude. For Gaussian processes, the rms and 1σ levels are synonymous. This leads to the common (if not strictly accurate) tendency among engineers to consider 1σ performance to be "average," while 3σ behavior represents worst case. The reader who has labored through this discussion will now appreciate both the utility and limitations of such characterizations.

Jitter in a spacecraft must be accepted; by its definition, it is the error for which we do not compensate. It may, however, be reduced or controlled through proper mechanical, configuration, and structural design, as well as through attention to use of low-noise subsystems in the vehicle. If this proves insufficient, a more sophisticated control system design is required to compensate for disturbances at a finer level. In many spacecraft, minimizing attitude jitter becomes a shared, and nearly all-consuming, task for the attitude control and structural engineers on the project.

7.2.3 Rotational Kinematics and Celestial Sphere

Figure 7.7 depicts a celestial sphere centered in the origin of a coordinate frame. As we have discussed, length scales do not influence attitude determination and control, and so we may consider the sphere to be of unit radius.

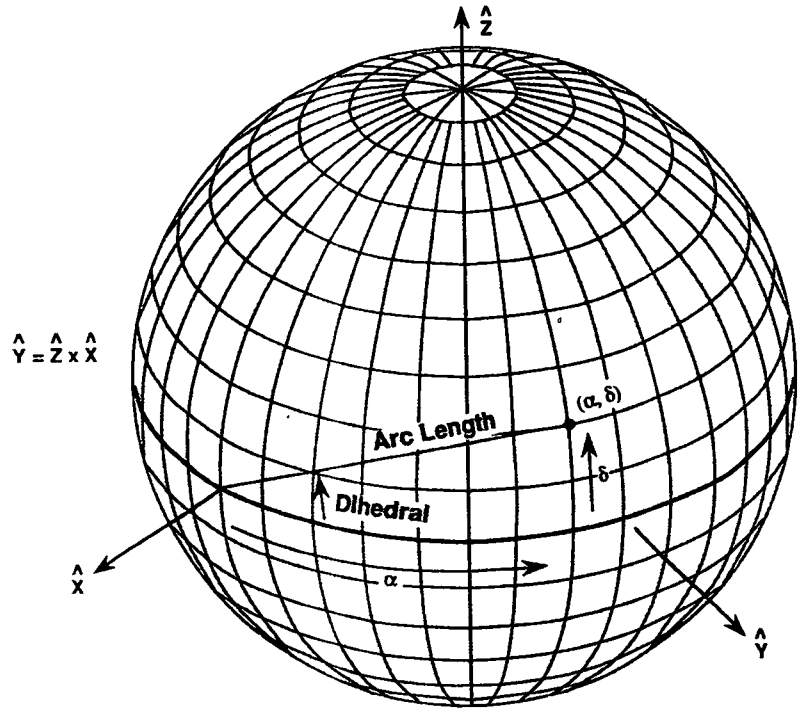


Fig. 7.7 Attitude measurements on the unit celestial sphere.

Directions may be specified in several ways on the celestial sphere. Possibly the most obvious is to use the Cartesian (x, y, z) coordinates of a particular point. Since

$$x^2 + y^2 + z^2 = 1 \tag{7.1}$$

only two of the three coordinates are independent. It is common in astronomy to use the right ascension α and the declination δ, defined as shown in Fig. 7.7, to indicate direction.[4] Note

$$x = \cos\alpha\cos\delta \tag{7.2a}$$

$$y = \sin\alpha\cos\delta \tag{7.2b}$$

$$z = \sin\delta \tag{7.2c}$$

The use of Euler angles to describe body orientation is common in rotational kinematics. An Euler angle set is a sequence of three angles and a prescription for rotating a coordinate frame through these angles to bring it into alignment with another frame. Figure 7.8 shows a typical Euler rotation sequence, specifically a 2-1-3 set, meaning that the rotation is first about the y axis, then about the new x axis, then about the new z axis. Other choices are often encountered as well;

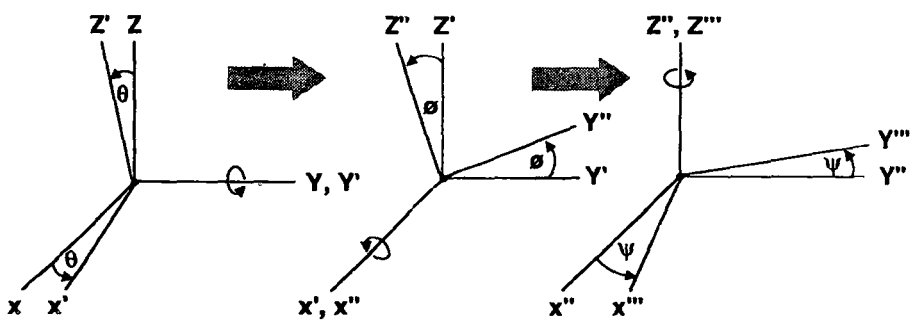


Fig. 7.8 Euler angle rotation sequence.

we used a 3-1-3 sequence to obtain the orbital state vector rotation matrix in Chapter 4.

The most common set of orientation angles used in spacecraft attitude determination and control is the *roll*, *pitch*, and *yaw* system shown in Fig. 7.9. This system derives from nautical, and later aeronautical, practice. Like all three-parameter orientation systems, it is singular at certain angles. The utility of this system derives in part from the fact that the singularities occur at $\pm 90°$ angles that are essentially not encountered in nautical and aeronautical applications, and not commonly encountered with space vehicles.

We define an Euler angle set (ϕ, θ, ψ) corresponding to the roll, pitch, and yaw angles of the spacecraft body frame relative to a rotating local vertical frame, which for our purposes we take to be an inertial frame. Note that this frame is often referenced to the spacecraft velocity vector, and not necessarily to the local horizontal. Using $S\theta$ and $C\theta$ to represent $\sin\theta$ and $\cos\theta$, the transformation matrix that rotates the inertial frame into the body frame via sequential elementary 2-1-3

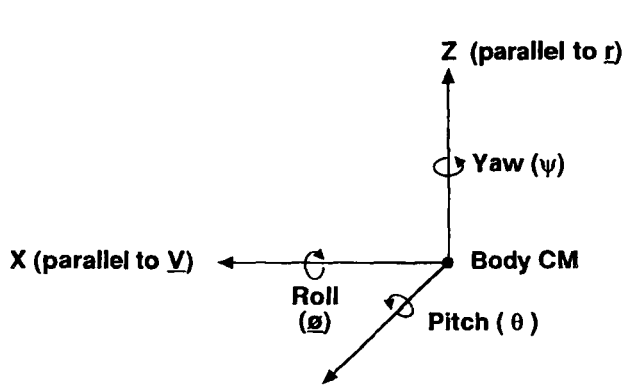


Fig. 7.9 Pitch, roll, and yaw angles.

, rotations in pitch, roll, and yaw can be written as

$$T_{I\to B} = \underbrace{\begin{bmatrix} C\psi & S\psi & 0 \\ -S\psi & C\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{(yaw)}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & C\phi & S\phi \\ 0 & -S\phi & C\phi \end{bmatrix}}_{\text{(roll)}} \underbrace{\begin{bmatrix} C\theta & 0 & -S\theta \\ 0 & 1 & 0 \\ S\theta & 0 & C\theta \end{bmatrix}}_{\text{(pitch)}} \tag{7.3}$$

or, in combined form,

$$T_{I\to B} = \begin{bmatrix} C\psi C\phi + S\psi S\theta S\phi & S\psi C\theta & -C\psi S\phi + S\psi S\theta C\phi \\ -S\psi C\phi + C\psi S\theta S\phi & C\psi C\theta & S\psi S\phi + C\psi S\theta C\phi \\ C\theta S\phi & -S\theta & C\theta C\phi \end{bmatrix} \tag{7.4}$$

Transformation matrices possess a number of useful properties. They are orthonormal, and so the inverse transformation (in this case, from inertial to body coordinates) is found by transposing the original:

$$T_{B\to I} = T_{I\to B}^{-1} = T_{I\to B}^{T} \tag{7.5}$$

Recall that matrix multiplication is not commutative; thus, altering the order of the rotation sequence produces a different transformation matrix. This is reflective of the fact that an Euler angle set implies a prescribed sequence of rotations, and altering this sequence alters the final orientation of the body if the angles are of finite size. It is readily shown for small angles that the required matrix multiplications are commutative, corresponding to the physical result that rotation through infinitesimal angles is independent of order.

Euler angle representations of spacecraft rotation are important in attitude analysis because they are easily visualized; they are suited to the way in which humans think. They can be computationally inconvenient because all such formulations implicitly contain a singularity corresponding exactly to the mechanical engineer's "gimbal lock" problem in multiple-gimbal systems. As noted, the Euler angle set chosen here (from among 12 possible sets) is among the more convenient, in that the singularity can often be kept out of the working range of rotations. However, it cannot be eliminated altogether in any three-parameter attitude representation, just as a mechanical engineer cannot avoid the possibility of gimbal lock using a three-gimbal set.

Relief is possible, however. Euler's theorem in rotational kinematics states that the orientation of a body may be uniquely specified by a vector giving the direction of a body axis and a scalar parameter specifying a rotation angle about that axis. A redundant fourth parameter is now part of the attitude representation. As a fourth gimbal allows a mechanical engineer to eliminate the possibility of gimbal lock, so too this analytical redundancy avoids coordinate singularities. From this result is derived the concept of quaternion, or Euler parameter, representation of attitude.[2] Hughes[1] considers the Euler parameter formulation to be, on balance, the most suitable choice for practical work.

The overview of attitude kinematics given here is sufficient only to acquaint the reader with the nature of the problem. More detailed discussions of attitude representations and rotational kinematics are given by Wertz,[2] Kaplan,[5] or Hughes.[1]

7.3 Review of Rotational Dynamics

A goal of attitude determination and control analysis is to describe the rotational behavior of a spacecraft body frame subject to the forces imposed upon it. This requires the use of Newton's laws of motion and the tools of calculus for the formulation and solution of such problems. From sophomore physics we recall that time-differentiation in a rotating (hence noninertial) coordinate system produces extra terms, and so we are prepared for some additional complication in attitude analysis.

Figure 7.10 shows the essential geometry. We have a vector $\boldsymbol{\rho}$ given in a rotating body frame, whereas Newton's laws describe motion in an inertial frame and require the use of second derivatives. Recalling the basic rule for time differentiation in a rotating frame, we write

$$\left(\frac{d\boldsymbol{\rho}}{dt}\right)_i = \left(\frac{d\boldsymbol{\rho}}{dt}\right)_b + \boldsymbol{\omega} \times \boldsymbol{\rho} \tag{7.6}$$

where $\boldsymbol{\omega}$ is the angular velocity vector of the rotating frame *in body coordinates.*

Newtonian dynamics problems involve the position vector $\boldsymbol{r}$ and its derivatives velocity $\boldsymbol{v}$ and acceleration $\boldsymbol{a}$. If $\boldsymbol{\rho}$ is a position vector in a body frame having angular velocity $\boldsymbol{\omega}$, it is given in the inertial frame as

$$\boldsymbol{r} = \boldsymbol{R} + \boldsymbol{\rho} \tag{7.7}$$

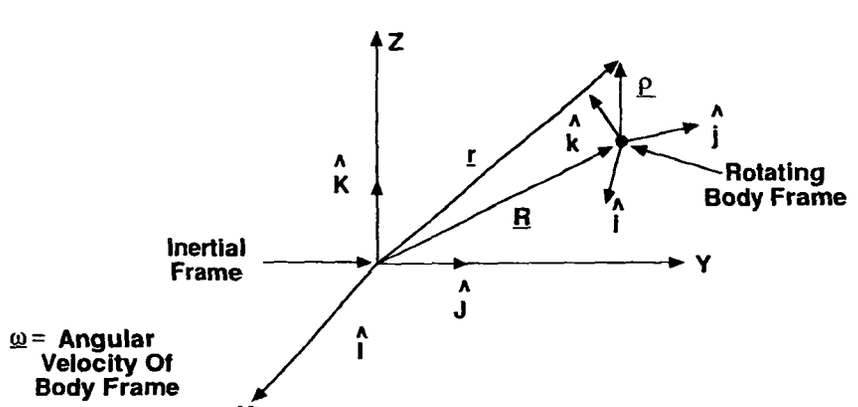


Fig. 7.10 Time differentiation in rotating frame.

hence

$$\boldsymbol{v} = \left(\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}\right)_i = \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} + \left(\frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}\right)_b + \boldsymbol{\omega} \times \boldsymbol{\rho} \tag{7.8}$$

and

$$\boldsymbol{a} = \left(\frac{\mathrm{d}^2\boldsymbol{r}}{\mathrm{d}t^2}\right)_i = \frac{\mathrm{d}^2\boldsymbol{R}}{\mathrm{d}t^2} + \left(\frac{\mathrm{d}^2\boldsymbol{\rho}}{\mathrm{d}t^2}\right)_b + 2\boldsymbol{\omega} \times \left(\frac{\mathrm{d}\boldsymbol{\rho}}{\mathrm{d}t}\right)_b + \frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t} \times \boldsymbol{\rho} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{\rho}) \tag{7.9}$$

The third term on the right is commonly called the Coriolis force, while the last term on the right-hand side is the centrifugal force.

The fundamental quantities of interest in Newtonian translational dynamics are mass, momentum, and kinetic energy. Conservation laws for these quantities provide the basis for the description of dynamical systems in classical physics. In rotational dynamics, the analogous quantities are the moment of inertia, angular momentum, and rotational kinetic energy.

The angular momentum of a mass is the moment of its linear momentum about a defined origin. From Fig. 7.10, the angular momentum of mass m_i about the origin in the inertial frame is

$$\boldsymbol{H} = \boldsymbol{r}_i \times m_i \boldsymbol{v}_i \tag{7.10}$$

and for a collection of point masses, the total angular momentum is

$$\boldsymbol{H}_t = \Sigma \boldsymbol{r}_i \times m_i \boldsymbol{v}_i \tag{7.11}$$

If we apply Eqs. (7.7) and (7.8) with $\boldsymbol{V} = \mathrm{d}\boldsymbol{R}/\mathrm{d}t$, and if we assume that 1) the origin of the rotating frame lies at the body center of mass ($\Sigma m_i \boldsymbol{\rho}_i = 0$), and 2) the position vectors $\boldsymbol{\rho}_i$ are fixed in the body frame (i.e., we have a rigid body, with $\mathrm{d}\boldsymbol{\rho}/\mathrm{d}t = 0$), we obtain

$$\boldsymbol{H}_t = (\Sigma m_i)\boldsymbol{R} \times \boldsymbol{V} + \Sigma m_i \boldsymbol{\rho}_i \times \frac{\mathrm{d}\boldsymbol{\rho}_i}{\mathrm{d}t} = \boldsymbol{H}_{\mathrm{orb}} + \boldsymbol{H}_b \tag{7.12}$$

The first term on the right is the angular momentum of the rigid body due to its translational velocity $\boldsymbol{V}$ in the inertial frame. The second term is the body angular momentum due to its rotational velocity about its own center of mass. If we consider the body to be an orbiting spacecraft, the first term is the orbital angular momentum introduced in Chapter 4, while the second is the angular momentum in the local center-of-mass frame, which is of interest for attitude dynamics analysis.

Equation (7.12) gives the important result that, for a rigid body, it is possible to choose a coordinate frame that decouples the spin angular momentum from the orbital angular momentum. Clearly, this is not always possible, and so-called spin-orbit coupling can at times be an important consideration in attitude control. However, unless stated otherwise, we employ the rigid body assumption in the

discussion to follow and will be concerned only with the body angular momentum.

Subject to the rigid body assumption, Eq. (7.6) yields

$$\frac{d\boldsymbol{\rho}_j}{dt} = \boldsymbol{\omega} \times \boldsymbol{\rho}_i \tag{7.13}$$

and from Eq. (7.12) the body angular momentum is (dropping the subscript),

$$\boldsymbol{H} = \Sigma m_i \boldsymbol{\rho}_i \times \frac{d\boldsymbol{\rho}_i}{dt} = \Sigma m_i \boldsymbol{\rho}_i \times (\boldsymbol{\omega} \times \boldsymbol{\rho}_i) = \boldsymbol{I}\boldsymbol{\omega} \tag{7.14}$$

where $\boldsymbol{I}$ is a real, symmetric matrix called the inertia matrix, with components

$$I_{11} = \sum m_i(\rho_{i2}^2 + \rho_{i3}^2) \tag{7.15a}$$

$$I_{22} = \sum m_i(\rho_{i1}^2 + \rho_{i3}^2) \tag{7.15b}$$

$$I_{33} = \sum m_i(\rho_{i1}^2 + \rho_{i2}^2) \tag{7.15c}$$

$$I_{12} = I_{21} = -\sum m_1 \rho_{i1} \rho_{i2} \tag{7.15d}$$

$$I_{13} = I_{31} = -\sum m_i \rho_{i1} \rho_{i3} \tag{7.15e}$$

$$I_{23} = I_{32} = -\sum m_i \rho_{i2} \rho_{i3} \tag{7.15f}$$

The diagonal components of the inertia matrix are called the moments of inertia, and the off-diagonal terms are referred to as the products of inertia. Because $\boldsymbol{I}$ is a real, symmetric matrix, it is always possible to find a coordinate system in which the inertia products are zero, i.e., the matrix is diagonal.[5] The elements of the inertia matrix may then be abbreviated I_1, I_2, and I_3, and are referred to as the principal moments of inertia, while the corresponding coordinates are called principal axes. These are the "natural" coordinate axes for the body, in that a symmetry axis in the body, if it exists, will be one of the principal axes.

Because of the generality of this result, it is customary to assume the use of a principal axis set in most attitude analysis. Unless otherwise stated, we assume such in this text. This convenient analytical assumption is usually violated in the real world. Spacecraft designers will normally select a principal axis coordinate frame for attitude reference purposes. However, minor asymmetries and misalignments can be expected to develop during vehicle integration, leading to differences between the intended and actual principal axes. When this occurs, attitude error measurements and control corrections intended about one axis will couple into others. Such coupling is seen by the attitude control system as an unwanted disturbance to be removed; therefore, there will normally be interface control document specifications limiting the allowable magnitude of the products of inertia in the defined coordinate frame.

A force F_i applied to a body at position ρ_i in center-of-mass coordinates produces a torque about the center of mass defined by

$$T_i = \rho_i \times F_i \tag{7.16}$$

The net torque from all such forces is then

$$T = \sum \rho_i \times F_i = \sum \rho_i \times m_i \frac{d^2 r_i}{dt^2} \tag{7.17}$$

After expanding $d^2 r_i / dt^2$ as before, we obtain

$$T = \frac{dH}{dt} = \left(\frac{dH}{dt}\right)_{body} + \omega \times H \tag{7.18}$$

The total kinetic energy of a body consisting of a collection of lumped masses is given by

$$E = \frac{1}{2}\sum m_i \left(\frac{dr_i}{dt}\right)^2 = \frac{1}{2}\sum m_i \left(\frac{dR_i}{dt} + \frac{d\rho_i}{dt}\right)^2 \tag{7.19}$$

If center-of-mass coordinates are chosen for the ρ_i, then the cross terms arising in Eq. (7.19) vanish, and the kinetic energy, like the angular momentum, separates into translational and rotational components,

$$E = \frac{1}{2}\Sigma m_i \left(\frac{dR_i}{dt}\right)^2 + \frac{1}{2}\Sigma m_i \left(\frac{d\rho_i}{dt}\right)^2 = E_{trans} + E_{rot} \tag{7.20}$$

If the rigid body assumption is included, such that $H = I\omega$, we may, after expanding Eq. (7.20), write

$$E_{rot} = \frac{1}{2}\omega^T I \omega \tag{7.21}$$

Equations (7.14) and (7.21) define angular momentum and kinetic energy for rotational dynamics and are seen for rigid bodies to be completely analogous to translational dynamics, with ω substituted for v and I replacing m, the body mass. Equation (7.18) is Newton's second law for rotating rigid bodies.

A particularly useful formulation of Eq. (7.18) is obtained by assuming a body-fixed principal axis center-of-mass frame in which to express H, T, and ω. In this case, we have

$$\left(\frac{dH}{dt}\right)_{body} = T - \omega \times I\omega \tag{7.22}$$

which becomes, on expansion into components,

$$\dot{H}_1 = I_1\dot{\omega}_1 = T_1 + (I_2 - I_3)\omega_2\omega_3 \tag{7.23a}$$

$$\dot{H}_2 = I_2\dot{\omega}_2 = T_2 + (I_3 - I_1)\omega_3\omega_1 \tag{7.23b}$$

$$\dot{H}_3 = I_3\dot{\omega}_3 = T_3 + (I_1 - I_2)\omega_1\omega_2 \tag{7.23c}$$

These are the Euler equations for the motion of a rigid body under the influence of an external torque. No general solution exists for the case of an arbitrarily specified torque. Particular solutions for simple external torques do exist; however, computer simulation is usually required to examine cases of practical interest.

7.4 Rigid Body Dynamics

An understanding of the basic dynamics of rigid bodies is crucial to an understanding of spacecraft attitude dynamics and control. Although in practice few if any spacecraft can be accurately modeled as rigid bodies, such an approximation is nonetheless the proper reference point for understanding the true behavior. The Euler equations derived in the previous section can in several simple but interesting cases be solved in closed form, yielding insight not obtained through numerical analysis of more realistic models.

The most important special case for which a solution to the Euler equations is available is that for the torque-free motion of an approximately axisymmetric body spinning primarily about its symmetry axis, i.e., a spinning top in free fall. Mathematically, the problem is summarized as

$$\omega_x,\ \omega_y \ll \omega_z = \Omega \tag{7.24}$$

$$I_x \cong I_y \tag{7.25}$$

With these simplifications, the Euler equations become

$$\dot{\omega}_x = -K_x\Omega\omega_y \tag{7.26a}$$

$$\dot{\omega}_y = K_y\Omega\omega_x \tag{7.26b}$$

$$\dot{\omega}_z \cong 0 \tag{7.26c}$$

where

$$K_x = \frac{I_z - I_y}{I_x} \tag{7.27a}$$

$$K_y = \frac{I_z - I_x}{I_y} \tag{7.27b}$$

The solution for the angular velocity components is

$$\omega_x(t) = \omega_{x0} \cos \omega_n t \tag{7.28}$$

$$\omega_y(t) = \omega_{y0} \sin \omega_n t \tag{7.29}$$

where the natural frequency ω_n is defined by

$$\omega_n^2 = K_x K_y \Omega^2 \tag{7.30}$$

The conceptual picture represented by this solution is that of a body with an essentially symmetrical mass distribution spinning rapidly about the axis of symmetry, which we have defined to be the z axis. This rapid rotation is at essentially constant speed $\omega_z = \Omega$. However, a smaller x–y plane component of angular velocity, time varying in its orientation, also exists. This component rotates periodically around the body z axis at a natural or "nutation" frequency ω_n determined by the body's inertia ratios. This results in a circular motion of the body z axis around the angular momentum vector $\boldsymbol{H}$ at the nutation frequency. (Recall $\boldsymbol{H}$ is fixed in inertial space because no torques are present.) The motion can have one of two general patterns, depending on the ratio of I_z to I_x or I_y. Figure 7.11 shows the two cases. The angle ν between the body z axis and the inertially fixed $\boldsymbol{H}$ vector is called the nutation angle.

The space cone in Fig. 7.11 refers to the fact that $\boldsymbol{H}$ is fixed in inertial space. Conversely, Eqs. (7.27) are expressed in body coordinates, and so the body cone is defined relative to the body principal axis frame. The space cone may lie inside or outside the body cone, depending on whether the spinning body is "pencil shaped" ($I_z < I_x \cong I_y$) or "saucer shaped" ($I_z > I_x \cong I_y$).

The preceding discussion can be generalized[1,5] to include the case where $I_x \neq I_y$. We then have the possibility that the spin axis inertia I_z is intermediate between I_x and I_y. In such a case, Eqs. (7.27) and (7.30) show that $\omega_n^2 < 0$; i.e., the

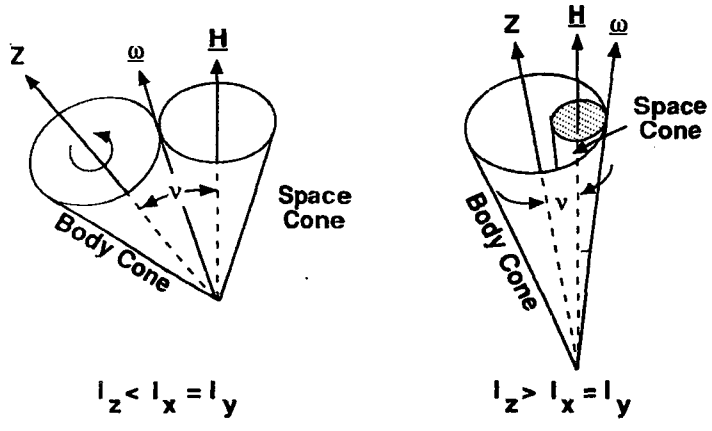


Fig. 7.11 Possible cases for torque-free motion of a symmetric rigid body.

nutation frequency is imaginary. The previous sinusoidal solution for ω_x and ω_y becomes an exponential solution, expressible if desired in terms of hyperbolic functions. Thus, the body cannot have a fixed nutation angle when spinning primarily about an axis of intermediate inertia moment. Elementary stability analysis shows that if a rigid body is initially spinning perfectly about such an axis, any perturbing torque will result in the growth of the nutation angle until the body is spinning about either the maximum or minimum inertia axis, depending on initial conditions. We thus have the important result that a rigid body can rotate about its extreme inertia axes, but not the intermediate axis.

This important conclusion is further modified if flexibility is considered. If the body is not rigid, energy dissipation must occur as it flexes. Because total system energy must be conserved, this energy, dissipated as heat, must be derived from the rotational kinetic energy, Eq. (7.21). Thus, a flexible spinning body causes E_{rot} to decrease. At the same time, however, the angular momentum for the torque-free system must be constant. Ignoring for the moment the vector nature of these quantities, we have then the constraints

$$E_{\text{rot}} = \tfrac{1}{2} I\omega^2 \tag{7.31}$$

$$H = I\omega \tag{7.32}$$

hence

$$E_{\text{rot}} = \tfrac{1}{2}\frac{H^2}{I} \tag{7.33}$$

Clearly, if energy is to be dissipated (and the second law of thermodynamics guarantees that it will), the moment of inertia must increase. If the body is not to be permanently deformed, the spin axis must shift (in body coordinates, of course; the spin axis is fixed in space by the requirement that $\boldsymbol{H}$ be constant) to the maximum-inertia axis. To visualize this, we imagine an energy-dissipating body with its angular momentum vector initially aligned with the minimum-inertia axis. As energy is lost, the nutation angle grows to satisfy Eq. (7.33). Eventually, the nutation has grown to 90°, the maximum possible. The body will be spinning at a slower rate about the maximum-inertia principal axis. This is colloquially referred to as a *flat spin*, a name instantly evocative of the condition.

Thus, in the absence of external torques a real body can spin stably only about the axis of maximum moment of inertia. This so-called major axis rule was discovered, empirically and embarrassingly, following the launch of Explorer 1 as the first U.S. satellite on 1 February 1958. The cylindrically shaped satellite was initially spin stabilized about its long axis, and had four flexible wire antennas for communication with ground tracking stations. Within hours, the energy dissipation inherent in the antennas had caused the satellite to decay into a flat spin, a condition revealed by its effect on the air-to-ground communications link. This initially puzzling behavior was quickly explained by Ron Bracewell and Owen Garriott, then of Stanford University (Garriott later became a Skylab and shuttle astronaut).[6]

This result can be illustrated most graphically by considering the homely example of a spinning egg. It is well known that a hard-boiled egg can be readily distinguished from a raw egg by attempting to spin it about its longitudinal axis. Because of very rapid internal energy dissipation by the viscous fluid, the raw egg will almost immediately fall into a flat spin, while the boiled egg will rotate at some length about its minor axis.

It should be carefully understood that the arguments just made, while true in the general terms in which they are expressed, are heuristic in nature. Equations (7.31) and (7.32) must hold, leading to the behavior described. However, it is equally true that Newton's laws of motion must be satisfied; physical objects do not move without the imposition of forces. The energy-momentum analysis outlined earlier is incomplete, in that the origin of forces causing motion of the body is not included. Nonetheless, "energy sink" analyses based on the arguments outlined earlier can be quite successful in predicting spacecraft nutation angle over time.

Further refinements of these conclusions exist. For example, the major axis rule strictly applies only to simple spinners. Complex bodies with some parts spinning and others stationary may exhibit more sophisticated behavior. In particular, a so-called dual spin satellite can be stable with its angular momentum vector oriented parallel to the minor principal axis. We will address the properties of dual spin satellites in a later section.

By assuming a flexible (e.g., nonrigid) body, we have violated a basic constraint under which the simplified results of Eq. (7.12) and those following were derived. Specifically, the spin motion and orbital motion are no longer strictly decoupled. Much more subtle behavior can follow from this condition.

Even if rigid, an orbiting spacecraft is not in torque-free motion. An extended body will be subject to a number of external torques to be discussed in the following section, including aerodynamic, magnetic, and tidal or gravity-gradient torques. The existence of the gravity-gradient effect, discussed in Chapter 4, renders a spinning spacecraft asymptotically stable only when its body angular momentum vector is aligned with the orbital angular momentum vector, i.e., the orbit normal.

The topic of stability analysis is a key element of spacecraft attitude dynamics. Even when the equations of motion cannot be solved in closed form, it may be determined that equilibria exist over some useful parametric range. If stable, such equilibria can be used as the basis for passive stabilization schemes, or for reducing the workload upon an active control system.

7.5 Space Vehicle Disturbance Torques

As mentioned, operating spacecraft are subject to numerous disturbance forces which, if not acting through the center of mass, result in a net torque being

imparted to the vehicle. Assessment of these influences in terms of both absolute and relative magnitude is an essential part of the ADCS designer's task.

7.5.1 Aerodynamic Torque

The role of the upper atmosphere in producing satellite drag was discussed in Chapter 4 in connection with orbit decay. The same drag force will, in general, produce a disturbance torque on the spacecraft due to any offset that exists between the aerodynamic center of pressure and the center of mass. Assuming $\boldsymbol{r}_{cp}$ to be the center-of-pressure (CP) vector in body coordinates, the aerodynamic torque is

$$\boldsymbol{T} = \boldsymbol{r}_{cp} \times \boldsymbol{F}_a \tag{7.34}$$

where, as in Chapter 4, the aerodynamic force vector is

$$\boldsymbol{F}_a = \left(\frac{1}{2}\right)\rho V^2 S C_D \frac{\boldsymbol{V}}{V} \tag{7.35}$$

and

ρ = atmosphere density
V = spacecraft velocity
S = spacecraft projected area $\perp \boldsymbol{V}$
C_D = drag coefficient, usually between 1 and 2 for free-molecular flow.

It is important to note that $\boldsymbol{r}_{cp}$ varies with spacecraft attitude and, normally, with the operational state of the spacecraft (solar panel position, fuel on board, etc.). As we discussed in Chapter 4, major uncertainties exist with respect to the evaluation of Eq. (7.35). Drag coefficient uncertainties can easily be of order 50%, while upper atmosphere density variations approaching an order of magnitude relative to the standard model are not uncommon. Thus, if aerodynamic torques are large enough to be a design factor for the attitude control system, they need to be treated with appropriate conservatism.

As an example, let us consider a satellite with a frontal area $S = 5\ \text{m}^2$ and a drag coefficient $C_D = 2$ orbiting at 400 km, with standard model atmospheric density $\rho = 4 \times 10^{-12}\ \text{kg/m}^3$. Assuming circular velocity at this altitude, the magnitude of the disturbance torque is $T/r_{cp} = 1.2 \times 10^{-3}$ N. This seems small, and is, but to put it in perspective, let us assume it to be the only torque acting on the spacecraft, through a moment arm of $r_{cp} = 1$ cm, and that the spacecraft moment of inertia about the torque axis is $I = 1000\ \text{kg} \cdot \text{m}^2$. Equations (7.23) simplify to

$$T = \frac{\mathrm{d}H}{\mathrm{d}t} = I\frac{\mathrm{d}\omega}{\mathrm{d}t} = I\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} \tag{7.36}$$

with initial conditions on angular position and velocity

$$\theta(0) = 0 \tag{7.37}$$

$$\omega(0) = \frac{d\theta}{dt} = 0 \tag{7.38}$$

We find for this example

$$\theta(t) = \left(\frac{T}{I}\right)t^2 = (1.2 \times 10^{-5}\,\mathrm{N \cdot m}/1000\,\mathrm{kg \cdot m^2})t^2$$

$$= 1.2 \times 10^{-8}\,s^{-2}t^2 \tag{7.39}$$

The angular displacement predicted by Eq. (7.39) certainly seems small. However, left uncorrected, the spacecraft will drift about 0.012 rad, or 0.7 deg, after 1000 s. This is in most cases a large error for an attitude control system. Even worse, the error growth is quadratic, so this aerodynamic torque applied over 10,000 s, about 3 h or 2 orbits, would produce a 1 rad pointing error! This is unacceptable in almost any conceivable application. Thus, even small disturbance torques will be problematic if corrections are not applied.

7.5.2 Gravity-Gradient Torque

Planetary gravitational fields decrease with distance r from the center of the planet according to the Newtonian $1/r^2$ law, provided higher order harmonics as discussed in Chapter 4 are neglected. Thus, an object in orbit will experience a stronger attraction on its "lower" side than its "upper" side. This differential attraction, if applied to a body having unequal principal moments of inertia, results in a torque tending to rotate the object to align its "long axis" (minimum inertia axis) with the local vertical. Perturbations from this equilibrium produce a restoring torque toward the stable vertical position, causing a periodic oscillatory or "librational" motion. Energy dissipation in the spacecraft will ultimately damp this motion.

The gravity-gradient torque for a satellite in a near-circular orbit is

$$\boldsymbol{T} = 3n^2\hat{\boldsymbol{r}} \times \boldsymbol{I} \cdot \hat{\boldsymbol{r}} \tag{7.40}$$

where

$\hat{\boldsymbol{r}} = \boldsymbol{r}/r$ = unit vector from planet to spacecraft
$n^2 = \mu/a^3 \cong \mu/R^3$ = orbital rate
μ = gravitational constant (398,600 $\mathrm{km^3/s^2}$ for Earth)
$\boldsymbol{I}$ = spacecraft inertia matrix

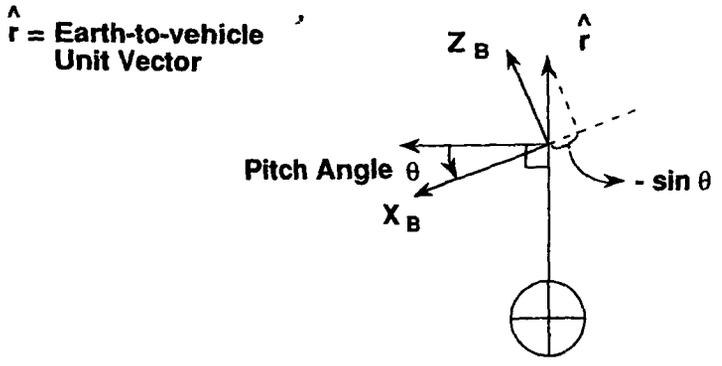


Fig. 7.12 Pitch plane geometry for gravity gradient torque.

In the spacecraft body frame with pitch, roll, and yaw angles given as in Eqs. (7.4), the unit vector to the spacecraft from the planet is, from Fig. 7.12,

$$\hat{r} = (-\sin\theta,\ \sin\phi,\ 1 - \sin^2\theta - \sin^2\phi)^T \cong (-\theta,\ \phi,\ 1)^T \tag{7.41}$$

with the approximation holding for small angular displacements. Then the gravity gradient torque vector may be expressed in the body frame as

$$T = 3n^2[(I_z - I_y)\phi,\ (I_z - I_x)\theta,\ 0]^T \tag{7.42}$$

It is seen that the spacecraft yaw angle ψ does not influence the gravity-gradient torque; this is intuitively reasonable, because yaw represents rotation around the local vertical. We also note from Eq. (7.42) that the gravity-gradient influence is proportional to $1/r^3$. The torque magnitude clearly depends upon the *difference* between principal moments; thus, spacecraft that are long and thin are more affected than those that are short and fat.

To get an idea of a typical gravity-gradient torque magnitude, consider a low-orbiting spacecraft with $n \cong 0.001$ rad/s and an inertia moment difference in the relevant axis of 1000 kg · m^2. Then $T = 5 \times 10^{-5}$ N · m/deg.

7.5.3 Solar Radiation Pressure Torque

Solar radiation pressure and its effect upon spacecraft orbital dynamics was discussed in Chapter 4. As with aerodynamic drag, solar radiation pressure can produce disturbance torques as well as forces, which may require compensation from the attitude control system (ACS). The solar radiation pressure torque is, in body coordinates,

$$T = r \times F_s \tag{7.43}$$

where

r = vector from body center of mass to spacecraft optical center of pressure
$F_s = (1 + \mathrm{K})p_s A_\perp$
K = spacecraft surface reflectivity, $0 < K < 1$
$A_\perp$ = spacecraft projected area normal to sun vector
$p_s = I_s/c$
$I_s = 1358\ \mathrm{W/m^2}$ at 1 AU
$c = 2.9979 \times 10^8\ \mathrm{m/s}$

Solar radiation torque is independent of spacecraft position or velocity, as long as the vehicle is in sunlight, and is always perpendicular to the sun line. It will, in many cases, thus have no easily visualized relationship with the previously considered aerodynamic and gravity-gradient disturbance torques. For an order of magnitude comparison, consider typical values to be $A_\perp = 5\ \mathrm{m^2}$, $K = 0.5$, $r = 0.1$ m, and the spacecraft to be in Earth orbit. Then the torque magnitude is $T = 3.5 \times 10^{-6}\ \mathrm{N \cdot m}$. This would be about two orders of magnitude below the representative aerodynamic torque computed earlier for a satellite orbiting at 400-km altitude. As noted, however, the solar torque is independent of position, while the aerodynamic torque is proportional to atmospheric density. Above 1000-km altitude, solar radiation pressure usually dominates the spacecraft disturbance torque environment.

At geostationary orbit altitude, solar radiation pressure can be the primary source of disturbance torque, and designers must take care to balance the geometrical configuration to avoid center-of-mass to center-of-pressure offsets. The useful lifetime of a geostationary satellite is often controlled by the mass budget available for stationkeeping and attitude control fuel. Poor estimates of the long-term effect of disturbance torques and forces can and do result in premature loss of on-orbit capability.

Solar radiation pressure can also be important for interplanetary missions. While its strength obviously drops off rapidly for outer planet missions, it may in the absence of internally generated disturbances be essentially the only torque acting upon the vehicle during interplanetary cruise. For missions to Venus and Mercury, solar torques will often define the spacecraft disturbance torque limits, and can even be used in an active control mode. This was first accomplished on an emergency basis following the shutdown of an unstable roll-control loop on the Mariner 10 mission to Venus and Mercury. In this case, differential tilt between opposing sets of solar panels was used to introduce a deliberate offset between the center of mass and center of pressure in such a way as to effect roll control.

7.5.4 Magnetic Torque

Earth and other planets such as Jupiter that have a substantial magnetic field exert yet another torque on spacecraft in low orbits about the primary. This is

given by

$$\boldsymbol{T} = \boldsymbol{M} \times \boldsymbol{B} \tag{7.44}$$

$\boldsymbol{M}$ is the spacecraft magnetic dipole moment due to current loops and residual magnetization in the spacecraft. $\boldsymbol{B}$ is the Earth magnetic field vector *expressed in spacecraft coordinates*; its magnitude is proportional to $1/r^3$, where $\boldsymbol{r}$ is the radius vector to the spacecraft.

Few aerospace engineers are intimately involved with electromagnetic equipment, and so a brief discussion of measurement units for $\boldsymbol{M}$ and $\boldsymbol{B}$ is in order. Magnetic moment may be produced physically by passing a current through a coil of wire; the larger the coil, the greater the moment produced. Thus, in the SI system, $\boldsymbol{M}$ has units of ampere-turn-m^2 (Atm^2). $\boldsymbol{B}$ is measured in tesla (T) in the SI system. With $\boldsymbol{M}$ and $\boldsymbol{B}$ as specified, $\boldsymbol{T}$ of course has units of N · m.

An older but still popular system of units in electromagnetic theory is the centimeter-gram-second (CGS) system. In CGS units, $\boldsymbol{M}$ and $\boldsymbol{B}$ are measured in pole-cm and gauss (G), respectively, with the resultant torque in dyne-cm. Conversion factors between the two systems are:

$$1\,\text{Atm}^2 = 1000\,\text{pole cm} \tag{7.45a}$$

$$1\,\text{T} = 10^4\,\text{G} \tag{7.45b}$$

Earth's magnetic field at an altitude of 200 km is approximately 0.3 G or 3×10^{-5} T. A typical small spacecraft might possess a residual magnetic moment on the order of 0.1 Atm^2. The magnetic torque on such a spacecraft in low orbit would then be approximately 3×10^{-6} N · m.

Magnetic torque may well, as in this example, be a disturbance torque. However, it is common to reverse the viewpoint and take advantage of the planetary magnetic field as a control torque to counter the effects of other disturbances. We shall discuss this in more detail in a later section.

7.5.5 Miscellaneous Disturbance Torques

In addition to torques introduced by the spacecraft's external environment, a variety of other sources of attitude disturbance exist, many of them generated by the spacecraft during the course of its operation.

Effluent venting, whether accidental or deliberate, is a common source of spacecraft disturbance torque. When such venting must be allowed, as for example with propellant tank pressure relief valves, "T-vents" are typically used to minimize the resulting attitude perturbations. Jettisoned parts, such as doors or lens covers, will produce a transient reaction torque when released.

All of the effects we have discussed so far involve an actual momentum exchange between the spacecraft and the external environment, resulting from the application of an external torque. The momentum change in the spacecraft is

the integral of this torque. Of major significance also in spacecraft attitude control are internal torques, resulting from momentum exchange between internal moving parts. This has no effect on the overall system angular momentum, but can and does influence the orientation of body-mounted sensors and hence the attitude control loops that may be operating. Typical internal torques are those due to antenna, solar array, or instrument scanner motion, or to other deployable booms and appendages. As these devices are articulated, the rest of the spacecraft will react to keep the total system angular momentum constant.

A major portion of the spacecraft ADCS designer's effort may be devoted to the task of coping with internally generated disturbances. If at sufficiently low bandwidth, these will be compensated by the ACS. Typically, however, internal torques are transient events with rather high-frequency content relative to the ADCS passband limits. When this is so, the ACS can remove only the low-frequency components, leaving the remainder to contribute to the overall system jitter. Control of such jitter can be a major problem in the design and operation of observatory or sensor spacecraft.

7.6 Passive Attitude Control

The concept of passive attitude control follows readily from the discussion of the preceding sections. Passive stabilization techniques take advantage of basic physical principles and naturally occurring forces by designing the spacecraft to enhance the effect of one force while reducing others. In effect, we use the previously analyzed disturbance torques to control the spacecraft, choosing a design to emphasize one and mitigate the others.

An advantage of passive control is the ability to attain a very long satellite lifetime, not limited by onboard consumables or, possibly, even by wear and tear on moving parts. Typical disadvantages of passive control are relatively poor overall accuracy and somewhat inflexible response to changing conditions. Where these limitations are not of concern, passive techniques work very well. An excellent example was furnished by the now-obsolete Transit radio navigation satellite system,[7] for which the main operational requirement was a roughly nadir-pointed antenna. These satellites are gravity-gradient stabilized, with several having operational lifetimes of over 15 years.

A spacecraft design intended to provide passive control does not necessarily guarantee stability in any useful sense, and indeed we have seen that environmental and other effects can induce substantial unwanted attitude motion in a passively "stabilized" vehicle. For this reason, most such spacecraft include devices designed to augment their natural damping. Such "nutation dampers" can take a variety of forms, as we will discuss, and include eddy current dampers, magnetic hysteresis rods, ball-in-tube devices, and viscous fluid dampers.

7.6.1 Spin Stabilization

A basic passive technique is that of spin stabilization, wherein the intrinsic gyroscopic "stiffness" of a spinning body is used to maintain its orientation in inertial space. If no external disturbance torques are experienced, the angular momentum vector remains fixed in space, constant in both direction and magnitude. If a nutation angle exists, either from initial conditions or as the result of a disturbance torque, a properly designed energy damper will quickly (within seconds or minutes) remove this angle, so that the spin axis and the angular momentum vector are coincident.

An applied torque will, in general, have components both perpendicular and parallel to the momentum vector. The parallel component spins the spacecraft up or down, i.e., increases or decreases $\boldsymbol{H}$. The perpendicular torque component causes a displacement of $\boldsymbol{H}$ in the direction of $\boldsymbol{T}$. This is illustrated in Fig. 7.13, where the external force $\boldsymbol{F}$ causing the torque $\boldsymbol{T}$ is perpendicular to the plane containing $\boldsymbol{H}$. Note then that $\Delta \boldsymbol{H}$, while parallel to $\boldsymbol{T}$, is perpendicular to the actual disturbance force $\boldsymbol{F}$, since $\boldsymbol{T} = \boldsymbol{r} \times \boldsymbol{F}$. The magnitude of the angular momentum displacement is found from

$$\frac{dH}{dt} = T = rF \cong \frac{\Delta H}{\Delta t} \tag{7.46}$$

where, from the geometry,

$$\Delta H = 2H \sin\left(\frac{\Delta\theta}{2}\right) \cong H\Delta\theta = I\omega\Delta\theta \tag{7.47}$$

hence

$$\Delta\theta \cong \frac{rF\Delta t}{H} = \frac{rF\Delta t}{I\omega} \tag{7.48}$$

The gyroscopic stability to which we have alluded shows up in Eq. (7.48) with the appearance of the angular momentum in the denominator. The higher this

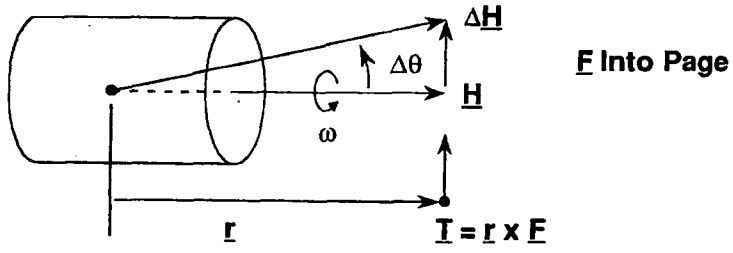


Fig. 7.13 Response of spin-stabilized spacecraft to external torque.

value, the smaller the perturbation angle $\Delta\theta$ that a given disturbance torque will introduce.

Spin stabilization is useful in a number of special cases where reliability and simplicity are more important than operational flexibility. Satellites intended for geostationary orbit, for example, are usually spin stabilized for the two required transfer orbit burns (see Chapter 4). Some missions utilize spin stabilization as the best means of meeting scientific objectives. Notable examples in this regard are Pioneers 10 and 11, the first spacecraft to fly by Jupiter and Saturn. The primary scientific goal of these spacecraft was the investigation of interplanetary electromagnetic fields and particles; this was most easily done from a spinning platform.

Long-term stability of a spinning spacecraft requires, as we have said, a favorable inertia ratio. In visual terms, the vehicle must be a "wheel" rather than a "pencil." Also, most spin-stabilized satellites will require nutation dampers as mentioned earlier to control the effect of disturbance torques on the spin axis motion. Furthermore, if it is desired to be able to alter the inertial orientation of the spin axis during the mission, the designer must provide the capability for control torques to precess the spin axis. This is commonly done with magnetic coils or small thrusters.

7.6.2 Gravity-Gradient Stabilization

From our previous discussion, it is clear that a spacecraft in a reasonably low orbit will tend to stabilize with its minimum-inertia axis in a vertical orientation. This property can obviously be used to advantage by the designer when a nadir or zenith orientation is desired for particular instruments. The principal design feature of such a satellite again involves the inertia ratio; the vehicle must possess an axis such that $I_z \ll I_x, I_y$. As noted previously, even when the spacecraft is designed in this fashion, the control torques are small, and additional damping is required to remove pendulum-like oscillations due to disturbances. These oscillations, or librations, are typically controlled through the use of magnetic hysteresis rods or eddy current dampers. Active "damping" (really active control) is also possible and, as might be expected, typically offers better performance.

The usual way of obtaining the required spacecraft inertia properties (i.e., long and thin) is to deploy a motor-driven boom with a relatively heavy (several kilograms or more) end mass. The "boom" will often be little more than a reel of prestressed metallic tape, similar to the familiar carpenter's measuring tape, which when unrolled springs into a more or less cylindrical form. Such an "open stem" boom will have substantial (for its mass) lateral stiffness, but little torsional rigidity. The possibility of coupling between easily excited, lightly damped torsional modes and the librational modes then arises, and often cannot be analytically dismissed. Again, careful selection of damping mechanisms is required.

Pure gravity-gradient attitude control provides no inherent yaw stability; the spacecraft is completely free to rotate about its vertical axis. When this is unacceptable, additional measures must be taken. One possibility is to add a momentum wheel with its axis perpendicular to the spacecraft vertical axis, as shown in Fig. 7.14. A stable condition then occurs with the wheel angular momentum aligned along the positive orbit normal.[8] Such a configuration has been flown on numerous satellites, though not with uniform success. Large amplitude librations are sometimes observed, often during particular orbital "seasons" (i.e., sun angles). Oscillations of sufficient magnitude to invert the spacecraft have occasionally occurred. These have been linked to long-period resonances in the spacecraft gravity-gradient boom that are excited by solar thermal input under the right conditions.[9]

Gravity-gradient stabilization is useful when long life on orbit is needed and attitude stabilization requirements are relatively broad. Libration amplitudes of 10–20 deg are not uncommon, although better performance can be obtained with careful design. An example is the GEOSAT spacecraft, a U.S. Navy radar altimetry satellite launched in 1984. Vertical stabilization to within 1° (1σ) was achieved through the use of a very stiff boom having an eddy current damper as its tip mass. In general, though, it will be found that gravity-gradient stabilization is too inflexible and imprecise for most applications.

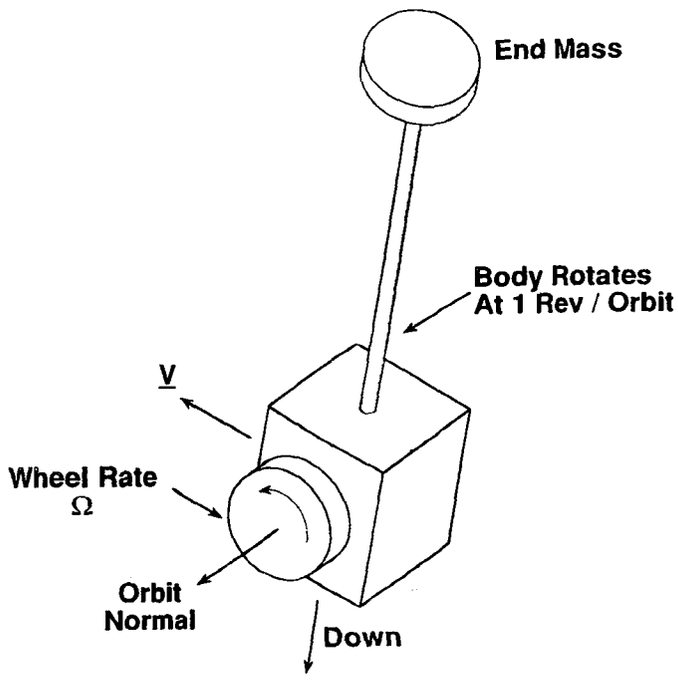


Fig. 7.14 Gravity-gradient stabilization with momentum wheel.

7.6.3 Aerodynamic and Solar Pressure Stabilization

As with gravity gradient, the existence of aerodynamic and solar radiation pressure torques implies the possibility of their use in spacecraft control. This has in fact been accomplished, although the flight history is considerably reduced compared to the gravity-gradient case. The most prominent example of aerodynamic stabilization occurred with MAGSAT, a low-altitude spacecraft intended to map the Earth's magnetic field.[10] This vehicle used an aerodynamic trim boom to assist in orienting the spacecraft.

The first use of solar radiation pressure to control a spacecraft occurred during the Mariner 10 mission, during which a sequence of Mercury and Venus flybys were executed. Nearly three months into its cruise phase between Earth and Venus, instability in the attitude control loop was encountered during a sequence of spacecraft roll and scan platform articulation maneuvers.[11] The resulting oscillations depleted 0.6 kg, about 16%, of the spacecraft's nitrogen control gas over the course of an hour, prior to shutdown of the roll loop by mission controllers. Subsequent analysis showed the problem to be due to an unforeseen flexible-body effect, driven by energy input from the scan platform and the roll/yaw thrusters.

The roll/yaw thrusters were mounted on the tips of the solar panels to take advantage of the greater moment of force produced in this configuration. Preflight analysis had been done to alleviate concerns over potential excitation of the solar panels by the thrusters; the judgment was that only minimal interaction was possible due to the substantial difference between the ACS bandwidth and the primary solar array structural modes. Under certain flight conditions, however, it was found that higher order modes could be excited by the thrusters and that energy in these modes could couple into lower frequency modes that would alter the spacecraft body attitude. This would, of course, result in further use of the thrusters to correct the attitude error, followed by additional disturbances, etc., in a classic example of an unfavorable interaction between the structural and attitude control system designs.

In any event, various system-wide corrective measures were taken, and among them was a scheme to implement roll control by differentially tilting the separately articulated solar panels when necessary to implement a maneuver. The scheme worked well, albeit through intensive ground-controller interaction, and allowed sufficient fuel to be hoarded to carry the spacecraft through three encounters with Mercury.

7.7 Active Control

7.7.1 Feedback Control Concepts

The basic concept of active attitude control is that the satellite attitude is measured and compared with a desired value. The error signal so developed

is then used to determine a corrective torque maneuver T_c, which is implemented by the onboard actuators. Because external disturbances will occur, and because both measurements and corrections will be imperfect, the cycle will continue indefinitely. Figure 7.15 illustrates the process conceptually for a very simple single-input, single-output (SISO) system.

This is not a text on feedback control; the subject is too detailed to be treated appropriately here. Excellent basic references include texts by Dorf,[12] Saucedo and Schiring,[13] and Kwakernaak and Sivan.[14] Kaplan[5] and Wertz[2] include brief reviews of basic feedback control concepts oriented toward the requirements of spacecraft attitude control. Nonetheless, a cursory overview of control system design concepts is appropriate before discussing the various types of hardware that might be used to implement them onboard a space vehicle.

The reader will recognize that most of the system-level blocks in Fig. 7.15 are fixed either by mission requirements (e.g., desired attitude at a given time) or by the vehicle hardware itself. The control system designer can expect to have a major role in the selection of attitude control actuators and attitude measurement devices, but once this is done, he must live with the result. The only flexibility remaining lies in the "gain" block. At the undergraduate level, a course in feedback control is nothing more than an introduction to various methods for determining the appropriate gain K and analyzing the resulting performance of the system.

The gain block, or compensator, is a control law that specifies the magnitude of the correction torque to be applied in response to a given error measurement. Conceptually, this could be a correction factor that is a constant multiple of the error magnitude, i.e., a 1° error requires 2 N · m of correction torque, while a 2° error calls for twice as much restoring torque. Reality is rarely this simple, and often the required compensator is somewhat more complex. Nonetheless, useful insight can be obtained even assuming constant gain, as we will see in a subsequent section.

There are several basic performance parameters commonly of interest to the designer. Figure 7.16 shows the temporal response of a typical closed-loop control system to a unit step input. This illustrates the system's behavior in

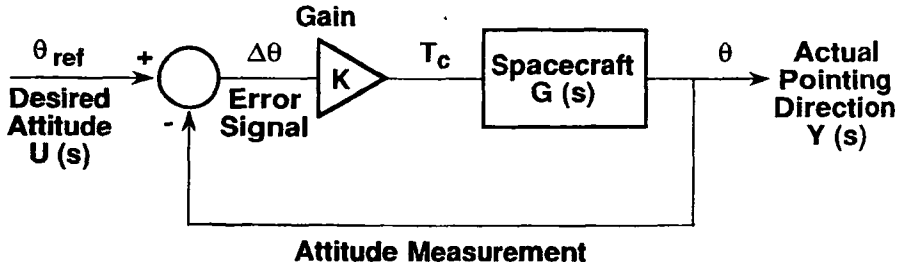


Fig. 7.15 Basic closed-loop control system block diagram.

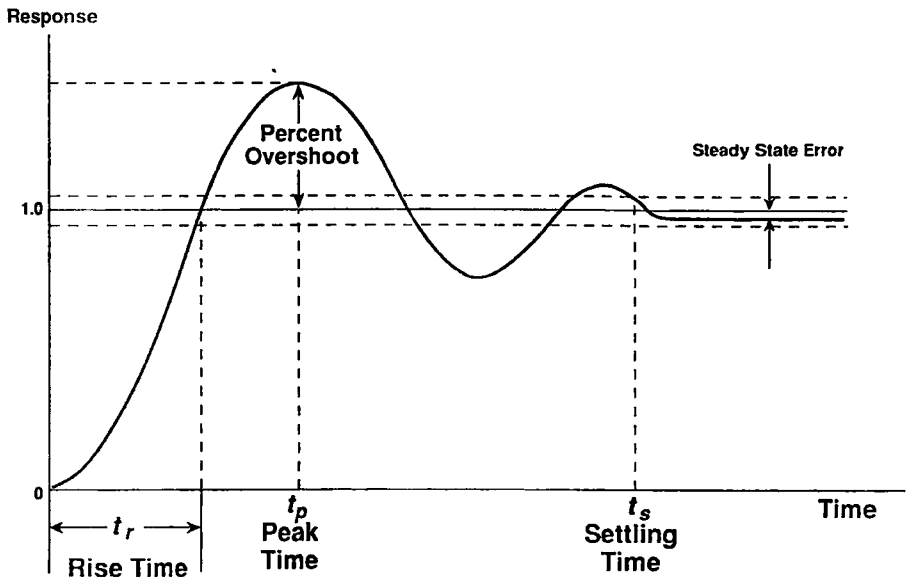


Fig. 7.16 Closed-loop control system response to step-function input.

response to a sudden disturbance at the input, such as an instantaneous shift in the desired attitude angle. The designer might have specifications for rise time, settling time, allowable overshoot, or allowable steady-state error. If the controller is expected to track a time-varying attitude profile, it would be of interest to examine the response to a ramp input. If it were necessary to follow an accelerating track, the system response to a parabola would be important. The requirement to follow more complex inputs requires control systems of correspondingly increased complexity.

The key to elementary control system analysis is that an electrical, hydraulic, or mechanical system (for us, the spacecraft) can usually be modeled over a useful operating range as a linear time invariant (LTI) system. Such systems can be represented mathematically by linear ordinary differential equations (ODE) having constant coefficients, an extremely useful property. Doebelin[15] provides an excellent treatment of the methods for and pitfalls of mathematically describing common physical systems. When there is only one variable to be controlled, such as the attitude of a single spacecraft axis, the system may be both LTI and SISO, and the design and analysis are relatively straightforward.

The advantage of describing a system with linear constant-coefficient ODEs lies in the utility of the Laplace transform[16] in solving such equations. The transformed differential equation is a polynomial, allowing the solution to be obtained with algebraic rather than integro-differential manipulations. The subsequent analysis of the input/output relationship for the system is greatly simplified in the transform domain.

Referring to Fig. 7.15 and employing standard notation, we define the Laplace-transformed "output," the actual pointing direction, as $Y(s)$ and the desired pointing direction or "input" as $U(s)$. The spacecraft or "plant" dynamics are represented as polynomial $G(s)$. The input/output relationship is then

$$H(s) \equiv \frac{Y(s)}{U(s)} = \frac{G(s)K(s)}{1 + G(s)K(s)} \tag{7.49}$$

where $H(s)$ is called the system *transfer function*. The time-domain signal $h(t)$, the inverse Laplace transform of $H(s)$, is the *impulse response* of the system. The denominator of the transfer function, $1 + G(s)K(s)$ as written here, is called the *characteristic equation*. Most performance characteristics of LTI SISO systems are determined by the locations of the system poles in the complex s domain (i.e., $s = \sigma + i\omega$). These poles are the roots of the characteristic equation, leading to the use of root locus techniques in the design and analysis of control systems. The polynomial degree of the characteristic equation is referred to as the *order* of the system. For example, the damped simple harmonic oscillator used to model many basic physical systems, such as a simple pendulum with friction, a mass-spring-dashpot arrangement, or a resistive-capacitive-inductive circuit, is the classic second-order system.

7.7.2 Reaction Wheels

Reaction wheels are a common choice for active spacecraft attitude control, particularly with unmanned spacecraft. In this mode of control an electric motor attached to the spacecraft spins a small, freely rotating wheel (much like a phonograph turntable), the rotational axis of which is aligned with a vehicle control axis. The spacecraft must carry one wheel per axis for full attitude control. Some redundancy is usually desired, requiring four or more wheels. The electric motor drives the wheel in response to a correction command computed as part of the spacecraft's feedback control loop. Reaction wheels give very fast response relative to other systems. Control system bandwidths can run to several tens of hertz.

Reaction wheels are fairly heavy, cumbersome, expensive, and are potentially complex, with moving parts. They are capable of generating internal torques only; the wheel and spacecraft together produce no net system torque.

With such a system, the wheel rotates one way and the spacecraft the opposite way in response to torques imposed externally on the spacecraft. From application of Euler's momentum equation, the integral of the net torque applied over a period of time will produce a particular value of total angular momentum stored onboard the spacecraft, resident in the rotating wheel or wheels, depending on how many axes are controlled. When it is spinning as fast as it can with the given motor drive, the wheel becomes "saturated," and cannot further compensate external torques. If further such torques are applied, the spacecraft will tumble. In practice it is desirable to avoid operation of a reaction wheel at

speeds near saturation, not only because of the limited control authority but also because of the substantial jitter that is typically generated by an electric motor operating at maximum speed.

Because reaction wheels can only store, and not remove, the sum of environmental torques imposed on the spacecraft, it is necessary periodically to impose upon the spacecraft a counteracting external torque to compensate for the accumulated onboard momentum. Known as "momentum dumping," this can be done by magnetic torquers (useful in LEO) or by control jets (in high orbit or about planets not having a magnetic field). Magnetic torquing as a means of momentum dumping is greatly to be preferred, because when jets are used, the complexities of a second system and the problems of a limited consumable resource are introduced. Indeed, in many cases when jets must be used, reaction wheels will lose much of their inherent utility, and the designer must weigh their drawbacks against their many positive features, among which are precision and reliability, particularly in the newer versions that make use of magnetic rather than mechanical bearings.

A reaction wheel operating about a given spacecraft axis has a straightforward control logic. If an undesirable motion about a particular axis is sensed, the spacecraft commands the reaction wheel to rotate in a countervailing sense. The correction torque is computed as an appropriately weighted combination of position error and rate error. That is, the more the spacecraft is out of position, and the faster it is rotating out of position, the larger will be the computed correction torque.

As long as all of the axes having reaction wheels are mutually orthogonal, the control laws for each axis will be simple and straightforward. If full redundancy is desired, however, this approach has the disadvantage of requiring two wheels for each axis, bringing a penalty in power, weight, and expense to operate the system. A more common approach today is to mount four reaction wheels in the form of a tetrahedron, coupling all wheels into all spacecraft axes. Any three wheels can then be used to control the spacecraft, the fourth wheel being redundant, allowing failure of any single wheel while substantially increasing momentum storage when all wheels are working. Thus, the system can operate for a longer period before needing to dump momentum.

Although reaction wheels operate by varying wheel speed in response to the imposition of external torques, that does not mean that the average speed of the wheels must necessarily be zero. The wheels can also be operated around a nominal low speed (possibly a few rpm) in what is called a momentum-bias system. The momentum-bias configuration has several advantages. It avoids the problem of having the wheel go through zero speed from, say, a minus direction to a plus direction in response to torques on the spacecraft. This in turn avoids the problem of sticking friction (stiction) on the wheel when it is temporarily stopped.

Because of the nonlinearity of the stiction term, the response of wheels to a control torque will be nonlinear in the region around zero speed, imposing a

jerking or otherwise irregular motion on the spacecraft as it goes through this region.[17] If this poses a problem in maintaining accurate, jitter-free control of the spacecraft, then the system designer may favor a momentum-bias system, which avoids the region around zero. As a disadvantage, the momentum-bias system lowers the total control authority available to the wheel before the saturation torque limit is reached, forcing momentum to be dumped from the spacecraft more frequently.

7.7.3 Momentum Wheels

When a reaction wheel is intended to operate at a relatively high speed (perhaps several tens of revolutions per minute), then a change of both terminology and control logic is employed. The spacecraft is said to possess a momentum wheel; a tachometer-based control loop maintains wheel speed at a nominally constant value with respect to the spacecraft body. This speed is adjusted slightly up or down in response to external torques. When the range of these adjustments exceeds what the control-system designer has set as the limit, momentum dumping allows the wheel speed to be brought back into the desired range. When magnetic coils are used to unload the wheel, this is done more or less continuously so that the tachometer circuit can operate around an essentially constant nominal value.

Use of a momentum wheel on a spacecraft offers the advantage of substantial gyroscopic stability. That is, a given level of disturbance torque will produce a much smaller change in desired nominal position of the spacecraft because of the relatively small percentage change it makes in the total spacecraft angular momentum vector. For this reason momentum-wheel systems are generally confined to use on spacecraft requiring a relatively consistent pointing direction. An example might be a low-orbit satellite where it is desired to have the vehicle angular momentum vector directed more or less continuously along the positive orbit normal, and to have the body of the spacecraft rotate slowly (i.e., 0.000175 Hz) to keep one side always facing the Earth. Use of a momentum wheel on the spacecraft aligned with its angular momentum vector along the orbit normal would be a common approach to such a requirement. The tachometer wheel control loop would function to keep the slowly rotating body facing correctly toward Earth.

The momentum-wheel system described here represents an attitude-control design referred to as a dual-spin configuration (Fig. 7.17). This configuration exists whenever a spacecraft contains two bodies rotating at different rates about a common axis. Then the spacecraft behaves in some ways like a spinner, but a part of it, such as an antenna or sensor on the outer shell, can be pointed more or less continuously in a desired direction.

As with simple spinners, dual-spin spacecraft require onboard devices such as jets or magnetic torquers for control of the overall momentum magnitude and direction. A reaction-wheel system offers the advantages of high-precision,

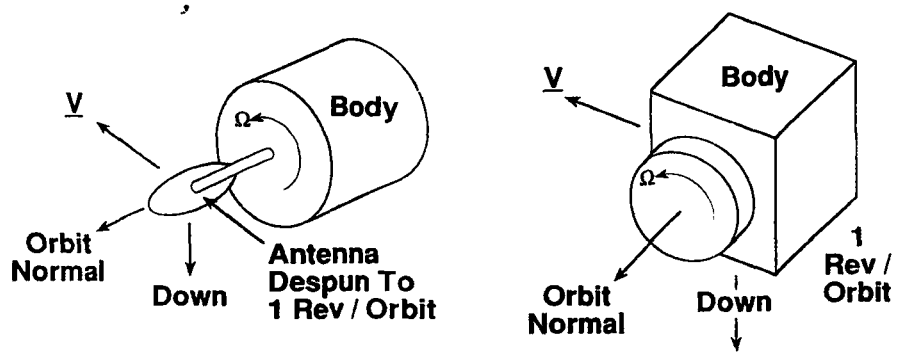


Fig. 7.17 Dual-spin spacecraft configuration concept.

independent control about all three spacecraft axes, whereas a simple spinner will have an extremely straightforward control-system design but minimal flexibility for pointing sensors or other devices on the spacecraft. Dual-spin design offers a combination of features from each—the dynamic advantages of simple spinners plus some of the precision pointing capability of a three-axis control system. Dual spin offers an even more important advantage under certain conditions: it allows relaxation of the major axis spin rule if the despun platform has more damping than the spinning portion of the spacecraft.[18,19]

As discussed in Section 7.4, no spacecraft can be characterized exactly as a rigid body. All objects will have some inherent flexibility, and under stress will dissipate energy. Objects with nonzero angular momentum thus tend to stabilize in a minimum-energy, flat-spin condition, i.e., rotating about the axis of maximum moment of inertia. But if the designer deliberately provides more energy dissipation on the despun platform than in the spinning portion of the spacecraft, the flat-spin condition may not be the only low-energy position of stable equilibrium. This can allow spinning the spacecraft about the minor axis of inertia with the assurance that, at least over some range, nutation angles resulting from disturbance torques will be damped rather than grow.

This important finding—arrived at independently by Landon[18] and Iorillo in the early 1960s—has resulted in many practical applications, particularly to geostationary communication satellites, because it ameliorates the configuration limitations imposed by common launch vehicles. In terms of its control dynamics, a spacecraft is better when pancake, as opposed to pencil-shaped. In contrast, most launch vehicles (prior to the space shuttle) foster a pencil-shaped spacecraft configuration. With the realization that the major-axis spin stability rule could be relaxed with the dual-spin design, appropriate distribution of damping mechanisms allowed a better match between launch vehicle shroud and spacecraft configuration requirements.

7.7.4 Control Moment Gyros

Momentum wheels can be used in yet another configuration, as control moment gyros (CMG). The CMG is basically a gimballed momentum wheel, as shown in Fig. 7.18, with the gimbal fixed perpendicular to the spin axis of the wheel.[20] A torque applied at the gimbal produces a change in the angular momentum perpendicular to the existing angular-momentum vector $\boldsymbol{H}$, and thus a reaction torque on the body. Control moment gyros are relatively heavy, but can provide control authority higher by a factor of 100 or more than can reaction wheels. Besides imposing a weight penalty, CMGs tend to be relatively noisy in an attitude control sense, with resonances at frequencies that are multiples of the spin rate.

In many applications not requiring the ultimate in precision pointing, however, CMGs offer an excellent high-authority attitude control mechanism without the use of consumables such as reaction gas. The most notable use of control moment gyros in the U.S. space program has been the Skylab spacecraft launched in 1973 and occupied by three crews of Apollo astronauts during 1973 and 1974. In more recent times, the Russian Mir space station program has used similar devices for vehicle stabilization, as does the International Space Station.

7.7.5 Magnetic Torquers

A spacecraft orbiting at relatively low altitude about a planet with an appreciable magnetic field can make effective use of magnetic torquers, particularly for initial attitude acquisition maneuvers and for dumping excess angular momentum from reaction wheels. They prove particularly advantageous when the burden of carrying consumables, such as fuel for reaction jets, would be an impediment in spacecraft design or when exhaust gas flowing from such jets

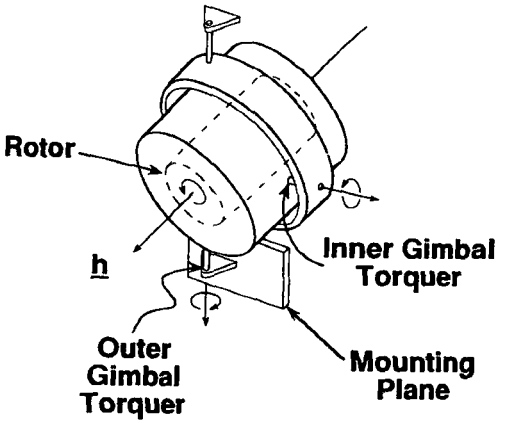


Fig. 7.18 Control moment gyro concept.

might contaminate or otherwise harm the spacecraft. A classic example in this regard, the HST, must have its primary mirror kept as clean as possible. As drawbacks, magnetic torquers have relatively low control authority and can interfere with other components on the spacecraft.

7.7.6 Reaction Jets

Reaction-control jets are a common and effective means of providing spacecraft attitude control. They are standard equipment on manned spacecraft because they can quickly exert large control forces. They are also common on satellites intended to operate in relatively high orbit, where a magnetic field will not be available for angular-momentum dumping. Offsetting these advantages, reaction-control jets use consumables, such as a neutral gas (e.g., Freon or nitrogen) or hydrazine in either monopropellant or bipropellant systems. Normally on/off operated, they do not readily lend themselves to proportional control, although that is possible by using pulse lengths of varying duration or a mix of control jets, not all of which need to be used in every situation. It is usually not acceptable to have only one jet functioning for a given control axis, because its failure will leave the spacecraft disabled in that axis. Thus, jet control systems usually require redundant thrusters, which leads to complex plumbing and control. Also, when attitude jets are used, there will likely be some coupling between the attitude and translation control systems. Unless a pure couple is introduced by opposing jets about the spacecraft's center of mass, the intended attitude control maneuver will also produce a small component through the spacecraft's center of mass. This will result in an orbital perturbation.

7.7.7 Summary

Table 7.1 summarizes several different methods of spacecraft control. The column labeled Accuracy should not be taken literally; the intent is to provide a one-significant-digit comparison among the various methods, rather than a definitive statement of achievable accuracies.

Table 7.1 Attitude control technique

Method	Accuracy, deg	Remarks
Spin stabilization	0.1	Passive, simple, low cost, inertially oriented
Gravity gradient	1–3	Passive, simple, low cost, central-body oriented
Reaction jets	0.1	Quick, high authority, costly, consumables
Magnetic torquers	1–2	Near-Earth usage, slow, lightweight, low cost
Reaction wheels	0.01	Quick, costly, high precision
Control moment gyros	0.1	High authority, quick, heavy, costly

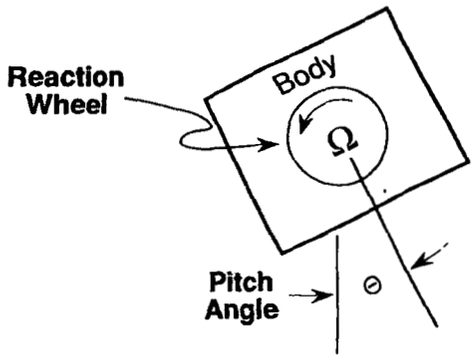


Fig. 7.19 Reaction wheel control of spacecraft pitch axis.

We conclude this section with a simple example, control of the pitch angle θ on a spacecraft through the use of a reaction wheel. As shown in Fig. 7.19, the reaction wheel is assumed to be aligned with the pitch axis and spinning at a relatively slow angular velocity Ω and to have moment of inertia J about its rotational axis. The spacecraft has moment of inertia I about the pitch axis and is assumed to be stabilized in roll and yaw, i.e., $\omega_\phi \cong \omega_\psi \cong 0$. Unknown external disturbance torques are presumed to act on the spacecraft. The rate of change of body angular momentum can be written as the sum of the wheel reaction torque and any external torques. Euler's equation under these circumstances is

$$\dot{H} = I\dot{\omega}_\theta = I\ddot{\theta} = T_{\text{wheel}} + T_{\text{ext}} \tag{7.50}$$

The wheel torque profile is chosen by the attitude control system designer. Obviously, the designer seeks a stable, controlled response to external torques as well as satisfaction of certain performance criteria such as mentioned earlier. Based on intuition and experience, he might choose a feedback control law for the wheel that is the sum of position- and rate-error terms,

$$T_{\text{wheel}} = -K_p\theta - K_r\omega_\theta = -J(\ddot{\theta} + \dot{\Omega}) \tag{7.51}$$

The constants K_p and K_r represent appropriately chosen position- and rate-feedback gains. Combining Eqs. (7.52) and (7.53) yields

$$\ddot{\theta} + \left(\frac{K_r}{I}\right)\dot{\theta} + \left(\frac{K_p}{I}\right)\theta = \frac{T_{\text{ext}}}{I} \tag{7.52}$$

which has the form of the classical damped simple-harmonic oscillator, with a driving or forcing term on the right-hand side. Upon taking the Laplace transform of Eq. (7.54), the characteristic equation is found to be

$$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0 \tag{7.53}$$

where ζ and ω_n are the damping ratio and natural frequency, given by

$$\omega_n^2 = \frac{K_p}{I} \tag{7.54}$$

$$\zeta = \frac{K_r}{2I\omega_n} \tag{7.55}$$

The selection of a particular damping ratio ζ and natural frequency ω_n forces the choice of gains K_r and K_p. This choice can be made to a certain extent at the discretion of the control-system designer, depending on which performance criteria are most important in a given circumstance. However, for such simple systems there are long-standing criteria by which the feedback gains are chosen to obtain the most appropriate compromise among the various parameters such as overshoot, settling time, etc.[12]

To complete the example, suppose we desire the settling time to be less than 1 s following a transient disturbance. The settling time is[12]

$$T_s \equiv 4\tau = \frac{4}{\zeta\omega_n} \leq 1\,\text{s} \tag{7.56}$$

hence $\zeta\omega_n \geq 4$ rad/s. According to the ITAE performance index (integral over time of the absolute error of the system response), optimal behavior is obtained for a second-order system when $2\zeta\omega_n = \sqrt{2}\zeta\omega_n$. Thus we find $\zeta = \sqrt{2}/2 \equiv 0.707$, and $\omega_n \geq 4/\zeta = 4\sqrt{2}$ rad/s. The system gains K_p and K_r are found from Eqs. (7.54) and (7.55). Higher values of ω_n than the minimum would produce a shorter settling time and might be desirable provided stability can be attained.

7.8 Attitude Determination

7.8.1 Attitude Determination Concepts

We now consider spacecraft attitude determination, the process of deriving estimates of actual spacecraft attitude from measurements. Note that we use the term "estimates." Complete determination is not possible; there will always be some error, as discussed in the introductory sections of this chapter.

ADCS engineers treat two broad categories of attitude measurements. The first, single-axis attitude determination, seeks the orientation of a single spacecraft axis in space (often, but not always, the spin axis of either a simple spinner or a dual-spin spacecraft). The other, three-axis attitude determination, seeks the complete orientation of the body in inertial space. This may be thought of as single-axis attitude determination plus a rotational, or clock, angle about that axis.

Single-axis attitude determination results when sensors yield an arclength measurement (see Fig. 7.7) between the sensor boresight and the known reference

point. The reference point may be the sun, the Earth nadir position, the moon, or a star. The crucial point is that only an arc-length magnitude is known, rather than a magnitude and direction. Specification of the axis orientation with respect to inertial space then theoretically requires three independent measurements to obtain a sufficient number of parameters for the measurement. In practice, the engineer often selects two independent measurements together with a scheme to choose between the true solution and a false (image) solution caused by the underspecification of parameters. The most common scheme entails using an a priori estimate of the true attitude and choosing the measurement that comes closest to the assumed value. Figure 7.20 illustrates the concept.

To effect the three-axis attitude determination requires two vectors that can be measured in the spacecraft body frame and have known values in the inertial reference frame. Examples of such potentially known vectors include, again, the sun, the stars, and the Earth nadir. The key lies in the type of sensor used to effect the measurement rather than in the nature of the reference point. The sensor must measure not merely a simple boresight error, as in single-axis attitude determination, but two angular components of the error vector. The third vector component is known since only unit vectors need be considered in spacecraft attitude control.

Consider that $\boldsymbol{u}$ and $\boldsymbol{v}$ are measured line-of-sight vectors to known stars in spacecraft body coordinates. We can define a unitary triplet of column vectors $\boldsymbol{i}, \boldsymbol{j}$,

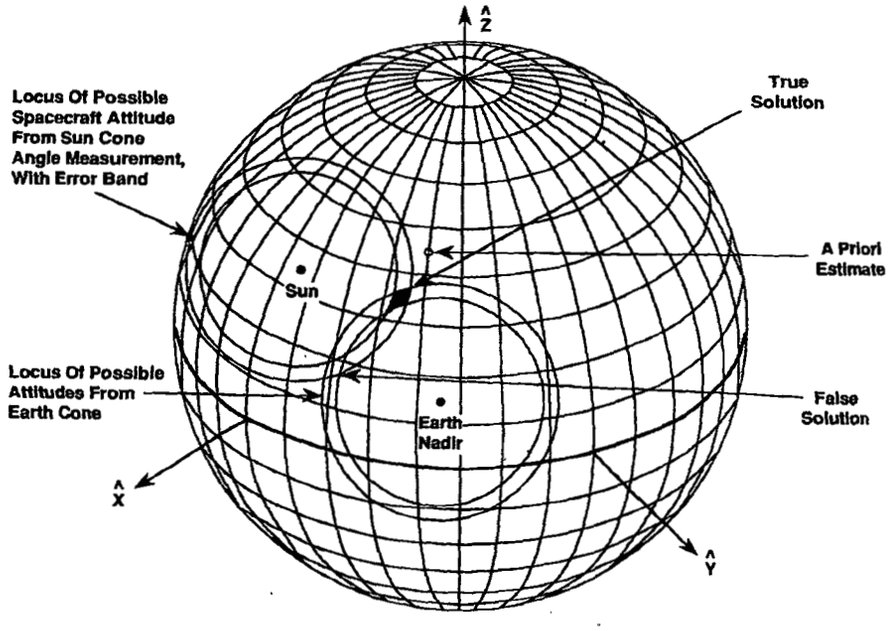


Fig. 7.20 Single-axis attitude determination.

and k, as

$$i = \frac{u}{|u|} \tag{7.57a}$$

$$j = \frac{u \times v}{|u \times v|} \tag{7.57b}$$

$$k = i \times j \tag{7.57c}$$

The unit vectors are measured in the spacecraft body frame but are also known, for example from star catalogs, in the inertial reference frame. The attitude matrix $T_{I \to B}$ introduced in Eq. (7.4) rotates the inertial frame into the body frame, and can be obtained as

$$[i\ j\ k]_b = T_{I \to B}[\, i\ j\ k]_i \tag{7.58}$$

or

$$T_{I \to B} = [i\ j\ k]_b [i\ j\ k]_i^{-1} \tag{7.59}$$

Once the attitude matrix $T_{I \to B}$ is available, Eq. (7.5) can be used to obtain the individual pitch, roll, and yaw angles. Recall from Eq. (7.5) that $T_{I \to B}$ contains redundant information on spacecraft orientation, and so the preceding system of equations is overdetermined. This allows a least-squares or other estimate[25] for spacecraft attitude, rather than a simple deterministic measurement, as the preceding equations would apply. Nonetheless, Eqs. (7.58) and (7.59) are useful to demonstrate the conceptual approach to three-axis attitude determination.

7.8.2 Attitude Determination Devices

The analytical approaches just discussed require measurements to be made as input data for the calculations. Attitude measurements are commonly made with a number of different devices, including sun sensors, star sensors, magnetometers, gyroscopes, and Earth-horizon scanners. These will be discussed briefly in the sections to follow.

Although much simpler devices are available, the sun sensors most commonly used on spacecraft are digital sensors, an example of which is shown in Fig. 7.21. A given sensor measures the sun angle in a plane perpendicular to the slit entrance for the sunlight. These typically will be used in orthogonally mounted pairs to provide a vector sun angle in body coordinates, as discussed earlier. Sun sensors can be used on either spinning spacecraft of despun three-axis stable spacecraft. The sensor depicted in Fig. 7.21 has nine bits of resolution. A variety of choices are possible in trading off this resolution between total dynamic range of the sensor and the precision of the measurement associated with the least-significant bit. For example, in designing a coarse acquisition sensor it might be useful to specify a range of $\pm 64°$ and an accuracy of $0.125°$ in the least-

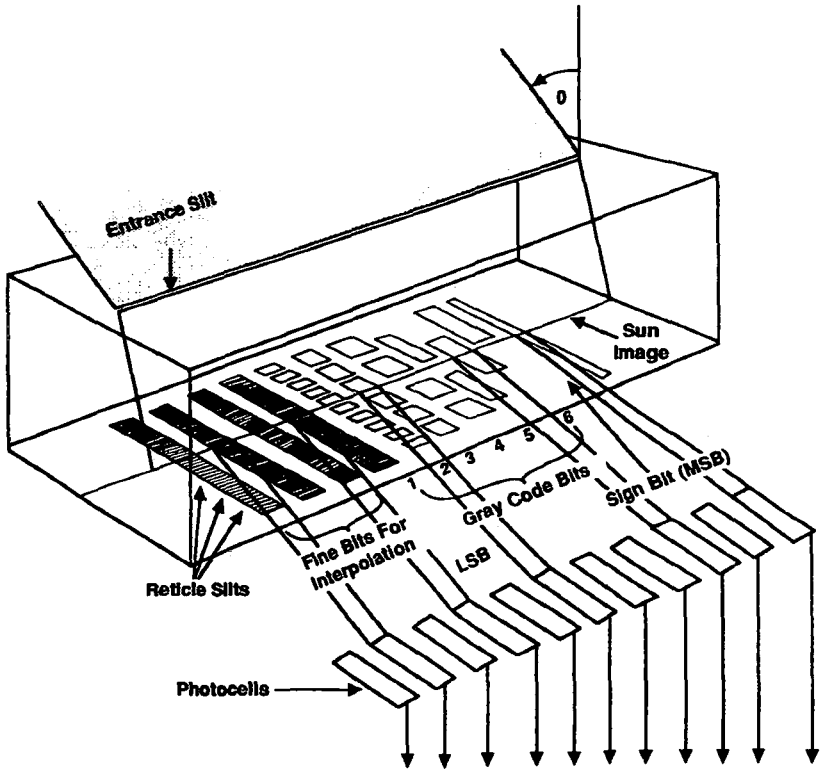


Fig. 7.21 Digital sun sensor.

significant bit. Such a sensor would not be useful for precision attitude determination measurements, but would provide a means of determining the spacecraft's attitude from a wide variety of initially unknown configurations. Four such sensors mounted around the spacecraft can provide essentially hemispherical coverage in terms of the ability to find the sun from an initially unknown position. The importance of such a capability to mission designers and to the operations team is obvious.

At the other end of the scale, sun sensors can be designed to yield a precision of a few arc-seconds in the least-significant bit, but at the price of a compromise in the overall dynamic range available with a given sensor.

For a satellite orbiting at an altitude below 1000 km, the Earth will subtend a cone angle greater than 120° as viewed from the spacecraft. The size of the target thus makes the Earth a tempting reference point for most LEO spacecraft attitude determination requirements. This is all the more true when, as is often the case, the fundamental purpose of placing the satellite in such an orbit is to observe a target on the Earth's surface or in its atmosphere. Earth-referenced attitude

determination schemes therefore make substantial engineering sense in such cases.

The most common means of determining the Earth nadir vector employs horizon scanners. With the position of the horizon defined on each side of the spacecraft, and to its fore and aft positions, the subsatellite point or nadir will be readily defined. Although a variety of sensors have been used in this application, the most common operate in the 14–16 μm infrared band. This is the so-called CO_2 band, characteristic of the carbon dioxide layer in the Earth's upper atmosphere. This relatively well-defined atmospheric band is usable both day and night, irrespective of the cloud layer. For these reasons the CO_2 band makes an especially good attitude reference within the Earth's atmosphere.

The 15 μm horizon varies by as much as 20 km from point to point on the Earth's surface, or between daylight and darkness, or at different times of the year. For low-orbiting spacecraft this alone produces an angle accuracy limit of around 0.05° in either pitch or roll. When the various known and normally modeled effects, including the Earth's oblateness (discussed in Chapter 4) are taken into account, angular accuracies can be on the order of 0.02–0.03° using Earth horizon sensors.

The most common types of Earth horizon sensors feature small scanners attached to a wheel oriented to rotate around the spacecraft's pitch axis. The scanners angle outward somewhat from Earth's nadir direction but not so far that they miss the outer edge of the Earth's horizon.

On each rotational scan in the pitch axis, therefore, each sensor will record a rising pulse and a falling pulse as the Earth's horizon is encountered going from cold space, across the Earth, and then back into cold space. Because the wheel speed is known and controlled by a tachometer circuit, the timing of these pulses, when compared against the reference time at which the pulses should occur, can be used to measure spacecraft pitch angle.

Additionally, if the scanners are mounted symmetrically on the spacecraft and if the roll angle is zero, then each scanner will have exactly the same duration between pulses. Any difference between the periods of pulse separation on either side of the spacecraft can be used to deduce the spacecraft roll angle. Figure 7.22 illustrates the geometry.

Spacecraft yaw angle, the rotational position around the radius vector from the Earth, cannot be determined using the measurements described earlier, because the Earth appears circular from the spacecraft no matter what the yaw angle might be. In inertial space, however, the yaw angle at a given moment is the same angle that will be observed as a roll angle one-fourth of an orbit later. Thus, spacecraft roll information can be used to estimate yaw, albeit on a very low bandwidth basis.

We have discussed horizon sensors for use in pairs, but a single scanner can provide excellent pitch information and adequate roll information. This allows graceful degradation in the event of a failure in one of the pair. However, extraction of roll information from one scanner requires knowledge of the orbital

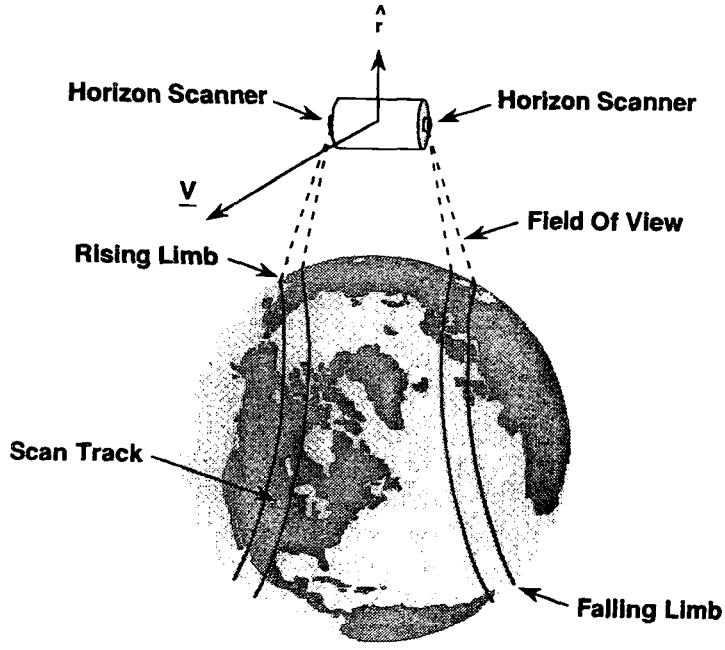


Fig. 7.22 Earth horizon scanner attitude determination concept.

altitude by the onboard control logic. This is required because the reference Earth width as seen by the scanner depends upon this altitude.

The use of sun sensors and Earth horizon scanners together can provide a very powerful attitude determination and control system for a LEO spacecraft. A system can be configured using these sensors in a scheme that uses a succession of single-axis attitude determination measurements. If the spacecraft carries a slightly more sophisticated computer, with sufficient capacity to store spacecraft orbital ephemeris data periodically uplinked from the ground, then more sophisticated processing is possible. The Earth nadir vector will be known in inertial space, as will the position of the sun at any time during the year. From the sensors, the two unit vectors necessary for three-axis attitude determination will be available in the spacecraft body frame. The spacecraft attitude can then be determined in the inertial frame and, because Earth's position is known, relative to the Earth as well. Of course, most low-orbiting satellites will have the sun available as a reference point during only part of the orbit around the Earth. For this reason, any onboard logic designed to use the sun vector as a reference must also be designed to cope with periods when the sun is not available.

A useful approach to compensating for the lack of a sun vector is to use an onboard magnetometer. Measurements are made of the three mutually orthogonal components of the ambient magnetic field. These components are then compared with the known reference components for that point in the orbit, as determined by standard magnetic field models. The difference between the measured

components in spacecraft body coordinates and the known components yields the spacecraft's rotational attitude. Various models of Earth's magnetic field describe minute variations from point to point over Earth's surface. These high-order models are mainly of scientific interest because the variability of the field from one time to another and because of small perturbing magnetic effects onboard the spacecraft render their use in attitude determination somewhat problematic. The U.S. repository for such models is the National Geophysical Data Center (NGDC) in Boulder, Colorado, and the repository is updated every five years. The version current as this is written, with an epoch of 2000, may be downloaded from that center's website.

The most common magnetic field model for use in spacecraft attitude determination is the so-called tilted-centered dipole model (Fig. 7.23). This can be expressed as

$$\begin{bmatrix} B_{\text{North}} \\ B_{\text{East}} \\ B_{\text{Down}} \end{bmatrix} = -(6378\,\text{km/r})^3 \begin{bmatrix} -C\phi & S\phi C\lambda & S\phi S\lambda \\ 0 & S\lambda & -C\lambda \\ -2S\phi & -2C\phi C\lambda & -2C\phi S\lambda \end{bmatrix} \times \begin{bmatrix} 29900 \\ 1900 \\ -5530 \end{bmatrix} nT \tag{7.60}$$

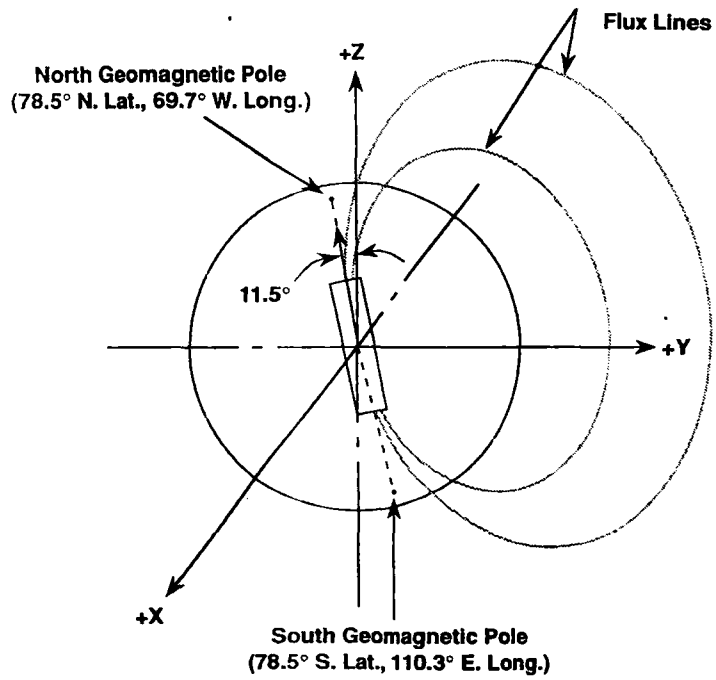


Fig. 7.23 Tilted-centered dipole model for Earth's magnetic field.

where r is the magnitude of the radius vector and (ϕ, λ) are the subsatellite latitude and longitude.

Equation (7.60) provides a vector known in inertial space as a function of the satellite's position in orbit and against which a measurement made in spacecraft body coordinates may be compared. The typical precision of a magnetic field based attitude determination measurement is on the order of $1-2°$.

The most accurate source for a reference vector to use in spacecraft attitude determination is a fixed star of known catalog position. Star trackers offer the potential of absolute attitude determination accuracy down to the order of approximately an arc-second, roughly the precision of most star-catalog data. However, this is obtained at relatively high cost, not only in dollars but also in power, weight, and onboard processing required to use the information returned by the tracker. Star trackers impose additional operational penalties in that they are obviously sensitive to light from the sun and reflected light from the Earth, the moon, and stray objects that may appear in the field of view.

Many different types of star trackers have been built. Gimballed trackers point at a star and maintain the star in a centered position. The star angles in body coordinates are then read from the gimbals. Such trackers offer great precision and a very wide effective field of regard, but use many moving parts and are quite cumbersome. They are thus rarely seen in current use. More commonly used today, so-called fixed-head star trackers scan the star field either electronically or by means of the spacecraft's motion. These trackers use no moving parts but have a relatively narrow field of view, on the order of $5-10°$.

The operational problem of star identification and verification is not trivial. Modern trackers employ processing logic internal to the tracker and can catalog stars according to both brightness and spectral type. The ADCS designer typically prefers to use only the brighter stars, both to minimize tracking error due to noise and to minimize potential confusion between similar stars in the catalog. However, to use stars sufficiently bright (third or fourth visual magnitude) to minimize confusion, the time gaps between appropriate star observations may be as much as 15 min to a half-hour, depending on the spacecraft orbit and the number and orientation of star trackers. If dimmer stars—down to seventh or eighth magnitude—are used, then there are nearly always appropriate stars in the field of view, but by the same token their identification and verification presents a problem. Moreover, even if the stars can be unambiguously identified, a much larger catalog will be required for onboard storage, at a substantially greater investment of time in loading and debugging the attitude determination software.

Long gaps between suitable star observations pose problems for an attitude control system not unlike those associated with sun outage in attitude determination using sun sensors. To address this problem, star trackers having a very wide field of view (e.g., 60° cone angle) have recently been developed. These trackers use only the brighter stars (e.g., third or fourth magnitude or brighter), but compensate for the existence of fewer such stars with the large field of view (FOV). Instead of merely supplying the coordinates of individual stars

within the FOV, companion software identifies specific "constellations" or known groups of stars (not the classical constellations of human experience). Using these star groups, a single tracker can identify both a boresight pointing vector and a rotational angle about that vector, thus providing full three-axis attitude determination with a single sensor.

As a practical matter, if the attitude control designer must handle gaps of several minutes or more between suitable stars, then gyroscopes will be needed on the spacecraft to provide an attitude reference during periods when no star is in the field of view of one of the trackers. Indeed, in such a case the preferred control algorithm will normally feature the use of rate gyros to measure the rotational rates around the various spacecraft axes. These rates will then be integrated over a period of time to establish the spacecraft's rotational position starting from a known baseline. When a suitable star is in the field of view of one of the trackers, or if Earth horizon scanners or sun sensors are also used, angular position updates are available to recalibrate the gyros. The system operates in this fashion more or less continuously, using gyros for high-bandwidth vehicle control and other attitude sensors to update the angular position reference. With this procedure, one is actually flying the spacecraft from a gyro reference platform and updating the navigational accuracy of the platform as targets of opportunity for attitude reference come into the field of view of one or more spacecraft sensors.

A classical gyroscope makes use of the fact that the angular momentum vector of a spinning body is constant unless an external torque is imposed on it. The gyroscope may be suspended in a nested set of gimbals and the whole assembly inserted into the spacecraft. The rotational motion of the spacecraft with respect to the gyroscope, which is fixed in space, is noted by measuring the change in gimbal angles with respect to the spacecraft. As with gimballed star trackers, this approach provides the highest accuracy available.

However, it is more common and nearly as accurate to use a "strapdown" platform. In such a platform the gyroscope is kept in more or less the same position relative to its assembly and to the spacecraft body by means of a control loop that supplies the torque necessary to keep the rotor in the correct position. The applied torque needed to maintain equilibrium is known by the electronic logic and used as a measure of the torques that have actually been imposed on the spacecraft body. Gyros may be configured to measure either rate or integrated rate (position), and may provide either one or two degrees of freedom for making a measurement.

Gyroscopes can provide extremely high bandwidth and extremely sensitive attitude or attitude-rate information. In many ways they make ideal components in an attitude-determination and control system. As noted earlier, they do require periodic recalibration from star trackers, sun sensors, or Earth horizon scanners, to remove the effects of drift and other systematic errors resulting from the fact that the gyro rotor cannot be maintained in truly torque-free motion. The models of highest accuracy are heavy, and all models have traditionally been rather expensive.

Gyro systems have been designed to last many years on orbit. However, it is fair to say that individual unit failures are not uncommon, and that system level robustness is generally obtained through the use of individually redundant units.

The limitations of conventional electromechanical gyroscopes have been well understood for many decades. For this reason, since the mid-1960s, there has been interest in replacing conventional gyros with inertial rate-measuring devices that use different underlying principles. These include ring-laser, laser fiber-optic, and hemispherical-resonator gyroscopes (the name "gyroscope" applied to such devices is a misnomer, but has wide currency).

Fiber-optic and ring-laser gyros provide a closed path around which laser light can be sent in opposite directions from a laser source. When the closed path (loop) is rotated relative to inertial space, the counter-rotating beams of light experience slightly different path lengths depending on which direction they have traveled. When the two beams are brought together at the end of the path, one will be slightly out of phase with respect to the other. The amount of the phase difference depends on the rotational rate applied to the closed loop.

The advantages of laser gyros have sparked development of them for various airborne and spacecraft applications. They have few moving parts, very high levels of reliability, and are inherently robust, capable of withstanding much rougher treatment than most mechanical systems. Fiberoptic gyros in particular can be made quite compact. For these and other reasons, aircraft and spacecraft designers have for some time been moving away from electromechanical gyros and toward such newer concepts. This is exemplified by the standard use of ring-laser gyros in aircraft such as the Boeing 757 and 767 series. Laser gyros were included as the baseline inertial navigation platform on the Orbital Sciences Corporation Transfer Orbit Stage as early as the late 1980s. The Near-Earth Asteroid Rendezvous (NEAR) mission to the asteroid Eros featured the first use of solid-state gyros on an interplanetary mission. Nevertheless, despite these successes, laser and other innovative gyro types have yet to see full acceptance in the aerospace industry, and electromechanical gyros can be expected to remain in heavy application for a long time.

The advent of Global Positioning System (GPS) navigation (see Chapter 11) offers the possibility, especially for large spacecraft in Earth orbit, of very precise and economical (in terms not only of money, but also of mass, power, and volume) attitude determination, along with basic spacecraft navigation. In this concept, several GPS antennas are mounted at different points on the spacecraft, dispersed as widely as possible. Using the techniques of differential GPS position estimation, the location (to within a centimeter or so of accuracy) of each antenna relative to the Earth-centered, Earth-fixed (ECF) coordinate frame is determined. These position estimates are immediately convertible to spacecraft attitude angles in the ECF frame, and because the GPS navigator also supplies the spacecraft ephemeris, transformation to the GCI frame is trivial. This technique is described more fully in standard references,[21] and, while not yet reduced to common practice, would seem to offer many advantages for future systems.

Table 7.2 Attitude determination techniques

Sensor	Accuracy, deg	Remarks
Sun sensor	0.01–0.1	Simple, reliable, cheap, intermittent use
Horizon scanner	0.02–0.03	Expensive, orbit-dependent, poor in yaw
Magnetometer	1	Cheap, low altitude use, continuous coverage
Star tracker	0.001	Expensive, heavy, complex, high accuracy
Gyroscope	0.01/h	Best short-term reference, costly

Table 7.2 summarizes the range of traditional attitude determination approaches discussed in this chapter and provides a rough quantitative level of accuracy obtainable for each.

7.9 System Design Considerations

Attitude determination and control system design does not end, nor does the designer's responsibility end, with the selection of measurement and control methods and components to implement them; that may indeed be the easiest part of the task. Many system-level considerations confront the ADCS designer. These must be addressed before the spacecraft can be ready to fly, or even before the ADCS can be properly related to the remaining subsystems aboard the spacecraft.

Throughout this text we have mentioned the difficulty of coping with the errors that are inevitable in translating any idealized approach into practice, or which arise as a consequence of simplifying the real world to a model for use in design and analyses. Nowhere is this truer than in attitude determination and control systems. For example, throughout this chapter it has been assumed that the spacecraft coordinate system will be anchored in the spacecraft's center of mass. This may be true in principle, because the coordinate system can always be redefined if necessary to fit the actual center-of-mass location; but the attitude control system actuators and the attitude determination system sensors will have been located with respect to the *intended* spacecraft-body-axis frame *supposedly* located in the center of mass.

Because offsets between the center of mass and the geometric center will always exist, it follows that there will inevitably be a coupling between attitude control and translational maneuvers. That is, thrusters intended to effect translational motion (to impart ΔV) will inevitably alter the attitude. This must be compensated by the attitude control system during ΔV maneuvers, and indeed it may well be that the attitude control system's overall control authority is determined on the basis of expected center-of-mass and center-of-thrust offsets for the ΔV system. Similarly, the use of attitude thrusters to control the rotational

position of a spacecraft will usually impart a residual ΔV to the spacecraft, probably causing an undesired orbital perturbation.

Another assumption throughout this chapter (and usually in the real world) is that of an essentially rigid spacecraft body to which are attached the various parts and components that make up the complete system. In practice no spacecraft structure or component attached to it will ever be perfectly rigid. This has obvious and well-understood effects with regard to the major-axis spin rule, as discussed earlier. Additional and more subtle effects are present, the most important being the potential for interaction between structural flexibility and the attitude determination and control system.[22]

This situation has been depicted conceptually in Fig. 7.24, showing a sensor mounted through a spring-like attachment to the spacecraft body. As the spacecraft moves in response to attitude control maneuvers, there will be relative motion between the attitude determination sensor and the spacecraft body. The sensor will not register the correct angle between the spacecraft's body axis and the desired reference source, and its error will be fed back into the spacecraft control system and used to cause an additional maneuver, which will again deflect the sensor relative to the spacecraft. The process continues ad infinitum. In short, the feedback signal has become corrupted by the relative deflection between the spacecraft body and the attitude determination sensor. The phasing of this error can cause the entire loop to become unstable. In practice, the spacecraft designer desires to remove the effects of structural flexibility to the maximum extent possible. This is done by using the most rigid mounting possible between various spacecraft components and also by using attachments that will provide as much damping as possible.

However, as spacecraft primary structures become larger and larger, the fundamental modal frequency inevitably becomes lower and lower, to a point where not uncommonly the fundamental structural modes fall near or within the

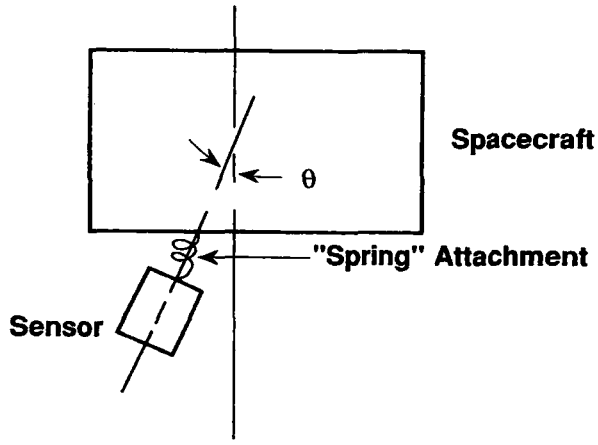


Fig. 7.24 Conceptual model for a spacecraft control-structure interaction.

desired attitude control system bandwidth. This is not an acceptable situation. The spacecraft fundamental structural modes and the attitude control system bandwidth must be separated to the maximum extent practical; this usually means a factor of five or more (and more is definitely better) separation between the ADCS passband and the lowest-order structural mode. Thus, a 2-Hz ADCS passband would require that the structural modes be at 10 Hz or above, and even greater separation is desirable when it can be achieved. If this cannot be done either by stiffening the spacecraft or lowering the control bandwidth, then the spacecraft structural modes themselves must be modeled as part of the overall attitude control loop. Development of control systems for flexible spacecraft has been an important research topic for several decades, and as spacecraft have become larger, it has acquired great practical moment. The literature is replete with examples of both theoretical and applied developments relating spacecraft attitude control and structural interactions.

So far, this chapter has mainly concerned design of systems to measure spacecraft attitude and to control it about some nominal point, which has implicitly been assumed to be constant. This in fact may be the easiest part of the mission operator's problem. In contrast, the question of how and in what way to maneuver the spacecraft between one desired attitude and another can be a major issue. Provision in the spacecraft design for such attitude maneuvers strongly effects spacecraft control through development of both onboard and ground operations software. At the very least, provisions must be made for initial acquisition maneuvers following deployment of the spacecraft from its launch vehicle, and delicate sensors or instruments cannot usually be allowed to point at the sun, or in many cases at the Earth or even the moon.

Furthermore, if the spacecraft has a very flexible operating profile, with more than one intended target of observation (such as would be the case, for example, with an astronomical telescope), then the designer's goal must include the ability to execute an equally flexible set of attitude maneuvers. The attitude-maneuver design must consider other spacecraft subsystems as well as the attitude system. A solar power system may need to have the arrays pointed toward the sun, and the arrays may need articulation individually to compensate for spacecraft maneuvers. Thermal radiators may need to be oriented toward dark space, or at least not toward the sun, while antennas may be required to point continuously toward Earth. As an excellent example of Murphy's Law, ground controllers want uninterrupted communication with the spacecraft at the exact time, during an attitude maneuver, when it is most difficult to achieve.

Similarly, the vehicle's scientific sensors or other systems often cannot look at the Earth or the sun without at least interrupting operation or, in the worst case, being irrepairably damaged. Furthermore, attitude maneuvers usually must be conducted with reasonable fuel efficiency. In other cases it may be required to minimize the time to execute a maneuver. Thus, design of optimal attitude maneuvers subject to known constraints on the spacecraft has been the subject of much theoretical and applied interest.[23,24]

We close this chapter with a few comments on testing. Few tasks are more challenging to the ADCS designer than testing his system. To be as realistic as possible, the system must be tested in essentially its operational configuration. However, the spacecraft is intended to operate in $0g$ and vacuum, not at $1g$ in the atmosphere. Air bearing tables can be useful in simulating $0g$ behavior, but cannot replicate the extremely low damping of the vacuum environment. It is also difficult (though not impossible for small systems) to devise an air bearing arrangement to allow simultaneous testing of all three axes. Attitude sensor inputs are also difficult to replicate on the ground, even in space simulation chambers. Given the uncertainties that are present, it is not surprising that getting the attitude determination and control system to behave properly is often the major task following initial deployment of the spacecraft in orbit.

References

[1]Hughes, P. C., *Spacecraft Attitude Dynamics*, Wiley, New York, 1986.

[2]Wertz, J. R. (ed.), *Spacecraft Attitude Determination and Control*, D. Reidal, Boston, MA, 1978.

[3]Noether, G. E., *Introduction to Statistics*, 2nd ed., Houghton Mifflin, Boston, MA, 1976.

[4]Taff, L. G., *Computational Spherical Astronomy*, Wiley, New York, 1981.

[5]Kaplan, M., *Modern Spacecraft Dynamics and Control*, Wiley, New York, 1976.

[6]Bracewell, R. N., and Garriott, O. K., "Rotation of Artificial Earth Satellites," *Nature*, Vol. 82, No. 4638, 1958, pp. 760–762.

[7]Mobley, F. F., and Fischell, R. E., "Orbital Results from Gravity-Gradient Stabilized Satellites," NASA SP-107, 1966.

[8]Thomson, W. T., "Spin Stabilization of Attitude Against Gravity Gradient Torque," *Journal of the Astronautical Sciences*, Vol. 9, 1962, pp. 31–33.

[9]Goldman, R. L., "Influence of Thermal Distortion on Gravity Gradient Stabilization," *Journal of Spacecraft and Rockets*, Vol. 12, No. 7, 1975, pp. 406–413.

[10]Tossman, B. E., Mobley, F. F., Fountain, G. H., Heffernan, K. J., Ray, J. C., and Williams, C. E., "MAGSAT Attitude Control System Design and Performance," AIAA Paper 80–1730, Aug. 1980.

[11]Dunne, J. A., and Burgess, E., *The Voyage of Mariner 10: Mission to Venus and Mercury*, NASA SP-424, 1978.

[12]Dorf, R. C., *Modern Control Systems*, 3rd ed., Addison-Wesley, Reading, MA, 1984.

[13]Saucedo, R., and Schiring, E. E., *Introduction to Continuous and Digital Control Systems*, MacMillan, New York, 1968.

[14]Kwakernaak, H., and Sivan, R., *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[15]Doebelin, E. O., *System Modeling and Response*, Wiley, New York, 1980.

[16]Wylie, C. R., and Barrett, L. C., *Advanced Engineering Mathematics*, 5th ed., McGraw-Hill, New York, 1982.

[17]Dahl, P. R., "A Solid Friction Model," Aerospace Corp., TOR-0158 (3107–18)1, El Segundo, CA, May 1968.

[18]Landon, V., and Stewart, B., "Nutational Stability of an Axisymmetric Body Containing a Rotor," *Journal of Spacecraft and Rockets*, Vol. 1. Nov.–Dec. 1964, pp. 682–684.

[19]Likins, P. W., "Attitude Stability for Dual-Spin Spacecraft," *Journal of Spacecraft and Rockets*, Vol. 4, Dec. 1967, pp. 1638–1643.

[20]O'Connor, B. J., and Morine, L. A., "A Description of the CMG and Its Application to Space Vehicle Control," *Journal of Spacecraft and Rockets*, Vol. 6, March 1969, pp. 225–231.

[21]Parkinson, B. W., and Spilker, J. J., Jr., *Global Position System: Theory and Applications I & II*, Vols. 163 and 164, Progress in Astronautics and Aeronautics, AIAA, Reston, VA, 1996.

[22]"Effects of Structural Flexibility on Spacecraft Control Systems," NASA SP-8016, April 1969.

[23]Li, F., and Bainum, P. M., "Numerical Approach for Solving Rigid Spacecraft Minimum Time Attitude Maneuvers," *Journal of Guidance, Control, and Dynamics*, Vol. 13, Jan.–Feb. 1990, pp. 38–45.

[24]Byers, R. M., Vadali, S. R., and Junkins, J. L., "Near-Minimum Time, Closed-Loop Slewing of Flexible Spacecraft," *Journal of Guidance, Control, and Dynamics*, Vol. 13, Jan.–Feb. 1990, pp. 57–65.

[25]Gelb, A., *Applied Optimal Estimation*, MIT Press, Boston, MA, 1974.

Problems

7.1 A rigid body has angular velocity $\boldsymbol{\omega} = (10, 20, 30)^T$ rad/s in body coordinates. The inertia matrix is

$$I = \begin{bmatrix} 20 & -10 & 0 \\ -10 & 30 & 0 \\ 0 & 0 & 40 \end{bmatrix} \text{kg} \cdot \text{m}^2$$

(a) What is the angular momentum of the body about its center of mass?
(b) What is the rotational kinetic energy about the center of mass?
(c) What are the principal-axes moments of inertia?

7.2 Given that the rotational kinetic energy of a rigid body about its center of mass is

$$T_{\text{rot}} = (25\omega_x^2 + 34\omega_y^2 + 41\omega_z^2 + 24\omega_y\omega_z)/2$$

where x, y, and z are a known body-fixed frame,
(a) What are the principal moments of inertia?
(b) Calculate the angles between (x, y, z) and the principal axes (1, 2, 3).
(c) What is the magnitude of the angular momentum?

7.3 Consider a rigid body with (in body coordinates) $\boldsymbol{h} = (200, 200, 400)^T$, $\boldsymbol{\omega} = (10, 10, 10)^T$, and inertia matrix

$$I = \begin{bmatrix} 30 & -I_{xy} & -I_{xz} \\ -10 & 20 & -I_{yz} \\ 0 & -I_{zy} & 30 \end{bmatrix} \text{kg} \cdot \text{m}^2$$

(a) What are the inertia moments I_{xy}, I_{xz}, I_{yz}, and I_{zy}?
(b) What is the moment of inertia about the spin axis (axis of $\boldsymbol{\omega}$)?
(c) What is the rotational kinetic energy?
(d) What are the principal moments of inertia?

7.4 An earlier version of the International Space Station featured as one assembly milestone the so-called man-tended capability (MTC) configuration. In this version, the station would have flown in a highly asymetrical configuration, with a long truss having a habitation module at one end and a solar array at the other. Assume the body z axis to be from the habitation module outward along the truss, the y axis in the plane of the solar array, and the x axis to complete a right-handed set. Unit vectors for the body (x, y, z) frame are $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$. The station inertia matrix at MTC was to be

$$I = \begin{bmatrix} 2.5 & 0 & 0 \\ 0 & 2.6 & 0 \\ 0 & 0 & 0.32 \end{bmatrix} \times 10^7 \text{ kg} \cdot \text{m}^2$$

When the solar array is perpendicular to the velocity vector, the vector to the aerodynamic center of pressure was $\boldsymbol{r}_{cp} = 22 \text{ m } \boldsymbol{k}$. Assume the station to be in a 400-km circular orbit, with standard atmospheric density at this altitude of $\rho = 2.8 \times 10^{-12}$ kg/m^3. The projected area of the station with its array normal to the velocity vector is 600 m^2, with a drag coefficient of $C_D = 2$ assumed.

(a) What is the nominal orientation of the station in the absence of aerodynamic or solar radiation pressure torques?
(b) In actuality, aerodynamic drag would somewhat offset the ideal orientation attained in (a). What is the stable torque-equilibrium attitude (TEA) that results in this case? For simplicity, you may assume the yaw and roll angles to be zero, i.e., motion is in the pitch plane only.

7.5 A drum-shaped Earth orbiting spacecraft (see Fig. 7.13) of radius 1 m is spin stabilized with $H = 2000$ Nms and has a spin-axis moment of inertia $I_z = 500 \text{ kg} \cdot \text{m}^2$. It is nominally aligned with the spin axis along either the

positive or negative orbit normal direction, depending on the considerations discussed in the following. The spacecraft is designed to radiate heat to dark space out of the "bottom" side. Because of the nodal regression of the orbit, and the consequent time-varying angle between the sun vector and the orbit plane, it is necessary several times per year to "invert" the spacecraft, i.e., to precess the spin axis around to the opposite orbit normal direction, so that the sun does not shine into the bottom side. The spin axis attitude is controlled by four 50-N thrusters mounted on the rim of the spacecraft. The thrusters are operated in pulse mode, with a pulse width of 40 ms, followed by a 60-ms off-period. How many thruster pulses are required to accomplish this? How long does the process take? Because of the short pulse, we may ignore any cosine losses associated with finite pulse width.

7.6 The spacecraft outlined in problem 7.5 has thrusters that are slightly misaligned, a consequence of which is that a coning motion develops during the described yaw inversion maneuver. The spacecraft is symmetric about the spin axis, with $I_x = I_y = 300\,\text{kg} \cdot \text{m}^2$ and, as stated in problem 7.5, $I_z = 500\,\text{kg} \cdot \text{m}^2$. A precession rate of 11 rad/s is developed; what is the precession angle?

7.7 When Voyager 2, a three-axis stabilized spacecraft with 0 angular momentum (see Fig. 8.1), was in cruise phase just past the orbit of Saturn (10 A.U. from the sun), controllers began to notice a small but persistent attitude motion drift away from its nominal solar-inertial attitude. (Voyager was flown with its dish antenna oriented essentially toward the sun; this automatically kept Earth within the antenna beamwidth.) The spacecraft showed a consistent pattern of drifting away from an initially stationary position at one side of its control "deadband" to the control limit on the other side of the deadband. At this point, a small thruster would be fired to bounce the spacecraft attitude back toward the other side of the deadband. The deadband was 4° wide, and the drift motion required two hours to go from the stationary position at one limit to the opposite limit. What was the net external torque on the spacecraft if its moment of inertia in the plane of the motion was $500\,\text{kg} \cdot \text{m}^2$? It was assumed that solar radiation torque was the culprit, because of Voyager's unsymmetrically mounted booms. If so, and assuming a projected area of $10\,\text{m}^2$ with 50% reflectivity, what was the effective center-of-mass to center-of-pressure (CM-CP) offset? At 1 AU, $p_s = 4.4 \times 10^{-6}\,\text{N/m}^2$.

7.8 For readers with a background in undergraduate-level control systems design: A missile-tracking spacecraft in low Earth orbit is to use reaction wheels to slew the spacecraft as it follows its target in flight. In the plane of the maneuver, the slewing control scheme can be represented by the block diagram of Fig. 7.15, with $G(s) = 1/Is^2$, and $I = 100\,\text{kg} \cdot \text{m}^2$.

It is desired that, after an initial slew to track the target, the control system should achieve the following goals with regard to settling time and peak overshoot:

$$t_s < 0.1 \text{ s (5\% criterion)}$$
$$M_p < 0.05$$

With performance specifications such as these, we elect to use a proportional-plus-derivative compensator of the form

$$\kappa(s) = \kappa_r s + \kappa_p$$

where κ_p and κ_r are position and rate feedback gains, respectively.

(a) Write the system transfer function in terms of gains (κ_p, κ_r) as well as damping ratio and natural frequency (ζ, ω_n).
(b) What are the values of (κ_p, κ_r) and (ζ, ω_n) needed?
(c) Is the system stable for the values of (ζ, ω_n) chosen? [If you did not solve part (b), use values of (0.75, 45) for (ζ, ω_n).] Why?
(d) Sketch a root locus diagram for the compensated system. For what values of gain is the system unstable? (Hint: Look for shortcuts.)
(e) What are the steady-state position, velocity, and acceleration errors (not the static error constants)?
(f) Sketch a Bode plot of the uncompensated and compensated systems. What are the approximate phase and gain margins of the compensated system?
(g) Assuming the natural frequency of the system to be an effective measure of the bandwidth, what lower bound on spacecraft modal frequencies would you, as the control system designer, specify to the structural engineer? Why? Assume $\omega_n = 45$ rad/s if you did not solve part (b).
(h) What are the open-loop and system transfer functions if phase-lead compensation is chosen instead of proportional-plus-derivative?
(i) Where would you place the dominant poles for a phase-lead compensator design using a root-locus approach?

7.9 A torque-free semirigid spacecraft is in a circular orbit in a nominally stable attitude (i.e., rotating at $\Omega = 1$ rev/orbit, axis of rotation aligned with the positive orbit normal, and minimum-inertia axis aligned with the radius vector). The spacecraft maximum inertia axis is also an axis of symmetry. Its attitude is described in terms of roll, pitch, and yaw angles (ϕ, θ, ψ). These angles are defined relative to a quasi-inertial, local-vertical, local-horizontal (LVLH) frame (i, j, k) that rotates at rate Ω. Assume $\omega_1(t_o = 0) = 0$. Further assuming that (ϕ, θ, ψ) as defined here are small and also that all angular rates are small, what is the time history of (ϕ, θ, ψ) *in this coordinate frame*?

7.10 A spacecraft executes two sequential attitude maneuvers in a roll-pitch-yaw (ϕ, θ, ψ) Euler angle sequence, starting from $\phi = \theta = \psi = 0$. The first maneuver is a roll of 0.1 rad, a pitch of 0.05 rad, and a yaw of 0.02 rad. The second is a maneuver of 0.04 rad in roll, 0.1 rad in pitch, and 0.03 rad in yaw. What is the approximate final attitude relative to the inertial frame?

7.11 A spacecraft is designed to fly in low Earth orbit with a conventional roll-pitch-yaw (ϕ, θ, ψ) attitude reference system. Roll is about the x axis, aligned with the velocity vector. Pitch is about the y axis, aligned with the positive orbit normal. Yaw is about the z axis, approximately aligned with the local vertical, since $\boldsymbol{k} = \boldsymbol{i} \times \boldsymbol{j}$. The rotation sequence taking the inertial frame into the body frame is always assumed to be in the roll-pitch-yaw (RPY) order. Assume the body frame is initially aligned with the inertial or reference frame, after which we perform an attitude maneuver $(\phi, \theta, \psi)_1 = (20, -15, 10)°$, followed by a second rotation with $(\phi, \theta, \psi)_2 = (15, -20, 10)°$. What is the equivalent RPY single maneuver for the compound rotation?

7.12 The shuttle is in a low circular orbit such that, in body coordinates, a line-of-sight (LOS) vector to a reference star is $\boldsymbol{u}_b = (0, 1, 0)$, and to the sun is $\boldsymbol{v}_b = (0, 0, 1)$. (Shuttle body coordinates have the x axis positively aligned with the nose, the y axis positive along the left wing, and the z axis positive out of the cargo bay.) Meanwhile, for the particular orbit and position that the shuttle occupies at the moment of the observation, the same star and the sun have LOS vectors of $\boldsymbol{u}_i = (s, c, 0)$ and $\boldsymbol{v}_i = (0, 0, 1)$ in the inertial or reference frame, where $c = \cos 28.5°$ and $s = \sin 28.5°$. (The reference frame is centered in the shuttle, with x along the velocity vector, y along the positive orbit normal, and z in the direction of the cross product of x and y.) What are the yaw, pitch, and roll angles (ϕ, θ, ψ) for the shuttle at this time?

7.13 Assume for the sake of this problem that the North Star, Polaris, is the star being utilized for the observations in Problem 7.12, and that Polaris is in fact directly above the Earth's North Pole. The attitude observation is made exactly at the moment of vernal equinox. What is Ω, the longitude of the ascending node, for this orbit? What is the orbital inclination?

7.14 The Shuttle orbiter inertia matrix is given by

$$I = \begin{bmatrix} 1.3 & 0 & 0 \\ 0 & 9.7 & 0 \\ 0 & 0 & 10.1 \end{bmatrix} \times 10^6 \text{ kg} \cdot \text{m}^2$$

The shuttle coordinate system has its origin at the CM and the positive x axis toward the nose, the positive y axis in the plane of the right-hand wing, and the positive z axis downward through the belly. What are the two stable attitudes for the shuttle in low circular orbit, and why?

7.15 Estimate the various disturbance torques on the space shuttle in a 300-km altitude circular orbit. Assume the payload carried in the cargo bay results in an overall CM-CP offset of -0.3 m in the x axis, or longitudinal, direction, when the shuttle is flying with the nose oriented to the local vertical and the belly into the wind, the worst-case aerodynamic drag configuration. Assume a 1.0 m CM-CP offset for worst-case solar radiation pressure analysis.

8
Configuration and Structural Design

8.1 Introduction

Of all the subsystem areas discussed in this book, configuration design may most closely approximate systems engineering as a whole. The configuration designer must be involved in detail with every other subsystem in the spacecraft. The configuration must accommodate all the disparate requirements and desires of the various subsystems and, where those are in conflict, reach a suitable compromise. For a complex spacecraft, the wide variety of requirements, desires, and constraints and the conflicts that inevitably arise among them provide a substantial challenge. This fact is not new to mechanical and structural designers, but space applications do present a set of challenges extending well beyond the reach of conventional ground-based techniques.[1] A variety of innovative solutions have evolved in various projects. These will be discussed in some detail, not as final answers, but simply as examples of working solutions.

8.2 Design Drivers

Before discussing solutions, we need to understand the factors that drive the design. It can be stated as an axiom that configuration design is *always* a compromise. A variety of requirements, which invariably involve some conflict, drive the design of every spacecraft. As with any complex system, there usually exists a variety of solutions, each of which can result in a more or less satisfactory design. As a result, this section will discuss the design drivers and some of the considerations involved in developing solutions.

8.2.1 Mission Goals

A variety of typical missions can be listed to illustrate the types of missions that can be carried out by any spacecraft. The common generic classes of spacecraft missions and goals are 1) communications relay; 2) Earth observation, which includes civilian, military, high-altitude, and low-altitude observations; 3) solar observation; 4) astronomical; 5) fields and particles; and 6) planetary observation, including flybys, orbiters, and landers. It may be possible to think of

missions that do not precisely fit this list, but the general characteristics that are encompassed in the list will cover most cases.

8.2.1.1 Communications satellites. Communications satellites have historically been located almost exclusively in geostationary orbit because of the wide area of coverage available and the simplicity of communicating with an object that remains stationary in the sky. Because ground stations require no tracking capability, construction and operating costs are substantially reduced. The need merely to point accurately in one direction and perform only a relatively simple relay function also simplifies the spacecraft. On the other hand, the very large number of channels handled by a modern communications satellite effectively complicates the avionics design, while the large investment involved and the importance of the function dictates very high reliability and long life.

Recently, interest has grown in using networks of low-altitude satellites that replace the geostationary type. The low-altitude constellations may offer lower unit costs, but require a very large number of satellites and some increase in operational complexity. The major advantage of this approach is in its robustness. Loss of a substantial percentage of the satellites will result in a degraded but functional system, whereas loss of a single large satellite will shut down the entire system.

Because much of its territory lies at the high latitudes poorly served from geostationary orbit, Russia has evolved the Molniya communications spacecraft. These spacecraft operate in highly elliptic synchronous orbits oriented so that the apoapsis is located over the regions of interest. Thus, the spacecraft spends most of its time above the horizon as viewed from Russia. As discussed in Chapter 4, by selecting an inclination of 63.5 deg it is possible to prevent the line of apsides from precessing. Thus the desired orientation relative to the Russian landmass is maintained throughout the year.

8.2.1.2 Earth observation satellites. With the possible exception of communications satellites, Earth observation satellites are probably the most common general type of spacecraft currently in existence. Both military and civilian versions exist. Operating orbits range from low circular through elliptic to geostationary and beyond. Even though they all carry out the same generic function (i.e., they observe the Earth and its near environment), the variety of spacecraft is huge, with many types of sensors operating in a variety of wavelengths. Most are passive, i.e., they conduct their observations using naturally emitted radiation. A few conduct active observations using radar or lidar (laser radar).

The type of Earth observation spacecraft most familiar to the casual observer is the weather satellite. Both military and civilian agencies operate networks of these spacecraft. The military and civilian spacecraft are generally similar, although specific requirements may result in some differences in sensors or operations. Two generic types of weather satellites exist: those located in geostationary orbit to provide wide area coverage (almost 40% of the Earth from

one satellite) and those in low circular polar orbit that provide high-resolution data, but over smaller viewing arcs. The latter are usually in sun-synchronous orbit (see Chapter 4) so that a given locality is viewed at the same local time (hence sun angle) each day. The low-orbit spacecraft are generally nadir-pointing, or nearly so (see Chapter 9), unless their mission is to scan the upper regions of the Earth's atmosphere, in which case their primary field of regard will likely be the Earth's limb. The geosynchronous types are likely nadir-pointing, or spin-stabilized with despun platforms or spin-scan instruments.

Military reconnaissance satellites constitute a large percentage of Earth observation spacecraft. Some of these are at high altitude for wide area surveillance, whereas others operate at low altitude to obtain the best resolution. Among the latter are some of the highest-resolution spacecraft imagers yet flown. Actual performance is classified, but open literature discusses cases in which specific individual aircraft have been identified by tail number. In some cases, the spacecraft descend to relatively low altitudes to improve resolution. Circular orbits at such altitudes would not be stable; therefore, the spacecraft operate in elliptic orbits with very low periapsis altitudes to allow time to raise the apogee periodically, thus compensating for the drag that would result in a quick reentry from a circular orbit. This same strategy was used by NASA's Atmospheric Explorer series of satellites in the 1970s. Perigee altitudes below 150 km were used to allow direct sampling of the upper atmosphere, with the lower limit set by the allowable heating rate and the need to control drag sufficiently well to avoid premature reentry.

In recent years, a number of private commercial companies have been formed for the purpose of offering imaging of 1-m or better resolution, the limit (for U.S. companies) set by the U.S. government, and comparable to the resolution of military reconnaissance satellites. The commercial availability of images of such high resolution has caused some consternation among the military of various nations, because other nations without space capability can now purchase military-quality reconnaissance data. This has led to the desire to prevent imaging of certain critical areas, a concept that is essentially impossible to enforce on a global basis, especially as spaceflight capability becomes more broadly available.

Earth resources satellites such as the U.S. Landsat and the French Satellite Probatoire d'Observation de la Terre (SPOT) are invaluable for the study of the surface composition of the Earth. Both scientific and commercial interests are served by the data from these spacecraft, which generally employ sensors operating in a variety of spectral bands. Again, near-polar sun-synchronous orbits are most commonly used.

Also of interest has been the release, beginning in the late 1990s, of much formerly classified imagery from early strategic and military reconnaissance programs, especially the National Reconnaissance Office (NRO) Corona program of the early 1960s. These data have been and will continue to be of great value in assessing global change over a multidecade span.

8.2.1.3 Solar observation. Solar observation is among the oldest disciplines in space science, going back to the sounding rocket observations that began just after World War II. The advantages to solar observation of eliminating atmospheric filtering are obvious. For some observations it is desirable to get away from the Earth altogether; thus, many solar observation spacecraft have been in solar rather than Earth orbit. The sun emits huge amounts of energy in all wavelengths from infrared to x-ray, plus considerable particulate radiation. Thus, the sensors for solar observation are by no means restricted to optical wavelengths. Such phenomena as the decay of solar-emitted neutrons make it necessary to approach as close to the sun as possible if those particles are to be detected. To date, no spacecraft has come much closer than the orbit of Mercury, but several mission concepts have been studied for grazing or impact missions. An interesting possibility, applied to the International Sun-Earth Explorer (ISEE) mission and various subsequent spacecraft, is to place the spacecraft in a "halo" orbit about the libration point (see Chapter 4), thus locating the spacecraft near the line between the Earth and the sun but slightly offset from it. The halo orbit about the libration point allows Earth-based antennas to view the spacecraft without the sun, an overwhelming noise source, in the antenna field of view. As viewed from Earth, the spacecraft appears to circle around the sun, thus the name "halo orbit."

8.2.1.4 Astronomical. With few exceptions (such as the 1970s International Ultraviolet Explorer and the recent Chandra x-ray telescope), astronomical spacecraft have operated in low Earth orbit. Observations in the infrared, visible, and ultraviolet are of interest. Some instruments used for broad sky surveys will have relatively generous pointing constraints, whereas others designed for detailed observation will have extremely tight constraints. The Hubble Space Telescope is a case in point, requiring the most difficult pointing accuracy (approximately 0.01 arcseconds) and stability (approximately 10^{-5} arcseconds) yet flown. The reason for discussing this topic, seemingly more relevant to attitude control, here is that pointing accuracy and stability constraints translate into alignment accuracy and control requirements on the spacecraft structure, which will be strong drivers on configuration, structural design, and material choice and, above all, cost.

8.2.1.5 Fields and particles. Spacecraft devoted to the observation of magnetic fields and particulate radiation are generally less concerned with accurate pointing than are other types of spacecraft. In many cases, a rotating spacecraft is desired to allow widespread coverage of the sky. Spacecraft designed to conduct this type of investigation (as well as those requiring high-accuracy pointing) often have some difficulty meeting multiple and competing requirements and desires. For example, during interplanetary cruise the three-axis

stabilized Voyager spacecraft were occasionally commanded to perform a roll-and-tumble sequence to provide the fields and particles payload with a survey of the celestial sphere.

8.2.1.6 Planetary observation. Spacecraft designed for planetary observation from orbit differ little from their counterparts at Earth except for requirements edicted by differing environments. Some planetary spacecraft will be on flyby rather than orbital missions. In such a case, a scan platform for narrow field of view instruments is highly desirable if not mandatory. This allows multiple scans and photomosaic generation, which would be very difficult to accomplish by maneuvering the entire spacecraft during the few minutes available in a typical encounter. Planetary landers, of course, require aerodynamic deceleration and/or rocket propulsion for descent and landing.

8.2.2 Payload and Instrument Requirements

The requirements that may be levied on the spacecraft by the payload are 1) location, 2) pointing accuracy, 3) temperature, 4) magnetic field, 5) radiation, and 6) field of view. This list primarily addresses a payload of observational instruments, but many of the requirements are typical of essentially any payload.

Payload items may demand a specific location on the spacecraft to meet the other requirements listed. This can often be a problem when more than one instrument wants the same piece of spacecraft "real estate," or when the requirement conflicts with those of other subsystems.

Pointing accuracy requirements can drive configuration and structural design far more substantially than might appear to the casual observer. For example, stringent requirements may dictate extreme rigidity and temperature stability to minimize distortion in alignment between the instrument mount and the attitude control reference. This can in turn dictate structural design, material choice, and configuration design.

Many payload elements have delicate components with relatively tight temperature constraints. This will require attention but is usually not a major design driver. However, when a particular sensor requires very low temperature, as is often the case with infrared sensors, the need to provide a clear view of space while eliminating the sun, planetary surfaces, or illuminated or hot spacecraft parts from the radiator field of view can pose a major design problem.

In some cases a magnetically sensitive component can simply be shielded from spacecraft-generated magnetic fields and thus will not offer any particular configuration problem. However, if the component is a sensor for detecting and measuring planetary magnetic fields, it must be isolated from the spacecraft fields without compromising its function. Generally, the answer is distance, often a fairly large distance. This in turn usually dictates some sort of deployable structure.

Many components are sensitive to radiation dosage. Although shielding is possible, it requires added mass, the anathema of the space systems engineer. Clever configuration design may be called upon to minimize exposure to radiation sources such as radioisotope-based power generators (RTG) and heaters.

Field of view requirements on configuration are obvious because the payload has to be able to see its target without interference from other parts of the spacecraft. This requirement is more easily stated than satisfied and will often tax the designer's ingenuity to achieve an acceptable compromise.

8.2.3 Environment

Environmental drivers on configuration and structural design are fairly obvious: solar distance, atmosphere, radiation, thermal, vibration, and acoustic. The variable intensity of solar energy with distance is primarily of concern for thermal control and solar to electric conversion. In a discussion of spacecraft, one might assume that atmosphere would be of concern only for planetary landers and entry systems. Recall, however, that all spacecraft have to survive in the Earth's atmosphere first, and concerns regarding chemical attack (oxygen and water vapor), temperature and pressure fluctuations, wind, etc., must be considered. Of particular concern is the rapid pressure drop during ascent and passage through the pressure regime conducive to corona discharge.

Environmental radiation is usually not a major concern in configuration and structural design, except that on occasion it may be necessary to accommodate shielding of sensitive components. In severe environments, where one might shield the entire spacecraft, the configuration may be driven toward a very compact design to maximize self-shielding and minimize the external area that must be shielded. It was noted in Chapter 3 that long-term radiation exposure may cause degradation in the properties of composite structures.

The impact of the spacecraft's local thermal environment can range from minimal to substantial. For operations in deep space, the sun is essentially the entire thermal environment, and, unless it is very close, it is relatively easy to deal with. On the other hand, a spacecraft in low orbit about Mercury not only experiences solar intensity on the order of 10 times that of Earth, but is also exposed to radiation from the hot surface of the planet. The temperature of the hot side of Mercury (up to 700°C) is such that the re-radiation of the absorbed solar energy takes place in the infrared. Because spacecraft are usually designed to radiate in the infrared to dispose of absorbed and internally generated heat, they are also fairly good infrared absorbers. Thus, the surface of Mercury, radiating infrared at a rate nearly comparable to the sun itself, is a major source of thermal input. Very clever configuration and mission design is required to maintain a spacecraft within acceptable thermal limits in this environment.

The design requirements for vibration and acoustics are sufficiently obvious to require little comment. However, the engineer should keep in mind that *which*

environment is the driver may be less clear. Launch or atmospheric entry may be the most severe; however, they are brief compared to, for example, a four-hour cross-country flight or a four-day truck ride. Designing for the mission without considering how the hardware is to be handled on the ground frequently causes major problems. In fact, the in-flight difficulty experienced in deploying the Galileo high-gain antenna was ultimately attributed to damage sustained during multiple cross-country trips resulting from repeated launch delays.

8.2.4 Power Source

Various types of power sources that may impact configuration and structure are 1) solar photovoltaic, 2) radioisotope thermoelectric generators, and 3) new technology such as reactor based, solar dynamic, and radioisotope dynamic. Notably absent are batteries and fuel cells, which, except for the requirement to accommodate a certain mass and volume, pose few constraints as a rule. The same cannot be said for the other types of power sources listed. Solar photovoltaic systems require large areas with an essentially unobstructed view of the sun and, at least in the case of large flat arrays, the ability to maintain the array surface normal to the sun. Drum-shaped, spin-stabilized craft require even larger areas because only a part of the area can be exposed to the sun. None of the preceding factors would not cause great problems except that it is also necessary to mount and point accurately various antennas, science instruments, attitude control sensors, etc. These requirements are often in conflict regarding which item of hardware occupies a particular area on the vehicle, and because of possible shadowing, field of view interference, etc.

RTGs generally relieve some of the location problem and the demand for large area; however, they bring their own set of problems. Because of the need to reject heat from the outer surface of the RTG and because of the radiation from the decaying isotopes, it is usually not practical to mount them inside the spacecraft or even extremely close to it. In most applications, RTGs are boom mounted at some distance from the spacecraft to reduce the effect of both nuclear and thermal radiation. Launch volume constraints dictate that the mounting structure be deployable. Examples of this sort of installation will be seen later.

The newer technology systems that are listed have not flown (except for one experimental reactor) on U.S. spacecraft. The radiation output from a nuclear reactor is far more energetic and damaging than that from an RTG. Also, because reactors emit far more power, the waste heat to be disposed of is greater. This latter requirement leads to very large radiator areas, with all the predictable problems in launch stowage, thermal input, view of space, etc. The radiation requires great distances and/or massive shielding. In proposed designs using reactors a compromise is usually reached, placing the reactor as far from the spacecraft as practical and then shielding to reduce the radiation flux at the spacecraft distance to an acceptable level. Because of the thickness and great weight of the shield, "shadow shielding" is employed. That is, the shield is placed

between the reactor and spacecraft rather than shielding the full 4π sr as is done for Earth installations. The very long boom with large masses at either end and possibly in the middle (as would be the case with ion propulsion units) introduces some major challenges in structure and mechanism design as well as in dynamics and control.

Solar dynamic systems require the same solar field of view and pointing control requirements as photovoltaic arrays, possibly with somewhat tighter accuracy constraints depending on the type of collector used. Most dynamic conversion systems have very large waste heat radiators because the low-temperature end of the thermodynamic cycle must be relatively cool to achieve good efficiency. The radiators require a good view of space and a minimum view of the sun, nearby planets, etc.

An additional problem involving any unit using dynamic energy conversion is the introduction into the structure of a "hum" at the frequency of the rotating machinery. It may be necessary to design the structure and select materials to damp out the vibration as much as possible to reduce the impact on attitude control.

Radioisotope dynamic systems are, as one would expect, a hybrid of the problems of the solar dynamic system and the RTG. The radiation problem persists and is combined with the need for large radiator area and the potential vibration problems inherent in dynamic conversion. The much greater efficiency of these dynamic units compared to RTGs or solar photovoltaic arrays is the incentive to use them. However, they do introduce some challenges to the configuration and structure designer.

8.2.5 Launch Vehicles

Launch vehicle constraints exist for mass, dimension, vibration and acoustic energy, and safety. Of these, the first and most obvious constraint forced on the designer is that of launch mass capability. Next is payload dimension, not only the length and diameter of the payload volume available but also the dimensions of the attachment interfaces. The mass and volume available dictate the size of the basic structure, drive the selection and design of deployable structures, and, in many cases, strongly influence the choice of materials.

The acoustic and vibration environment imposed by the launch vehicle is generally the most intense that the spacecraft will encounter, although as mentioned earlier, because of the relative brevity of the powered flight, the cumulative effect of prelaunch environments may be of equal or greater severity. The measured or calculated launch environments are used to define qualification test criteria. These criteria are defined by drawing curves that envelop the actual environment. Factors are then applied to these curves to define flight qualification or flight acceptance (FA) criteria. Higher factors are used to define type acceptance (TA) criteria. A factor of 1.5 applied to the actual environmental stress might be used to define FA levels, and a factor of 2.0 used to define TA.

Actual flight articles would be tested to FA levels to demonstrate workmanship and margin over expected values. Flight articles would be expected to withstand FA without damage or unacceptable response. TA levels are used to demonstrate qualification of the basic design, and may push the structure to near failure. Structural yielding or other responses may occur that would render the article unacceptable for flight use, and so TA levels are only applied to nonflight prototypes. Severe budget constraints in recent years have often resulted in programs where only one spacecraft is built and flown, with no prototypes or test articles. In such cases, compromise test levels (protoflight levels) between FA and TA may be defined. Chapter 3 provides representative environmental data for several launch vehicles.

Launch safety constraints are generally not a major driver for structure and configuration design when expendable launch vehicles are used. The primary constraint is that the spacecraft not fail, a criterion to which everyone involved will subscribe. Occasionally, as when radioactive or other hazardous materials are carried, there will be issues revolving around a launch abort. In the event of an errant launch vehicle being destroyed by range safety, there may be a requirement that the spacecraft break up in certain ways or not break up at all upon reentry. This situation might obtain in the case of a spacecraft bearing RTGs or a nuclear reactor, for example. For the most part, the constraints will be on pad operations involving personnel. The most common example is imposition of minimum safety factors and possibly fracture mechanics criteria on pressure vessels that must be pressurized in the presence of personnel.

Spacecraft destined for launch on the space shuttle come under much more severe scrutiny. Because the shuttle is manned and because there are few of them, strict safety constraints are levied to ensure that no problem with the payload will cause a hazard to the shuttle or crew. With the exception of higher safety factors and more emphasis on fracture mechanics, there is not a great difference in the engineering aspects of designing for the shuttle vs an expendable launcher. The real difference is in the extensive review and certification process designed to prove compliance.

8.2.6 Communication

The spacecraft communications system must consider antenna size, pointing accuracy, and radiated power. The primary impact of the spacecraft communication system on configuration lies in antenna size and required location relative to the major attitude control references. Relatively small lower-gain antennas usually do not present a major problem, whereas a large high-gain dish, especially one that is required to move during the course of a mission, can require considerable attention. In the latter case, the antenna will generally be latched down during launch and deployed after placement on orbit, placing additional requirements on the structure design and on related mechanical deployment devices.

Not only is pointing direction a concern to the configuration designer, but the required pointing accuracy may be as well. Pointing accuracy requirements may drive structural design and the choice of structural materials in an effort to minimize the effect of distortion from thermal effects or structural loads. Such distortions, by introducing bias errors between the attitude control reference and the antenna mount, can cause problems in accurate pointing of very tight beams, and similarly for other spacecraft instruments.

As we have mentioned, pointing stability (the magnitude and frequency spectrum of the instantaneous variation about the mean pointing direction) can be of equal concern. In many applications familiar to the authors, relatively generous mean pointing accuracy constraints may apply; however, the allowed variation, or jitter, about this mean position must be very tightly controlled. This is accomplished through passive or active vibration isolation, the use of artificial vibration damping in structural materials, and the enforcement of rigid vibration control specifications on spacecraft subsystems and instruments.

Mission requirements for reasons other than instrument or pointing control may exist. For example, many if not most materials processing experiments to be conducted on the International Space Station (ISS), or elsewhere, are rendered useless by even low average levels of vibration (e.g., 10^{-6} g). It is often stated that space materials research requires a "microgravity" environment, which is true. However, what is usually required is in actuality the absence of forces of any kind, a matter often not fully appreciated. We will summarize this issue by noting that few aspects of space vehicle design will be more demanding than the requirement to maintain a very "quiet" spacecraft when necessary to satisfy particular mission goals.

Finally, the radiated power of the communications system may impose certain requirements. In the case of very high power systems, it may be necessary to avoid placing components where they can be illuminated by sidelobes and backlobes of the antenna. Also, if a very large amount of power is being radiated from the antenna, then even more is being dissipated in the form of waste heat within the spacecraft. The configuration design must be able to accommodate the conduction of this internally generated heat to the appropriate radiating surface of the spacecraft for rejection to space, or on some fortuitous occasions as arranged by a clever thermal control engineer, to a place elsewhere on the spacecraft where the additional heat is wanted.

8.3 Spacecraft Design Concepts

This section presents several spacecraft design concepts as an illustration of how various design teams have dealt with the design drivers discussed previously. Both overall configuration and internal packaging concepts are presented. Some of the pros and cons of each concept are discussed. Concepts for deployable booms and scan platforms are also discussed.

8.3.1 Spacecraft Configuration

8.3.1.1 Voyager. Figure 8.1 shows the Voyager spacecraft. Two of these vehicles, built and tested by the Jet Propulsion Laboratory (JPL) for NASA, were

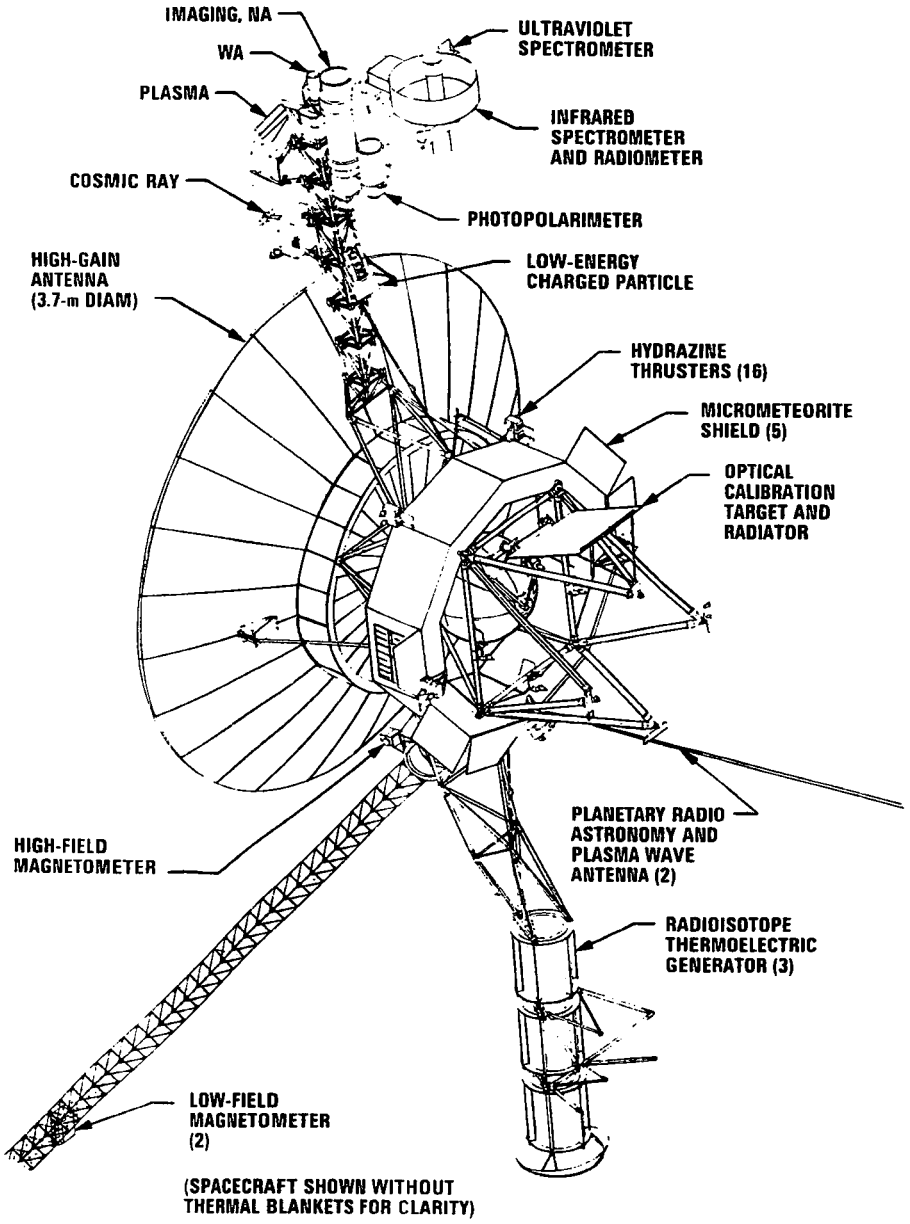


Fig. 8.1 Voyager spacecraft. (Courtesy of Jet Propulsion Laboratory.)

launched in 1977 to explore the outer planets of the solar system. The baseline mission was to be a four-year trip involving a flyby of Jupiter and Saturn by each spacecraft. In the event, by use of planetary gravitational assists as discussed in Chapter 4, Voyager 2 was redirected to extend its mission to encounters of Uranus and Neptune. The latter encounter, in August 1989, was some 12 years after launch. Voyager 1, its trajectory bent out of the plane of the ecliptic by its encounter with Saturn, encountered no more planets but continues to send back data concerning regions of space not previously visited. Both spacecraft have considerably exceeded solar system escape velocity and will continue indefinitely into interstellar space. As this is written, both spacecraft are being tracked periodically. While there is no longer sufficient power to run the imaging instruments, which in any case have nothing to see, the fields and particles instruments still measure the local environment. It is hoped that the Voyagers will soon encounter the elusive interface between the solar-dominated environment and true interstellar space.

The great distances from the sun at which the Voyager spacecraft are designed to operate dictate two of the most prominent features of the configuration. Because solar power is not available (e.g., it is reduced to only about 1.6 W/m^2 at the orbit of Neptune), power is provided by RTGs. Because great distance from the sun also connotes great distance from Earth, a large antenna is required to support the data rates desired.

However, the sun does provide a useful attitude reference. A sun sensor peers through an opening in the antenna dish. Because the antenna must point at the Earth and it is necessary to keep the sun off the axis of the antenna to prevent overheating of the subreflector, there is a slight offset between the antenna boresight and that of the sun sensor. However, this bias was designed for the baseline Jupiter/Saturn mission and is no longer necessary at the huge distance from the sun at which both spacecraft now operate. A star tracker is used to provide the reference for the third axis.

The RTGs are mounted on a rigid hinged boom that was latched to the final launch stage and then swung into final position after stage burnout. The in-line arrangement is used to minimize radiation to the bus, because the inboard RTG, although itself a source of radiation, is also a shield for the radiation from the other two.

With one exception, the science instruments are mounted on the boom radially opposite to the RTG boom. This has the desirable feature of placing them as far as possible from the RTGs. The fields and particles instruments, concerned mostly with the sun, solar wind, and planetary trapped radiation, are mounted on the boom outside the shadow of the antenna. The visual imaging, infrared, and ultraviolet instruments are located on a two-axis scan platform located at the end of the boom. These instruments are primarily concerned with the planets and satellites and require accurate pointing and quick retargeting capability.

The one science instrument not located on the science boom is the magnetometer. This unit needs to be as far as possible from the spacecraft

magnetic field. It is mounted on a deployable Astromast™ boom, which at full deployment is 15 m (50 ft) in length but which retracts for launch into a can less than 1 m in length. Two magnetometers are mounted: one at the outer end of the boom and the other halfway out. The outer one is used for obtaining science data, whereas the other is primarily intended to help evaluate the residual spacecraft field in support of data analysis.

With both the RTG boom and the science boom folded aft for launch, these items are clearly much closer together than when deployed. Both are latched to the solid-propellant final stage, and the case and propellant provide considerable shielding. Nevertheless, a higher radiation dose was accumulated in the few weeks between final assembly and launch than in months of spaceflight.

The 10-sided bus contains the majority of engineering subsystems. The large spherical tank contains the hydrazine propellant for the combined attitude control and propulsion subsystem. This tank is located in the center of the bus not only to maintain the hydrazine at a satisfactory temperature but also to aid in shielding the instruments from the RTGs. Equally important is the fact that, as propellant is used, the vehicle center of mass remains essentially unchanged, greatly simplifying attitude control requirements.

The four legs descending from the bottom of the spacecraft, often taken for landing gear, are actually the truss that supported the solid-propellant rocket motor, which was the final launch stage. These structures form the basis of a story that is repeated here not with the intent to embarrass anyone, but to provide an object lesson concerning how a seemingly innocuous decision in one subsystem can have an unexpected detrimental effect in another area.

Originally, the separation plane was to be at the bottom of the bus and the truss was to be jettisoned along with the rocket motor. However, some concern was expressed regarding the effect of shock generated by the explosive release nuts on nearby electronics. An easy solution was adopted: separation would occur at the motor, leaving the truss legs attached to the bus. The small additional weight of the truss would have no significant effect on the mission, and this approach would save much effort in shock isolation, etc.

All was well until, a few days after the launch, the aft-facing array of thrusters was fired to correct the trajectory. The thrusters fired for the proper amount of time and appeared to have operated properly; however, the maneuver ΔV was low by a substantial percentage. The technical detective work will not be detailed here, but the conclusion was that the exhaust jets from the thrusters, expanding rapidly in the vacuum, impinged upon the nearby truss members. The effect was generation of drag on the truss by the supersonic flow, thus reducing the effective thrust. Spurious forces normal to the thrust line were also generated, perturbing the spacecraft attitude, corrections for which required the expenditure of yet more fuel.

Because the propellant supply was calculated with a substantial margin, and because launch vehicle injection accuracy was so good, the effect of this mistake was not a catastrophe but rather the relatively minor annoyance of a reduction in propellant margin for the mission. However, one can easily see how a similar

situation could be catastrophic to another mission with less margin. Again, the lesson to be learned is that no subsystem is an island. Any decision made in one area must be assessed for its impact upon others. This is the essence of systems engineering.

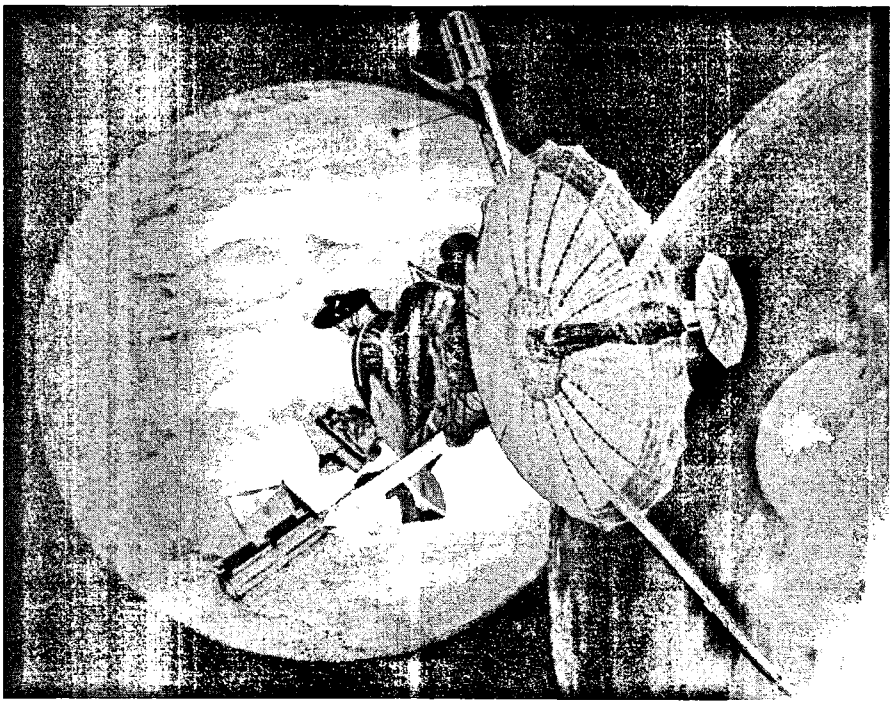
The primary Voyager structure is made of aluminum, with composites used in various locations. The largest composite structure is the antenna dish. The Astromast^{TM} boom makes use of very thin fiberglass members with parallel orientation of the fibers to obtain the strength and elasticity to allow it to coil into its canister and still maintain reasonable rigidity when deployed.

8.3.1.2 Galileo. The JPL Galileo spacecraft, shown in Fig. 8.2, has a history of frustration almost unparalleled in the history of the space program. Originally scheduled for launch in 1982, the program was buffeted by shuttle development delays, early changes in specified upper stage capability, further delays due to the *Challenger* accident, cancellation of the planned shuttle-Centaur upper stage, and a variety of other problems. It was finally launched in 1989 on an upper stage of much lower capability than originally intended. This greatly extended the flight time to Jupiter, because a gravity assist flyby of Venus and two such flybys of Earth were required to give the spacecraft sufficient energy to reach Jupiter.

The mission plan was for the spacecraft to drop a probe into Jupiter's atmosphere and record the probe data for later playback to Earth. Galileo would then fire its rocket engine for insertion into a highly elliptic orbit about Jupiter. From this orbit, the planet and most of its satellites would be closely studied, the latter in a series of close flybys. Each flyby would use the gravity of the satellite to modify Galileo's orbit for its next encounter. The mission plan was successfully executed, although data return was hampered by the antenna failure discussed in the following.

Galileo is unique among planetary spacecraft to date in that it is a dual spinner. That is, one portion of the spacecraft spins while the other spins at a different rate or not at all, a concept that is discussed in some detail in Chapter 7. The idea of using it on Galileo was to incorporate the best aspects of spin-stabilized and three-axis-stabilized spacecraft. The spinning portion would provide the global coverage desired by the fields and particles instruments, and the fixed portion would provide a stable base for the high-resolution imaging that a spinning spacecraft cannot provide. The spinning portion of the spacecraft would provide attitude stability with minimal expenditure of attitude control propellant. The concept has been used successfully in a variety of Earth-orbiting spacecraft, most notably communication satellites built by Hughes Aircraft Corporation (now Boeing).

Dual spin attitude control, however, turned out to be less adapted to the Galileo application. Large amounts of power and data passed across the spin-bearing interface, greatly complicating the design. The impossibility of properly shielding the data channels from noisy power conductors as they crossed the slip-ring assembly was a major problem from early development.



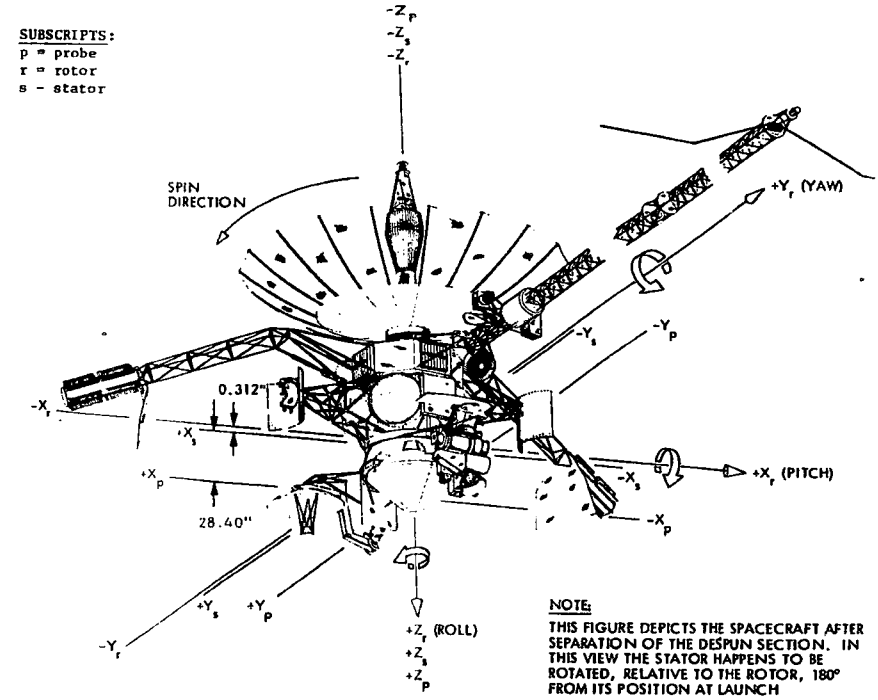


Fig. 8.2 Galileo spacecraft. (Courtesy of Jet Propulsion Laboratory.)

The spinning portion of the spacecraft includes the multisided bus, similar to that of Voyager, the communications antenna, the booms supporting the RTGs, and the magnetometer. The bus contains the engineering subsystems of the spacecraft. The antenna, of metal mesh, is 15 ft in diameter when deployed. For launch it is folded, using its rigid ribs, around the center feed support. A major setback occurred shortly after launch, when the antenna failed to deploy properly.

Failure of the antenna to deploy fully has rendered it entirely useless. All data were sent back via the low-gain antenna at very low bit rates. The failure appears to have been caused by abrasive removal of the coating on the ribs and stowage slots during three truck trips across the United States [one from JPL to Kennedy Space Center (KSC) for the original launch date, back to JPL after the *Challenger* failure, then again to KSC for launch]. The antenna was left stowed after launch for several months while the spacecraft flew by Venus and then by Earth for the first time. It is considered probable that about three ribs cold-welded (see Chapter 3) into their slots and would not deploy.

The lower portion, as depicted in the illustration, is the fixed portion. It mounts the scan platform for science, the small dish antenna for receiving data from the Jupiter probe, and the probe itself. The probe is the cone-shaped object on the centerline of the vehicle. When the probe was separated from the spacecraft some 150 days prior to encountering Jupiter, it was spinning in order to provide stability. This required that the nonspinning portion of the spacecraft be spun for probe separation and then despun again.

The main orbit insertion rocket motor is located on the centerline of the vehicle behind the probe. If the probe had failed to separate, orbit insertion would have been impossible. Also mounted on this section, on outriggers, are attitude control thrusters.

The original idea for this approach to the Galileo configuration was to simplify and to reduce cost while, as noted earlier, obtaining the advantages of two different types of spacecraft. From the preceding discussion it is not at all clear that things worked out as intended.

Some changes, e.g., sunshades, were required for the Venus swingby that results from the low-energy stage discussed earlier. These are not shown in Fig. 8.2b but do appear in Fig. 8.2a.

In general, the structure of the spacecraft is quite similar to that of Voyager.

8.3.1.3 Cassini. Cassini is a program intended to perform, at Saturn, a mission similar to that of Galileo at Jupiter. As this it written, it is planned that Cassini, shown in Fig. 8.3, will enter an elliptic orbit about Saturn in 2004. An atmosphere probe, Huygens, will be carried, but this probe is targeted for the large satellite Titan, rather than Saturn itself. Titan is the largest satellite in the solar system and the only one possessed of a substantial atmosphere.

In an effort to reduce cost, Cassini has no scan platform but rather has all instruments body fixed. This means that the spacecraft itself must be maneuvered

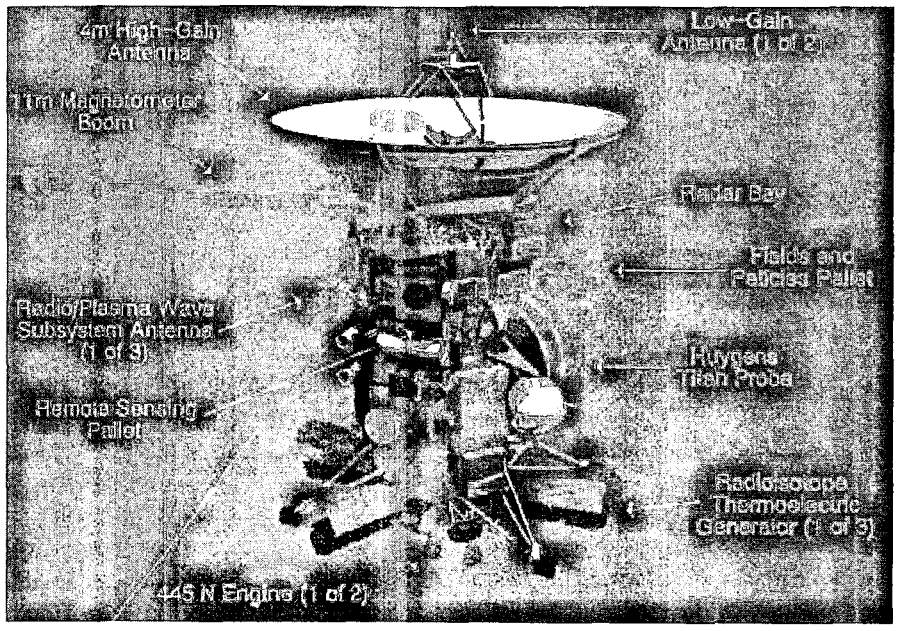


Fig. 8.3 Cassini spacecraft. (Courtesy of Jet Propulsion Laboratory.)

to point the instruments for data acquisition and then maneuvered to point the high-gain antenna toward Earth for transmission. This approach was successfully used on the Magellan radar mapping mission to Venus, which used a spare Voyager antenna fixed to the bus as both the radar antenna and the data transmission antenna.

The RTGs that power Cassini are mounted at the bottom (when in launch orientation) of the bus rather than on a boom. This increases the radiation dose to some degree.

As discussed in Chapter 2, Cassini was launched in October 1997 on a Venus-Earth-Jupiter flyby orbit designed to allow the spacecraft to reach Saturn in July 2004.

8.3.1.4 Deep Space 1. Figure 8.4 depicts Deep Space 1 (DS1). The spacecraft was primarily designed as a technology demonstration mission but has carried out scientific investigations of two near-Earth asteroids and a short-period comet. Technology demonstrations by DS1 include the use of an electrostatic ion thruster as main propulsion. Such thrusters are widely used for stationkeeping. They have been proposed for decades as primary propulsion for missions such as multiple near-Earth asteroid flyby, but this is the first such use.

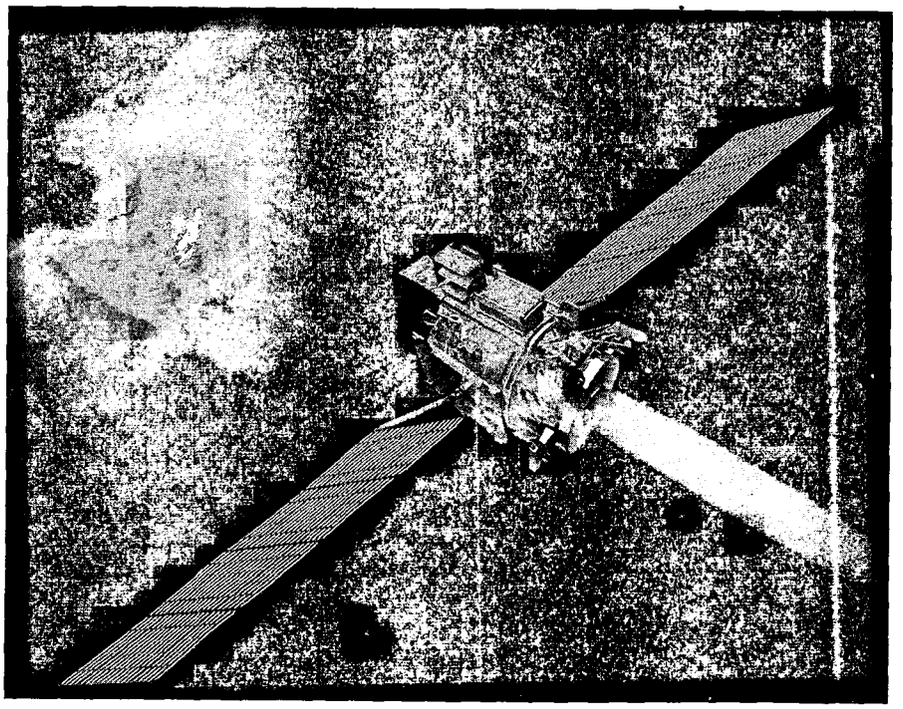


Fig. 8.4 Deep Space 1 spacecraft. (Courtesy of Jet Propulsion Laboratory.)

To supply the power needed for the ion thruster, DS1 uses an innovative solar concentrator approach. Fresnel lenses are used to concentrate sunlight on gallium arsenide solar cells. This is discussed in more detail in Chapter 10.

Launched in 1998, DS1 has completed two near-Earth asteroid flybys and a comet flyby and is nearing propellant exhaustion.

8.3.1.5 FLTSATCOM. Turning from planetary spacecraft to geostationary communications satellites, Fig. 8.5 depicts the FLTSATCOM. This is a satellite made by TRW for the U.S. Navy for communications purposes. Of fairly conventional aluminum construction, this configuration is interesting in that it consists of two identical hexagonal buses mounted one above the other. The lower bus contains all of the engineering subsystems (note that the solar panels are mounted on that portion) that perform all of the functions required by the spacecraft (e.g., power, attitude control, etc.). The second bus contains all of the payload related hardware (e.g., transponders, etc.). The antennas related to its communications relay function are mounted on this bus.

Because the antennas point to the Earth, the solar array arms are oriented normal to the orbit plane and the arrays rotate to maintain sun pointing at all times. For launch, the arrays fold so that their long axes are parallel to the antenna

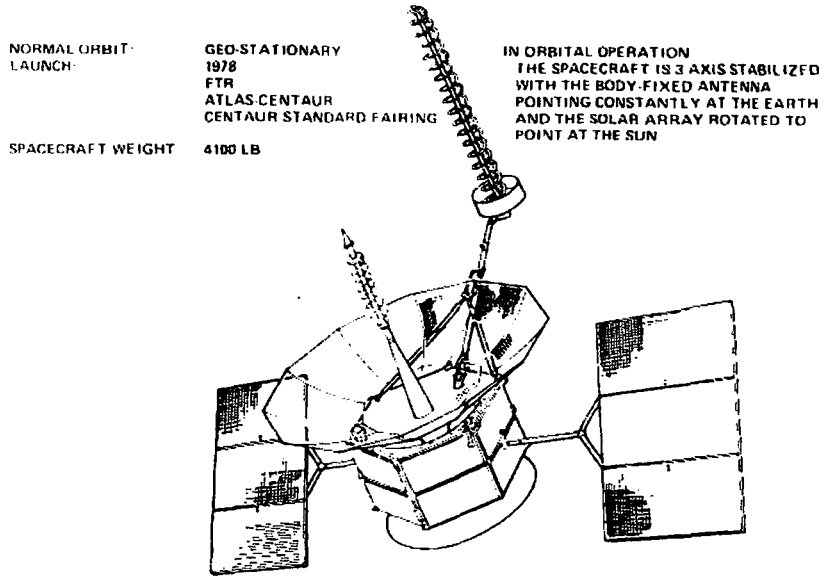


Fig. 8.5 FLTSATCOM spacecraft. (Courtesy of TRW.)

boresight and fold between segments to form a hexagonal cylinder around the spacecraft. This configuration is maintained until after orbit insertion so that the deployed arrays do not have to withstand the insertion *g* loads. Only a portion of the arrays is illuminated in this configuration, but power requirements are so low in cruise mode compared to the operational relay mode that ample power is available.

The advantage of the two-bus configuration is adaptability. If an entirely different payload is desired, the communications relay bus can be replaced by a different package while still retaining the tested and reliable engineering spacecraft essentially unchanged. For example, this same spacecraft was proposed as a low-altitude Mars Polar Orbiter with minimal changes in engineering subsystems. One change was a reduction in size of the solar arrays, because, even at the greatly reduced illumination at Mars, the small power usage of a science payload vs the massive communications relay made the full-size arrays unnecessary.

8.3.1.6 HS-376. Figure 8.6 shows a different approach to geosynchronous communications satellites. The HS-376 series by Hughes Aircraft Corporation (now Boeing) is typical of the drum-shaped dual-spin satellites that were the mainstay of that company for many years. The dual-spin terminology refers to the fact that the main bus spins at a moderate rate while the antenna assembly spins once per orbit, in other words, maintaining orientation toward the Earth at all times. The electronics is mounted on shelves within the rotating bus,

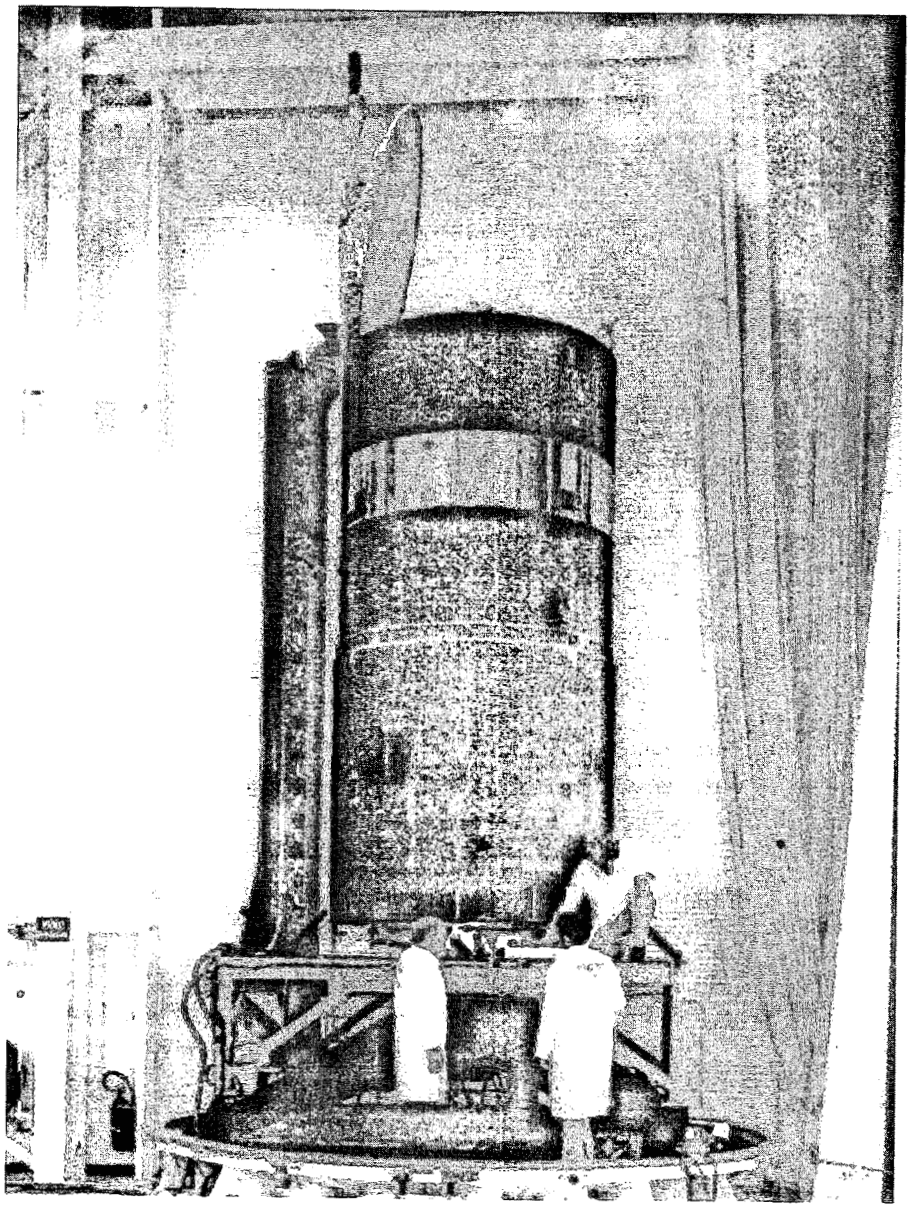


Fig. 8.6 HS-376 spacecraft. (Courtesy of Hughes Aircraft Company.)

as is the attitude control propellant and other equipment. Only radio frequency energy crosses the spin bearing between the bus and the antennas. (Note the difference between this and Galileo, which must send both power and data signals through slip rings or equivalent devices at the spin bearing.)

The orbit-insertion or apogee kick solid-propellant rocket motor is mounted in the bus firing out the anti-antenna end. Waste heat rejection is also primarily from this end of the bus.

Earlier spacecraft of similar configuration were of fixed geometry. However, these spacecraft quickly encountered one of the major weaknesses of the rotating, drum-shaped configuration. Because only 30–40% of the exterior of the spacecraft can be effectively illuminated at any time, high-powered spacecraft of fixed geometry begin to experience power limitations. There is simply not enough fixed surface area on the exterior of the spacecraft to provide sufficient solar array area to generate the required power. Launch volume constraints preclude simply making the spacecraft larger in diameter or longer. There are several possible solutions to this problem involving deployable vanes, paddles, etc. The solution chosen in this instance maintains the general configuration of the vehicle. Additional solar panel area is incorporated in a cylindrical shell that surrounds the main bus. This provides power during cruise to orbit. Once the apogee motor has fired, the shell is deployed down, exposing the solar cells on the main bus structure. This approximately doubles the length of the spacecraft and the available solar array area. The fact that this can be done without compromising the stability of the spacecraft results from the fact that dual-spin spacecraft are not bound by the axis of maximum moment of inertia requirement that dominates the simple spinner (Chapter 7).

8.3.1.7 Defense Meteorological Support Program. The Defense Meteorological Support Program (DMSP) spacecraft built by RCA (now Lockheed Martin) is a weather satellite designed to operate in low circular, sun-synchronous orbit. The civilian Television and Infrared Observation Satellite (TIROS), although not identical, is sufficiently similar that most comments made here apply with equal force to both. The spacecraft is depicted in Fig. 8.7. The main electronics bus is the large, boxlike structure covered with circular temperature-control devices. These are thin, polished sheets with pie-shaped cutouts that rotate under control of a bimetallic element to expose insulated or uninsulated skin areas depending on the amount of heat to be rejected. The spacecraft is nadir pointing in operation. The flat face not visible in the illustration faces the planet, and instrumentation is mounted in that area. Instruments are also located in the large, transversely mounted structure on the end of the main body.

Two versions of the spacecraft are shown, one with a large, solid-propellant motor and auxiliary monopropellant system, and the other with a monopropellant-only system, a large hydrazine tank replacing the solid rocket motor. The solid-propellant version was launched using a refurbished Atlas missile as the lower stage, whereas the other version was intended to be launched in the shuttle. The TIROS/DMSP is a very capable spacecraft, providing guidance and control functions for its own launch vehicle in the expendable launch case.

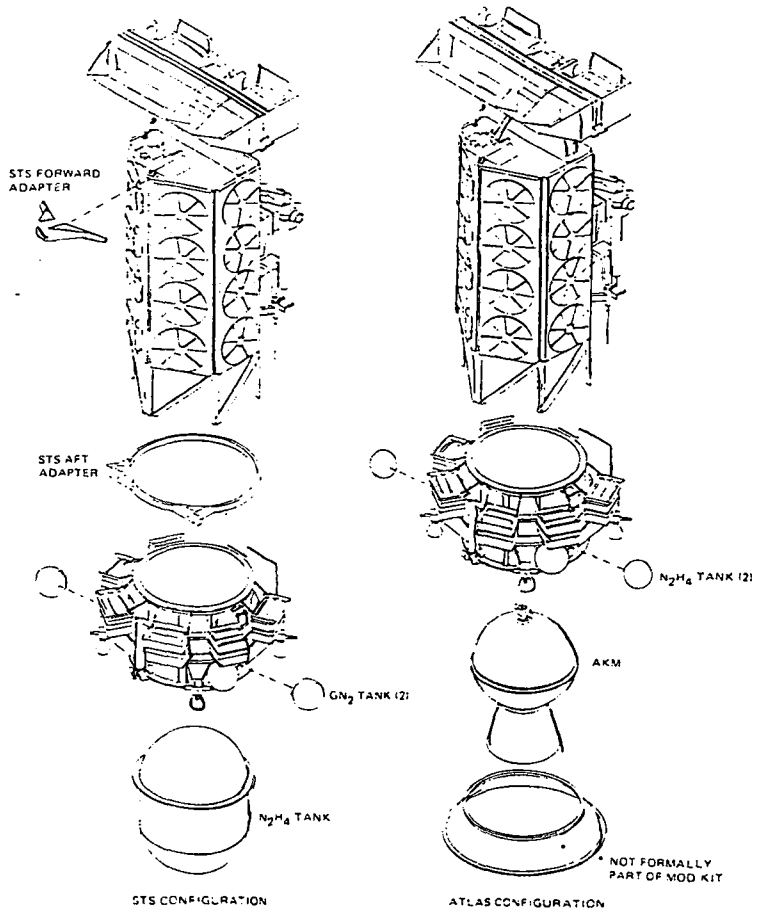


Fig. 8.7 TIROS/DMSP spacecraft. (Courtesy of General Electric Astronautics Division.)

One virtue of this concept is ease of internal access. The large, Earth-facing side opens like a door, exposing equipment mounted on that side as well as that mounted on the other walls. With the spacecraft in a vertical position as shown, the size and configuration make internal access quite easy compared to some.

This spacecraft was also evaluated for adaptation to planetary missions and showed excellent promise. In fact, the TIROS/DMSP bus was originally chosen for the Mars Observer mission. However, well into the program, the decision was made to adopt a smaller bus. This change by itself resulted in numerous problems and delays, an example of the desirability of a large bus with ample room and straightforward access. (Mars Observer was ultimately lost in 1993, following the attempted pressurization of its propellant tanks prior to its planned injection into Mars orbit. The role of the decision to change the bus in this case is a matter for speculation.)

8.3.1.8 HS-702. The HS-702 spacecraft is a large geostationary communications satellite. This three-axis stabilized spacecraft was developed when the power limitations of the HS-376 spacecraft became too constraining. This spacecraft uses trough-like concentrators on its solar arrays to reduce the required array size.

8.3.2 Design and Packaging Concepts

A variety of internal structural design and electronic packaging concepts have evolved in conjunction with the configuration designs discussed previously. Table 8.1 describes three basic types along with some of the good and bad features of each. These points are discussed further in the following paragraphs.

It should be noted that various organizations will use their own variations of these approaches, and the names applied to the various concepts will not necessarily agree between companies or with the terminology of this book.

Table 8.1 Structural/packaging design concepts

Concept	Features
Dual shear plate	Bus frame
	Shear plates close frame inside and out
	Custom electronic modules or mounting plates tie to shear plates
	Pros/cons
	Strong, rigid structure
	Good thermal contact
	Requires custom electronics packaging and cabling
	Efficient volumetric packaging
	Examples: Mariner, Viking, Voyager
Shelf	Shelf structure inside spacecraft skin
	Electronic packages mount on shelf
	Pros/cons
	Can use standard "black boxes"
	Less efficient volumetric packaging
	More difficult heat-transfer-path
	Example: HS-376
Skin panel/frame	Bus frame
	Large skin panels (often hinged) close frame
	Electronics mounted on skin
	Pros/cons
	Can use standard "black boxes
	Good heat-transfer contact
	Easy access
	Examples: FLTSATCOM, TIROS/DMSP

8.3.2.1 Dual shear plate. This approach, most prominently used by JPL in the Mariner, Viking, and Voyager family of spacecraft, mounts the electronics on flat honeycomb plates or, in the case of components such as gyroscope packages, in especially tailored boxes compatible with the overall packaging scheme. The plates are then bolted to inner and outer shear plates as shown in Fig. 8.8. These shear plates are inserted into the bus frame from the outside, and both shear plates are bolted to the bus. The shear plates provide closure to the bus frame, resulting in a final structure that is very rigid and sturdy. Because the electronics is customized for the application, the packaging density can be very efficient and quite high, probably the best among the concepts discussed.

Because the electronics is distributed over the aluminum honeycomb sheets that are then tightly mounted to the shear plates, which are also heat rejection surfaces, thermal transfer capability is generally good, although special provisions may be required for high-power dissipation items.

The negative aspects of this concept are the relatively large part count: shear plates, honeycomb sheets, very large fastener count, etc. This implies high manufacturing cost and labor-intensive operations. Furthermore, if one wishes to use an already existing electronics subsystem, it must be repackaged to be compatible with this approach. Cabling also may be more complex.

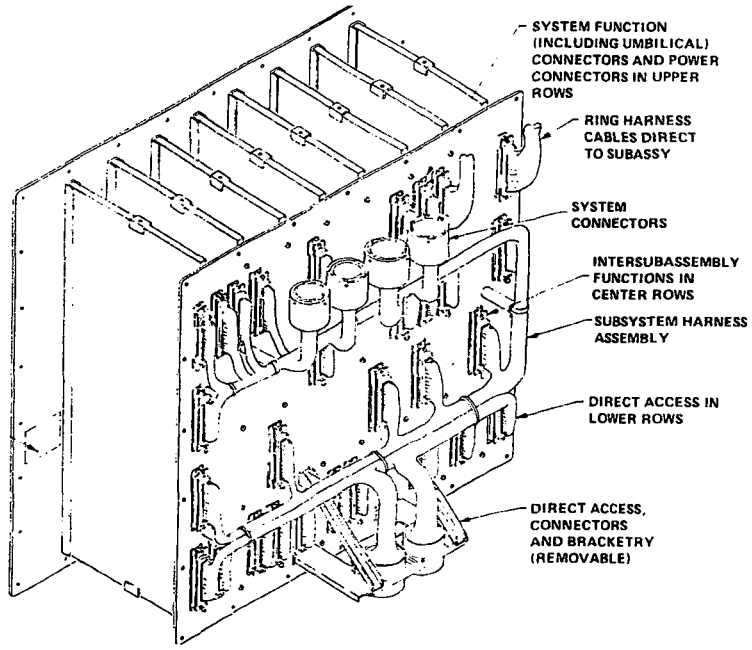


Fig. 8.8 Dual shear plate packaging.

In summary, this is an expensive approach compared to other schemes. However, the virtues of structural strength, rigidity, and high-density custom packaging may be worth the cost in some applications.

8.3.2.2 Shelf. The arrangement referred to as shelf-type packaging could equally well be characterized by other descriptions. It refers to an arrangement wherein shelves or bulkheads mounted orthogonal to the axis of a cylindrical spacecraft provide support for the electronics and other internal systems. This arrangement is typical, for example, of the interior structure of the dual-spin spacecraft described earlier.

This approach generally is less volumetrically efficient in terms of the amount of electronics per unit volume of spacecraft than some others. On the other hand, this is often not a major disadvantage, since the volume of the spacecraft is driven by the required solar array area and more internal volume is available than is required. The use of a basic flat mounting structure is more adaptable to the use of standard electronics in existing "black box" configuration without requiring customizing.

Rejection of large amounts of internally generated heat can be a problem, because components mounted near the centerline are far from the walls of the cylinder. If heat is rejected from these walls, the conduction path may be rather lengthy. In the case of the HS-376, the end opposite the antennas is essentially open for heat rejection. This works well for items with a clear view of the open end but will be less satisfactory if a stack of shelves or bulkheads is used.

8.3.2.3 Skin panel/frame. This concept, of which examples are shown in Figs. 8.9 and 8.10, uses a basic structural frame or bus. The faces of this structure are closed with plates or panels that may in some cases form part of the load-bearing structure, as do the shear plates in the other configuration discussed earlier. In the examples shown, FLTSATCOM and TIROS/DMSP, the panels are hinged along one side to swing open for easy access. The panels provide mounting structure for electronic equipment, cabling, and other hardware.

The ability to use standard, uncustomized electronics assemblies is an advantage of this configuration. Emplacement of the boxes directly on the plates that can directly reject heat to space is an advantage for thermal control. This approach is in general somewhat less rigid and structurally efficient than the dual shear plate concept.

8.3.2.4 Factors in structural concept selection. All of the structural concepts discussed earlier have significant virtues as well as some undesirable features. Which one is chosen will depend on a variety of factors, including overall configuration, mission, payload, and, occasionally, organizational prejudice. However, there are a number of factors that should be considered in any basic choice and in the subsequent implementation of that choice.

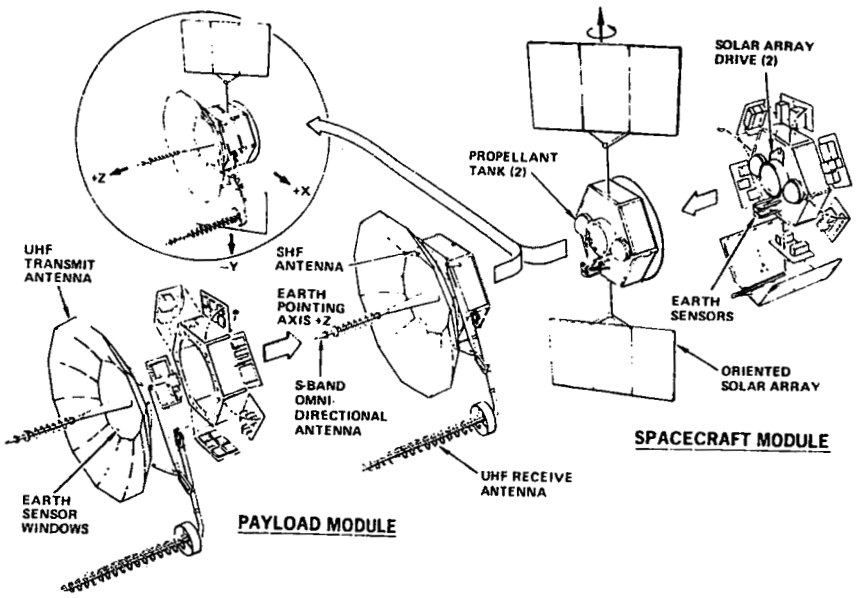


Fig. 8.9 FLTSATCOM skin panel frame packaging.

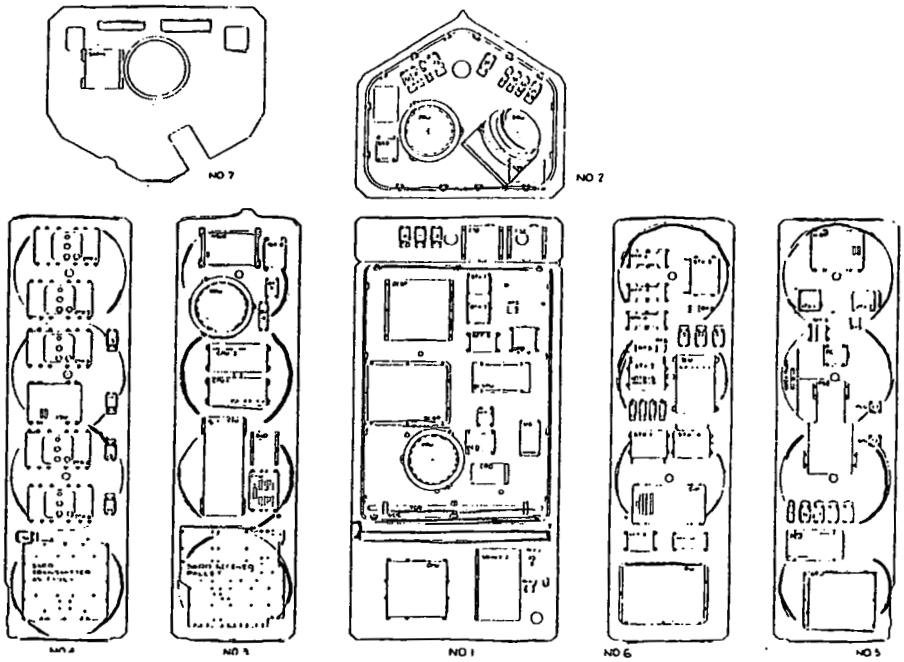


Fig. 8.10 TIROS/DMSP skin panel frame packaging.

Almost axiomatically, it is desirable to minimize parts count. The larger the number of individual parts, particularly small ones such as fasteners, the higher the cost is likely to be. This is true not only of manufacturing but also of test and operations, which are likely to be much more labor intensive and time consuming. It must be noted, however, that the desire to minimize fasteners will be viewed with alarm by structural analysts, who tend to prefer many small fasteners to a few large ones in order to improve structural load transmission and distribution. As always, the design will be a compromise.

With rare exceptions, there will be pressure to minimize structural mass. This is usually a tradeoff against cost of materials and qualification testing. A very sophisticated structure designed with tight margins to achieve minimum mass will often require expensive materials and will be expensive to design and fabricate because of the more detailed and sophisticated analysis required. Even less obvious, and therefore often a cost trap for the unwary, is the fact that such structures are often more expensive to test for qualification and workmanship certification.

In most cases, the spacecraft structural design will be driven by a requirement for structural stiffness rather than strength. This is because excessive deflection under load, even though there is no permanent yielding, usually cannot be tolerated.

The degree of understanding and the maturity of the concept are important in minimizing development cost, as is minimization of complexity. A well-understood, reasonably simple structure with characteristics that can be firmly predicted is highly desirable to minimize both risk and cost.

The operational aspects of the design are also important, yet are sometimes overlooked. Whatever the mission, it is first necessary that the spacecraft be integrated and tested on the ground. Ease of subsystem integration, work access, means of handling, and ease of disassembly and repair in the event of damage should all be considered. Finally, with the growing importance of on-orbit repair and servicing, the special needs of extravehicular activity (EVA) and concomitant manned flight safety constraints will be important for some missions, even though they may not be launched on the shuttle.

8.3.3 Deployable Structures

The requirement for deployable structures arises from the limitations on dimensions and geometry of the launch vehicle payload volume as compared with the need for large antennas and solar arrays, the requirements for instrument field of view and isolation, and the requirement to isolate radiation producing objects such as RTGs and reactors. A variety of concepts have been developed, of which a representative few will be discussed here.

8.3.3.1 Solar arrays. Deployable solar arrays have for the most part been flat rigid panels. The simplest have been the single-hinged type, which are

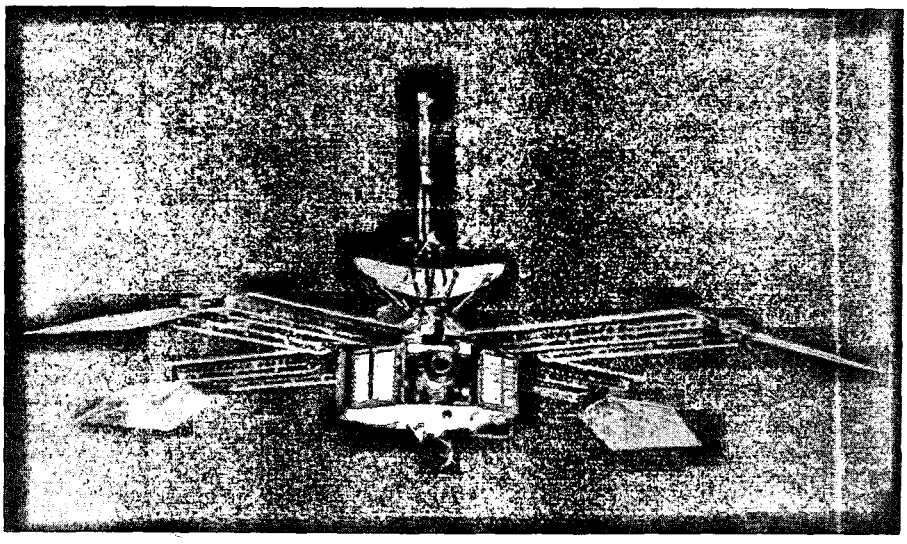


Fig. 8.11 Mariner IV spacecraft. (Courtesy of Jet Propulsion Laboratory.)

launched pinned to the spacecraft structure and/or to one another to form a compact rigid assembly. The JPL Ranger and Mariner series of spacecraft typify this approach, which is illustrated in Fig. 8.11.

More complex flat panel folding schemes such as those for FLTSATCOM discussed earlier have also been applied. An approach used on a variety of spacecraft and favored by the Soviets for their various manned craft is an extendable linkage concept.

Arrays made up of a series of flat plates that fold into a long narrow box or arrays flexible enough to roll up like an old-fashioned window shade have been designed and demonstrated and have seen some operational use, most notably in the International Space Station. These designs are deployed by an extendable boom of the Astromast™ type discussed next and can be retracted for high-g maneuvers or entry.

This by no means exhausts the variety of concepts. These are limited only by the ingenuity of the designer. From the viewpoint of the spacecraft systems engineer, the concerns include realizing the required area, obtaining reliable deployment (and retraction if required), and ensuring that the lightweight flexible structure does not detrimentally interact with the vehicle attitude control subsystem.

8.3.3.2 Deployable booms. A number of concepts exist and have been successfully flown. The first, simplest, cheapest, and (usually) heaviest is the hinged rigid boom. This is simply a long rigid boom, usually of tubular construction, with one or more hinged joints. The boom is folded and latched in

place for launch. Upon release, springs deploy the boom until it latches into the proper configuration. The virtues of low cost and simplicity make this boom attractive where there is room to stow it (stowed length is typically one-half to one-third deployed length) and the weight is acceptable. Many booms of this type have been flown. Use of stiff composite materials to keep the natural frequency as high as possible to minimize attitude control interactions is attractive. Close attention to joint design, particularly with regard to rigidity, is required for precise location and natural frequency control.

The Astromast^{TM} boom is an extremely sophisticated deployable structure. The illustration of the Voyager spacecraft in Fig. 8.1 depicts such a boom. The full deployed length is 15 m, but it is contained for launch in a canister about 1 m long. The boom consists of three fiberglass longerons stiffened at intervals by fiberglass intercostals and beryllium-copper cables. Stowed, the longerons are coiled into the canister, taking on the appearance of a coil spring. The intercostals and cables stack in the interior of the canister. The boom provides its own deployment force using the extensive amount of strain energy stored in the coiled longerons. The problem is to restrain the deployment to prevent damage from too rapid movement. This is done by a cable running up the center. The cable is attached to a motor through an extremely high ratio, anti-backdrive gear system that pays out the cable at the desired deployment rate. This motor can also be used to retract the boom as desired. The boom has an excellent deployed-to-stowed-length ratio and is very light in weight for its length. Considering the length and light weight, it is also fairly rigid, but typically cannot be deployed horizontally in a 1-g gravity field without support. As one would expect for such a sophisticated device, these booms are fairly expensive compared to the hinged rigid types. Deployment and retraction cycles may be limited because of the large amounts of strain experienced by the longerons when retracted.

Stem-type booms come in several variants. In general they consist of two metal or composite strips formed to a particular cross section. These may be welded along the edges or may mechanically join via a series of teeth along the edges. In any case, the two strips are stowed by rolling up on a reel where they are deformed to a flat shape and thus stow rather like tapes. As they are reeled out, the two strips return to their originally formed shape to provide a cross section for stiffness. The edges, if they are not already welded, interlock during this process. These booms are capable of many cycles. Beryllium-copper is a favored material to allow high cycle life and precise repeatability. Such a boom was used as the manipulator arm on the Viking Lander. Length limitations on booms of this type depend on the loads. They are fairly rigid even at fairly high length-to-cross-section ratios. Some experiments have been done by JPL and the World Space Foundation with lower cost variants using stainless steel shim stock. These are less precise and have lower cycle life, but may be satisfactory for applications.

8.3.3.3 Articulating platforms. For many missions involving high-resolution imaging and/or where rapid retargeting of instruments is required, a

scan platform is desirable. Platforms of this type are usually free to move in two orthogonal axes, although simple single-axis platforms have been flown. These platforms have been used on the various Mariner, Viking Orbiter, and Voyager spacecraft have been highly effective in obtaining maximum coverage during brief flybys, and in providing detailed mosaics in areas of interest.

The usual practice has been to move the platform using a precision stepper motor. Attitude reference is that of the spacecraft, and the platform motion is controlled relative to that reference frame. Occasionally this has required special effort in control of spacecraft attitude and platform pointing to achieve high-precision results.

Predictably, the techniques just described have become inadequate for some applications. A more recent high-precision approach involves providing the scan platform with its own attitude references, including a gyroscope package and celestial sensors. This eliminates any error that might be introduced between the reference frame and the platform, e.g., in the joints, angular position sensors, etc., in the older system.

In essence the platform points itself relative to its own reference frame by reacting against the greater mass and inertia of the spacecraft. The spacecraft then stabilizes itself using its own attitude references.

For many instruments, the lower precision of the spacecraft-referenced platform is quite satisfactory. The Mariner Mark II spacecraft concept, depicted in Fig. 8.12, has two scan platforms on diametrically opposed booms. One is a high-precision platform with inertial references and a star tracker, and the other is a conventional low-precision type.

8.4 Mass Properties

Spacecraft mass properties that are usually of interest to the spacecraft designer are vehicle mass, center of mass, moment of inertia, and moment-of-inertia ratio. Depending on the spacecraft type and mission, some may be of less interest than others, but the first two are always important.

8.4.1 Vehicle Mass

Spacecraft mass is always of substantial importance. Even if weight is not particularly critical in the absolute sense, given ample launch vehicle performance margins, it is still important to know the payload mass accurately. Generally, however, mass is critical, and control must be exercised to ensure that acceptable values are not exceeded. This requires maintenance of a detailed list at least to the major component level. Table 8.2 is a typical example of such a list.

The list will change constantly all the way up to launch. Early in the program, the list will be composed mostly of estimates and may contain some factors that are "to be determined" (TBD) or tentative allocations. As the design matures, the

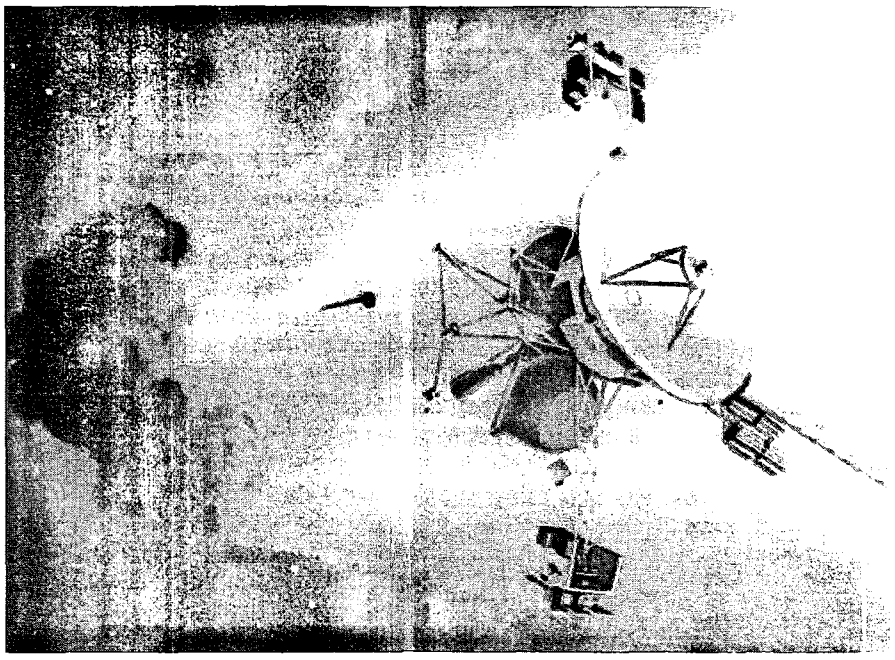


Fig. 8.12 Mariner Mark II spacecraft. (Courtesy of Jet Propulsion Laboratory.)

estimates will improve and the TBDs will disappear. Eventually prototype hardware will become available and actual values will appear in the list. As the hardware is refined toward actual flight units, the values will continue to change somewhat, but the uncertainties are much smaller. Occasionally, test results will mandate a hardware modification that will cause a substantial change, but such occurrences should be rare, and the mass list should remain fairly stable in the later phases of the program.

Some of the reasons that an accurate knowledge of mass is required include launch vehicle performance, propellant loading requirements for maneuvers, and determination of the other mass properties as needed for attitude control algorithms.

The significance of spacecraft mass relative to launch vehicle performance is obvious and is the one that usually springs to mind in discussing spacecraft mass, but other factors are important as well.

To ensure that adequate propellant is loaded for propulsion maneuvers, knowledge of the spacecraft mass is essential. This is especially true when a solid-propellant rocket is used, because the total impulse is fixed once the propellant is cast and trimmed. The only control over velocity change is total mass, which is achieved by ballasting. This requires accurate knowledge of the basic mass.

Table 8.2 Galileo subsystem mass allocations

Configuration code		Subsystem	Allocated mass, kg			
			Orbit module	Upper Spacecraft adapter	Lower Spacecraft adapter	Airborne support equipment
2001	STRU	Structure[a]	237.4	38.3	61.8	0
2002	RFS	Radio frequency	45.9	0	0	0
2003	MDS	Modulation/ demodulation	9.4	0	0	0
2004	PPS	Power/pyro	154.2	2.3	4.2	0
2006	CDS	Command and data	34.4	0	0	0
2007	AACS	Attitude and articulation control[b]	113.9	0	0	0
2009	CABL	Cabling	60.4	4.4	1.4	0
2010	RPM	Propulsion (RPM burnout)	215.68	0	0	0
2011	TEMP	Temperature control	37.6	2.0	4.3	0
2012	DEV	Mechanical devices	38.5	5.0	1.6	0
2016	DMS	Data memory	8.9	0	0	0
2017	SXA	S/X band antenna	6.1	0	0	0
2042	XSDC	X/S downconverter	2.5	0	0	0
2070	BAL	Ballast	20.0	0	0	0
2071	OPE	Orbiter purge equipment	0.9	1.6	0.1	0
2080	SAH	System assembly hardware	3.9	1.5	1.1	0
2023	PWS	Plasma wave[b]	7.62	0	0	0
2025	EPD	Energetic particles	9.37	0	0	0
2027	PPR	Photopolarimeter	4.91	0	0	0
2029	DDS	Dust detector	4.15	0	0	0
2032	PLS	Plasma	11.99	0	0	0
2034	UVS	Ultraviolet spectrometer	5.16	0	0	0
2035	MAG	Magnetometer[b]	5.86	0	0	0
2036	SSI	Solid state imaging	27.71	0	0	0
2037	NIMS	Near infrared mapping spectrometer	18.23	0	0	0
2040	SCAS	Science calibration	3.43	0	0	0
2002	USO	Ultra stable oscillator	2.05	0	0	0
2052	RRH	Relay radio hardware[c]	23.3	0	0	0

(*continued*)

Table 8.2 Galileo subsystem mass allocations (continued)

Configuration code	Subsystem	Allocated mass, kg: Orbit module	Upper Spacecraft adapter	Lower Spacecraft adapter	Airborne support equipment
2060	Probe adapter[c]	7.0	0	0	0
2072 PPE	Purge purification equipment	0	0	0	38.0

[a]Includes HGA Structural elements and RHUs.
[b]Includes RHUs.
[c]Includes System Mass Contingency.

Note: In addition to the subsystem mass allocations given in the table, the following system mass contingency breakdown exists:

Orbiter engineering 11.6 kg
Orbiter science 1.62 kg
Upper Spacecraft adapter 4.9 kg
Lower Spacecraft adapter 10.5 kg
Airborne support equipment 2.0 kg

Finally, the mass of the total vehicle and its major subassemblies must be known to compute the other mass properties, which will be discussed subsequently.

The final check on mass is usually a very accurate weighing of the entire spacecraft. This may also be done once or twice during assembly and test to verify the mass list as it then stands. The final weighing will be done shortly before launch, with the spacecraft as complete as possible. An accurate knowledge of what components are on the spacecraft and a list of deviations (i.e., missing parts, attached ground support equipment) is mandatory. Weighing is usually done with highly accurate load cells.

8.4.2 Vehicle Center of Mass

For any space vehicle, accurate knowledge of the location of the center of mass is vital. It is essential for attitude control purposes, because, in space, all attitude maneuvers take place around the center of mass. Placement of thrusters, size of thrusters, and the lever arms upon which they act are all designed relative to the center of mass. When thrusters are used for translation, it is important that the effective thrust vector pass as nearly as possible through the center of mass to minimize unwanted rotational inputs and the propellant wasted in correcting such inputs.

Launch vehicles frequently impose relatively tight constraints on the location of payload center of mass to limit the moment that may be imposed on the payload adapter by the various launch loads.

From the preceding discussion, it is clear that the payload center of mass must be both well controlled and accurately known. From the beginning, the configuration designer works with the design to place the center of mass within an acceptable envelope and locates thrusters, etc., accordingly. It is often necessary to juggle the location of major components or entire subsystems to achieve an acceptable location. This will sometimes conflict with other requirements such as thermal control, field of view, etc., resulting in some relatively complex maneuvering to achieve a mutually acceptable arrangement.

As noted earlier, the center of mass is computed from the beginning of the design process using the best weights and dimensions available. As with the mass, the information is updated as the design matures and actual hardware becomes available. Actual measurement is used to verify the center-of-mass location of the complete assembly. This usually takes place in conjunction with the weighing process, with all of the same constraints and caveats regarding accurate configuration knowledge as discussed earlier. Often the center-of-mass location is measured in all three spacecraft axes. Sometimes, however, it will be acceptable to determine it only in the plane normal to the launch vehicle thrust axis (parallel to the interface plane in an expendable launch vehicle) and compute it in the third axis if the tolerance on accuracy is acceptable.

8.4.3 Vehicle Moment of Inertia

An accurate knowledge of vehicle moment of inertia is vital for design of attitude control effectors (e.g., thrusters, magnetic torquers, momentum wheels) to achieve the desired maneuver rates about the spacecraft axes. This, together with mission duration, expected disturbance torques, etc., is used to size the tank capacity in a thruster-based system.

Moment of inertia is computed based on knowledge of component mass and location. Reasonable approximations usually provide satisfactory accuracy. Examples of this include using point masses for compact items and rings, shells, or plates in place of more complex structures.

In most cases, moment of inertia is not directly measured, particularly for large, complex spacecraft, because experience has shown that careful calculations based on measured mass and location data provide satisfactory accuracy. Direct measurement of moment of inertia has occasionally been done on programs for which it was considered necessary. The calculated moment of inertia can be in error by as much as 20%. The decision of whether to measure it directly, or to depend upon analytical results, should be based upon an analysis of the impact of a potential error of this magnitude.

8.4.4 Moment-of-Inertia Ratio

For a spinning spacecraft, the moment-of-inertia ratio between the three major axes is usually more important than the actual values of moment of inertia. (However, knowledge of moment of inertia about the spin axis is certainly necessary in computing spin-up requirements.) The reason is that, for a spinning

body in free space, the spin is most stable about the axis of maximum moment of inertia (Chapter 7). A spacecraft set spinning about one of the other axes will eventually shift its spin axis until it is spinning about the maximum moment-of-inertia axis. If there are no significant energy-dissipating mechanisms (e.g., flexible structures such as whip antennas or liquids) in the spacecraft, then spin about the lesser moment axis may be maintained for an extended period, e.g., hours or maybe even a day or so in extreme cases. However, any physical object will dissipate internal strain energy in the form of heat, and the presence of such mechanisms will eventually cause the shift. The classic example is the Explorer 1 satellite, a long, thin spinner with four wire whip antennas. After a relatively short time on orbit, spin shifted from a bullet-like spin about the long axis to a flat or propeller-like spin. This was merely an annoyance in the Explorer case, but such a flat spin or the coning motion that occurs in the transition from one axis to another can prove fatal to the mission in some cases. Active nutation control can prevent the shift or delay its onset, but of course this increases mass and complexity. Knowledge and control of moment-of-inertia ratio is therefore a major factor in the design of spinning spacecraft.

8.4.5 Mass Properties Bookkeeping

It is common to maintain mass properties lists with a contingency allocation to allow for unforeseen mass growth or other uncertainties. When done, it is important to vary the contingency allocation to reflect the changing state of knowledge of the mass properties. For example, in the conceptual phase of space vehicle design, it will be common to assume an allocation of 20% or more of contingency mass. As the design matures, this allocation will be reduced, and may be 1–2% for components whose design is fixed and that may even have flight heritage. Note that it is perhaps inadvisable to assume no contingency mass at all, even for systems with flight heritage. Until spacecraft integration and test operations are complete, there remains the possibility that a deficiency will be found in a new application of even a well-characterized design, and that additional mass will be needed as part of the solution.

8.5 Structural Loads

8.5.1 Sources of Structural Loads

The primary sources of structural loads that may be imposed on a spacecraft are 1) linear acceleration, 2) structurally transmitted vibration, 3) shock, 4) acoustic loads, 5) aerodynamic loads, 6) internal pressure, and 7) thermal stress. Although most are concerned with launch and ground handling, some affect the vehicle throughout its operating lifetime.

Linear acceleration is usually a maximum at staging, often of the first stage, which often has a higher thrust-to-weight ratio than the upper stages. The exception to this would be a vehicle such as the three-stage Delta, where the solid

third stage as it approaches the end of burn probably causes the highest acceleration.

Even though it is the factor most associated with space launch in the eyes of the layman, linear acceleration is often not the most significant design driver. This is especially true for an all-liquid-propellant launch vehicle where acoustic and vibration loads may well overshadow linear acceleration as design factors. In the case of a vehicle that reenters the atmosphere in a purely ballistic mode, the loads imposed during entry may well exceed those for launch. Lifting entry substantially reduces such loads.

Structurally transmitted vibration is one of the major design drivers. Main propulsion is usually the primary source of such vibration during the launch phase, although aerodynamic and other forces may also contribute and may dominate in particular cases. For example, "hammerhead" payload fairings are notorious for the inducement of aerodynamic buffeting loads induced at the point where the more bulbous front end "necks down" to the vehicle upper stage diameter. The space shuttle, which was designed to minimize longitudinal loads, is especially bad in terms of structurally transmitted vibration because the payload is mounted immediately above the engines and without the isolation afforded by a long, flexible tank assembly in between.

In addition to flight loads, however, the more prosaic ground handling and transportation loads may be significant as well. Although typically less intense, these inputs will be of longer duration. The several hours or days of vibration experienced on a truck as compared with that encountered during 8–10 minutes of launch may well be the dominant factor.

Shock loads in flight are usually associated with such functions as firing of pyrotechnic devices, release of other types of latches, or engagement of latches. Ground handling again can be a contributor, because such activities as setting the spacecraft on a hard rigid surface even at relatively low speeds can cause a significant shock load. Ground-handling problems can be minimized by proper procedures and equipment design. In-flight shocks may require isolation or relocation of devices farther from sensitive components.

Acoustic loads are most severe at liftoff because of reflection of rocket engine noise from the ground. They may also be fairly high in the vicinity of maximum dynamic pressure because of aerodynamically generated noise. This is especially true of the space shuttle with its large, flexible payload bay doors and the proximity of the payload to the engines. Acoustic loads are especially damaging to structures fabricated with large areas of thin-gauge material such as solar panels.

Aerodynamic load inputs to the payload come about as a result of their effect on the launch vehicle, because the payload is enclosed during passage through the atmosphere. Passage through wind shear layers or aerodynamic loads due to vehicle angle of attack caused by maneuvering can cause abrupt changes in acceleration. They may also cause deflection of the airframe of the vehicle. Since, in general, payloads of expendable launchers are cantilevered off the forward end

of the vehicle, airframe deflection has little impact on the payload except in extreme cases. In the case of the space shuttle, long payloads are attached at points along the length of the cargo bay. Deflection of the airframe can therefore induce loads into the payload structure. Some load alleviation provision is built into the attach points, and in many cases it is possible to design a statically determinant attachment that at least makes the problem reasonably easy to analyze. Very large or complex payloads may require attachment at a number of points, leading to a complex analytical problem. In some cases airborne support equipment (ASE) is designed to interface with the shuttle and take the loads from the airframe and protect the payload. This can be costly in payload capability, because all such ASE is charged against shuttle cargo capacity.

Internal pressure is a major source of structural loads, particularly in tanks, plumbing, and rocket engines. It may also be a source of loads during ascent in inadequately vented areas. Early honeycomb structures, especially nose fairings, sometimes encountered damage or failure because pressure was retained inside the honeycomb cavities while the external pressure decreased with altitude. Weakening of the adhesive, caused by aerodynamic heating, allowed internal pressure to separate the face sheets. Careful attention to venting of enclosed volumes is important in preventing problems of this type.

Internal pressure or the lack of it can also be a problem during handling and transportation. Some operations may result in reduced pressure in various volumes that must then resist the external atmospheric pressure. A common but by no means unique example is in the air transport of launch vehicle stages, especially in unpressurized cargo aircraft. If the internal tank pressure is reduced during high-altitude flight, either deliberately or because of a support equipment malfunction, then during descent the pressure differential across the tank walls can be negative, resulting in the collapse of the tank. Prevention of this simply requires attention and care, but the concern cannot be ignored.

Thermal stress usually results from differential expansion or contraction of structures subjected to heating or cooling. It may also arise as a result of differential heating or cooling. The former effect can be mitigated to some degree by selection of materials with compatible coefficients of thermal expansion.

Once the vehicle is in space, the primary sources of heat are the sun and any internally generated heat. The latter is usually the smaller effect, but cannot be ignored, especially in design of electronic components, circuit boards, etc. Differential heating caused by the sun on one side and the heat sink of dark space on the other can result in substantial structural loads. These are most easily dealt with by thermal insulation or by simply designing the structure to withstand the stress. Note that in a rotating spacecraft the inputs are cyclic, possibly at a fairly high rate. In massive structures the thermal inertia of the system tends to stabilize the temperature. However, if the material being dealt with is thin, substantial cyclic stress can be generated, possibly leading to eventual failure.

For low-orbit spacecraft, entry into eclipse results in rapid cooling of external surfaces and low thermal mass extremities, which can quickly become quite cool

without solar input. Upon reemergence into the sunlight, the temperature rapidly increases. This can cause not only substantial structural loads, but also sufficient deformation that accurate pointing of sensors may be difficult.

Thermal inputs to long booms of various types can easily cause substantial deflection, often of a cyclic nature. This in turn can couple with the structural design, possibly depending on local shadow patterns, to cause cyclic motion of the boom, and can cause instability in spacecraft pointing or at least increase the requirements on the attitude control system.

The presence of cryogenic materials onboard the spacecraft for propulsion or sensor cooling is a major source of thermally induced stress. The problem is complicated by the need for thermal isolation of the cryogenic system from the spacecraft structure to minimize heat leakage.

8.5.2 Structural Loads Analysis

Detailed analysis of structural loads usually requires the use of complex, but well established and understood, computer software such as NASTRAN. Modern computer aided design (CAD) packages, such as IDEA-S™, AutoCAD™, ProEngineer™, and numerous others, include this and many other features, offering outstanding interactive design capability to the structural engineer.

For preliminary purposes, however, inputs can usually be approximated using factors and formulas empirically derived from previous launches. Structural elements may then be sized in a preliminary manner using standard statics techniques.[2,3] The resulting preliminary size and mass estimates and material choices may then be refined with more sophisticated techniques.

It should be borne in mind that, although the sources of structural loads were discussed separately, they generally act in combination and must be used that way for design purposes. As an example, a cryogenic tank, pressurized during launch, will be subjected to thermally induced loads, internal pressure loads, and the vibration, linear acceleration, and acoustic loads of launch. Similarly, a deployable structure may encounter release and latching shocks while still under differential thermal stress resulting from exiting the Earth's shadow. Design load assessment must incorporate reasonable assumptions regarding such composite loads, based on the requirements of the actual flight profile.

8.5.3 Load Alleviation

Various means are used to alleviate structural loads. For example, the shuttle main engines are throttled back to approximately 65% of rated thrust during passage through the period of maximum dynamic pressure in ascent flight. Although this is done out of concern for the structural integrity of the orbiter, it can be beneficial to the payload as well. Most expendable vehicles lack this capability, although solid motor thrust profiles and angle-of-attack control may be practiced to moderate aerodynamic loads during this critical period.

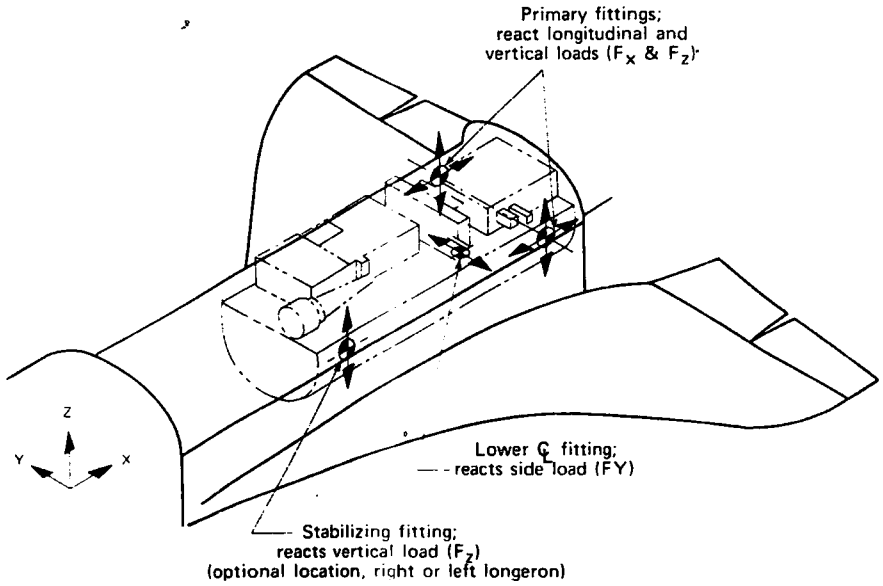


Fig. 8.13 Shuttle payload attachment.

Acoustic inputs can probably best be dealt with by design of the launch facility to minimize reflection of engine exhaust noise back to the vehicle. The payload must be designed to withstand whatever acoustic inputs the launch vehicle and launch facility impose. Use of stiffeners and/or dampening material on large, lightweight areas can help to minimize the structural response to these inputs.

The shuttle payload attachment system is designed to minimize input of airframe structural loads into the payload. Figure 8.13 presents the basic attachment concept. By providing one or more degrees of freedom at each attach point, a statically determinant attachment is created. However, for some payloads, which may be very long and flexible or otherwise not able to accept the loads, it will be necessary to design a structural support that interfaces to the orbiter attach points and isolates the payload itself from the orbiter airframe deflection.

8.5.4 Modal Analysis

Along with the loads analysis just discussed, it will be necessary to produce a structural dynamics model for use in launch vehicle coupled loads and attitude control analysis, as discussed in Chapters 3 and 7. This model, which is continually refined as the level of design definition increases, serves a variety of purposes.

The launch vehicle environment was discussed in Chapter 3, where it was seen that some launch vehicles are the source of considerable sine vibration, i.e., vibration at or near a specific frequency, and all are sources of random vibration.

It is necessary to ensure that the spacecraft has no resonant modes at or near any of those for the launch vehicle, or near any peaks of the random vibration spectrum. Usually there will be a basic specification that the first spacecraft mode must be higher than some threshold frequency, with other more specific concerns as noted. As mentioned earlier, preliminary analysis will be carried out assuming the launch vehicle and spacecraft are separate entities; later, it will be necessary to combine the rocket and space vehicle models and assess them as a single, fully coupled structure.

For launch vehicles themselves, and for some spacecraft, it will be required to verify that vehicle resonant modes do not closely couple to "slosh modes,"[4] which exist when propellant tanks are partially full. Launch vehicles tanks contain slosh baffles[5] and other design features to control these modes of oscillation, and spacecraft sometimes use "bladder" tanks to prevent it, but in all cases the issue must be addressed by the design.

Spacecraft structural modes are, as discussed in Chapter 7, also relevant to the attitude control system design. It is necessary either to keep the spacecraft primary mode well above the control system passband, or to include any offending modes as part of the "plant" to be controlled. This latter feature naturally complicates the design, but often cannot be avoided. Even then, failure to model the structure with sufficient accuracy can lead to difficulty, and, as Murphy's Law would have it, higher order modes are generally less accurately known than those of lower order. Often the worst problems are those associated with uncertainty in the structural damping ratio (see Chapter 7) to be assumed. Spacecraft structures are often quite lightly damped (e.g., $\zeta < 0.01$), and significant uncertainty in the actual value can lead to gross errors in estimating the settling time following maneuvers or other disturbances. A classic case in this regard is that of the original solar arrays on the Hubble Space Telescope;[6] unfortunately, however, this is far from the only such case.

Modal analysis can be performed via two basic methods.[7,8] The first is the so-called lumped mass model, in which the spacecraft structure is, for analytical purposes, modeled as a collection of discrete mass elements representing the various solar arrays, connecting booms, tanks, instruments, star trackers, primary structure, etc., which make up the complete vehicle. Each of these elements is assumed to be connected to its neighbors through a spring-and-dashpot arrangement that describes the stiffness of, and damping associated with, the individual connection. The result is a highly-coupled mass-spring-dashpot arrangement for which the motion of the elements is described by a coupled set of second-order ordinary differential equations,

$$\boldsymbol{M}\ddot{\boldsymbol{x}} + \boldsymbol{C}\dot{\boldsymbol{x}} + \boldsymbol{K}\boldsymbol{x} = \boldsymbol{F}(t) \tag{8.1}$$

where

$\boldsymbol{x} = (n \times 1)$ coordinate vector
$\boldsymbol{M} = (n \times n)$ mass matrix

$C = (n \times n)$ damping matrix
$K = (n \times n)$ stiffness matrix
$F = (n \times 1)$ forcing function vector
$n =$ degrees of freedom, $m \times d$
$m =$ number of discrete mass elements
$d =$ number of spatial dimensions

We cannot undertake the solution of Eq. (8.1) in this text; indeed, the treatment of vibration theory and modal analysis is the subject of numerous excellent texts.[9,10] The reader will not be surprised to find, however, that in close analogy with the classical one-degree-of-freedom (1-DOF) system, the solutions to Eq. (8.1) take the form of damped sinusoidal oscillations at the system modal frequencies.

It is also possible to obtain closed-form solutions for the vibrational behavior of numerous simple structures by means of continuum analysis. Among the structures for which solutions are known are strings, cables, rods, beams, torsional beams, plates, cylindrical shells, etc. Such results can be very useful in preliminary design. Blevins[11] provides an excellent compendium of techniques and results.

Historically, the approaches just outlined represented the only tenable ones for structural vibration analysis. The lumped-mass technique is still favored for relatively simple systems having few degrees of freedom. However, as stated earlier, the modern design engineer will almost always—and we are tempted to omit the word "almost"—have access to CAD programs. The ability to analyze the dynamical behavior of the structure in both free oscillation and as a result of applied loads is but one more feature of these state-of-the-art tools.

8.5.5 Fracture Mechanics

Fracture mechanics is a highly specialized field and will not be dealt with in any detail here. It is important, however, that the spacecraft designer be aware of the existence and purpose of the discipline.[12]

Although fracture mechanics analysis can be applied to any highly stressed part, its greatest application is to the design of pressure vessels. The most important characteristic of a pressure vessel, especially for man-rated applications, is the so-called *leak before burst* criterion. In other words, if a crack forms, it is desirable that it propagate *through* the tank wall before it reaches the critical crack length, which will result in the crack propagating *around* the tank. The leak thus provides warning and possibly pressure relief before catastrophic failure occurs.

Fracture mechanics analysis is used to compute the probability of failure and, if appropriate, leak before burst criteria based on numerous factors including the material, vessel size, wall thickness, pressure, contained fluid, environment, vessel history (particularly pressure cycles and exposure to various substances), and extensive empirical data on crack propagation under similar circumstances.

All of this information allows computation, to some level of confidence, of probability of failure and of leak before failure. Use of this technique is especially important for shuttle payloads. In most cases, program requirements for fracture mechanics analysis will be derived from, or essentially identical to, NASA standards in this area.[13]

8.5.6 Stress Levels and Safety Factors

In a great many cases, material choice and thickness of spacecraft structures will be driven by factors other than strength. The primary factors typically will be stiffness, i.e., minimizing deflection under load, and the minimum gauge of material that is available or that will allow it to be handled safely. In some cases, however, pressure vessels and some major structures being classic examples, the actual strength of the material to resist yielding or breakage is important. At this point safety factors become crucial.

Typical factors of safety will often be in the range of 1.2–1.5 for yield. That is, the structure is designed to yield only when subjected to loads 1.2–1.5 times the maximum expected to be encountered in service. Yield is defined in this case as undergoing a deformation in shape from which the structure does not recover when the load is removed. For all except very brittle materials, actual failure, i.e., structural breakage, takes place at stresses somewhat higher than yield. The ratio of yield stress to failure stress varies from one material to another, but typically if the factor of safety on yield is 1.5, the factor of safety to failure will be about 2.0. For some applications in manned spacecraft or man-rated systems, the factors of safety may be higher, especially for items critical to flight safety.

For noncritical components the safety factors may be lower than those discussed earlier. In the case of a component that is not safety related or critical to mission success, lower factors may be acceptable. In some cases a factor of 1.0 on yield might even be accepted, meaning that a small, permanent deformation is acceptable as long as the part does not break.

An important factor to be considered is the nature and duration of the load, particularly whether it is steady or cyclic. The factors previously discussed assume steady loads or very few cycles. If the load is cyclic, then the fatigue characteristic of the material is the major consideration. If many cycles are expected, then it is important to keep the stress in the material at a level that allows an acceptable fatigue life. Typically this will result in a structure substantially overdesigned compared to one which is required to withstand a static load of the same magnitude.

If the load is steady but will be applied for extremely long periods, the "creep" characteristics of the material may become important. An example might be a bolted joint, which is expected to maintain the same tension for years. However, the bolts might lose tension over long periods because of creep if subjected to a high level of stress, resulting in an inadequate creep life, even though there is no immediate danger of failure due to overload.

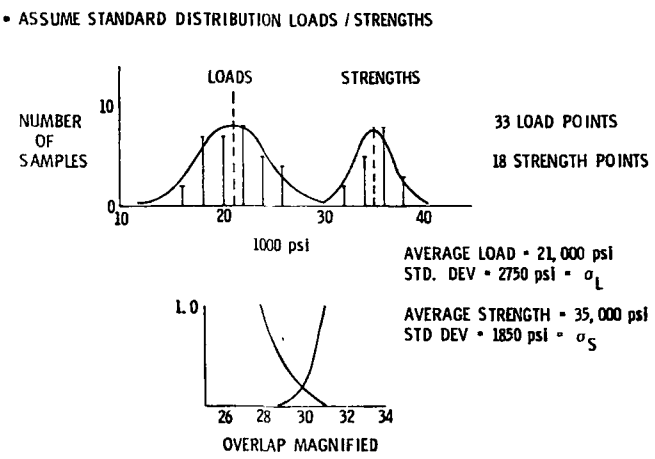


Fig. 8.14 Uncertainty distributions of loads and strength.

In considering safety factors, a frequently overlooked point is that all of the data, both loads and the material characteristics, have some associated uncertainty. This may be of secondary importance when designing ground equipment with safety factors of 5 or 10. It can be very critical, as we shall see, when designing for the small safety factors typical of aerospace hardware. This is most easily demonstrated by an example, depicted in Fig. 8.14.

In this example, we assume that we have a structural material with a quoted yield strength of 35,000 psi, such as might be typical of an aluminum alloy. Let us assume a load stress of 21,000 psi. If we simply apply a safety factor of 1.5, the allowable stress in the structure in question would then be 23,333 psi, and we would expect no problems with the 21,000 psi load. However, it is known that there exists a significant spread in the strength data available, the standard deviation being 1850 psi. Thus, to minimize the probability of failure, the 3σ low strength should be used. This amounts to $35{,}000-3(1850) = 29{,}450$ psi.

Some may be inclined to consider that "aluminum is aluminum" and use the handbook value;[14] however, there can be lot-to-lot variation or within-lot variations due to handling, processing, or environmental history that can be significant. In the case of many composite materials, the effect of the environment and the fabrication process is even more pronounced, and considerable attention must be paid to possible variations in characteristics. As a result of this, a composite structure is often designed with significantly higher factors of safety than metallic structures. This prevents achieving the full theoretical advantages attributed to composites.

Continuing with our example, we note that due to a variety of factors there is a deviation about the average load. We have an average value of 21,000 psi. Comparing that to the average strength yields an apparent factor of safety of $35{,}000/21{,}000 = 1.67$. Because our target factor of safety is 1.5, we might

naively be tempted to reduce the cross section, bringing the strength of the part down to 31,500 psi, and saving weight. This would be a dangerous error, because the standard deviation of the load value in this case is 2750 psi. A combination of the 3σ high load (29,250 psi) and the 3σ low strength (29,450 psi) essentially uses up the entire design safety factor; the original part will survive, but just barely. However, the "thinned down" part would have a 3σ low yield stress of only 25,950 psi, and would fail.

To be certain of a 1.5 safety factor in the worst-case combination of 3σ high load and 3σ low strength, the original value of a 1.67 safety factor based on the average must be maintained.

It can be seen from the preceding discussion that a relatively small increase in safety factor can have a substantial impact on probability of failure. Less apparent but equally true is the fact that increased safety factor can allow substantial cost saving. With a larger safety factor it may be possible to reduce the amount of testing and detailed analysis required with a resulting reduction in costs. Thus, availability of substantial mass margins can translate into a much lower cost program if program management is clever enough to take advantage of the opportunity thus offered. This requires management to avoid the pitfall of proceeding with a sophisticated test and analysis effort simply because "that's the way we have always done it." The JPL/Ball Aerospace Solar Mesosphere Explorer (SME) is a textbook example of a program that took advantage of ample mass margin to keep spacecraft costs low.

Figure 8.15 presents a handy means of estimating failure rate based on the average loads and strengths and the standard deviation about each. The vertical axis plots number of combined standard deviations, i.e., standard deviation in load plus standard deviation in strength times the number on the vertical axis. The horizontal axis plots number of failures per 10^7 load events. For example, a failure rate of one per million load events requires 3.5 combined standard

SF_{00} = SAFETY FACTOR BASED UPON MEANS =

$$\frac{1 + I\alpha_L}{1 - I\alpha_S}$$

SF_{03} = SAFETY FACTOR BASED UPON AVERAGE STRENGTH AND 3 σ LOAD =

$$\frac{SF_{00}}{1 + 3\alpha_L}$$

IN PREVIOUS EXAMPLE $SF_{33} = 1.67$

$$I = \frac{S_{avg} - L_{avg}}{\sigma_S + \sigma_L}$$ = NUMBER OF COMBINED DEVIATIONS SEPARATING LOAD MEAN FROM STRENGTH MEAN

$$\alpha_L = \frac{\sigma_L}{L_{avg}} \qquad \alpha_S = \frac{\sigma_S}{S_{avg}}$$

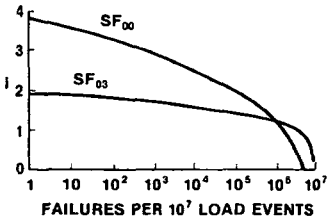


Fig. 8.15 Safety factor vs failure rate.

deviations separating the average values. Also plotted are the related safety factors where the upper curve SF_{00} is based upon the average values of both load and strength, whereas the lower curve is based on average strength divided by average load plus three standard deviations. Note that the horizontal axis is load events. This may be one per mission in some cases and in others it could be hundreds, thousands, or even millions per mission if the member is loaded in a cyclic or vibratory fashion.

This has been a very cursory treatment of a complex subject that is generally not well understood. The point is that a simple statement of a value as "safety factor" is meaningless without understanding the basis from which it is derived. Furthermore, safety factor and failure rate trade with mass and cost and should be considered in that light.

8.6 Large Structures

As space activities increase in variety and complexity, it is to be expected that there will be increased interest in very large structures. In fact, proposals have already been made for solar arrays and microwave antennas on a scale of kilometers to beam power from geostationary orbit to Earth. The popular view, encouraged by those of entrepreneurial bent, is that, in a weightless environment, structures can be arbitrarily large and light in mass. Although there is an element of truth in this, there are major practical limitations with which the designer of such structures must deal.

Most large structures such as solar arrays, antennas, telescopes, etc., must maintain shape to a fairly tight tolerance if they are to function effectively. This is often difficult with the small structures that are currently in use and becomes far more so on a scale of tens of meters or kilometers. Thermal distortion for a given temperature differential is a direct function of the dimensions of the structure. Bigger structures distort more in an absolute sense. In most applications, attitude control maneuvers are required. A very lightweight flexible structure will distort during maneuvers and in response to attitude hold control inputs because of the inertia of the structure. As the force is removed, the structure springs back, but the low-mass, low-restoring force and (probably) near-zero damping tend to give rise to low-frequency oscillations that die out very slowly and, in fact, may excite control system instability.

Obviously, the control force distortion concern can be partially alleviated by using relatively small forces and maneuvering very slowly. However, operational needs will dictate some minimum maneuver rate and settling time, and the system must be able to deal with anticipated disturbances. These requirements will set a lower limit to the control forces required.

Even the assumption of weightlessness is not entirely valid for very large structures. In such large structures, the forces caused by the radial gradient in the gravity field can cause distortion, at least in lower-altitude orbits.

Unfortunately, the accuracy requirements do not change as the size of the structure increases. A microwave antenna still requires surface accuracy on the order of a wavelength, whether it is 1 m or 1 km in diameter. Although solar arrays need not maintain the same degree of surface control as an antenna, it is still important to maintain shape with reasonable accuracy. In any case, excessive structural distortion makes accurate pointing almost impossible.

For very large space structures, it is simply not practical to maintain shape by designing strength and stiffness into the structure. The mass of such a structure would be enormous, increasing transportation costs to an intolerable level. The greater complexity of assembling a more massive structure is also a matter of concern. In any case, it is not at all clear that a brute force approach could solve the problem. New materials, particularly composites that offer the possibility of tailoring characteristics such as stiffness and thermal response, can contribute greatly but do not offer a total solution.

The concerns just listed clearly indicate that large space structures are by no means as simple as their proponents, enamored by the tremendous promise offered by such structures, have indicated. Worthwhile large structures must be relatively easy to deploy and assemble and must have predictable, repeatable, controllable characteristics. To properly design control systems, it is necessary to be able to model the response with satisfactory fidelity. This capability is now becoming available through the use of modern, high-capacity computers.

One way to deal with the problem is active shape control. This concept has been successfully utilized for large, Earth-based optical telescopes. The shape of the surface determined using laser range finders or by measuring the energy distribution in the beam leaving an antenna is used as input to an active control system that mechanically or thermally distorts the surface to compensate for structural irregularities. In the case of a phased array antenna, the phasing can be altered to accomplish the same end. Note that this is a potential solution to the surface control problem. The operation of the structure as a spacecraft, i.e., attitude control and gross pointing, remains and demands adequate modeling and solution to the control input, response, settling time, and flexibility problems.

The control of large, flexible structures is a complex issue involving optimization among material characteristic choices, structural design approach, and control system design. Adding to the complexity is the probable requirement to launch in many separate pieces that are themselves folded into a compact shape. Each piece must be deployed, checked out, and joined to its mating pieces in as straightforward and automatic a fashion as possible to create the structure that the system was designed to control.

8.7 Materials

8.7.1 Structural Materials

Most materials used in space applications to date have been the conventional aerospace structural materials. Properties of some representative materials may

be found in Appendix B. These will continue to dominate for the foreseeable future, although steady growth in the use of newer materials is to be expected.

Among the conventional structural materials, aluminum is by far the most common. A large variety of alloys exist, providing a broad range of such characteristics as strength and weldability. Thus, for applications at moderate temperature in which moderate strength and good strength-to-weight ratio are desirable, aluminum is still most often the material of choice. This popularity is enhanced by ready availability and ease of fabrication. A number of surface-coating processes exist to allow tailoring of surface characteristics for hardness, emissivity, absorbtivity, etc.

Magnesium is often used for applications in which higher stiffness is desired than can be provided by aluminum. It is somewhat more difficult to fabricate and, being more chemically active than aluminum, requires a surface coating for any extensive exposure to the atmosphere. Several coatings exist. Environmental constraints in recent years have limited the availability of certain desirable magnesium alloys containing zirconium.

Steel, in particular stainless steel, is often used in applications requiring higher strength and/or higher temperature resistance. A variety of steels may be used, but stainless steel is often preferred because its use eliminates concern about rust and corrosion during the fabrication and test phase. Additionally, if the part may be exposed to low temperature, the low ductile-to-brittle transition (DBT) temperature of stainless steel and similar alloys is an important factor.

Titanium is a lightweight, high-strength structural material with excellent high-temperature capability. It also exhibits good stiffness. Some alloys are fairly brittle, which tends to limit their application, but a number of alloys with reasonable ductility exist. Use of titanium is limited mostly by higher cost, lower availability, and fabrication complexity to applications that particularly benefit from its special capabilities. Pressure vessels of various types and external skin of high-speed vehicles are typical applications.

Beryllium offers the highest stiffness of any naturally occurring material along with low density, high strength, and high temperature tolerance. Thermal conductivity is also good. Beryllium has been used in limited applications where its desirable characteristics have been required. The main limitation on more extensive use of this apparently excellent material is toxicity. In bulk form, beryllium metal is quite benign and can be handled freely. The dust of beryllium or its oxide, however, has very detrimental effects on the human respiratory tract. This means that machining or grinding operations are subject to extensive safety measures to capture and contain dust and chips. This renders normal fabrication methods unusable without resorting to these intensive (i.e., expensive) measures.

Glass fiber-reinforced plastic, generically referred to as fiberglass, was the first composite material used for space structure and is probably still the most common. The matrix material may be epoxy, phenolic, or other material, and the glass can range from a relatively low-quality fiber all the way to highly processed quartz fiber. Fiberglass is desirable because of the relative ease with which

complex shapes can be fabricated. It also exhibits good strength and offers the ability to tailor strength and stiffness both in absolute value and direction in the material by choice of fiber density and orientation.

Graphite-epoxy is in very common use and may even have supplanted fiberglass in frequency of use. The use of high strength and stiffness graphite fiber in a matrix of epoxy or other polymer makes an excellent high-strength structural material. Proper selection of the cloth and/or unidirectional fibers offers the ability to tailor strength and stiffness directly and to the desired levels to optimize it for the purpose. The low density of graphite offers a weight advantage as well. High temperature characteristics are improved by use of graphite instead of glass, although the matrix is the final limiting factor. An increasing number of high-temperature polymers are available for higher temperature structures.

In addition to graphite, Kevlar® and other high-strength fibers are increasingly used.

The inconel family of alloys and other similar alloys based on nickel, cobalt, etc., are used for high-temperature applications. Typical application is as a heat shield in the vicinity of a rocket nozzle to protect the lower temperature components from thermal radiation or hot gas recirculation. These alloys are of relatively high density, equal to that of steel or greater so, weight can be a problem. However, inconel in particular lends itself to processing into quite thin foils, which allows its use as a shield, often in multiple layers, with minimum mass penalty.

New materials coming into use are mostly composites of various types, although some new alloys have also appeared. Among the alloys, aluminum-lithium is of considerable interest, because the addition of the lithium results in alloys of somewhat higher strength than the familiar aluminum alloys, but having equal or lower density. This material is already seeing extensive use in commercial aviation and in the most recent version of the space shuttle external tank.

High-temperature refractory metals have been available for many years but have seen limited use because of high density, lack of ductility, cost, and other factors. Tungsten, tantalum, and molybdenum fall into this category. These materials are actually somewhat less available than they were some years ago. A great many suppliers have dropped out of the field. This may in part be related to the collapse of the commercial nuclear power industry in the United States. One exception is niobium (formerly called columbium). This material is useful to temperatures as high as 1300 K but has a density only slightly higher than steel. It is available in commercial quantities. Like all the refractory metals, it oxidizes rapidly if heated in air, but a silicide coating offers substantial protection in this environment.

Metal matrix composites involve use of a metal matrix, e.g., aluminum, stiffened and strengthened by fibers of another metal or nonmetallic material. In aluminum, for example, fibers of boron, silicon carbide, and graphite have been used. Some difficulties have been encountered, such as the tendency of the molten

aluminum to react with the graphite during manufacture of the composite. Work on protective coatings continues. Boron-stiffened aluminum is well developed and is used in the tubular truss structure that makes up much of the center section of the shuttle orbiter. This entire area is one of enormous promise. As yet, we have hardly scratched the surface of the potential of this type of composite.

Carbon-carbon composite consists of graphite fibers in a carbon matrix. It has the ability to hold shape and resist ablation and even oxidation at quite a high temperature. For very high temperature use, an oxidation resistant coating, usually silicon carbide, is applied. At the present level of development, however, carbon-carbon is not suitable for a load-bearing structure. For example, it is used in the nose cap and wing leading edges of the shuttle orbiter where it must resist intense reentry heating, but it does not form a part of the load-bearing structure. Progress is being made in the development of structural carbon-carbon, and it is expected to have a bright future as a hot structure for high-speed atmospheric and entry vehicles.

Carbon–silicon-carbide, carbon fiber in a silicon-carbide matrix, is making considerable progress as a high-temperature material. It shows promise of being as good as or better than carbon-carbon in terms of offering a high use temperature with better oxidation resistance.

8.7.2 Films and Fabrics

By far the most commonly used plastic film material in space applications has been Mylar™. This is a strong, transparent polymer that lends itself well to fabrication into sheets or films as thin as 0.00025 in. Coated with a few angstroms of aluminum to provide reflectivity, Mylar™ is well suited to the fabrication of the multilayer insulation extensively used on spacecraft.

A newer polymeric film material with higher strength and the ability to withstand higher temperature than Mylar™ is the polyimide Kapton™. These characteristics have made Kapton™ a desirable choice for outer layers of thermal blankets. A problem has arisen with the discovery that, in low Earth orbits, polymer surfaces undergo attack and erosion by atomic oxygen, which is prevalent at these altitudes (see Chapter 3). Kapton™ seems to be more susceptible to this sort of attack than Mylar™. In any case, for long-life use in low orbit, metallization or coating with a more resistive polymer such as Teflon® will probably be required. The erosion rate is sufficiently low that, for shorter missions, the problem may not be serious.

Teflon® and polyethylene have been used extensively as bearings, rub strips, and in various protective functions because of their smoothness, inertness, and, particularly for Teflon®, lubricative ability.

Fiberglass cloth, which is strong and flexible, has been used as an insulator and as protective armor against micrometeoroids. A commercially available cloth of fiberglass coated with Teflon® called Betacloth™ has been used as the external surface of spacecraft thermal blankets for this purpose.

A variety of materials superficially similar to fiberglass but of much higher temperature capacity are available. These materials are made from fibers of high-temperature ceramic material and are available as batting, woven cloth, and thread. The most well-known application of such materials is as the flexible reusable surface insulation (FRSI) used on the upper surfaces of the later-model shuttle orbiters. They can also be useful as insulators of high-temperature devices such as rocket engines.

8.7.3 Future Trends

As has been the case in the past, future trends in materials will be characterized by a desire for increased specific strength and specific stiffness. The latter will tend to dominate because, as observed earlier, most space structure designs are driven by stiffness more than strength. Higher thermal conductivity with lower coefficient of thermal expansion is also highly desirable for obvious reasons. Figure 8.16 indicates desirable trends in stiffness and thermal characteristics. The currently available materials are grouped to the left with beryllium still showing an edge even over the composites. Graphite-aluminum offers the possibility of substantial improvement once its problems are solved, and graphite-magnesium shows even greater promise for the future. It is quite probable that other candidates will emerge as research continues.

Damping capability is also important as a means of reducing sensitivity to vibration and shock. Figure 8.17 rates damping ability vs density. The common aerospace alloys are generally poor, magnesium being the best. Excellent dampers are available as indicated toward the upper right hand; however, they tend to be heavy, dirty, and relatively weak and have a high DBT temperature. All of these characteristics make them unusable for space applications. The developing field of composites may offer the best hope of achieving the goal, although the present trend to use high-stiffness fibers may make this difficult.

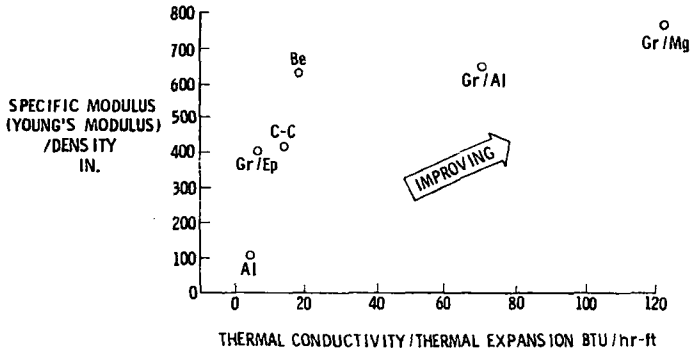


Fig. 8.16 Desired structural and thermal characteristics.

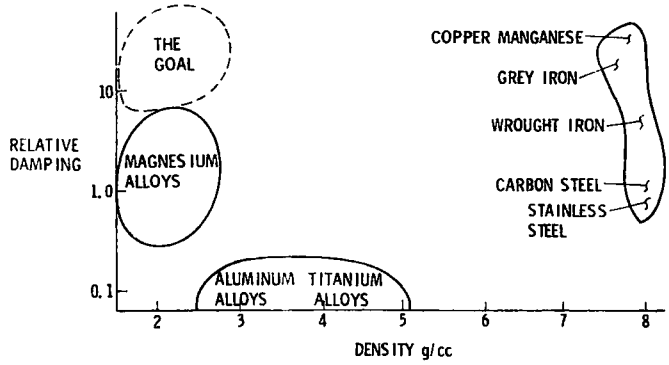


Fig. 8.17 Damping capability.

Refractory metals stiffened with high-temperature fibers, structural carbon-carbon, and other new material developments should open new avenues for entry thermal protection. This will allow replacement of the existing fragile shuttle tiles with hardier versions and offer improved capability in future entry systems.

References

[1]Shigley, J. E., and Mischke, C. R., *Mechanical Engineering Design*, 5th ed., McGraw-Hill, New York, 1989.

[2]Beer, F. P., and Johnston, E. R., Jr., *Mechanics of Materials*, 2nd ed., McGraw-Hill, New York, 1992.

[3]Boresi, A. P., Schmidt, R. J., and Sidebottom, O. M., *Advanced Mechanics of Materials*, 5th ed., Wiley, New York, 1993.

[4]"Propellant Slosh Loads," NASA SP-8009, Aug. 1968.

[5]"Slosh Suppression," NASA SP-8031, May 1969.

[6]Foster, C. L., Tinker, M. I., Nurre, G. S., and Till, W. A., "The Solar Array-Induced Disturbance of the Hubble Space Telescope Pointing System," NASA TP-3556, May 1995.

[7]"Natural Vibration Modal Analysis," NASA SP-8012, Sept. 1968.

[8]"Structural Vibration Prediction," NASA SP-8050, June 1970.

[9]Thomson, W. T., and Dahleh, M. D., *Theory of Vibration with Applications*, 5th ed., Prentice-Hall, Upper Saddle River, NJ, 1998.

[10]Chopra, A. K., *Dynamics of Structures*, Prentice-Hall, Upper Saddle River, NJ, 1995.

[11]Blevins, R. D., *Formulas for Natural Frequency and Mode Shape*, Krieger, Malabar, FL, 1995.

[12]Anderson, T. L., *Fracture Mechanics—Fundamentals and Applications*, 2nd ed., CRC Press, New York, 1995.

[13]"Fracture Control Requirements for Payloads Using the Space Shuttle," NASA STD-5003, Oct. 1996.

[14]Baumeister, T. F., Avallone, E. A., and Baumeister, T. F., III, *Marks Standard Handbook for Mechanical Engineers*, 8th ed., McGraw-Hill, New York, 1978.

9
Thermal Control

9.1 Introduction

The thermal control engineer's task is to maintain the temperature of all spacecraft components within appropriate limits over the mission lifetime, subject to a given range of environmental conditions and operating modes. Thermal control as a space vehicle design discipline is unusual in that, given clever technique and reasonable circumstances, the thermal "system" may require very little special-purpose spacecraft hardware. More demanding missions may require extra equipment such as radiators, heat pipes, etc., to be discussed in the following sections. In all cases, however, the required analysis will involve the thermal control engineer in the design of nearly all other onboard subsystems.

As with attitude control, thermal control techniques may be broadly grouped within two classes, passive and active, with the former preferred when possible because of simplicity, reliability, and cost. Passive control includes the use of sunshades and cooling fins, special paint or coatings, insulating blankets, heat pipes, and tailoring of the geometric design to achieve both an acceptable global energy balance and local thermal properties.

When the mission requirements are too severe for passive techniques, active control of spacecraft temperatures on a local or global basis will be employed. This may involve the use of heating or cooling devices, actively pumped fluid loops, adjustable louvers or shutters, radiators, or alteration of the spacecraft attitude to attain suitable conditions.

Most readers will recall the basic heat transfer mechanisms: conduction, convection, and radiation. Broadly generalizing, it may be said that the overall energy balance between a spacecraft and its environment is dominated by radiative heat transfer, that conduction primarily controls the flow of energy between different portions of the vehicle, and that convection is relatively unimportant in space vehicle design. As with all generalizations this is an oversimplification, useful to a point but allowing numerous exceptions. This will be seen in the following sections.

As always, our treatment of this topic will be very limited in its sophistication. Examples are provided for illustrative purposes, not as guidelines for detailed design. Wertz and Larson[1] provide a useful discussion for those requiring

additional detail, and Gilmore[2] offers an especially comprehensive treatment of spacecraft thermal design and engineering practice.

9.2 Spacecraft Thermal Environment

Comments on the space thermal environment were offered in Chapter 3 as part of our discussion of the overall space environment. However, it is useful to expand on our earlier discussion prior to considering the design features that are intended to deal with that environment.

The spacecraft thermal environment can vary considerably, depending upon a variety of naturally occurring effects. Orbital characteristics are a major source of variation. For example, most spacecraft orbits will have an eclipse period; however, as the orbit precesses, the time and duration of the eclipse will vary, particularly for a highly elliptic orbit. Obviously, for a spacecraft in interplanetary flight where the orbit is about the sun, the solar intensity will vary as the distance from the sun changes. As discussed in Chapter 4, even the solar intensity experienced in orbit around the Earth will vary seasonally (from an average value of 1388 W/m^2) because of the ellipticity of the Earth's orbit around the sun.

In addition to direct solar input to the spacecraft, there will be reflected solar input to the vehicle from whatever planet it orbits. This reflected solar energy input depends on the orbital altitude, the planetary reflectivity or albedo, and the orbital inclination. Reflected solar input decreases with altitude, as does the range of variation that must be accommodated. Planetary albedo varies with latitude and, depending on the planet and its surface features, possibly longitude and season as well. Values can range from a lower limit of roughly 5% to over 85%. Interestingly, the lunar surface, which appears quite bright from Earth, has a very low average albedo. The upper end of the range would be represented by reflection of sunlight from heavy cloud cover on Earth. The albedo will also be a strong function of wavelength. This can be a problem because it can be difficult to find surface materials or coatings that are good reflectors across a wide spectrum of wavelengths. As we will see, polished surfaces that are good reflectors in the visible spectrum may well be very efficient absorbers at infrared. A worst-case scenario in this regard might be low-altitude flight over the day side of the planet Mercury, where infrared irradiance from the surface will be a major factor in the design.

Operational activities alter the thermal environment as well. Very low orbital altitudes can produce heating due to free-molecular flow (see Chapter 6). Spacecraft attitude may change, resulting in exposure of differing areas and surface treatments to the sun and to space. Onboard equipment may be turned on or off, resulting in changes in the amount of internally generated heat. In the course of thruster firings, local cooling may occur in tanks or lines due to gas expansion at the same time as local heating may occur in the vicinity of hot gas

thrusters. Expenditure of propellant reduces the thermal mass of the tanks and the spacecraft as a whole, resulting in differences in the transient response to changing conditions.

As flight time in space increases, spacecraft surface characteristics change due to ultraviolet exposure, atomic oxygen attack, micrometeoroid/debris impact, etc. This will affect both the absorptivity and emissivity of the surfaces and must be considered in the design of long-life spacecraft.

Anomalous events provide an unpredictable source of change in the thermal environment. A failure in a wiring harness may cause loss of part of the solar array power, or a power-consuming instrument may fail, thus reducing internally generated heat. A sun shade or shield may fail to deploy, louvers may stick, etc. Although one cannot predict every possible problem, nor can a spacecraft be designed to tolerate every possible anomaly, it is desirable to provide some margin in the design to allow for operation at off-design conditions.

9.3 Thermal Control Methods

9.3.1 Passive Thermal Control

The techniques applied for passive thermal control include the use of geometry, coatings, insulation blankets, sun shields, radiating fins, and heat pipes. By "geometry" we imply the process of configuring the spacecraft to provide the required thermal radiating area, placing low-temperature objects in shadow, and exposing high-temperature objects to the sun or burying them deeply within the structure, and other similar manipulation of the spacecraft configuration to optimize thermal control.

Insulation blankets typically feature a multilayer design consisting of several layers of aluminized Mylar or other plastic, spaced with nylon or Dacron mesh. External coverings of fiberglass, Dacron®, or other materials may be used to protect against solar ultraviolet radiation, atomic oxygen erosion, and micrometeoroid damage.

Sun shields may be as simple as polished, or perhaps gold plated, aluminum sheet. More sophisticated reflectors may use silvered Teflon®, which essentially acts as a second-surface mirror with the silver on the back to provide visible-light reflectivity, with the Teflon® providing high infrared emissivity. Along the same line are actual glass second-surface mirrors, which are more thermally efficient, but have the cost of greater weight and possible problems with the brittle glass.

Fins are often used where it is necessary to dissipate large amounts of heat, or smaller amounts at low temperature, thus requiring a large cooling surface area. Large numbers of fins in circular configurations will have difficulty obtaining an adequate view factor to space. Very long fins may be limited in effectiveness by the ability to conduct heat through the fins.

Heat pipes are tubular devices containing a wick running the length of the pipe, which is partially filled with a fluid such as ammonia. The pipe is connected between a portion of the spacecraft from which heat is to be removed and a portion to which it is to be dumped. The fluid evaporates from the hot end, and the vapor is driven to condense (thus releasing its heat of vaporization) at the cold end. Condensed fluid in the cold end is then drawn by capillary action back to the hot end.

Some may question whether heat pipes belong in the passive category, because there is active circulation of fluid within the heat pipe driven by the heat flow. We consider heat pipes to be passive from the viewpoint of the spacecraft designer because there is no direct control function required, nor is there a requirement for the spacecraft to expend energy. The heat pipe simply conducts energy when there is a temperature differential and ceases to do so if the differential disappears. Control of heat pipes is possible by means of loaded gas reservoirs or valves. This of course reduces the advantages of simplicity and reliability that are inherent in the basic design.

Caution in using heat pipes is required to make sure that the hot end is not so hot as to dry the wick completely, thus rendering capillary action ineffective in transporting new fluid into that end. Similarly, the cold end must not be so cold as to freeze the liquid. Also, heat pipes work quite differently in 0*g* because of the absence of free convection, making interpretation of ground test results a problem unless the heat pipe is operating horizontally. It is customary to provide a 50% margin in energy transfer capacity when sizing a heat pipe for spacecraft applications.

9.3.2 Active Thermal Control

Active thermal control of spacecraft may require devices such as heaters and coolers, shutters or louvers, or cryogenic materials. Thermal transport may be actively implemented by pumped circulation loops.

Heaters usually are wire-wound resistance heaters, or possibly deposited resistance strip heaters. Control may be by means of ground command, or automatically with onboard thermostats, or both. For very small heaters where on/off control is not required, radioisotope heaters are sometimes used. The usual size is 1-W thermal output. It might be argued that such devices are passive, because they cannot be commanded and do not draw spacecraft power.

Various cooling devices have been applied or are under consideration. Refrigeration cycles such as those that are used on Earth are difficult to operate in 0*g* and have seen little or no use. Thermoelectric or Peltier cooling has been used with some success for cooling small, well-insulated objects. The primary application is to the cooling of detector elements in infrared observational instruments that are operated for long periods. The Villaumier refrigerator is of

considerable interest for similar applications, and development of such devices has been in progress for many years.

A straightforward device that has seen considerable use is the cryostat, which depends on expansion of a high-pressure gas through an orifice to achieve cooling. To achieve very low temperature, two-stage cryostats using nitrogen in the first stage and hydrogen in the second have been used. The nitrogen, expanded from high pressure, precools the system-to near liquid nitrogen temperature. The hydrogen, expanding into the precooled system, can then approach liquid hydrogen temperatures, thus cooling the instrument detectors to very low temperatures. Other gases may be used as well.

For long-term cooling to low temperature, an effective approach is to use a cryogenic fluid. The principal applications have been to spacecraft designed for infrared measurements, such as the infrared astronomy satellite (IRAS), launched in 1983, or the Cosmic Background Explorer (COBE) spacecraft, launched in 1989. In these spacecraft, cooling is achieved by expansion of supercritical helium (stored at 4.2 K) through a porous plug to as low as 1.6 K. This allows observations at very long infrared wavelengths with the minimum possible interference to the telescope from its own heat.

IRAS performed the first all-sky infrared (IR) survey, expiring after nearly 11 months of operation, upon depletion of its helium. The more sophisticated COBE spacecraft showed that the cosmic microwave background spectrum is that of a nearly perfect blackbody at a temperature of 2.725 ± 0.002 K, an observation that closely matches the predictions of the so-called Big Bang theory.[3] The COBE helium supply was depleted after approximately 10 months of operation.

The DoD/Missile Defense Agency's Mid-course Space Experiment (MSX) included an infrared telescope for the purpose of tracking missiles and reentry vehicles. It was launched in 1996 and, like COBE, operated for about 10 months. While this telescope was routinely used for military surveillance experiments, some observing time was also devoted to astronomical observations. MSX utilized a block of solid hydrogen as its fundamental coolant, offering a step up in sophistication from the IRAS/COBE experience.

The observational lifetime of each of these satellites was less than a year, at least for their far-infrared instruments, due to exhaustion of their onboard refrigerant, even though all other systems were still functioning. This provides a strong argument for the development of both cryogenic refrigerators and cost-effective on-orbit servicing techniques, neither of which has yet reached the required level of maturity.

Shutters or louvers are among the most common active thermal control devices. Common implementations are the louver, which essentially resembles a venetian blind, or the flat plate with cutouts.

The former may be seen in the Voyager spacecraft illustration in Chapter 8. A fixed outside plate with pie-slice cutouts is provided. Between that plate and the spacecraft itself is a movable plate with similar cutouts, which is rotated by a

bimetallic spring. When the spacecraft becomes warm, the plate moves to place the cutouts in registration, and thus expose the spacecraft skin to space. When the spacecraft becomes too cold, the movable plate rotates to close the cutouts in the fixed plate, thus reducing the exposure of the spacecraft skin to space.

The flat-plate variety is shown on the Television and Infrared Observation Satellite/Defense Meteorological Satellite Program (TIROS/DMSP) spacecraft illustration in Chapter 8. The flat plate is rotated by the bimetallic element. The plate has cutout sectors that are placed over insulated areas to decrease heat flow and rotate over uninsulated areas to increase heat flow and cool the spacecraft. The flat-plate variety is much simpler and less costly but allows less efficient use of surface area and fine tuning of areas on a given surface. Although the automatic control described is most common and usually satisfactory, it is obviously possible to provide commanded operation as well, either instead of the thermostatic approach or as an override to it.

Actively pumped fluid loops, conceptually identical to the cooling system in an automobile engine, have a long history of spaceflight applications. In this approach, a tube or pipe containing the working fluid is routed to a heat exchanger in the area or region to be heated or cooled. Heat transfer occurs via forced convection (see the following section) into the fluid. The fluid is circulated to an energy source or sink, where the appropriate reverse heat exchange takes place. Working fluids in typical applications include air, water, methanol, water/methanol, water/glycol, Freon, carbon tetrachloride, and others.

The most visible space application of this cooling technique is to the space shuttle, where the payload-bay doors contain extensive cooling radiators that, while on orbit, are exposed to dark space. Indeed, the doors must be opened shortly after orbital injection or the mission must be aborted and the shuttle returned to Earth. Other manned-flight applications of fluid loop cooling included the Mercury, Gemini, and Apollo programs. The Apollo lunar surface suits featured water-cooled underwear with a heat exchanger in the astronaut's backpack.

Active fluid cooling was also briefly mentioned in Chapter 5 in connection with regenerative engine nozzle cooling. This technique, while complex, is a primary factor enabling the design of high thrust-to-weight rocket engines.

9.4 Heat Transfer Mechanisms

Heat transfer mechanisms affecting spacecraft are of course the same as those with which we are familiar on Earth: conduction, convection, and radiation. The primary difference is that convection, which is very often the overriding mechanism on Earth, is usually nonexistent in space. Still, convection will be encountered on the surface of any planet with an atmosphere, during atmospheric flight, and inside sealed pressurized spacecraft and pumped fluid cooling loops.

All three mechanisms will be discussed in the sections that follow.

9.4.1 Conductive Heat Transfer

Conduction occurs in solids, liquids, and gases. It is usually the primary mechanism for heat transfer within a spacecraft (although radiation may be important in internal cavities). Because all electronic devices generate at least some heat while in operation, there exists a risk of overheating if care is not taken to provide adequate paths to conduct heat from the component to the appropriate heat rejection surface.

Of course, the same concern exists with ground-based equipment. However, thermal design of such equipment is usually much less of a problem because of the efficiency of free convection in providing heat relief. It is also largely self-regulating. In special cases, such as cooling the processor chip of a computer or the final amplifier stage of a radio transmitter, the ground-based designer can provide a small fan to ensure forced convection over a particular area. Free convection is unavailable in space, even in pressurized spacecraft, because of the lack of gravity, and fan cooling is generally found only in manned spacecraft. Deliberate provision of adequate conduction paths is therefore a key requirement for the spacecraft thermal engineer.

Design practice in providing thermal conduction involves more than selecting a material with suitable conductivity. For example, unwelded joints, especially in vacuum, are very poor thermal conductors. Worse yet, they may exhibit a factor of two or more variability in conduction between supposedly identical joints. This situation can be substantially improved by use of conduction pads, thermal grease, or metal-loaded epoxy in joints that are mechanically fastened. Obviously this is only done where high or repeatable conductivity is essential to the design.

Regarding materials selection, it is found that high thermal conductivity and high electrical conductivity normally are closely related. Therefore, a situation in which high thermal conductivity is required while electrical isolation is maintained is often difficult. One substance that is helpful is beryllium oxide (BeO), which has high thermal conductivity but is an excellent electrical insulator. Care must be taken in the use of BeO, which in powder form is highly toxic if breathed.

9.4.2 Fourier's Law of Heat Conduction

The basic mathematical description of heat conduction is known as Fourier's law, written one-dimensionally as

$$Q = -\kappa A\left(\frac{dT}{dx}\right) \tag{9.1}$$

and shown schematically in Fig. 9.1. Q is the power (energy per unit time), expressed in watts, British thermal units per second, or the equivalent. A is the area through which the heat flow occurs, and κ is the thermal conductivity in units such as watts per meter per Kelvin or British thermal units per hour per foot per

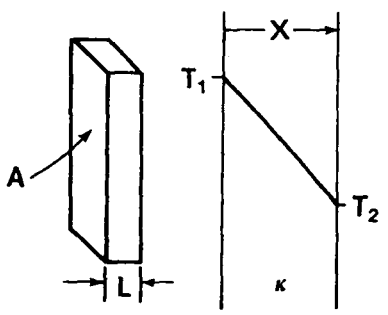


Fig. 9.1 Conduction in one dimension.

degree Fahrenheit. T is the temperature in absolute units such as Kelvins or degrees Rankine, and x is the linear distance over the conduction path. Qualitatively, Eq. (9.1) expresses the commonly observed fact that heat flows from hot to cold, as well as the fact that a more pronounced temperature difference results in a higher rate of energy transfer.

It is often more useful to consider the power per unit area, or energy flux, which we denote as

$$q \equiv \frac{Q}{A} = -\kappa\left(\frac{dT}{dx}\right) \tag{9.2}$$

with units of watts per square meter. Vectorially, Eq. (9.2) may be extended for isotropic materials to

$$\boldsymbol{q} = -\kappa \nabla \boldsymbol{T} \tag{9.3}$$

Equation (9.3) may be applied to the energy flux through an arbitrary control volume; invoking Gauss's law and the law of conservation of energy yields the conduction equation

$$\rho C \frac{\partial T}{\partial t} = \kappa \nabla^2 T + g(\boldsymbol{r}, t) \tag{9.4}$$

which allows the temperature in a substance to be calculated as a function of the position vector $\boldsymbol{r}$ and time. The source term $g(\boldsymbol{r}, t)$ accounts for internal heat generation (power per unit volume). C is the heat capacity of the substance, with units such as joules per kilogram per Kelvin, and ρ is its density. The term ∇^2 is the Laplacian operator, which in Cartesian coordinates is

$$\nabla^2 = \frac{\partial}{\partial x^2} + \frac{\partial}{\partial y^2} + \frac{\partial}{\partial z^2} \tag{9.5}$$

and is given for other coordinate systems of interest in standard references.[4]

The conduction equation is interesting mathematically for the range of solutions that are exhibited in response to differing initial and boundary

conditions. Except in simple cases, which are outlined in standard texts,[5] a numerical solution is usually required to obtain practical results. As always, a discussion of numerical techniques is outside the scope of this text.

Generally, one wishes to solve the conduction equation to obtain the temperature distribution in some region. This region will be defined by the coordinates of its boundary, along which certain conditions must be specified to allow a solution to be obtained. In the example of Fig. 9.1, the infinite slab is defined as a region by faces at $x = 0$ and $x = L$, with no specification on its extent in the y and z directions. (Equivalently, the slab may be considered to be well insulated at its edges in the y and z directions, so that no heat flow is possible.) One might wish to know the temperature at all points within (0, L) given knowledge of the slab's properties and the conditions on either face.

Boundary conditions for the conduction equation may be of two general types. Either the temperature or its derivative, the heat flux (through Fourier's law), may be specified on a given boundary. For a transient problem, the initial temperature distribution throughout the region must also be known. Let us consider the simple case of Fig. 9.1 and assume the faces at $x = (0, L)$ to have fixed temperatures T_0 and T_L. Then Eq. (9.4) reduces to

$$\frac{d^2T}{dx^2} = 0 \tag{9.6}$$

which has the general solution

$$T(x) = ax + b \tag{9.7}$$

Upon solving for the integration constants, we obtain

$$T(x) = (T_L - T_0)\frac{x}{L} + T_0 \tag{9.8}$$

and from Fourier's law, the heat flux through the slab is found to be

$$q = -\kappa\left(\frac{dT}{dx}\right) = -\kappa\frac{(T_L - T_0)}{L} \tag{9.9}$$

Note that, instead of specifying both face temperatures, we could equally well have specified the heat flux at one face (which in this constant-area steady-state problem must be the same as at the other face) and a single boundary temperature. Assuming that T_L and the heat flux q_w are known, we obtain, after twice integrating Eq. (9.6),

$$T(x) = -\left(\frac{q_w}{\kappa}\right)x + b \tag{9.10}$$

and upon solving for the constant of integration,

$$T(x) = (L - x)\frac{q_w}{\kappa} + T_L \tag{9.11}$$

It is seen that T_0 is now obtained as a solved quantity instead of a known boundary condition. Clearly, either approach can be used, but it is impossible to specify simultaneously both the face temperature and also the heat flux. Moreover, two boundary conditions are always required; specification of one face temperature, or the heat flux alone, is insufficient.

This is a simple but useful example to which we will return. In transient cases, or if two- or three-dimensional analysis is required, or when internal sources of energy are present, solutions to Eq. (9.4) rapidly become more complicated if they can be found at all, and are beyond the intended analytical scope of this book. The interested reader is referred to standard heat transfer texts[5,6] for treatment of a variety of useful basic cases.

One particularly useful transient case is that of the semi-infinite solid initially at temperature T_0 at time $t_0 = 0$, with a suddenly applied temperature T_w or flux q_w at $x = 0$ for $t > 0$. The geometry is that of Fig. 9.1, with $L \to \infty$. With no sources present, and conduction in one dimension only, Eq. (9.4) becomes

$$\frac{\partial T}{\partial t} = \left(\frac{\kappa}{\rho C}\right)\frac{\partial^2 T}{\partial x^2} = \alpha \frac{\partial^2 T}{\partial x^2} \tag{9.12}$$

where $\alpha = \kappa/\rho C$ is the thermal diffusivity. The solution for the suddenly applied wall temperature is[5]

$$T(x) = T_w + (T_0 - T_w)\ \text{erf}\ \eta \tag{9.13}$$

where

$$\eta = \frac{x}{2\sqrt{\alpha T}} \tag{9.14}$$

and erf η is the error function or probability integral, tabulated in standard texts,[17] and given formally as

$$\text{erf}\ \eta = \frac{2}{\sqrt{\pi}}\int_0^{\eta} e^{-\lambda^2} d\lambda \tag{9.15}$$

For convenience, Table 9.1 provides a few values of the error function.

When a sudden heat flux $q_w = -\kappa \partial T/\partial x$ is applied at $x = 0$, we have

$$T(x) = T_0 - 2\sqrt{\alpha t}\left(\frac{q_w}{\kappa}\right)\left(\eta\ \text{erfc}\ \eta - \frac{e^{-\eta^2}}{\sqrt{\pi}}\right) \tag{9.16}$$

where erfc $\eta = 1 - \text{erf}\ \eta$ is the complementary error function.

It should be appreciated that the solutions of Eqs. (9.13) and (9.16) are of more value than they might initially appear. Although the true semi-infinite solid is of course nonexistent, these solutions apply to the transient flow through a plate or slab where the time is sufficiently short that the far side of the plate remains essentially at the initial temperature.

Table 9.1 Error function

η	erf η	η	erf η
0	0	1.1	0.8802
0.1	0.1125	1.2	0.9103
0.2	0.2227	1.3	0.9340
0.3	0.3286	1.4	0.9523
0.4	0.4284	1.5	0.9661
0.5	0.5205	1.6	0.9763
0.6	0.6039	1.7	0.9838
0.7	0.6778	1.8	0.9891
0.8	0.7421	1.9	0.9928
0.9	0.7969	2.0	0.9953
1.0	0.8427	∞	1.0000

To obtain a closed-form analytic solution to Eq. (9.4) requires, at a minimum, that the boundary surfaces be constant-coordinate surfaces (in whatever coordinate system the problem is posed), and that $g(r, t)$ be of very simple form. When these conditions are not satisfied, a situation more common than otherwise in engineering practice, numerical solution of the governing equations is required. We will touch on this topic in later sections.

9.4.3 Convective Heat Transfer

Of all the heat transfer mechanisms, convection is the most difficult to analyze, predict, or control. This is because it is essentially a fluid dynamic phenomenon, with behavior dependent on many factors not easily measurable or predictable. Part of the problem arises because convection is in truth not a heat transfer mechanism at all. The energy is still transferred by conduction or radiation, but the conditions defining the transfer are highly modified by mass transport in the fluid. This is illustrated schematically in Fig. 9.2.

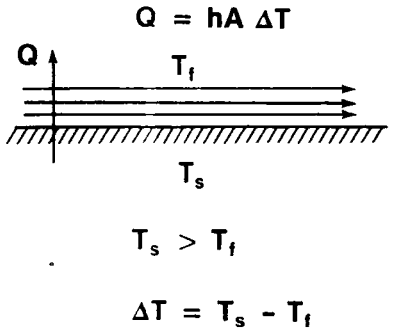


Fig. 9.2 Thermal convection.

So-called free convection is driven entirely by density differences and thus occurs only in a gravitational field. It does not occur in space except when the spacecraft is accelerating. However, it does occur unavoidably on Earth and thoroughly skews the application to vacuum conditions of any heat transfer data that might be obtained from testing the spacecraft in the atmosphere. This fact is a primary (but not the sole) reason for conducting spacecraft thermal vacuum tests prior to launch. It is literally the only opportunity available to the thermal control analyst to verify his results in something approximating a space environment.

If convective heat transfer is required in $0g$, it must be forced convection, driven by a pump, fan, or other circulation mechanism. The interior of a manned spacecraft cabin is one example. Another might be a propellant or pressurization tank where good thermal coupling to the walls is required. Forced convection is not commonly used as a significant means of unmanned spacecraft thermal control in U.S. or European spacecraft. However, Russian spacecraft have historically made extensive use of sealed, pressurized unmanned spacecraft with fans for circulation as a means of achieving uniform temperature and presumably to avoid the concern of operating some components in a vacuum. There is an obvious tradeoff here; design is much easier, but overall reliability may be lower because the integrity of the pressure hull is crucial to spacecraft survival.

A spacecraft having the mission of landing on a planet with an atmosphere, or operating within that atmosphere, must of course be designed to deal with the new environment, including free convection, as well as operation on Earth and during launch and interplanetary cruise. Although no such design problem can be viewed as trivial, the Mars environment presents unusual challenges. An atmosphere exists, but it is approximately equivalent to that of Earth at 30-km altitude. There is enough atmosphere to allow free convection to be significant, but not enough for it to be the dominant heat transfer mechanism that it is on Earth. Solar radiation is lower by a factor of two than on Earth, but is not so low that it can be ignored in the daytime, particularly at lower latitudes. The thin atmosphere does not retain heat once the sun has set, resulting in thermal extremes that approach those of orbital flight. Finally, windblown dust will settle on the lander surface, altering its thermal radiation properties and greatly complicating the analysis that must be done in the design phase. Although other planetary environments can be much harsher in particular respects, few if any offer as much variability as does Mars.

As discussed earlier, convection is important for space applications in various types of pumped cooling loops such as cold plates for electronics, regeneratively cooled rocket engines, and waste heat radiators. This of course is forced convection involving the special case of pipe or channel flow.

Convective heating is the critical mechanism controlling entry heating. It completely overpowers the radiative component until the entry velocity begins to approach Earth escape velocity. Even then, convection is still the more significant contributor. Similarly, it is the major mechanism in ascent aerodynamic heating. We have discussed this special case rather thoroughly in Chapter 6. In Table 9.2

Table 9.2 Thermal protection materials

Thermal protection system	Type	Advantages	Disadvantages
AVCOAT 5025	Low-density charring ablator	Low density ($\rho = 34$ lb/ft^3) Thoroughly tested Man rated Low thermal conductivity	Manual layup in honeycomb matrix Erosion capability estimated only
HTP-12-22 fibrous refractory composite insulation (FCRI)	Surface reradiation	Low density ($\rho = 12$ lb/ft^2) Does not burn Good thermal shock tolerance Can maintain shape and support mechanical loads Low thermal conductivity	May melt under certain flight conditions ($T_{melt} = 3100$°F) Uncertain erosion capability
ESM 1030	Low-density charring ablator	Low density ($\rho = 16$ lb/ft^3)	Erosion capability unknown
Carbon–carbon over insulator	Surface reradiation and heat sink	Erosion capability known	High conductivity Possible thermal expansion problems Requires silicon carbide coating for oxidation resistance
Silica phenolic	High-density charring ablator	Erosion capability known Low thermal conductivity	High density ($\rho = 105$ lb/ft^3)
Carbon phenolic	High-density charring ablator	Erosion capability known	High density ($\rho = 90$ lb/ft^3) Oxidation resistance uncertain

we include a summary of several common entry vehicle thermal protection materials.

9.4.4 Newton's Law of Cooling

For forced convection of a single-phase fluid over a surface at a moderate temperature difference, it was discovered by Newton that the heat transfer is proportional to both the surface area and the temperature difference. The convective heat flux into the wall may then be written according to Newton's law of cooling as

$$Q = h_c A \Delta T = h_c A (T_f - T_w) \tag{9.17}$$

where Q is the power, h the convection or film coefficient, A the area, and ΔT the driving temperature differential from T_f and T_w, the fluid and wall temperatures. As before, it is often more useful to deal with the heat flux,

$$q \equiv \frac{Q}{A} = h_c (T_f - T_w) \tag{9.18}$$

Equation (9.18) is the analog to Eq. (9.9) for one-dimensional heat conduction, with h_c assuming the role of κ/L, where we recall that L is the characteristic thickness of the slab through which the heat flows.

Recalling the one-dimensional transient heat conduction solution given earlier, we may have the case where a convective heat flux of the form of Eq. (9.18) is suddenly applied to the surface of a semi- infinite solid. Gebhart[5] gives the solution for temperature within the solid as

$$\frac{T(x) - T_0}{T_w - T_0} = \text{erfc}\ \eta - \exp\left[\frac{h_c}{\kappa}\left(x + \frac{h_c}{\kappa}\alpha t\right)\right]\text{erfc}\left(\eta + \frac{h_c}{\kappa}\alpha t\right) \tag{9.19}$$

The crucial element in Eq. (9.18) is the coefficient h_c. Values for h_c are for the most part both empirical and highly variable. Engineering handbooks[8] publish charts or tables giving ranges of values for the h_c under varying sets of conditions, but the variance is usually significant, and tests under the specific conditions being considered may be required if the necessary accuracy is to be obtained. Because convective heat transfer is a mass transport phenomenon as well as a thermal one, the coefficient depends strongly on whether the flow is laminar or turbulent, with the turbulent value being much higher. Thus, a laminar-to-turbulent transition along the surface of an entry body may result in a substantial increase in heating downstream of the transition point.

In most cases, convective heat transfer will result in a higher flux than with conduction. Forced convection is in turn more effective than free convection, which is driven entirely by the difference in density caused by the heat transfer in the presence of gravity. This relationship is illustrated qualitatively in Fig. 9.3. The film coefficient for free convection depends strongly on the orientation of the surface relative to the local vertical and, as noted earlier, does not occur in 0g.

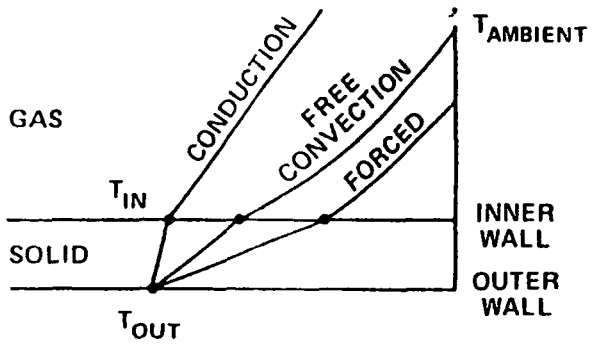


Fig. 9.3 Comparison of heat transfer mechanisms.

Newton's law of cooling is of course an approximation. The problem of heat transfer from a moving fluid to a boundary wall is a fluid dynamic problem, sometimes one that may be analyzed by means of particular approximations. If the fluid is a coolant in a pipe or tube, it may often be idealized as axisymmetric or one-dimensional incompressible viscous flow, for which closed-form solutions exist.[18]

At the other extreme is the flow of high-speed air along an exterior wall, for which we may apply the approximations of boundary-layer theory, which again yields numerous practical results. These have been discussed in Chapter 6, in connection with reentry vehicle heating. In either case, the analytical solution of a problem allows us to compute the value of h_c for use in the convection law. However, all of our comments elsewhere in this text concerning the intractability of fluid dynamics problems apply here as well; thus, direct solution for the film coefficient is restricted to a few special cases such as those just described.

It is both customary and advantageous in fluid dynamics to work in terms of non-dimensional parameters. In convection analyses, the appropriate parameter is the Nusselt number, defined as the ratio of convective energy transfer to conductive energy transfer under comparable conditions. For example, in the one-dimensional case just discussed, assume a wall is heated by a slab of fluid having thickness L and mass-averaged temperature T_f. If the fluid is stagnant, then from Eq. (9.9) the heat flux into the wall is

$$q_{\text{cond}} = (T_f - T_w)\frac{\kappa}{L} \tag{9.20}$$

whereas if the fluid is moving, convection occurs and the heat flux is

$$q_{\text{conv}} = h_c(T_f - T_w) \tag{9.21}$$

The ratio of convective to conductive heat transfer would then be

$$Nu = \frac{q_{\text{conv}}}{q_{\text{cond}}} = \frac{h_c L}{\kappa} \tag{9.22}$$

Thus, heat transfer at low Nusselt number, of order one, is essentially conductive; the slow flow of fluid through a long pipe offers a good example. High Nusselt number (100–1000) implies efficient convection; in the pipe example, this would correspond to rapid, turbulent flow in the pipe. Convective heat transfer experiments (or computations) are very frequently expressed in terms of the Nusselt number.

Equation (9.22) allows us to rewrite Newton's law of cooling in terms of Nusselt number and thermal conductivity,

$$q = Nu\left(\frac{\kappa}{L}\right)(T_f - T_w) \tag{9.23}$$

In this example, L was the thickness of the fluid slab. In a more general situation, L is a characteristic length scale for the particular case of interest. In the important special case of axisymmetric pipe flow, the pipe diameter D would be the natural choice. In the more general case of flow in a duct of arbitrary cross section, D is commonly taken to be the hydraulic diameter, given by

$$D_h = 4\frac{A}{P} \tag{9.24}$$

where A is the cross-sectional area of the flow, and P is the wetted perimeter of the duct.

To illustrate the application of the Nusselt number in heat transfer analysis, we continue with our circular pipe-flow example. For fully developed laminar flow (i.e., low-speed flow several pipe diameters downstream from the entrance), the Nusselt number is found to be[8]

$$Nu = 3.66 \text{ (constant pipe wall temperature)} \tag{9.25a}$$

$$Nu = 4.36 \text{ (constant pipe wall heat flux)} \tag{9.25b}$$

whereas for fully developed turbulent flow we have in both cases

$$Nu = 0.023 Re_x^{4/5} Pr^{1/3} \tag{9.25c}$$

valid for $0.7 < Pr < 160$, $Re_x > 10{,}000$, and $1/D > 60$. The Reynolds and Prandtl numbers are given by

$$Re_x = \frac{\rho V x}{\mu} \tag{9.26}$$

and

$$Pr = \frac{\mu Cp}{\kappa} \tag{9.27}$$

with

ρ = fluid density
V = flow velocity

x = downstream length from duct entrance
μ = fluid viscosity
Cp = fluid heat capacity

A feel for the uncertainty inherent in the use of empirical correlations such as Eq. (9.25c) may be gained by recognizing that this result is not unique. Various refinements have been published; for example, it has been found that using $Pr^{0.3}$ for cooling and $Pr^{0.4}$ for heating yields slightly more accurate results.

Results such as those just presented can be used to estimate the power per unit area, or flux, that can be extracted via forced convection in pipes or tubes, and are given here primarily for illustrative purposes. However, it should be understood that many other questions remain to be answered in the design of a practical cooling system. For example, a pump will be needed to move fluid through the system. Fluid flow in lengthy pipes will be subject to substantial friction; bends in the pipe as needed to realize a compact design add to this friction, which affects the size and power required of the pump. We ignore all such issues in favor of the more specialized references cited earlier.

9.4.5 Radiative Heat Transfer

Radiation is typically the only practical means of heat transfer between a vehicle in space and its external environment. Mass expulsion is obviously used as a spacecraft coolant when open-cycle cryogenic cooling is performed, as already discussed for IRAS, COBE, and MSX, but this should be regarded as a special case. As noted previously, radiation becomes important as a heat transfer mode during atmospheric entry at speeds above about 10 km/s. Even at entry speeds of 11.2 km/s (Earth escape velocity), however, it still accounts for only about 25% of the total entry heat flux. At very high entry speeds, such as those encountered by the Galileo atmospheric probe at Jupiter, radiative heat transfer dominates.

Radiative energy transfer can strongly influence the design of certain entry vehicles, particularly those where gliding entry is employed. Because convective heating is the major source of energy input, the entry vehicle surface temperature will continue to grow until energy dissipation due to thermal radiation exactly balances the convective input. This illustrates the reason for and importance of a good insulator (such as the shuttle tiles) for surface coating of such a vehicle. It is essential to confine the energy to the surface, not allowing it to soak back into the primary structure. Tauber and Yang[9] provide an excellent survey of design tradeoffs for maneuvering entry vehicles.

Radiative heat transfer is a function of the temperature of the emitting and receiving bodies, the surface materials of the bodies, the intervening medium, and the relative geometry. The intensity, or energy per unit area, is proportional to $1/r^2$ for a point source. If the distance is sufficient, almost any object may be considered a point source. An example is the sun, which subtends a significant arc

in the sky as viewed from Earth but may be considered a point source for most purposes in thermal control.

The ability to tailor the aborptivity and emissivity of spacecraft internal and external surfaces by means of coatings, surface treatment, etc., offers a simple and flexible means of passive spacecraft thermal control. Devices such as the louvers and movable flat-plate shades discussed previously may be viewed as active means of varying the effective total emissivity of the spacecraft.

It will be seen that the heat flux from a surface varies as the fourth power of its temperature. Thus, for heat rejection at low temperature a relatively large area will be required. This may constitute a problem in terms of spacecraft configuration geometry, where one must simultaneously provide an adequate view factor to space, compact launch vehicle stowage, and minimal weight.

9.4.6 · Stefan-Boltzmann Law

Radiative heat transfer may be defined as the transport of energy by electromagnetic waves emitted by all bodies at a temperature greater than 0 K. For purposes of thermal control, our primary interest lies in wavelengths between approximately 200 nm and 200 μm, the region between the middle ultraviolet and the far infrared. The Stefan-Boltzmann law states that the power emitted by such a body is

$$Q = \varepsilon \sigma A T^4 \tag{9.28}$$

where T is the surface temperature, A the surface area, and ε the emissivity (unity for a blackbody, as we will discuss later). The Stefan-Boltzmann constant σ is $5.67 \times 10^{-8}\ \mathrm{W/m^2 \cdot K^4}$.

Notation conventions in radiometry are notoriously confusing and are often inconsistent with those used in other areas of thermal control. To the extent that a standard notation exists, it is probably best exemplified by Siegel and Howell,[10] and we will adopt it here. Using this convention, we define the hemispherical total emissive power e as

$$e \equiv q = \frac{Q}{A} = \varepsilon \sigma T^4 \tag{9.29}$$

The name derives from the fact that each area element of a surface can "see" a hemisphere above itself. The quantity e is the energy emitted, including all wavelengths, into this hemisphere per unit time and per unit area.

9.4.7 The Blackbody

The blackbody, as the term is used in radiative heat transfer, is an idealization. By definition, the blackbody neither reflects nor transmits incident energy. It is a perfect absorber at all wavelengths and all angles of incidence. As a result,

provable by elementary energy-balance arguments, it also emits the maximum possible energy at all wavelengths and angles for a given temperature. The total radiant energy emitted is a function of temperature only.

Although true blackbodies do not exist, their characteristics are closely approached by certain finely divided powders such as carbon black, gold black, platinum black, and Carborundum. It is also possible to create structures that approximate blackbody behavior. For example, an array of parallel grooves (such as a stack of razor blades) or a honeycomb arrangement of cavities can be made to resemble a blackbody. Such structures may be used in radiometers.

The actual emissivity ε and absorptivity α that characterize how real bodies emit and absorb electromagnetic radiations often differ in value and are dissimilar functions of temperature, incidence angle, wavelength, surface roughness, and chemical composition. These differences can be used by the spacecraft designer to control its temperature. As an example, a surface might be chosen to be highly reflective in the visible light band to reduce absorption of sunlight and highly emissive in the infrared to enhance heat rejection. Silver-plated Teflon® was mentioned earlier as one material having such properties. Figure 9.4 shows α/ε values for a variety of common thermal control materials.

For analytical convenience, real bodies are sometimes represented as blackbodies at a specific temperature. The sun, for example, is well represented

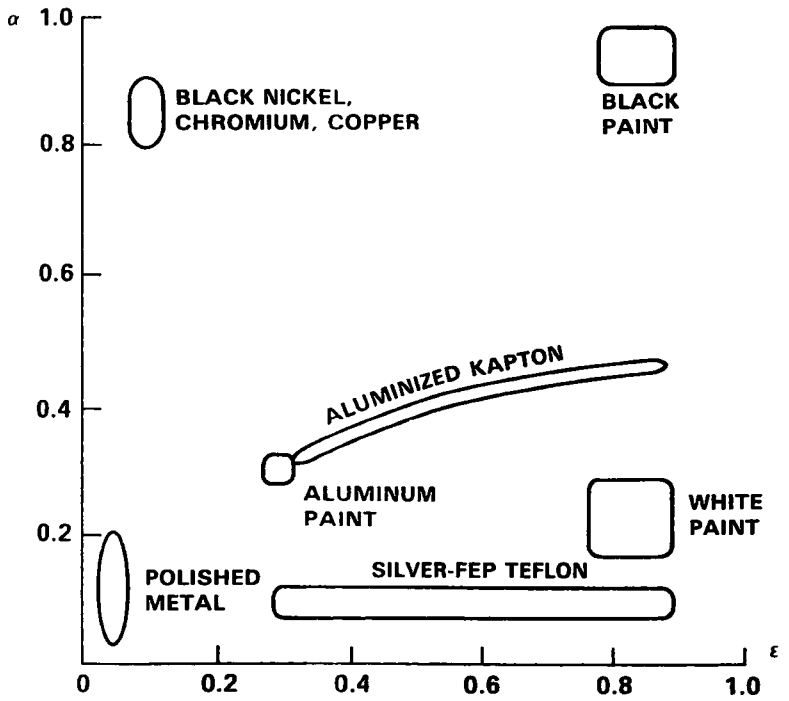


Fig. 9.4 Typical solar absorptivity and emissivity.

for thermal control purposes by a blackbody at 5780 K, and the Earth can be modeled as a blackbody at 290 K.

The equation describing blackbody radiation is known as Planck's law, after the German physicist Max Planck, who derived it in 1900. Because this development required the deliberate introduction by Planck of the concept of energy quanta, or discrete units of energy, it is said to mark the initiation of modern, as opposed to classical, physics. Planck's law is

$$e_{\lambda b} = e_{\lambda b}(\lambda, T) = \frac{2\pi hc^2}{\lambda^5(e^{hc/\lambda kT} - 1)} \tag{9.30}$$

where

$h = 6.626 \times 10^{-34}$ J·s = Planck's constant
$k = 1.381 \times 10^{-23}$ J/K = Boltzmann's constant
$c = 2.9979 \times 10^{8}$ m/s = speed of light

The subscript b implies blackbody conditions, and e_λ denotes the hemispherical spectral emissive power, i.e., the power per unit emitting surface area into a hemispherical solid angle, per unit wavelength interval. Care with units is required in dealing with Eq. (9.30) and its variations. Dimensionally, $e_{\lambda b}$ has units of power per area and per wavelength; however, one should take care that wavelengths are expressed in appropriate units, such as micrometers or nanometers, whereas area is given in units of m^2 or cm^2. If care is not taken, results in error by several orders of magnitude are easily produced.

Planck's law is for emission into a medium with unit index of refraction, i.e., a vacuum. It must be modified in other cases.[10]

Planck's law as given finds little direct use in spacecraft thermal control. However, it is integral to the development of a large number of other results. Included among these is Wien's displacement law, readily derivable from the Planck equation, which defines the wavelength at which the energy emitted from a body is at peak intensity. This may be considered the principal "color" of the radiation from the body, found from

$$(\lambda T)_{\max} = \left(\frac{1}{4.965114}\right)\frac{hc}{k} = 2897.8\ \mu\text{m} \cdot \text{K} \tag{9.31}$$

The Earth's radiation spectrum is observed to have a peak at $\lambda = 10\ \mu$m. Applying this fact and Eq. (9.31) yields the result given earlier that Earth is approximately a blackbody at a temperature of 290 K.

The important fourth-power relationship empirically formulated by Stefan and confirmed by Boltzmann's development of statistical thermodynamics may be derived by integrating Planck's law over all wavelengths. When this is done, one

obtains

$$e_b = \int_0^\infty e_{\lambda b}(\lambda, T)\mathrm{d}\lambda = \sigma T^4 \tag{9.32}$$

It is usually of greater practical interest to evaluate the integral of Eq. (9.32) between limits λ_1 and λ_2. This is most readily done by noting from Planck's law that an auxiliary function $e_{\lambda b}/T^5$ can be defined that depends only on the new variable (λT). Tables of the integral of $e_{\lambda b}/T^5$ may be compiled and used to evaluate the blackbody energy content between any two points $\lambda_1 T$ and $\lambda_2 T$. A few handy values for the integral over $(0, \lambda T)$ are found in Table 9.3.

9.4.8 Radiative Heat Transfer Between Surfaces

The primary interest in radiative heat transfer for spacecraft thermal control is to allow the energy flux between the spacecraft, or a part of the spacecraft, and its surroundings to be computed. This requires the ability to compute the energy transfer between arbitrarily positioned pairs of "surfaces"; the term is in quotes because often one surface will be composed totally or partially of deep space. The key point is that any surface of interest, say A_i, radiates to and receives radiation from all other surfaces A_j within its hemispherical field of view. All of these surfaces together enclose A_i and render a local solution impossible in the general case; the coupling between surfaces requires a global treatment. The problem is relatively tractable, though messy, when the various surfaces are black. When they are not, a numerical solution is required in all but the simplest cases. Fortunately, a few of these simple cases are of great utility for basic spacecraft design calculations.

9.4.9 Black Surfaces

Figure 9.5 shows two surfaces A_1 and A_2 with temperatures T_1 and T_2 at an arbitrary orientation with respect to each other. If both surfaces are black, the net

Table 9.3 Blackbody emissive fraction in range (0, λT)

λT, μm·K	$e_{0,\lambda T}/e_b$
1448	0.01
2191	0.10
2898	0.25
4108	0.50
6149	0.75
9389	0.90
23,220	0.99

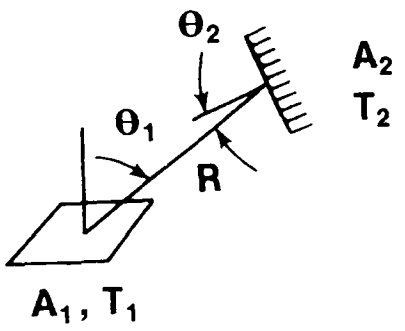


Fig. 9.5 Radiative heat transfer between black surfaces.

radiant interchange from A_1 to A_2 is

$$Q_{12} = \sigma(T_1^4 - T_2^4)A_1F_{12} = \sigma(T_1^4 - T_2^4)A_2F_{21} \tag{9.33}$$

where F_{ij} is the view factor of the jth surface by the ith surface. Specifically, F_{12} is defined as the fraction of radiant energy leaving A_1 that is intercepted by A_2. Note the reciprocity in area-view factor products that is implicit in Eq. (9.33). View factors, also called configuration or angle factors, are essentially geometric and may be easily calculated for simple situations. In more complex cases, numerical analysis is required. Extensive tables of view factors are available in standard texts.[10]

When the surfaces of an enclosure are not all black, energy incident on a nonblack surface will be partially reflected back into the enclosure; this continues in an infinite series of diminishing strength. The total energy incident on a given surface is then more difficult to account for and includes contributions from portions of the enclosure not allowed by the view factors F_{ij} for a black enclosure. Moreover, nonblack surfaces can and generally will exhibit variations in absorptivity, reflectivity, and emissivity as a function of the azimuth and elevation angle of the incident beam relative to the surface. Variations in all these characteristics with color will also exist. These complications render an analytical solution essentially impossible in most cases of interest. Excellent computational methods exist for handling these cases, mostly based on or equivalent to Hottel and Sarofim's net radiation method.[11]

9.4.10 Diffuse Surfaces

The simplest nonblack surface is the so-called diffuse gray surface. The term "gray" implies an absence of wavelength dependence. A "diffuse" surface offers no specular reflection to an incident beam; energy is reflected from the surface with an intensity that, to an observer, depends only on the projected area of the surface visible to the observer. The projected area is the area normal to the

observer's line of sight:

$$A_\perp = A \cos \theta \tag{9.34}$$

where θ is the angle from the surface normal to the line of sight. Thus, the reflected energy is distributed exactly as is energy emitted from a black surface; it looks the same to viewers at any angle. Reflected energy so distributed is said to follow Lambert's cosine law; a surface with this property is called a Lambertian surface. A fuzzy object such as a tennis ball or a cloud-covered planet such as Venus represents a good example of a diffuse or Lambertian reflector. Surfaces that are both diffuse and gray may be viewed conceptually as black surfaces for which the emissivity and absorptivity are less than unity.

The energy emitted by a gray surface A_1 is given by Eq. (9.28). The portion of this energy that falls upon a second surface A_2 is given by

$$Q = \varepsilon_1 \sigma A_1 F_{12} T_1^4 \tag{9.35}$$

This radiation, incident on a nonblack surface, can be absorbed with coefficient α, reflected with coefficient ρ, or transmitted with coefficient τ. From conservation of energy,

$$\alpha + \rho + \tau = 1 \tag{9.36}$$

If a surface is opaque ($\tau = 0$), then Kirchoff's law states that the surface in thermal equilibrium has the property that, at a given temperature T, $\alpha = \varepsilon$ at all wavelengths. This result, like all others, is an idealization. Nonetheless, it is useful in reducing the number of parameters necessary in many radiative heat transfer problems and is frequently incorporated into gray surface calculations without explicit acknowledgment.

A case of practical utility is that of a diffuse gray surface A_1 with temperature T_1 and emissivity ε_1 and which cannot see itself ($F_{11} = 0$, a convex or flat surface), enclosed by another diffuse gray surface A_2 with temperature T_2 and emissivity ε_2. If $A_1 \ll A_2$ or if $\varepsilon_2 = 1$, then the radiant energy transfer between A_1 and A_2 is[5]

$$Q_{12} = \varepsilon_1 \sigma A_1 (T_1^4 - T_2^4) \tag{9.37}$$

The restrictions on self-viewing and relative size can be relaxed at the cost of introducing the assumption of uniform irradiation. This states that any reflections from a gray surface in an enclosure uniformly irradiate other surfaces in the enclosure. With this approximation,

$$Q_{12} = \sigma A_1 \frac{(1 - F_{11})(T_1^4 - T_2^4)}{[1/\varepsilon_1 + (1 - F_{11})(1/\varepsilon_2 - 1)A_1/A_2]} \tag{9.38}$$

Equations (9.37) and (9.38) are important practical results in radiant energy transfer, easily specialized to include geometries such as parallel plates with spacing small relative to their size, concentric cylinders, or spheres. Many basic

spacecraft energy-balance problems can be treated using the results of this section.

9.4.11 Radiation Surface Coefficient

The foregoing results are obviously more algebraically complex than the corresponding expressions for conductive and convective energy transfer. This should not be taken to imply greater physical complexity; as we have mentioned, the complex physics of convective mass transfer is buried in the coefficient h_c, which may be difficult or impossible to compute. Nonetheless, there is great engineering utility in an expression such as Eq. (9.17), and for this reason we may usefully define a radiation surface coefficient h_r through the equation

$$Q_{12} = h_r A_1 (T_1 - T_2) \tag{9.39}$$

It is clear that h_r is highly problem dependent; indeed, even for the simple case of Eq. (9.37), if we are to put it in the form of Eq. (9.39), it must be true that

$$h_r = \varepsilon_1 \sigma (T_1^2 + T_2^2)(T_1 + T_2) \tag{9.40}$$

Though solving for T_1 or T_2 may be part of the problem, thus implying doubtful utility for Eq. (9.40), this result is more useful than it might at first appear. The coefficient h_r is often only weakly dependent on the exact values of T_1 and T_2, which in any case may have much less variability than the temperature difference $(T_1 - T_2)$. For example, when T_1 or $T_2 \gg (T_1 - T_2)$, then

$$4T_{\text{avg}}^3 \cong (T_1^2 + T_2^2)(T_1 + T_2) \tag{9.41}$$

Hence, we may write

$$h_r \cong 4\varepsilon_1 \sigma T_{\text{avg}}^3 \tag{9.42}$$

which has the advantage of decoupling h_r from the details of the problem.

The use of the radiation surface coefficient is most convenient when radiation is present as a heat transfer mechanism in parallel with conduction or convection. As we shall see, parallel thermal conductances add algebraically, thus allowing straightforward analysis using Eq. (9.40) or (9.41) together with a conductive or convective flux.

9.5 Spacecraft Thermal Modeling and Analysis

9.5.1 Lumped-Mass Approximation

For accurate thermal analysis of a spacecraft, it is necessary to construct an analytical thermal model of the spacecraft. In the simplest case, this will take the form of a so-called lumped-mass model, where each node represents a thermal

mass connected to other nodes by thermal resistances. This requires identification of heat sources and sinks, both external and internal, such as electronics packages, heaters, cooling devices, and radiators. Nodes are then defined, usually as the major items of structure, tanks, and electronic units. The thermal resistance between each pair of thermally connected nodes must be determined. This will involve modeling the conductive, radiative, and perhaps convective links between nodes. This in turn requires modeling the conductivity of the various materials and joints, as well as the emissivity and absorptivity of the surfaces. The analogy to lumped-mass structural models, introduced in Chapter 8, with mass element nodes connected by springs and dashpots, should be clear. Once constructed, the model can be used to solve steady-state problems; we will shortly illustrate with an example.

Often the model proceeds in an evolutionary manner, with the nodes initially being relatively few and large and the thermal resistances having broad tolerance. At this stage the model may be amenable to hand-calculator analysis or the use of simple codes for quick estimates. As the design of the spacecraft matures, the model will become more complex and detailed, requiring computer analysis. No matter how detailed the analysis becomes, however, a thermal vacuum test of a thermal mock-up or prototype will almost certainly be required, since the model requires a host of assumptions unverifiable by any other means. Also, as previously observed, the influence of the atmosphere as a convective medium and a conductor in joints renders thermal testing in atmosphere problematic. It is usually desirable to do an abbreviated test on flight units as well as a final verification. The following example demonstrates—in very basic terms—this approach to steady-state thermal modeling.

Example 9.1

Consider the insulated wall of a vertically standing launch vehicle liquid oxygen (LOX) tank, illustrated schematically in Fig. 9.6. The LOX is maintained at a temperature of 90 K in the tank by allowing it to boil off as necessary to accommodate the input heat flux; it is replaced until shortly before launch by a propellant feed line at the pad. It is desired to estimate propellant top-off requirements, for which the key determining factor is the heat flux into the tank.

The tank is composed of an aluminum wall of $\Delta_{al} = 5$ mm thickness and an outer layer of cork with $\Delta_{co} = 3$ mm. The ground and outside air temperatures are both approximately 300 K, and the sky is overcast with high relative humidity. The booster tank diameter of 8 ft is sufficient to render wall curvature effects negligible, and its length is enough to allow end effects to be ignored. What is the steady-state heat flux into the LOX tank?

Solution. The statement of the problem allows us to conclude that radiation from ground and sky at a temperature of 300 K to the wall, as well as free

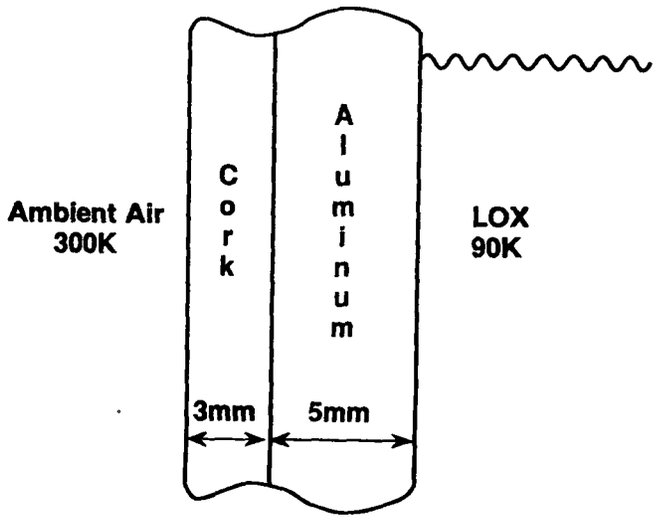


Fig. 9.6 Schematic of LOX tank wall.

convection from the air to the vehicle tank, will constitute the primary sources of heat input. The LOX acts as an internal sink for energy through the boil-off process; heat transfer to the LOX will be dominated by free convection at the inner wall.

Reference to standard texts yields for the appropriate thermal conductivities[8]

$$\kappa_{al} = 202\,\mathrm{W/m \cdot K}$$
$$\kappa_{co} = 0.0381\,\mathrm{W/m \cdot K}$$

and the free convection coefficients outside and inside are approximated as[8]

$$h_{co} = 5\,\mathrm{W/m^2 \cdot K}$$
$$h_{ci} = 50\,\mathrm{W/m^2 \cdot K}$$

We assume the cork to have $\varepsilon \cong \alpha \cong 0.95$ and the Earth and sky to have $\varepsilon \cong 1$. Because the outer tank wall is convex, it cannot see itself; thus, $F_{t,t} = 0$. The tank has a view of both sky and ground, in about equal proportions; and so $F_{t,s} \cong F_{t,g} \cong 0.5$; however, because we have assumed both to be blackbodies at 300 K, the separate view factors need not be considered. We therefore ignore the ground and take $F_{t,s} = 1$ in this analysis.

For clarity, we at first ignore the radiation contribution, considering only the free convection into and the conduction through the booster wall. The heat flux is unknown, but we know it must in the steady state be the same at all interfaces. The problem is essentially one-dimensional; therefore, the slab conduction result

of Eq. (9.9) is directly applicable. Thus, we may write

$$T_{\text{air}} - T_1 = \frac{q}{h_{\text{co}}}$$

$$T_1 - T_2 = \frac{q\Delta_{\text{co}}}{\kappa_{\text{co}}}$$

$$T_2 - T_3 = \frac{q\Delta_{\text{al}}}{\kappa_{\text{al}}}$$

$$T_3 - T_{\text{LOX}} = \frac{q}{h_{\text{ci}}}$$

Adding these results together yields

$$T_{\text{air}} - T_{\text{LOX}} = q\left[\frac{1}{h_{\text{co}}} + \frac{\Delta_{\text{co}}}{\kappa_{\text{co}}} + \frac{\Delta_{\text{al}}}{\kappa_{\text{al}}} + \frac{1}{h_{\text{ci}}}\right]$$

$$\cong \frac{q}{U} = \frac{Q}{UA}$$

The coefficient U defined here is called the universal heat transfer coefficient between the air and the LOX. As can be seen, the conductive and convective coefficients add reciprocally to form U. This leads to the definition, previously mentioned, of thermal resistance, analogous to electrical resistance. In this problem,

$$R_{\text{al}_{\text{conv}}} = \frac{1}{h_{\text{ci}}}$$

$$R_{\text{al}_{\text{cond}}} = \frac{\Delta_{\text{al}}}{\kappa_{\text{al}}}$$

$$R_{\text{co}_{\text{conv}}} = \frac{1}{h_{\text{co}}}$$

$$R_{\text{co}_{\text{cond}}} = \frac{\Delta_{\text{co}}}{\kappa_{\text{co}}}$$

and we see that

$$\frac{1}{U} = R_{\text{al}_{\text{conv}}} + R_{\text{al}_{\text{cond}}} + R_{\text{co}_{\text{conv}}} + R_{\text{co}_{\text{cond}}}$$

i.e., thermal resistances in series add. For this problem, we find

$$U = 3.35\ \text{W/m}^2 \cdot \text{K}$$

hence,

$$q = 703\,\mathrm{W/m^2}$$

Now that the heat flux is known, we can substitute to find the temperature at any of the interface points if desired. Each interface is a "node" in the terminology used, connected through appropriate thermal resistances to other nodes. For later use, we note that the outer wall temperature satisfies

$$T_{\mathrm{air}} - T_1 = \frac{q}{h_{\mathrm{co}}} = 141\,\mathrm{K}$$

hence,

$$T_1 = 159\,\mathrm{K}$$

Consider now the addition of the radiative flux. From Eq. (9.39), the radiative flux from the tank to the air is

$$q_r = \varepsilon_{\mathrm{tank}}\sigma(T_1^4 - T_{\mathrm{air}}^4) = h_r(T_1 - T_{\mathrm{air}})$$

where, from Eq. (9.32),

$$h_r = \varepsilon_{\mathrm{tank}}\sigma(T_{\mathrm{air}}^2 + T_1^2)(T_{\mathrm{air}} + T_1) \cong 2.85\,\mathrm{W/m^2 \cdot K}$$

Of necessity, we take T_1 from the convective solution to use in computing the radiation surface coefficient. If improved accuracy is required, the final result for T_1 obtained with radiation included can be used iteratively to recompute h_r, obtain a new result, etc. This is rarely justified in an analysis at the level exemplified here.

Changing the sign of the radiative flux to have it in the same direction (into the tank) as previously, we see that a second, parallel heat flux path has been added to the existing convective flux at the outer wall. This will result in a higher wall temperature than would otherwise be found.

At the wall, the flux is now

$$q_{\mathrm{total}} = q_{\mathrm{conv}} + q_r = (h_r + h_{\mathrm{co}})(T_{\mathrm{air}} - T_1)$$

which is substituted for the previous result without radiation. Thus, conductances in parallel add, whereas the respective resistances would add reciprocally. When the problem is solved as before, we obtain with the given data

$$q_{\mathrm{total}} = q = 929\,\mathrm{W/m^2}$$

and

$$T_1 = 182\,\mathrm{K}$$

Notice that the radiation surface coefficient method is only useful when the temperature "seen" by the radiating surface is approximately that "seen" by the convective transfer mechanism.

9.5.2 Spacecraft Energy Balance

One of the most important preliminary tasks that can be performed in a spacecraft program is to obtain a basic understanding of the global spacecraft energy balance.

Figure 9.7 shows a generic spacecraft in Earth orbit and defines the sources and sinks of thermal energy relevant to such a spacecraft. Not all features of Fig. 9.7 are appropriate in every case. Obviously, for a spacecraft not near a planet, the planet-related terms are zero. Similarly, in eclipse, the solar and reflected energy terms are absent. In orbit about a hot, dark planet such as Mercury, reflected energy will be small compared to radiated energy from the planet, whereas at Venus the opposite may be true. Solar energy input of course varies inversely with the square of the distance from the sun and can be essentially negligible for outer-planetary missions. These variations in the major input parameters will have significant impact upon the thermal control design of the spacecraft.

The energy balance for the situation depicted in Fig. 9.7 may be written as

$$Q_{\text{sun}} + Q_{\text{er}} + Q_i = Q_{ss} + Q_{se} \tag{9.43}$$

where we have neglected reflected energy contributions other than those from Earth to the spacecraft. This renders the enclosure analysis tractable. In effect, we have a three-surface problem (Earth, sun, and spacecraft) where, by neglecting certain energy transfer paths, a closed-form solution can be achieved. We define

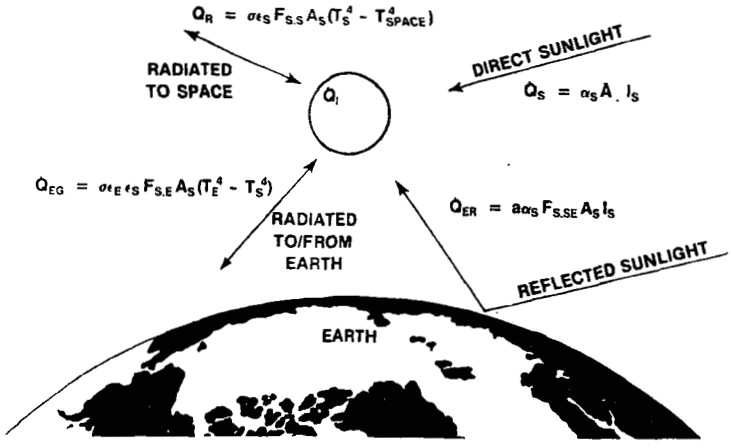


Fig. 9.7 Energy balance for an Earth orbiting spacecraft.

$Q_{\text{sun}} = \alpha_s A_\perp I_{\text{sun}}$ = solar input to spacecraft
$Q_{\text{er}} = a\alpha_s F_{s,se} A_s I_{\text{sun}}$ = *Earth-reflected solar input*
a = Earth albedo (ranges from 0.07 to 0.85)
α_s = spacecraft surface absorptivity
ε_s = spacecraft surface emissivity
Q_i = internally generated power
$Q_{se} = \sigma_s A_s F_{s,e}(T_s^4 - T_e^4)$ = net power radiated to Earth
$Q_{ss} = \sigma_s A_s F_{s,s}(T_s^4 - T_{\text{space}}^4)$ = net power radiated to space

If certain simplifications are made, such as assuming $T_{\text{space}} \cong 0$, and if we note that the sum of view factors from the spacecraft to Earth and space satisfies

$$F_{s,s} + F_{s,e} = 1 \tag{9.44}$$

then, in the equilibrium condition the energy balance equation becomes

$$\varepsilon_s \sigma A_s T_s^4 = \varepsilon_s \sigma A_s F_{s,e} T_e^4 + Q_{\text{sun}} + Q_{\text{er}} + Q_i \tag{9.45}$$

Equation (9.45) may be used to estimate the average satellite temperature. Note that this result provides no information on local hot or cold spots and does not address internal temperature variations. However, it is useful in preliminary design to determine whether the spacecraft is operating within reasonable thermal bounds.

Example 9.2

Consider Eq. (9.45) applied to an Earth-orbiting spacecraft. The spacecraft is a 1-m diameter sphere in a 1000-km altitude circular orbit. At the time under consideration the spacecraft is in full sunlight, yet essentially over the dark portion of the Earth. (This might occur in a near-polar sun-synchronous orbit over or near the terminator. Such a dawn-dusk orbit does see some portion of the sunlit Earth, but only near the limb, with consequently little energy input.) What is the average temperature of the spacecraft?

Solution. Assuming reasonable values for internally generated heat and for absorptivity and emissivity, we use

$$Q_i = 50\,\text{W}$$
$$\alpha = 0.7$$
$$\varepsilon = 0.9$$

From a 1000-km altitude orbit, Earth's disk subtends 120 deg, and hence a solid angle of π sr, or 25% of the celestial sphere. Thus,

$$F_{s,e} = 0.25$$

With the approximation discussed earlier that $F_{s,se}$, the view factor to the sunlit Earth, is essentially zero; hence,

$$Q_{\text{er}} \cong 0$$

The area intercepting sunlight and radiation from the Earth is the projected area of the spherical spacecraft, and is thus a disk; hence,

$$A_{\perp} = \frac{\pi D^2}{4} = \frac{\pi}{4}\,\text{m}^2$$

Using a solar intensity value of 1400 W/m^2 yields for the solar input to the spacecraft

$$Q_{\text{sun}} = (1400\,\text{W/m}^2)\alpha A_{\perp} = 770\,\text{W}$$

The total surface area of the spherical spacecraft is

$$A_s = 4\pi R^2 = 3.14\,\text{m}^2$$

To solve for the temperature of the spacecraft, the thermal equilibrium equation may be rewritten as

$$T_s^4 = F_{s,e}T_e^4 + \frac{Q_{\text{sun}} + Q_{\text{er}} + Q_i}{\varepsilon_s \sigma A_s}$$

Using the known values from the preceding equations and $T_e = 290$ K for the Earth, we obtain

$$T_s = 288\,\text{K}$$

This solution shows the dependence of spacecraft temperature on the surface α/ε ratio that is implicit in Eq. (9.45).

9.5.3 Thermal Analysis Tools

The preceding examples, while important, are really useful only for preliminary analysis under fairly simple conditions. When more accuracy is required, when transient conditions are of interest, when complicated boundary conditions obtain, then it will be necessary to construct a more accurate model and to apply different solution techniques. Needless to say, such requirements will be an inevitable part of almost every spacecraft design and development program as it moves through the sequence from conceptual design to launch and orbital operations.

Although some companies and institutions still maintain internal, proprietary thermal analysis tools, the overwhelming majority of thermal engineering design and analysis is performed using very sophisticated software packages that have become, essentially, industry standard engineering tools. Several computer-aided design (CAD) packages originally developed to perform finite element analysis

for structural engineering purposes (e.g., IDEA-S™ or ProEngineer™) offer excellent thermal analysis capability also.

The most popular dedicated thermal analysis package is the SINDA code, a very mature engineering tool with a history of some four decades of refinement and use.[12] SINDA (or CINDA in its earliest versions) utilizes a numerical finite-difference analysis engine, together with elaborate pre- and post-processing software to allow the user to develop a nodal mesh, or grid, appropriate to the case at hand, apply desired boundary conditions, and display the computed result in a variety of ways.

FLUINT is a code developed for the specialized analysis of internal one-dimensional fluid flows, e.g., the pumped fluid loops discussed earlier.[13] Although it can handle phase transitions (e.g., liquid to gas or vice versa) within a fluid loop, it is otherwise restricted to the low-speed incompressible flow of a single viscous fluid.

Modern versions of SINDA incorporating FLUINT are also available.

9.5.4 Thermal Analysis Accuracy

Even with the best available analysis tools, accurate and refined spacecraft design information, and carefully specified materials properties, spacecraft thermal analysis does not allow the level of precision customary in other disciplines discussed in this text. Experience shows that carefully developed models, correlated with preflight thermal test data, offer a 2σ accuracy band of only about $\pm$ 10 K.[14] This is the basis for the MIL-STD- 1540B[15,16] requirement for a band of $\pm$ 11 K to achieve 95% confidence that flight experience will be within predicted preflight tolerances. When accurate thermal-balance tests cannot be performed, a tolerance of $\pm$ 17 K is recommended.

Historically, both NASA and commercial spacecraft developers have commonly used a narrower tolerance band, typically $\pm$ 5 K, corresponding roughly to the 1σ confidence level recommended by MIL-STD-1540B.

It is usually best to view design margin requirements such as these as being functions of the program life cycle.[17] Thus, at the concept design stage, it might be expected that the thermal system be capable of handling a heat load of up to 50% greater than analytically predicted. This allows substantial change in the spacecraft design without having an inevitably adverse effect on the thermal control system. Because such changes rarely are in a favorable direction, an initially comfortable 50% margin will decrease as launch is approached, at which point a 20% margin may well be deemed adequate.

References

[1]Wertz, J. R., and Larson, W. (eds.), *Space Mission Analysis and Design*, 3rd ed., Kluwer Academic, Dordrecht, The Netherlands, 1999.

[2]Gilmore, D. G. (ed.), *Satellite Thermal Control Handbook*, Aerospace Corp. Press, El Segundo, CA, 1994.

[3]Mather, J., and Boslaugh, J., *The Very First Light*, Basic Books, New York, 1996.

[4]Wylie, C. R., and Barrett, L. C., *Advanced Engineering Mathematics*, McGraw–Hill, New York, 1982.

[5]Gebhart, B., *Heat Transfer*, 2nd ed., McGraw–Hill, New York, 1971.

[6]Holman, J. P., *Heat Transfer*, 6th ed., McGraw–Hill, New York, 1986.

[7]Selby, S. M., *Standard Mathematical Tables*, 22nd ed., CRC Press, Cleveland, OH, 1974.

[8]Baumeister, T. F., Avallone, E. A., and Baumeister, T. F., III, *Marks Standard Handbook for Mechanical Engineers*, 8th ed., McGraw–Hill, New York, 1978.

[9]Tauber, M. E., and Yang, L., "Performance Comparisons of Maneuvering Vehicles Returning from Orbit," *Journal of Spacecraft and Rockets*, Vol. 25, July–Aug. 1988, pp. 263–270.

[10]Siegel, R., and Howell, J. R., *Thermal Radiation Heat Transfer*, 2nd ed., Hemisphere, New York, 1981.

[11]Hottel, H. C., and Sarofim, A. F., *Radiative Transfer*, McGraw–Hill, New York, 1967.

[12]Gaski, J. D., *SINDA 1987/ANSI*, Network Analysis Associates, Chandler, AZ, 1992.

[13]Cullimore, B. A., "FLUINT: Generalized Fluid System Analysis with SINDA 85," AIAA Paper 87-1466, 1987.

[14]Stark, R. D., "Thermal Testing of Spacecraft", TR TOR-0172 (244-01-4), Aerospace Corp., El Segundo, CA, 1971.

[15]"Test Requirements for Space Vehicles," MIL-STD-1540B, DoD/USAF, 1982.

[16]"Application Guidelines for MIL-STD-1540B," MIL-HDBK-340, DoD/USAF, 1985.

[17]Anderson, B. J., Justus, C. G., and Batts, G. W., "Guidelines for the Selection of Near-Earth Thermal Environment Parameters for Spacecraft Design," NASA TM-2001-211221, Oct. 2001.

[18]White, F. M., *Viscous Fluid Flow*, McGraw–Hill, New York, 1974.

Problems

9.1 It is desired to place a satellite in an elliptic orbit with a 1000-km apogee and a very low perigee, to allow the upper atmosphere to be sampled and thus help to establish its characteristics. Ample propellant for drag make-up over the planned lifetime of the spacecraft is included, and so the limiting perigee will be governed by thermal considerations. It is desired to limit the thermal input from aerodynamic heating to 10% of the solar illumination of 1400 W/m^2. Using the result for heat transfer due to free molecular flow given in Chapter 6, and the standard atmosphere model of Chapter 3, what is the lowest altitude at which a satellite perigee can be allowed?

9.2 For the spacecraft in problem 9.1, assume that the aerodynamic heat flux limit of 140 W/m^2 is reached. The front of the spacecraft that encounters

the free molecular flow in the ram direction is made of 2-cm thick aluminum plate. Assuming conservatively that the heat flux is steady at the worst-case value, and ignoring any convective effects, which are not a factor in free-molecular flow, how much time is required to raise the temperature by 5 K at a depth of 1 cm into the plate?

9.3 What fraction of solar energy lies in the visible range, which we will define as being 0.75–0.35 μm?

9.4 What is the average spacecraft temperature for the situation in Example 9.2, if the spacecraft is in a noon-midnight orbit? Make reasonable assumptions as required.

9.5 A solar panel on a GEO satellite tracks the sun; the back of the panel faces dark space. The cells have a 90% packing factor and an energy conversion efficiency of 12%. The effective front-surface solar absorptivity is $\alpha = 0.90$, and the infrared emissivity is $\varepsilon = 0.94$. The anodized aluminum back surface panel has IR emissivity $\varepsilon = 0.80$. What is the steady-state operating temperature of the array?

9.6 For the spacecraft in Example 9.2, assume a total mass of 100 kg and an average heat capacity equal to that of aluminum, $C = 961$ J/kg·K. A piece of onboard equipment fails, causing the internal power generation to drop to 35 W. Treating the spacecraft as isothermal, as in the example, what is the new steady-state temperature, and approximately how long does it take to reach it?

9.7 A radiator on a LEO spacecraft will be oriented toward dark space while in use, and must dissipate 200 W on average. The radiator uses a pumped fluid loop containing water-glycol and operates at a nominal temperature of 310 K. The blackbody efficiency is $\eta = 0.85$ and the emissivity is $\varepsilon = 0.94$. What is the required radiator area?

9.8 For the radiator of problem 9.7, the 50/50 water-glycol mixture freezes at about 230 K. To allow an appropriate safety margin, the radiator must be maintained at or above 250 K. What is the minimum power that must be dissipated by the radiator to maintain safe operation? If for any reason this level of power usage in the spacecraft cannot be maintained, what operational strategy might be used to avoid freezing the radiator?

9.9 Using the parameters of Example 9.2, with the noon-midnight orbit of problem 9.4, and the spacecraft mass and heat capacity of problem 9.6, what temperature is reached by the spacecraft immediately prior to exiting its eclipse period? What temperature is reached after the equipment failure of problem 9.6?

10
Power Systems

10.1 Introduction

Constraints on available spacecraft power have imposed major limitations on space vehicle design since the beginning of the space age. The earliest orbiting vehicles flown by both the United States and Russia depended on batteries. The limited energy storage capabilities of the batteries then available prevented operations of more than a few days. This was not satisfactory for missions of the duration required for detailed scientific observations or military reconnaissance, and solar power arrays quickly appeared on the scene. Although not highly efficient in turning sunlight into electricity, solar arrays (or solar panels) were in many ways admirably suited to powering spacecraft. Because no consumables were used in generating electrical power, the life expectancy of the power system was limited only by degradation of the components of which it was composed. Spacecraft operating lifetimes of several years became feasible with the development of these photoelectric arrays, with batteries used to handle peak load requirements and to provide energy storage for those periods when the spacecraft was in eclipse.

Solar panels and batteries in combination have powered the majority of unmanned spacecraft so far launched. Exceptions include a few short-lived battery-powered systems, some outer-planet missions using radioisotope thermoelectric generators (RTGs), and some spacecraft (mostly Russian radar imaging satellites) powered by nuclear reactors. Early manned spacecraft, including Mercury, some Gemini spacecraft, and the Russian Vostok/Voshkod vehicles (which were essentially the same design) used batteries. The later Gemini spacecraft and the Apollo command and service module (CSM) and Lunar Module (LM) used hydrogen/oxygen fuel cells, as does the space shuttle, while the Russian Soyuz employs solar cells and batteries in a fashion similar to a typical unmanned spacecraft. The space stations so far built, including Salyut, Skylab, Mir, and the International Space Station, have all used solar arrays for prime power generation, with batteries for loadleveling and eclipse periods.

Solar power systems are unsatisfactory for missions beyond the asteroid belt, where the sun's energy becomes unacceptably diffuse. As interest developed in outer-planet missions, a new power source was required. At the same time, certain military spacecraft missions required a sturdy compact power source. Both requirements were met by the development of RTGs. These devices convert

the heat energy produced by radioisotope decay into electricity via the thermoelectric effect. Power output is independent of the sun, and lifetime is limited only by component degradation and the half-life of the radioisotope. RTGs are also useful for operations on planetary surfaces where extended dark periods may be encountered. Thus, outer solar system spacecraft such as Pioneer, Voyager, Galileo, and Cassini, as well as the Viking Mars Landers and the Apollo Lunar Surface Experiment Packages have all been RTG powered, as have some Earth orbiting spacecraft.

Nuclear reactor systems offer very high power in a compact package for quite a long duration and tend to be highly independent of the external environment. After an extensive development program in the 1960s, all U.S. space reactor work, for both power and propulsion, was terminated as a result of space program funding reductions in the early 1970s. Only recently has there been a revival of interest in power plants of this type. Although Russia continues to fly relatively short-lived reactor power systems on an operational basis, the United States has flown only a single reactor test mission, the SNAP-10A in 1972. The joint DoD/NASA/DoE SP-100 project of the mid-to-late 1980s was intended to remedy this matter; however, the program was delayed and eventually canceled because of its high cost and limited mission applicability. As this is written, interest in nuclear-powered systems has again arisen, because they are the only practical means of generating relatively high power for long periods in the absence of adequate sunlight.

As can be seen, the power system is a major driver in any spacecraft design and is in turn strongly driven by a variety of mission, system, and subsystem considerations. It interfaces directly with almost every other subsystem and, as a result, requires considerable attention from the systems engineer. Power system technology continues to evolve, especially in the application of automation to routine functions (e.g., battery reconditioning, to be discussed), and the development of more efficient power conditioning and control circuitry.

10.2 Power System Functions

The obvious functions of a spacecraft power system are to generate and store electric power for use by the other spacecraft subsystems. Other subsystems may have various specific requirements for voltage, frequency, stability, noise limits, or other characteristics, and the power system may be called upon to supply them. A significant system-level tradeoff underlies the decision as to whether to require the power system to meet these various individual requirements, or to supply all subsystems with the same basic power and let each subsystem meet its specific power conditioning requirements. For example, a requirement for a very high voltage in a particular scientific instrument might be supplied by the spacecraft power system, or by a dedicated high-voltage supply within the instrument that

operates off the basic power bus. Similar tradeoffs exist for special cleanliness requirements (e.g., absence of ripple on a dc line, ac harmonic suppression, etc.).

Regardless of the conclusion of these tradeoffs, the power system must control, condition, and process the raw power received from the primary source to comply with the needs of the spacecraft system. The system must supply stable, uninterrupted power for the design life of the system. Failures in many other subsystems can be tolerated, with solutions often found through operational compromises. However, if the power system does not work essentially as planned, the mission is lost.

To maintain the long-term reliability of the system, the power system must provide protection to other subsystems against reasonably likely failures either external to or within the power system itself. For example, no short circuit in another subsystem should be allowed to drag the main bus voltage down to the point of inducing failure elsewhere in the spacecraft. Similarly, failure protection should be implemented in the power system itself to allow for continued functioning of the system (perhaps in a degraded mode) following some degree of malfunction.

In the course of normal operation, the power system must accept commands from onboard and external sources and provide telemetry data to allow monitoring of its operation and general health.

Finally, it may be necessary to meet highly specialized power requirements for particular functions such as firing ordnance.

10.3 Power System Evolution

The evolution of spacecraft power systems has been characterized by growth from subsystems delivering a few watts to those delivering tens of kilowatts or more. The International Space Station required about 75 kWe initially with growth currently planned to 220 kWe or more. Line losses and other efficiency factors, including the desire to minimize the mass of spacecraft wire harnesses, have resulted in a trend toward higher voltages as power demands have increased. Figure 10.1 illustrates this trend and projects broadly what may be anticipated in the near future.

The design lifetime of space systems tends to increase along with required power levels as spacecraft become more complex and expensive. As the power level and lifetime change, the choice of a primary power source may change as well. Figure 10.2 illustrates the general operating regimes of various types of power sources. There is a substantial overlap between the regimes, and various other considerations may dictate use of some power source at a location in the power vs endurance space that may not otherwise appear to be optimum. Figure 10.2 provides a basis for preliminary concept design in regard to power source choices.

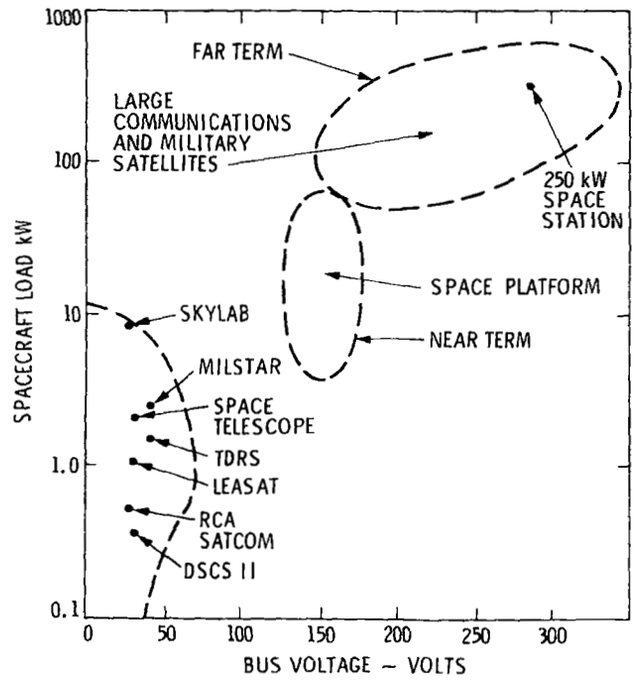


Fig. 10.1 Trends in spacecraft power.

10.4 Power System Design Drivers

A variety of considerations may affect the design of the power system. Table 10.1 presents a number of these considerations. Not all will be applicable to each system design, and, conversely, some designs may involve considerations not listed here. However, most cases of common interested will be treated. The designer should view Table 10.1 as a checklist, to be used as a reminder to cover all points in the initial design and, as the design matures, to assess the impact of changes.

Discussing the checklist items briefly, the customer or user may have specific requirements such as size, observability, or operational constraints that will limit the choices in regard to the primary power source or other subsystem elements. The target planet, whether Earth or another planet, and the resultant distance from the sun will in some cases limit design flexibility because of restrictions on available solar energy per unit area or, conversely, the requirement to control the temperature of exposed surfaces.

Lifetime requirements in a given operating environment may also drive the power system design. Solar array degradation due to radiation exposure may prevent use of these devices on long-lived spacecraft operating in the Van Allen

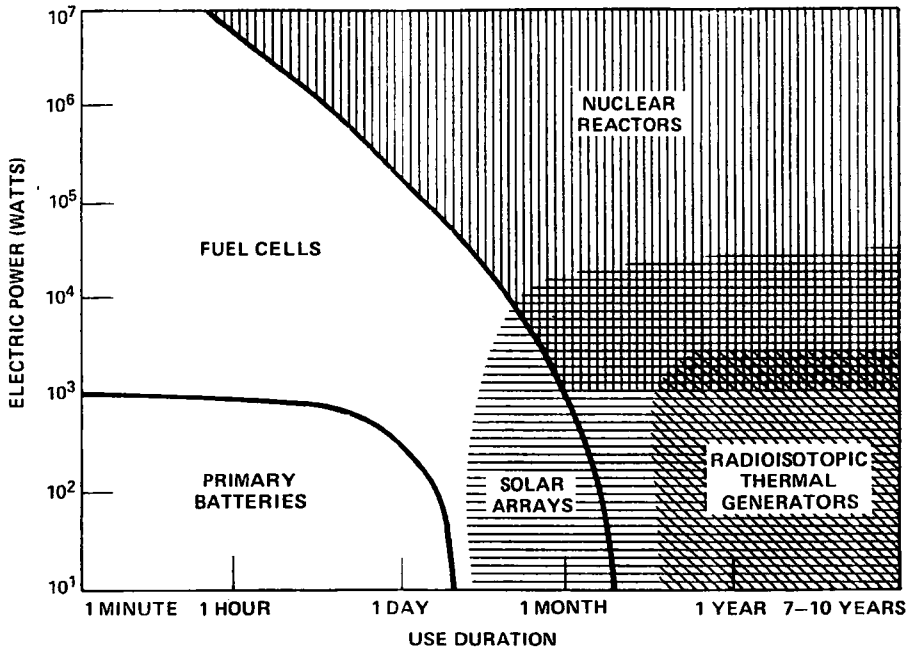


Fig. 10.2 Operating regimes of spacecraft power sources.

belts, for example. As will be seen, gallium arsenide solar cells offer improved radiation tolerance compared to silicon-based cells. Even so, the high radiation flux, particularly from inner-belt protons, place serious limits on array lifetime. Many spacecraft have a variety of operating modes requiring different power levels. The percentage of time in each mode is of great significance and may indicate a hybrid system using more than one power source or type of energy storage.

The attitude control concept employed will affect the power system both in terms of configuration constraints from solar arrays, waste heat radiators, and other elements, and in responding to specific power needs of the attitude control devices. The flexibility and frequency response of large arrays can, in turn, dictate the choice of attitude control effectors. Space mission history offers several notable examples of undesired control-structure interactions due to poorly modeled solar array flexibility effects. Finally, the attitude control system engineer will usually be involved in the design of whatever scheme is used to orient the solar arrays toward the sun.

Orbital parameters will strongly affect the choice of the primary power source and its configuration, as well as onboard energy storage requirements. However, despite the difficulties posed by some unique orbits and space mission requirements, system operation on a planetary surface will often be the most

Table 10.1 Power system design considerations

Customer/user
Target planet, solar distance
Spacecraft configuration
Mass constraints
Size
Launch vehicle constraints
Thermal dissipation capability
Lifetime
Total
Percentage in various modes, power levels
Attitude control
Spinner
Three-axis stabilized
Nadir pointing
Thrusters
Momentum wheel
Gravity gradient
Pointing requirements
Orbital parameters
Altitude
Inclination
Eclipse cycle
Payload requirements
Power type, voltage, current
Duty cycle, peak loads
Fault protection
Mission constraints and requirements
Maneuver rates
g loads

environmentally demanding, the most difficult in terms of deploying large solar arrays, and the most challenging in regard to meeting energy storage demands.

Specific mission demands may also impact the power system design. For example, a spacecraft that must maneuver rapidly may not be able to tolerate large, flexible solar arrays. A low-observable spacecraft may preclude use of concepts requiring high-temperature operation, such as imposed by RTGs.

10.5 Power System Elements

Figure 10.3 presents a typical spacecraft functional block diagram that identifies the major elements in the power system. A substantial variety of options exist within each of these elements. Table 10.2 identifies the options most likely to be encountered in normal spacecraft design practice.

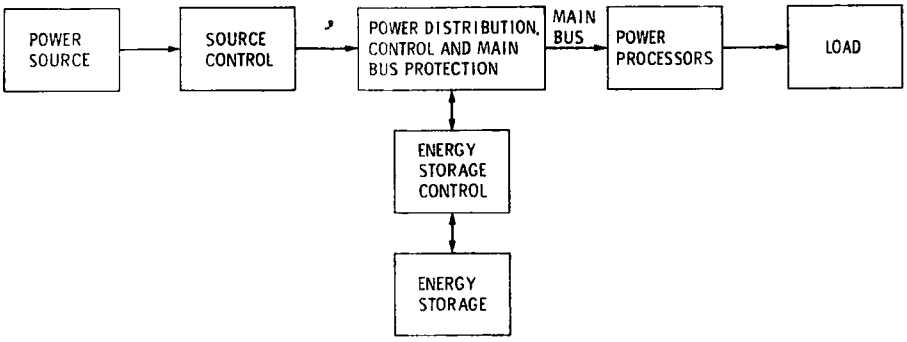


Fig. 10.3 Power subsystem functional block diagram.

10.6 Design Practice

Although details of design practice will vary from one organization to another, some broadly applicable rules can be articulated. These are discussed in this section.

10.6.1 Direct Current Switching

As a general rule, switches or relays should be in the positive line to a given element, with a direct connection to "ground" on the negative side. The purpose

Table 10.2 Power system elements

Power source
Solar photovoltaic
Radioisotope thermoelectric generator (RTG)
Nuclear reactor, static or dynamic energy conversion
Radioisotope dynamic
Solar dynamic
Fuel cells
Primary batteries
Source control
Shunt regulator
Series regulator
Shorting switch array
Energy storage control
Battery charge control
Voltage regulation
Power conditioning
DC-DC converters
DC-AC inverters
Voltage regulation

of this design practice is of course to allow power to be shut off in the event of a short circuit or other high-current flow failure within the element.

In a famous manned spaceflight emergency, the Gemini 8 mission was aborted less than a day into a planned three-day mission, following loss of flight control caused by a roll thruster stuck in the "on" position. Post-flight inspection of the spacecraft revealed a small solder ball shorting the thruster (which was switched in the negative, or return, line and thus was electrically "hot") to spacecraft ground, rendering the astronauts' hand controller inoperative with respect to that thruster.

Not surprisingly, astronauts Neil Armstrong and David Scott were unable to diagnose the problem while trying to cope with roll rates of up to 300 deg/s. Control could only be regained by shutting down the entire system and reverting to the backup reentry flight control system, after which individual reactivation of the primary thrusters, one by one, revealed the culprit. Disabling this thruster effectively solved the problem. However, according to the mission rules then in force, activation of the reentry flight control system required a mandatory abort, resulting in the loss of numerous mission goals.[1] Subsequent vehicles were rewired to have the thrusters switched in the positive control line, so that the thruster body was electrically inert when not firing.

10.6.2 Arc Suppression

To maximize effectiveness, arc suppression devices should be located as close to the source of the arc as possible. As discussed in Chapter 3 in connection with spacecraft charging, conductive cables, connectors, solar array edges, and other current-carrying elements on LEO spacecraft should not be exposed to the ambient plasma, particularly if the spacecraft operates at a bus voltage comparable to the arcing threshold of the conductor materials. High altitude spacecraft should be designed to ensure conductive paths between all spacecraft elements to preclude differential charging of isolated sections and subsequent arcing between them.

10.6.3 Modularity

Modular construction is desirable to simplify testing and to expedite replacement of failed or suspect units during system test or launch preparation. Spacecraft designers learned early, and at their peril, that the necessity to disassemble or remove several non-offending units to achieve access to a suspect system often resulted in "collateral damage" to the innocent parties, not to mention an excessive workload for numerous technicians. It was realized that the savings in mass and volume were more than offset by the delays and reductions in reliability caused by excessive "stacking" of components and subsystems. Thus, modern space systems design favors considerations of access and maintainability as well as conservation of mass and volume.

10.6.4 Grounding

Spacecraft grounding practices are often the subject of considerable debate among practitioners. This text presumes no single approach to spacecraft grounding practice. Many successful satellite and spacecraft programs have used a variety of grounding techniques of varying levels of complexity. However, some design principles are widely accepted,[2] and we advocate those here.

Use of a common ground cable is generally preferable to individually grounding various components and circuit elements to the structure, because it is difficult to maintain high continuity between isolated structural elements. When such difficulty occurs, electrical resistance results between various parts of the spacecraft ground structure that are intended to be at a common potential i.e., to have no electrical resistance between them. Thus, different portions of the electrical system, intended to be at the same electrical potential above the common ground, will not be so, and may therefore not function as intended.

As a simple example, a semiconductor switch may be designed to "trip" upon application of a 5-V potential, but not at a 3-V potential difference to avoid spurious switching due to electrical noise on the line. If there is enough electrical resistance in the ground loop to generate a 2-V drop on the return line, the switch can never be activated. While this may seem to be an extreme example, a voltage drop of this magnitude can easily occur across a long ground loop such as might be found on a large spacecraft. Numerous examples, some of direct experience to the present authors, occurred during early shuttle operations, when experience with large payloads remotely mounted in a cargo bay aboard an even larger vehicle was then minimal.

For these reasons, such "ground loops" are obviously to be avoided whenever possible. This gives rise to the practice of providing a common, low-resistance grounding strap or cable to all subsystems throughout the spacecraft, which is then carefully grounded at a single point to the spacecraft structure. Any ground circuit current flow is less likely to disturb sensitive components if confined to a properly isolated and connected ground cable. When electrical noise is thought to be a problem for certain circuits or instruments, as is often the case, noisy power-switching ground lines will often be kept separate from "signal" ground lines, prior to structural grounding. For similar reasons, radio-frequency (RF) grounding requirements may be incompatible with the needs of other systems, except again for the general requirement to be ultimately tied back to the primary structure.

On some occasions it may be necessary to isolate completely a given instrument or subsystem from other sources of spacecraft electrical noise. When this is the case, some portions of the vehicle will be electrically isolated, or "floated," with respect to others. Provided that it is intentional, this practice is acceptable. However, it is then necessary to take care to ensure that the separately floated ground does not inadvertently contact the "common" ground, because current would then flow between them.

It is sometimes very awkward, if not impossible, to provide a single-point common structural ground to subsystems in all portions of a large or complex space vehicle. When a single-point ground is difficult to achieve, but a common ground plane remains necessary, a multipoint grounding architecture will be employed. Several different grounding points to the structure are provided, with every effort made to maintain very low resistance paths between them. Nonetheless, the single-point ground generally remains the design objective to be achieved.

10.6.5 Continuity

Good continuity should be maintained between structural elements, thermal blankets, etc., to minimize the probability of buildup of static electrical potential or other voltage differences.

10.6.6 Shield Continuity

Shield continuity must be maintained across all connections. A single-point shield ground is desirable to minimize the possibility of shield current flow. Most circuits, especially noise-sensitive or noise-generating circuits, will be shielded, with the shields sharing a common ground if possible, as previously discussed.

10.6.7 Complexity

In keeping with good general engineering practice, the spacecraft power system should be no more complex than is necessary to do the job. Excessive, unnecessary complication will increase design, fabrication, and test costs and increase the probability of failure. It is for exactly this reason that the choice is often made to offer a relatively simple menu of power supply bus voltages to the various spacecraft subsystems, and to allow each of them to deal separately with any special requirements. Such a solution is rarely the least massive, and never the most electrically efficient, but it usually offers gains in simplicity that should be ignored only when no reasonable alternative exists.

Particular circumstances may force a violation of any of the previously mentioned rules to meet some overriding requirement. In the absence of such a requirement, however, adherence to these rules is very much recommended and will generally have a desirable impact on the overall operation.

10.7 Batteries

Batteries have been and will continue to be for the foreseeable future the primary means of electrical energy storage onboard spacecraft. In the following

discussion, a variety of terms relating to batteries will be used. These are defined here to enhance understanding of the material to follow:

Charge capacity, C_{chg}	Total electric charge stored in the battery, measured in ampere hours (e.g., 40 A for 1 h = 40 Ah).
Energy capacity, E_{bat}	Total energy stored in the battery, equal to charge capacity (Ah) times the average discharge voltage, typically measured in units of Joules or watt hours.
Average discharge voltage, V_{avg}	Number of cells in series times cell discharge voltage (1.25 V for many commonly used cells).
Depth of discharge, DOD	Percent of battery capacity used in the discharge cycle (75% DOD means 25% capacity remaining, the DOD is usually limited to promote long cycle life).
Charge rate, R_{chg}	Rate at which the battery can accept charge (measured in amperes per unit time).
Energy density, e_{bat}	Energy per unit mass [J/kg or (W · h)/kg] stored in the battery.

A battery (strictly speaking, an individual cell of a battery) is a device that converts chemical energy directly to electrical energy. A single cell has a negative electrode, a conductive electrolyte, and a positive electrode. The electrolyte may be in liquid, paste, or solid form; potassium hydroxide (KOH) is a common choice. If the cell is connected to an external electrical load, electrons flow from the negative electrode, through the load, and back to the positive electrode. The chemical reaction essentially ceases when the load is removed; however, it should be noted that the battery will slowly degrade chemically over time, whether used or not. Thus, most batteries have a "shelf life" within which they must be used.

Batteries are divided into two major categories: primary and secondary. The former offer higher energy and power densities for a given battery chemistry but are by definition not rechargeable. This definition is sometimes stretched to include as primary batteries those which are rechargeable for only a few cycles. Primary batteries are especially well adapted to one-time events requiring substantial power and minimal mass, as with missiles and expendable launch vehicle stages.

10.7.1 Primary Batteries

In cases where extremely long installed storage is required, e.g., a missile in its silo or a planetary atmosphere probe that is inert during interplanetary transfer, the battery is often dry (i.e., without electrolyte) prior to activation. Upon activation, a pyrotechnic valve fires to allow the electrolyte to enter the battery

from a separate reservoir. This approach provides a highly reliable quick-reaction power source that is nevertheless protected from degradation and requires no maintenance during extended storage. Another quick-reaction, dry storage battery is the thermal battery. In this case, the electrolyte is solid at normal temperature. Ignition of a chemical heater, which melts the electrolyte and results in a fully charged battery, activates the battery. The battery stays active as long as the electrolyte is molten or until it is fully discharged.

A major application of these types of batteries in long-life space systems is to supply power to activate pyrotechnic charges and other deployment devices. Such devices typically are operated at the beginning of the mission or for relatively brief periods during a longer mission. Another application is to short-duration, high power-drain devices such as electromechanical actuators. For a variety of reasons, such as minimizing power drain or isolating noisy circuits from the main power bus, it may be desirable to operate these circuits from a primary battery that is completely isolated from the main power system.

During the early years of space systems development, the most common type of primary battery was the silver-zinc battery, usually abbreviated Ag-Zn. This battery has excellent energy density and is still the battery of choice in many cases. In recent years a variety of batteries based on lithium in combination with various other materials have come on the scene. Some of these batteries offer the highest energy density currently available. Certain types of lithium batteries experienced significant "teething problems" in early applications, showing a distressing tendency to explode in some situations. Leakage and corrosion problems have also been encountered. However, these problems have largely yielded to better understanding of battery characteristics and ensuing engineering development, and lithium batteries can today be reliably employed in many space vehicle applications.

10.7.2 Secondary Batteries

The rechargeable or secondary battery generally has a much lower energy density, which is further aggravated by limitations on the depth of discharge. Again, silver-zinc batteries were the most commonly used for a number of years and have demonstrated good energy density (which is nonetheless reduced as compared to the primary form of these batteries, due to the extra wrapping material used to isolate each cell in the battery when it is intended to be recharged). However, these batteries suffer from life limitations, especially in applications involving a large number of charge/discharge cycles. As a result, nickel-cadmium (Ni-Cd, or nicad) batteries have been for many years very nearly the standard for spacecraft applications. Certainly they have been the most common in LEO spacecraft designs.

A more recent development in battery technology is the nickel-hydrogen (Ni-H_2) design. This battery differs from other types in that a large amount of free hydrogen is generated as part of a charge/discharge cycle. As a result, quite high

pressures are generated and the battery case is, in fact, a pressure vessel. (Actually, other battery types, such as Ni-Cd, do generate some pressure and require a reasonably strong case to contain it. However, Ni-H_2 battery pressure exceeds that of nicads by a factor of 10.) Ni-H_2 batteries are capable of greater depth of discharge than nicads and, even with the penalty of the high pressure case, offer better energy density. Ni-H_2 batteries do not require reconditioning. In large part because of this advantage, Ni-H_2 batteries have been very competitive in recent years with nicads, particularly for GEO spacecraft, and as this is written may even be used in the majority of new spacecraft.

The pressure vessel cases of Ni-H_2 batteries are generally cylindrical with hemispherical ends. This makes close packing difficult. In an effort to avoid the pressure containment problem while retaining the other advantages of nickel-hydrogen systems, the nickel-metal hydride battery (Ni-MH) was developed. This battery depends on the ability of some metallic hydrides to contain large amounts of hydrogen in the structure at low pressure. This allows the battery cell case to be rectangular like most other batteries, allowing for more efficient packing. Ni-MH batteries are in very common commercial use in cell phones, laptop computers, etc. Unfortunately, space applications of Ni-MH batteries have been few, primarily because of the limited cycle life so far demonstrated for this technology.

Lithium-based secondary batteries are also available and offer excellent energy density. Some of the chemistries available do require reconditioning. Table 10.3 provides a list of various battery types and the characteristics of each.

As noted earlier, the battery average discharge voltage V_{avg} is the product of the individual cell average discharge voltage and the number of cells in series. As seen in Table 10.3, the cell voltage of most of the battery chemistries discussed here is approximately 1.25–1.50 V. All batteries will have a higher discharge voltage when fully charged than when nearly depleted. Indeed, the drop in output voltage below a specified threshold is the indication that it is time to recharge the battery. A nicad cell might have an average discharge voltage of 1.25 V, with a charging cycle mandated should the voltage drop below 1.1 V.

Most spacecraft systems flown to date by the United States have used 28 VDC as the nominal bus voltage; thus, most associated battery hardware has also been designed for 28 VDC. This practice reflects the heritage of early spacecraft avionics from aircraft systems in terms of electronic component design and usage. Generally speaking, this was satisfactory for the relatively small, low-power spacecraft flown in earlier decades. However, as larger and more powerful systems have become common, higher voltage systems have become more attractive. The future will undoubtedly see a continuing trend toward the use of higher voltage spacecraft power buses (see Fig. 10.1). This practice reduces the current-handling requirements of the spacecraft wire harness and thus the attendant weight of that harness. Also, because resistive losses (and heating) are proportional to the square of the current being carried, such inefficiencies are also minimized by the use of a higher bus voltage.

Table 10.3 Battery chemical types

Silver-zinc, (AgZn)
- Commonly used in early space systems; still popular
- Good energy density [175 (W · h)/kg primary, 120–130 (W · h)/kg secondary]
- Limited cycle life (2000, 400, 75 at 25, 50, 75% DOD)
- 1.50 V/cell

Silver-cadmium, (Ag-Cd)
- Better cycle life than Ag-Zn, better energy density than Ni-Cd
- Fair energy density [60–70 (W · h)/kg secondary]
- Fair cycle life (3500, 800, 100 at 25, 50, 75% DOD)
- 1.10 V/cell

Nickel-cadmium, (Ni-Cd)
- Most common secondary battery presently in use
- Low energy density [20–30 (W · h)/kg]
- Long cycle life (20,000, 3000, 800 at 25, 50, 75% DOD)
- Good deep discharge tolerance
- Can be reconditioned to extend life
- 1.25 V/cell

Nickel-hydrogen, (Ni-H_2)
- High internal pressure requires bulky pressure vessel configuration
- Good energy density [60–70 (W · h)/kg]
- Good cycle life (15,000, 10,000, 5000 at 25, 50, 75% DOD)
- No reconditioning required
- 1.30 V/cell

Nickel-metal hydride, (Ni-MH)
- Same chemistry as nickel-hydrogen
- Hydrogen adsorbed in metal hydride to reduce pressure
- Improved packaging relative to nickel-hydrogen
- Good energy density
- Limited cycle life
- 1.30 V/cell

Lithium batteries
- Several types (Li-$SOCl_2$, Li-V_2O_5, Li-SO_2)
- Both primary and secondary designs available
- Very high energy density [650 (W · h)/kg, 250 (W · h)/kg, 50–80 (W · h)/kg secondary]
- Higher cell voltage (2.5–3.4 V)

The use of higher bus voltages is not an unmitigated good. As discussed in Chapter 3, LEO spacecraft tend to accumulate negative charge from the ambient plasma, to a level about 90% of the maximum negative exposed-conductor voltage, relative to the plasma reference potential. If this level exceeds the arcing threshold of common conductors, problems will occur. Earlier, lower voltage bus levels were not vulnerable to this effect.

Depth of discharge limitations usually require a tradeoff between battery mass due to the unused capacity and battery degradation and lifetime reduction due to repeated deep discharge. Spacecraft in low-altitude, low-inclination orbits around the Earth or another planet typically experience the most severe usage in terms of charge/discharge cycles, because they experience eclipse on each orbit. In LEO a spacecraft battery will be discharged and charged some 12–16 times per day. This results in some 10,000 or more cycles in only a few years, yet modern spacecraft are normally expected to function for substantially longer periods of time.

Most battery chemistries so far developed cannot accept so many charge-discharge cycles; thus, for such applications, Ni-Cd batteries have been the system of choice, despite their low energy density. However, even with nicads, it is necessary to limit the depth of discharge to a relatively small amount, 15–25%, and to recondition the batteries periodically if the desired total lifetime is to be obtained. As experience has grown with Ni-H_2 batteries, their advantages in this regard have made them the system of choice in many cases, especially for large spacecraft.

Eclipse time in low orbit can be as high as 40% of the orbital period, or on the order of 35 min for Earth orbits. Spacecraft in synchronous equatorial (geostationary) orbits go for extended periods without encountering eclipse. However, GEO spacecraft encounter two eclipse seasons each year, with each period being 45 days long. During these periods the spacecraft encounters one eclipse each day ranging from momentary duration at the beginning and end of the period up to 72 min at the midpoint.

Some spacecraft in near-polar sun-synchronous orbits with the orbit plane aligned essentially along the terminator may never be in eclipse; these are often called dawn-dusk orbits. The same may be true for deep space vehicles. This does not usually mean batteries are not needed, however. It may be necessary to maneuver the spacecraft off the sun line to obtain proper thruster pointing for course correction. Even if this is not necessary, it may be more efficient to use a battery to handle intermittent peak loads rather than to oversize the solar arrays to cope with the peak load. For similar reasons batteries may be required even on spacecraft using power sources (such as RTGs) that do not depend on the sun.

Given the power usage of the spacecraft and the maximum allowable depth of discharge (DOD) for the design lifetime of the battery, the battery can be sized by the following equation:

$$\text{DOD} = \frac{\text{Energy required during eclipse}}{\text{Stored battery energy}} \tag{10.1a}$$

or

$$\text{DOD} = \frac{P_L t_d}{C_{\text{chg}} V_{\text{avg}}} = \frac{P_L t_d}{E_{\text{bat}}} \tag{10.1b}$$

where

P_L = load power in watts
t_d = discharge time in hours
C_{chg} = charge capacity in ampere hours
V_{avg} = battery average discharge voltage in volts
E_{bat} = total battery energy capacity

The charge rate also drives battery size; a power input level that is too high can result in overheating of the battery and, if carried to extremes, explosive destruction. Although strict mathematical guidelines do not exist, a good rule of thumb for the allowable charge rate is

$$R_{\text{chg}} = \frac{C_{\text{chg}}}{15\,\text{h}} = I_{\text{chg}} \tag{10.2}$$

where the charge capacity is given in ampere hours. A "trickle charge," used when it is desired to store the maximum amount of charge in a battery, might use a charge rate of $C_{\text{chg}}/45$ h. Note that the rate of charge has dimensions of current, in this case a charging current.

According to the empirical rule of Eq. (10.2), a battery can accept a charge equal to 1/15 its total capacity per hour. This can prove to be a significant constraint in many cases. In a typical LEO spacecraft, where 40% of the orbit is spent discharging and 60% charging, the depth of discharge during eclipse is limited by the rate at which the charge can be restored during the illuminated phase, with the allowable rate given roughly by Eq. (10.2). All of the expended energy must be restored during the charge cycle, or there will be a net drain on the battery, which will ultimately result in the need to reduce the operations load to avert failure of the power system. Thus, the size of the battery in this case is driven not by the maximum allowable DOD as governed by the battery chemistry, but by the charge rate, with the depth of discharge per orbit limited to 7–8%. On the other hand, a GEO spacecraft will encounter only a few hundred discharge cycles in a 10-year life, and will have more than ample recharge time. Much deeper discharge can be tolerated in this case.

Equation (10.2) is quite conservative. Substantially higher recharge rates may be acceptable for a given battery; the manufacturer's specifications should always be the ultimate guide. Also, Eq. (10.2) is rather simplistic, because a variety of environmental factors can influence the allowable rate of charge, most importantly the battery temperature. Additionally, it should be noted that a battery generally must be charged at a slightly higher voltage than V_{avg}, or a full charge cannot be restored. Typically, the charging voltage will be of order 20% higher than the average discharge voltage. This will have implications for solar array design, as we will see in the following.

We offer here a simple example to demonstrate the process of preliminary battery size definition.

Example 10.1

What is the required size of a nicad battery to support a 1500-W payload in geostationary orbit, given the following design data:

Bus voltage	28 VDC
Peak load	1500 W
Maximum load duration	1.2 h
Battery energy density	15 (W · h)/lb at 100% DOD
Average cell voltage	1.25 V
Maximum DOD	70%

Solution. The number of cells is

$$N_{\text{cell}} = \frac{V_{\text{bus}}}{V_{\text{cell}}} = 22.4$$

We can choose either 22 or 23 cells; selecting 22 cells saves mass and results in a perfectly acceptable bus voltage of 27.5 VDC. From Eq. (10.1b) the total charge capacity and battery energy capacity are

$$C_{\text{chg}} = \frac{P_L t_d}{(V_{\text{avg}}\ \text{DOD})} = \frac{(1500\,\text{W})(1.2\,\text{h})}{(0.7)(27.5\,\text{V})} = 93.5\,\text{Ah}$$

and

$$E_{\text{bat}} = C_{\text{chg}} V_{\text{avg}} = (93.5\,\text{Ah})(27.5\,\text{V}) = 2571\,\text{W} \cdot \text{h}$$

The battery mass is

$$m_{\text{bat}} = \frac{E_{\text{bat}}}{e_{\text{bat}}} = \frac{2571\,\text{W} \cdot \text{h}}{15\,(\text{W} \cdot \text{h})/\text{lb}} = 171\,\text{lb}$$

It may be desirable to split the battery into two or three individual battery packs for ease in packaging, placement, and balance. Each battery pack must contain 22 series-connected cells to maintain the proper voltage. Finally, redundancy management issues have been ignored in this example.

10.7.3 Nicad Reconditioning

As mentioned earlier, to obtain maximum life from a nicad battery subject to repeated discharge, a reconditioning process is required. Reconditioning consists of a very deep discharge to the point of voltage reversal, followed by recharge under carefully controlled conditions. Figure 10.4 shows the effect of this operation. In the absence of reconditioning, the battery voltage begins to decline and, after four or five eclipse seasons, declines fairly rapidly. Usable depth of discharge is also greatly diminished. On the other hand, with periodic reconditioning the voltage declines only slightly from the "new" level and reaches a steady-state level. (Note that "eclipse season" refers to the two periods

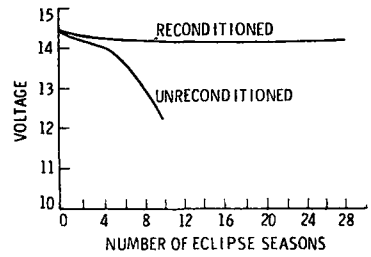


Fig. 10.4 Reconditioning of NiCd batteries.

per year when a geosynchronous spacecraft is eclipsed once per orbit. Thus, the scale on the abscissa converts to years when divided by two.) Because each eclipse season lasts for a few weeks at six-month intervals, it is easy to perform a reconditioning cycle during the several months of full solar exposure between eclipses. Other battery chemistries show a similar characteristic but are less widely used in space applications.

10.8 Primary Power Source

Once beyond the relatively small range of mission requirements for which batteries alone are suitable, choice of a prime power source includes several possibilities. The choice is governed by a variety of factors including the required power level, operating location, life expectancy, orientation requirements, radiation tolerance, and cost. Figure 10.2 depicts operating ranges that are deemed generally suitable for various prime power sources, based on power level and lifetime. Such curves are intended to be broadly indicative rather than specifically definitive. There is in any case substantial overlap between the various regions, indicating areas where more than one choice may be feasible. Also, other factors may bias the choice in a direction that would not be optimal from the viewpoint of simply meeting requirements on mass, power, and lifetime.

As previously observed, power requirements for spacecraft have tended to grow with time, and the related main bus voltage has risen accordingly in an effort to reduce conductor and component masses and resistive losses. Figure 10.1 presents data on past spacecraft as well as predictions concerning near- and far-term applications, both civil and military. The lifetime required of space assets has increased as well, a consequence of the large investment to build and operate a modern spacecraft and its associated ground equipment. In the case of scientific spacecraft, the required operational duration has increased because the targets are more distant, or the missions more complex, or both. In any case, life expectancy has become an increasingly important factor, especially because most spacecraft cannot readily be serviced or refurbished.

Environmental factors to be considered include the obvious issue of access to adequate solar illumination. If the spacecraft is too far from the sun (and Mars at 1.5 AU is at roughly the useful outer limit), solar arrays are not a viable choice. Concentrators may extend their capability to a limited degree, but eventually the inverse-square law renders solar energy simply too diffuse to be useful. Radiation resistance is another significant consideration; solar cells are seriously degraded by extensive exposure to radiation. This can be a major consideration for a spacecraft that must operate extensively in the Van Allen belts or other high-radiation environments.

In subsequent sections, various prime power sources are discussed, particularly in terms of the capabilities and limitations of each. Detailed technical descriptions of each are beyond the scope of this book, and the interested reader is referred to more specialized literature for such information.

10.9 Solar Arrays

Regardless of the size of the total array, each array is made up of a very large number of individual cells arranged on a substrate of some type. Although each cell puts out a relatively small current and voltage, proper series and parallel connection can provide any desired current and voltage within reasonable physical limitations. Individual cells are made in a variety of shapes and sizes. Probably the most common as this is written is the rectangular cell with dimensions on the order of 2 × 4 cm; however, cells in common use range from 2 × 2 cm to 2.5 × 6.2 cm. The rectangular shape allows for reasonably efficient packing, enabling array size and mass to be minimized. A well-designed array might have a cell packing density of 90%. Because some minimum spacing and allowance for connections must be provided, it is difficult to improve significantly on this, although innovative techniques may allow some gains to be made.

Because solar arrays can be quite large for higher power spacecraft, it quickly becomes impossible to find adequate area on the fixed spacecraft structure. Early low-powered spacecraft did in fact restrict the array area to the spacecraft skin. Most designs were drum-shaped spinning spacecraft, where only about 40% of the array was illuminated by the sun at any time. As power requirements grew, fixed arrays that were not specifically part of the spacecraft structural shell or skin were tried; however, launch vehicle nose fairing dimensions limit the utility of such an approach, and deployable solar arrays made an early appearance. Figure 10.5 depicts a variety of solar array designs.

Deployable solar arrays have typically been semirigid paddle-like structures that are deployed from the main structure after the spacecraft is injected into orbit. Keeping the array firmly locked to the spacecraft structure during launch allows the use of extremely lightweight structures. Such designs are constrained more by the need for rigidity than for strength; thus, structures having very thin

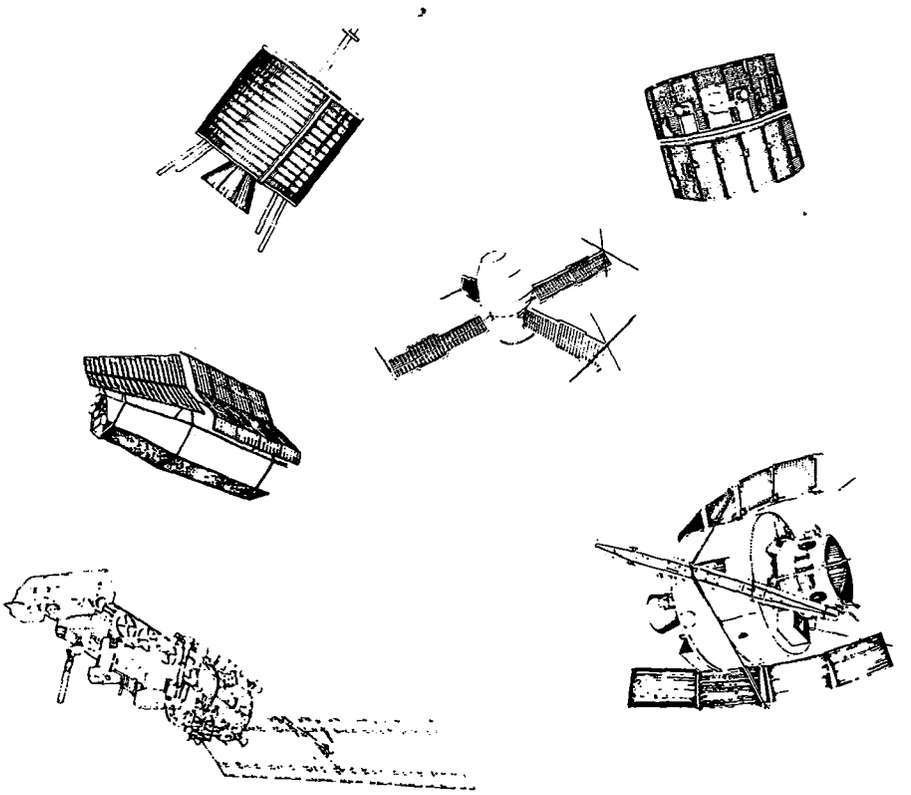


Fig. 10.5 Spacecraft solar array concepts.

cross sections are possible. Such structures are highly susceptible to handling damage, and often the primary criterion defining material thickness is the need to handle the assembly during installation. The development of highly rigid composite materials in recent years has greatly enhanced the possibilities for lightweight solar array design.

Because all photons incident upon a solar cell are absorbed within (at most) the first 10 μm, the active semiconductor material need not be at all thick. (Indeed, very thin cells allow red and infrared radiation to pass through the material without being absorbed, enhancing the conversion of shorter wavelength and hence higher energy photons, and increasing overall cell efficiency.) Similarly, the antireflective coating, cover glass, adhesive, and substrate that comprise the complete cell do not enhance the design by being thicker. Thus, the possibility and potential convenience of roll-up solar arrays was recognized early in the history of spacecraft design.

However, the early technology did not lend itself to such an approach; solar cells were too thick, connections were too stiff, and suitable substrates did not exist. Subsequently, however, advances in technology caught up with the

concept, and a variety of flexible roll-up and fold-up solar arrays are now in routine use. Perhaps the most spectacularly visible example was the 12.5-kW array demonstrated on space shuttle mission STS-10. This early demonstration led to the deployable arrays presently used on the International Space Station. This type of design, along with a large variety of deployable rigid array concepts, allows convenient packaging of very large arrays for launch.

Flexible solar arrays can introduce problems simply because of their flexibility. The primary resonant frequency of the roll-up arrays originally used on the Hubble Space Telescope (HST) allowed an undesirable interaction with the bandwidth of the attitude control system, causing difficulty in achieving the accurate pointing required. On-orbit replacement with stiffer arrays has largely corrected the problem. Another problem with very large arrays and the attendant high voltage and power levels they produce is the conductor mass and the required insulation between circuit elements. This can be particularly trying in flexible arrays, and represents one of the practical limits that solar array technology may impose on the spacecraft designer. Still, roll-up arrays offer the lowest mass approach currently available for providing large array areas.

To extend the capabilities of solar arrays to regions farther from the sun, reflective concentrators of various types have been proposed. Figure 10.6 shows two concentrator concepts. The flat concentrator array for use with silicon (Si) cells is basically a trough that increases the collection area relative to the cell area. This concept is useful at solar distances beyond 1.5 AU by increasing the energy available for conversion and keeping the cells from becoming excessively cold. Such concentrators can probably extend the useful range of Si cells out to 3–4 AU and possibly farther.

The other concentrator concept is particularly directed toward gallium arsenide cells. These cells, while coming into more common use, are nonetheless quite expensive relative to silicon arrays, and it is therefore desirable to minimize cell area. Also, these cells function best at a higher temperature than their silicon counterparts. The concept shown concentrates sunlight from a large collection area onto a small cell area. This reduces cell cost and brings the cells to a higher operating temperature than would otherwise be the case, thus providing a double benefit. An obvious disadvantage is that such concentrators are complex and expensive to manufacture. Any concentrator, even the relatively simple one shown for the conventional silicon array, clearly complicates stowage and deployment.

The highly successful Deep Space 1 technology demonstrator spacecraft used an array of Fresnel lenses to concentrate sunlight on the solar cells.

The HS-702 geosynchronous communications spacecraft used a trough-type concentrator for many years. However, problems began to develop, resulting in unacceptable power output degradation, and the concentrators were eventually eliminated. It appeared that the power loss was due to contamination, and it is possible that a change of material could have solved the problem while retaining the advantage of the concentrators.

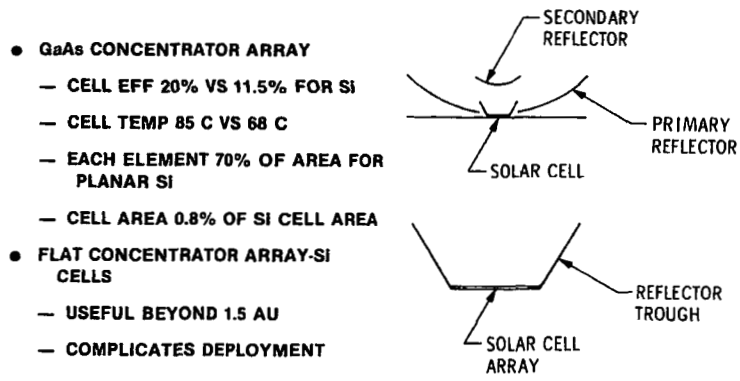


Fig. 10.6 Solar concentrators.

10.9.1 Solar Cell Characteristics

Certain technical characteristics of solar cells, such as temperature dependence and the current-voltage (*I-V*) curve, are of interest to the spacecraft designer.

As with other semiconductor devices, solar cell characteristics are temperature dependent. The first-order effect for a typical silicon cell operating within its normal range is a voltage decrease with temperature; the second-order effect is an increase of current flow with temperature. The effect on current is roughly 10% of that on voltage, so that the net result is a decrease in output power with temperature. Figure 10.7 illustrates this behavior; again, as voltage increases, current drops. The temperature coefficient for voltage, γ_V, will be in the range of -2 to -3 mV/K, while γ_I, the temperature coefficient for current, will be approximately 0.2–0.3 mA/K. The temperature corrections (from reference conditions) for voltage and current output are of the form

$$V = V_{\text{ref}} + \gamma_V(T - T_{\text{ref}}) \tag{10.3}$$

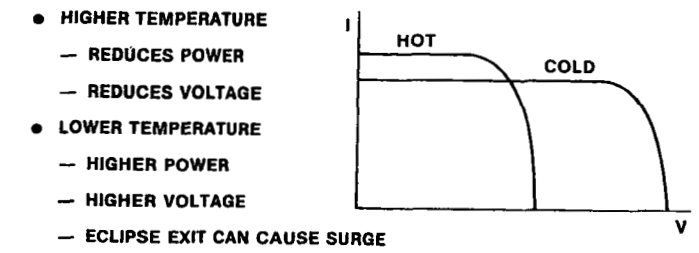


Fig. 10.7 Effect of temperature on solar cells.

and

$$I = I_{\text{ref}} + \gamma_I(T - T_{\text{ref}}) \tag{10.4}$$

Unless it is important to obtain the highest possible accuracy, it is common to ignore the separate variations of voltage and current and use instead a composite temperature coefficient for power in an equation of the same form as Eqs. (10.3) and (10.4). However, the effect of the voltage increase for cold panels, as with a spacecraft exiting an eclipse period, must be considered in the design of the power system, because a major power surge can occur under these circumstances.

The shape of the *I-V* curve shown in Fig. 10.8 is typical of solar cells and is important in the design of spacecraft power systems. To minimize mass and maximize efficiency, it is obviously desirable to operate the array at its maximum power point. Because power is the product of current and voltage,

$$P = IV \tag{10.5}$$

selection of the operating point to maximize the area under the *I-V* curve allows the maximum power point to be found. This is the point at which the maximum area rectangle that will fit within the *I-V* curve intersects the curve. As can be seen in Fig. 10.8, this lies on the knee of the *I-V* curve. Although some specific applications may dictate operation at some other point, the majority of systems will be designed for maximum power point operation. With the maximum power point for the cells defined, the current and voltage of individual cells is known. This information, in conjunction with the voltage and current requirements of the spacecraft, defines the series-parallel arrangement of the cells in the array.

It is common to specify the maximum power operating point, V_{mp} and I_{mp}, at a given temperature, often room temperature, as reference conditions in Eqs. (10.3) and (10.4), though of course this is not required. For a conventional silicon cell, V_{mp} will be 0.4–0.5 V at moderate temperatures. The maximum power operating point for current, I_{mp}, depends on the area of the cell for a given illumination level. For cell sizes in common use, I_{mp} will be in the range of 30–120 mA.

It occasionally results that a transient load or other problem will drive the operating point off the knee of the curve and down into the lower voltage range. The array may then not be able to return to the normal operating point. Having a battery in the system helps to stabilize it against such eventualities. Even systems

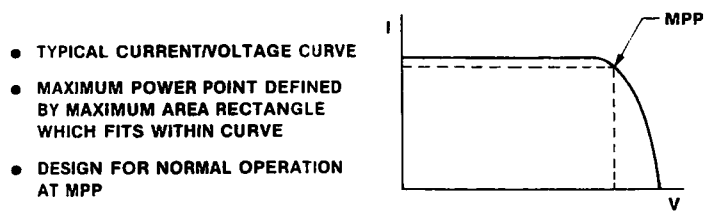


Fig. 10.8 Typical solar cell power characteristic.

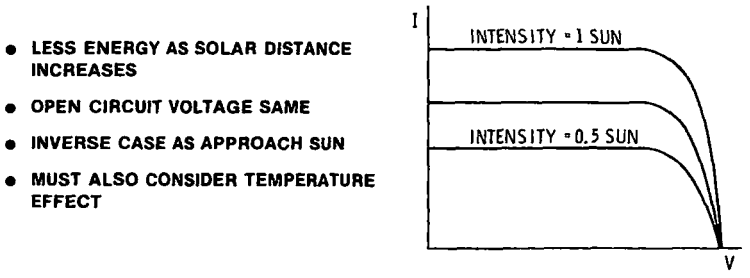


Fig. 10.9 Effect of distance from the sun.

that may not require a battery in normal operation (because they are never in eclipse) may require a battery for this reason. Off-line batteries or other devices that can be switched into the circuit to provide a temporary power boost may also be used. The operational alternative, of course, is controlled load-shedding to allow the bus voltage to recover.

As the solar array moves farther from the sun, the available current drops, while open circuit voltage stays the same, or increases if the temperature is lower. (This behavior is a consequence of the photoelectric effect, one of the first quantum-mechanical physical phenomena to be observed. Not widely realized is the fact that Albert Einstein received his Nobel prize for the explanation of the photoelectric effect, rather than for his development of the theories of special and general relativity, which remained controversial for several decades.) This leads to a family of curves similar to that shown in Fig. 10.9. A similar set could be drawn for an array moving toward the sun, with current increasing and voltage dropping slightly as solar distance decreases. Regarding Fig. 10.9, it will be noted that as a result of these changes the maximum power point moves slightly. This should be considered in spacecraft power system design for planetary missions. For example, in the case of a Mars orbiter, one would normally design for the maximum power point corresponding to Mars distance from the sun, because that is where the demand for power to operate the onboard instruments will be greatest, and because excess power will in any case be available while near the Earth.

10.9.2 Sun Tracking

It will be obvious that maximum power is available when the sun line is normal to the array. As the angle between the sun line and the array normal deviates from 0 deg, one expects that a cosine relationship between incidence angle and array output would obtain. This is indeed the case for angles up to about 60 deg. At larger angles, where the sun angle approaches parallel to the array face, the cosine relation begins to break down due to the finite thickness of the

cells and other effects such as specular reflection from the cover glass surface. Figure 10.10 gives an approximate curve for current vs sun angle.

Although early spacecraft were often designed with fixed-orientation panels (if separate panels were used at all), modern spacecraft are almost universally designed to allow sun tracking by the solar array. The required tracking accuracy is not particularly challenging, since even a 10-deg error yields a cosine loss of only 1.5%. However, it is usually better to minimize sudden attitude disturbances, and so typically the solar array will be articulated in small increments more or less continuously, rather than in a few large maneuvers.

To cause a solar panel attached to a spacecraft in a planetary orbit to track the sun requires, in general, two angular degrees of freedom (DOF). The first degree of freedom (often called the α angle) compensates for the apparent rotation of the sun vector in the orbit plane, as seen from the spacecraft in orbit. Clearly, α will range from 0 to 360 deg over the course of a single orbit, a factor that must be considered in making arrangements to transfer power across the rotating interface between the solar panel and the spacecraft body. (The power transfer harness cannot continue to be wound indefinitely around a fixed axle on the spacecraft.) The second degree of freedom, the β angle, is necessary to compensate for the component of the sun vector normal to the orbit plane. Unless the orbit is sun synchronous, the β angle will vary more or less slowly throughout the year, as determined by the particular orbit parameters (see Chapter 4).

It is often possible to use the spacecraft itself to supply one of these degrees of freedom; for example, many spacecraft deal with changes in β angle by means of body rotation, leaving a single solar array drive gear to cope with α angle articulation. However, when numerous other instruments on a spacecraft must also be oriented properly, and when thermal control requirements are taken into account, it will in the end often be simpler to use a full 2-DOF separately

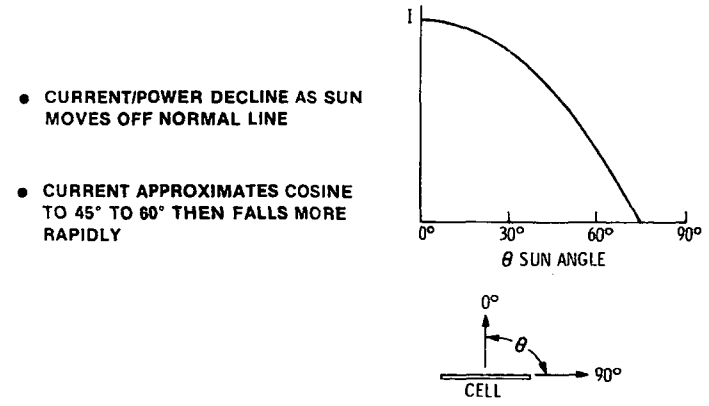


Fig. 10.10 Effect of sun angle on solar array power.

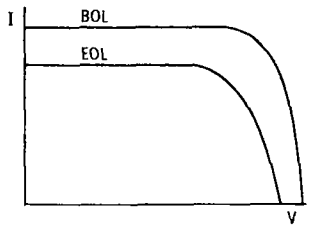


Fig. 10.11 Radiation effect on solar cells.

articulated solar array. These problems, and the decisions that arise from them, are usually the joint province of the attitude control engineer and the mechanical design engineer.

10.9.3 Radiation

As mentioned earlier, radiation has a detrimental effect on solar cells. The general effect of this degradation is shown in Fig. 10.11. Some loss of efficiency will take place during any mission of appreciable duration. If the operation takes place in more severe environments, e.g., the Van Allen belts, the rate of degradation will be more severe. Because the spacecraft normally requires as much power late in the mission as at the beginning, the solar array size must be based on end-of-life (EOL) capability rather than on beginning-of-life (BOL) performance characteristics. The radiation environment is discussed in general terms in Chapter 3, but for detailed design the specific environment should be assessed in the context of a particular mission, its orbit characteristics, its intended operational period relative to the 11-year solar cycle, etc.

10.9.4 Solar Cell Efficiency

Considerable effort has been expended on development of gallium arsenide (Ga-As) solar cells, which are more efficient and more radiation tolerant than those made of silicon. Figure 10.12 compares the radiation resistance of the two types of cells. Note that at some point the curves cross, and silicon may well be better again beyond that point. However, this effect occurs at very high radiation fluence and may not be of practical interest.

In terms of efficiency at the cell level, typical crystalline silicon cells at room temperature deliver a solar-to-electric conversion efficiency of 11–16% in production. As this is written, several commercial vendors have recently introduced production silicon cells with efficiencies in the 18–20% range.[3] Efficiency in the mid-20% range can be achieved in limited quantities. Gallium arsenide offers on the order of 18–20% efficiency in production cells, and close to 30% in special cases, but at a cost greater than that of silicon cells. As the cost

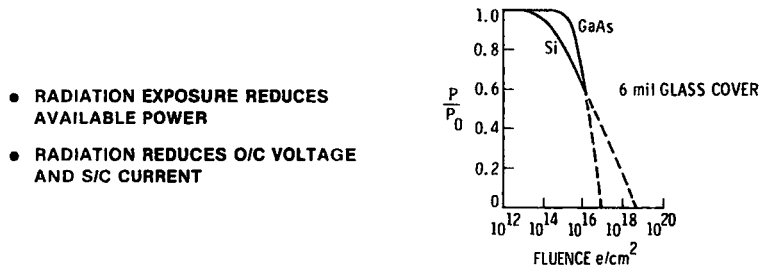


Fig. 10.12 Solar photovoltaic radiation effect.

of gallium arsenide cells has dropped, their use has become more common, particularly on long-lived GEO communications spacecraft, where the higher initial cost is easily overcome by the income potential of longer life. Concentrators may still be attractive in some applications to reduce the total expenditure for solar arrays, particularly Ga-As arrays.

A recent development, which offers the promise of very high efficiency, is the multijunction or multilayer solar cell. In this arrangement, the top or outer cell is optimized for conversion of light in the visible regime, where the solar output peaks. Beneath this top cell lie one or two layers of additional cells. These are optimized for energy conversion in the infrared range, thus using some of the energy that would ordinarily be radiated to space as waste heat. The conversion efficiency of the IR cells is much lower than that of the visible wavelength cells, but they nevertheless make a significant contribution. Cells of this type can deliver close to 30% conversion efficiency.

As would be expected, multilayer cells are heavier and more expensive than conventional cells. However, the cost penalty is not as great as might be imagined, because the additional layers of material are all deposited during a single manufacturing sequence in a vacuum chamber.

10.9.5 Preliminary Solar Array Sizing

Although detailed design of solar arrays is beyond the intended scope of this text, it is straightforward to perform preliminary sizing calculations to provide general characteristics of an array required for a given spacecraft. A simple example will illustrate these calculations.

Example 10.2

What is the size of a solar array necessary to support a 1500-W load, plus a suitable level of battery charging? If we assume 2 × 4 cm cells, how many are needed? The follow basic data may be used:

Cell efficiency	11.5% at 301 K
Maximum operating temperature	323 K
EOL degradation (10 years)	30%
Worst-case sun angle	6.5 deg off normal
Solar intensity	1350 W/m^2 at 1 A.U.
Temperature coefficient	−0.5%/K (power)
Packing factor	90% (10% loss for spacing)
Battery capacity	90 Ah

Solution. The array voltage must exceed the battery voltage for the battery to charge. For these voltage levels, a good rule of thumb is that the array must operate at a level 20% above the battery voltage. Assuming a 27.5-V battery as in Example 10.1, we have

$$V_{\text{array}} = (1.2)(27.5\,\text{V}) = 33\,\text{V} = V_{\text{chg}}$$

The EOL power requirement is equal to the 1500-W load plus the required battery charging power, which, from the empirical rule of Eq. (10.2), is found to be

$$P_{\text{chg}} = V_{\text{chg}} I_{\text{chg}} = V_{\text{chg}} R_{\text{chg}} = \frac{V_{\text{chg}} C_{\text{chg}}}{15\,\text{h}} = \frac{(33\,\text{V})(90\,\text{Ah})}{15\,\text{h}} = 198\,\text{W}$$

Thus, the total EOL power required of the array is

$$P_{\text{EOL}} = 1500\,\text{W} + 198\,\text{W} = 1698\,\text{W} \cong 1700\,\text{W}$$

We must now assess the various efficiency factors that cause the BOL power level to degrade to the EOL condition.

The effect of temperature at the hot operating point is to reduce efficiency by an amount proportional to the difference between the specified operating temperature and that at which maximum performance is obtained. Thus,

$$\eta_{\text{temp}} = 1 - \left(\frac{0.005}{\text{K}}\right)(323\,\text{K} - 301\,\text{K}) = 1 - 0.11 = 0.89$$

The degradation from radiation exposure is given as 30%, yielding an EOL efficiency due to radiation of

$$\eta_{\text{rad}} = 1 - 0.3 = 0.7$$

while the cosine loss due to the off-normal sun angle yields

$$\eta_{\text{angle}} = \cos(6.5\,\text{deg}) = 0.9766$$

The end-of-life power is the result of applying these losses to the beginning-of-life array power. Thus,

$$P_{\text{EOL}} = \eta_{\text{rad}} \eta_{\text{temp}} \eta_{\text{angle}} P_{\text{BOL}} = 1700\,\text{W}$$

and we then have

$$P_{\text{BOL}} = \frac{1700\,\text{W}}{0.619} = 2746\,\text{W} \cong 2750\,\text{W}$$

We are given a basic silicon solar array efficiency of $\eta_{\text{Si}} = 0.115$, and a solar illumination intensity of $I_s = 1350\ \text{W/m}^2$, so that in terms of total cell area A_{cell}, we have

$$P_{\text{BOL}} = \eta_{\text{Si}} I_s A_{\text{cell}} = (0.115)(1350\,\text{W/m}^2)A_{\text{cell}} = 2750\,\text{W}$$

hence

$$A_{\text{cell}} = 17.7\,\text{m}^2$$

The packing efficiency was given as $\eta_{\text{pack}} = 0.9$, and so the array area satisfies the relation

$$A_{\text{cell}} = \eta_{\text{pack}} A_{\text{array}} = 17.7\,\text{m}^2$$

and thus

$$A_{\text{array}} = \frac{17.7\,\text{m}^2}{0.9} = 19.7\,\text{m}^2 \cong 20\,\text{m}^2$$

For 2×4 cm cells, the area of a single cell is $8 \times 10^{-4}\,\text{m}^2/\text{cell}$; hence the number of cells is

$$N_{\text{cell}} = \frac{A_{\text{cell}}}{8 \times 10^{-4}\,\text{m}^2/\text{cell}} = 22{,}142\,\text{cells}$$

10.10 Radioisotope Thermoelectric Generators

The radioisotope thermoelectric generator (RTG) is a power source that renders the spacecraft independent of the sun. Although this is an advantage in many cases, it comes at a price that explains why these units have seen only limited use. The RTG functions by converting the heat energy generated by decay of a radioisotope into direct current electricity by means of the thermoelectric effect. In a typical RTG (Fig. 10.13), a central core of radioisotope material is surrounded by an array of thermocouples connected in series-parallel to obtain the desired voltage and current output. The hot side of the thermocouple junction is in contact with the canister containing the radioisotope, and the cold side is in contact with the external wall of the RTG, from which heat is radiated to space. The efficiency of the RTG is ultimately limited by the conversion efficiency of the thermoelectric elements.

Modern semiconductor thermoelectric devices, such as are used for Galileo and Cassini, can deliver a conversion efficiency of 10–11%, and research continues in an effort to improve this. Other limitations involve the internal thermal conductivity of the assembly. Considerable effort goes into designing a

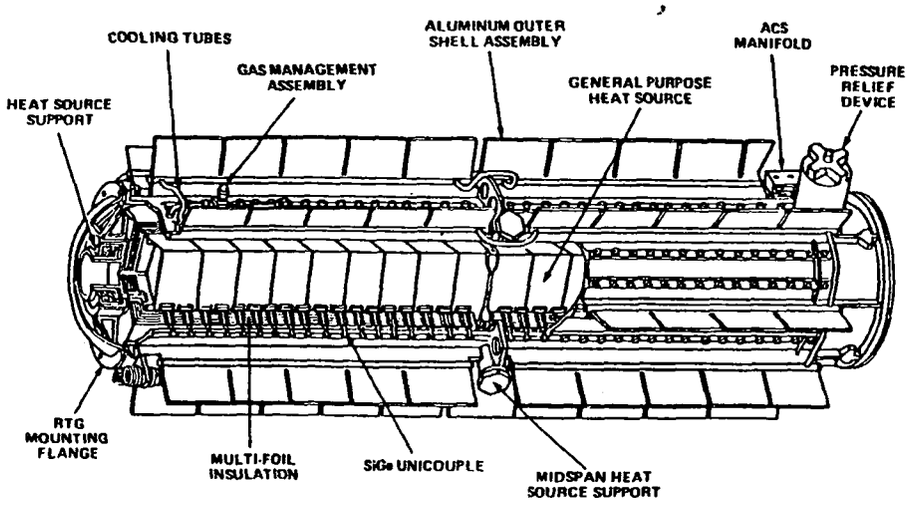


Fig. 10.13 Galileo RTG. (Courtesy of Jet Propulsion Laboratory.)

thermal path with minimum temperature drop from the isotope to the hot junction and from the cold junction to the outer case, while at the same time minimizing any energy leakage between these points that would bypass the thermoelectric conversion elements. Because the conversion efficiency of a given design depends on maximizing the temperature differential across the thermoelectric units, and the upper and lower limits are driven by material limits on the hot side and radiator size (usually) on the cold side, the importance of minimizing conductive drops or thermal leakage is obvious. These factors plus internal resistance and other losses explain why overall RTG efficiency is typically 6–7%, rather than that of the thermoelectric elements alone. All of this means, of course, that a very large amount of waste heat is produced for every unit of electrical energy produced.

The RTG used for the Galileo Jupiter Orbiter delivers 298 W $\pm$ 10% at the beginning of life from a mass of about 56 kg. The thermoelectric elements are doped silicon-germanium (Si-Ge). The radioisotope material used is plutonium-238 (^{238}Pu).

Although all radioisotopes exhibit a loss in energy output with time, the 86.7-year half-life of ^{238}Pu is not the major life-limiting mechanism of current RTGs. Degradation of the thermoelectric elements, caused mostly by dopant migration at the relatively high temperatures involved, is a more significant cause of performance loss. Breakdown of insulators because of temperature and radiation is also a factor.

Although ^{238}Pu has been most commonly used in RTGs, there are other candidates offering the combination of reasonably long half-life with high energy output that together qualify an isotope for RTG use. Table 10.4 lists some

Table 10.4 RTG material properties

Property	Po-210	Pu-238	Ce-144	Sr-90	Cm-242
Half-life, years	0.378	86.8	0.781	28.0	0.445
Watts/gram, thermal	141	0.55	25	0.93	120
$/Watt, thermal	570	3000	15	250	495

candidates. Note that, for long space missions, only strontium-90 (^{90}Sr) has a half-life adequate to be a viable candidate in addition to ^{238}Pu.

Each radioisotope has a particular form and energy of radiation that is given off in the decay process. ^{238}Pu gives off an alpha particle and a relatively low-energy beta particle, both of which are relatively easy to shield compared with the high-energy gamma radiation from ^{90}Sr. The long-term effect of RTG radiation upon electronics is definitely a factor in radioisotope selection.

By their nature, RTGs cannot be turned off in the conventional sense. That is, the radioisotope continues to decay and to generate heat regardless of any external action. Similarly, the thermoelectrics will generate electricity whenever a temperature differential exists and a load is placed across the output terminals. Because there is no practical way to control the generation of electricity at an essentially constant rate within the RTG, control must be external. In spacecraft applications control is typically accomplished by use of a shunt regulator to dispose of electrical energy in excess of the operational requirements at any given time. RTGs are usually stored in a shorted condition. This has the desirable characteristic of reducing the temperature of the RTG during storage because of the Peltier thermoelectric cooling effect.

Aside from their high cost, particularly for the radioisotope, RTGs have a variety of problems that have limited their use when other power sources will suffice. The high external temperature and the radiation from the radioisotope are a major problem in ground handling. Special equipment is required because the assembly crew cannot directly handle the units without thermal protection (e.g., heavy gloves). This complicates structural assembly and the making of electrical connections and significantly increases installation time. This in turn is problematic because it increases the exposure of the crew to radiation from the radioisotope. Such radiation can be of sufficient intensity to mandate the use of oversize work crews so that no individual exceeds allowable dose limits. It is also important to allow for contingencies. If the crew is sized to reach the exposure limit in the course of a normal installation, then there will be no one available in the event of a problem requiring the RTGs to be removed and reinstalled.

A further problem is the possibility of Earth contamination by the radioisotope in the event of a launch failure or decay from orbit. The radioisotopes in RTGs launched to date have been extensively protected against both destruction in the event of a launch failure and incineration upon reentry. The fuel itself is normally

the oxide of the radioisotope and is therefore reasonably strong and resistant to high temperature by itself. The lumps of oxide are then encased in graphite for atmospheric entry and impact protection. Extensive tests are performed to qualify the fuel elements for this environment, with notable success. The aborted Apollo 13 lunar landing mission eventually resulted in reentry of the RTG fuel element at lunar return velocity, followed by impact in the Pacific Ocean. There has been no subsequent evidence of any release of radioisotope material from this event.

Similarly, most reasonable launch failure scenarios can be accommodated by the internal protection built into the RTGs. It is possible, however, to postulate a launch failure of such severity that the fuel elements will be shattered and the radioactive material scattered into the atmosphere. This may be technically (if not politically) acceptable provided the material is thoroughly dispersed in the upper atmosphere in the form of very fine particles, so that the concentration at any point on the surface will be very low when the material settles out of the atmosphere.

It must be noted that an accident of this magnitude would represent a very improbable launch failure scenario, most of which are rather benign as explosive events are judged (i.e., they are deflagration as opposed to detonation events). For example, analysis indicates that the 1986 Challenger accident would not have created a hazard due to radioisotope dispersal had RTGs been aboard. In another noteworthy case, the launch abort and subsequent destruction of a military payload carrying an RTG was followed by recovery of the intact RTG from the water off Vandenberg Air Force Base, California. The unit was reused on a subsequent mission.

These examples aside, it remains necessary to plan for the worst possible case. Extensive analysis is necessary to determine the possible hazard environment and to devise protection adequate to ensure that the radioactive material comes down in a condition that minimizes dispersion and allows for recovery. This is especially of concern for plutonium, which, besides being radioactive, is extremely toxic.

The radiation from RTGs is detrimental to spacecraft electronics and instruments, making it necessary to mount the units on booms at some distance from the body of the spacecraft, and often to provide shielding as well. It must be noted that the spacecraft configuration as stowed for launch will not allow the RTG boom to be in its deployed configuration. Therefore, the spacecraft must be able to survive whatever radiation exposure will accrue during the stowed period by means of shielding and the inherent radiation tolerance of the onboard electronics.

10.11 Fuel Cells

Fuel cells are devices that allow direct conversion of chemical energy into electricity. In this they are like batteries, with the difference that fuel cells operate

much more efficiently. An oxidizer and a fuel are fed into the cell, which is roughly similar to a battery in its internal arrangement. Electricity is generated directly from the oxidation reaction within the cell, aided by the presence of a catalytic material, but without the high temperature and other complications associated with combustion. Space applications of fuel cells have been primarily to manned spaceflight, and were first used to power the later Gemini spacecraft as well as the Apollo CSM, the lunar module, and the space shuttle.

Although numerous fuel-oxidizer combinations are possible and have been used experimentally, only hydrogen and oxygen have so far been used as reactants in operational fuel cells for space applications. The output of the cells is essentially pure water, which is used for crew consumption with little or no treatment. Laboratory demonstrations have shown conversion efficiencies approaching 35%.

The overall mass of a fuel cell system is a function of the desired operating time, since the mass of the reactant must be included in the assessment. For a system of fixed mass, however, an energy density of 500 (W · h)/kg at a power level of 2.6 kW is a reasonable figure of merit. Lacking a substantial industrial production base, fuel cells remain a costly source of power, in the range of $3000/kW for commercial systems (essentially the same as for rechargeable batteries) and much higher for space-qualified systems.

Fuel cell development is in many ways still in its infancy. Fuel cells run most efficiently on pure hydrogen and oxygen, but because of the difficulty of storing and handling liquid hydrogen relative to most other fluids, there has been considerable interest in the use of other sources of hydrogen for commercial devices. Commercial fuel cell development has focused on the use of methane, methanol, ethanol, natural gas, and other sources rich in hydrogen. Catalytic conversion is necessary to render the hydrogen free for use in the fuel cell, and carbon-based impurities interfere with the desired operation of the cell. These factors have hindered the development of commercially viable production fuel cell technology.

A variant concept, the so-called direct methanol fuel cell (DMFC), offers the possibility of very small fuel cell power packs for portable applications and could be of great interest as a source of power for space suits and other applications where high energy density and small size are required. Although not as efficient as its larger brethren, the DMFC offers energy density of at least twice that of its lithium-battery competitors and can be "recharged" simply by adding methanol. The first use of DMFC power packs will undoubtedly be in commercial laptop computers, cellphones, and other consumer electronics devices. However, there is nothing in their nature that precludes use in space, and such applications can be expected to follow.

A tantalizing possibility for energy storage in large systems, such as the International Space Station, is to use regenerative fuel cells in lieu of batteries. In this scenario, fuel cells would use stored hydrogen and oxygen to generate electricity during eclipse periods. During the illuminated portion of the orbit,

solar arrays would generate electricity to power the spacecraft and to recharge the fuel cells by electrolyzing the water generated during operation. The resulting hydrogen and oxygen would then be stored to provide reactant to the cells for the next eclipse period. Regenerative fuel cells have been demonstrated in the laboratory but have not so far been reduced to engineering practice.

10.12 Power Conditioning and Control

The power conditioning or processing portion of the space power system carries the responsibility for many of the functions listed earlier in this chapter. Power conditioning is necessary because the voltage from the power source may vary substantially, especially with solar arrays, which are of course the most common source of primary spacecraft power, for a variety of reasons including load variability, array temperature, and other external environmental factors.

Broadly considered, the power conditioning subsystem must fulfill three functions on a spacecraft utilizing solar arrays. First, it must control the solar array output in response to changes in load requirements and to changes in array temperature and sun angle, which as we have seen significantly alter the source properties. Second, it must control the battery charge-discharge cycle, supplying the proper charging voltage and current and regulating the average discharge voltage. Finally, the power system must regulate the voltage supplied to the remainder of the spacecraft system to the specified level (within some tolerance), thus protecting the other subsystems from the fluctuations already cited. This last requirement is obviously present even on spacecraft using RTGs, fuel cells, or any other power source. The second requirement may be relevant as well, if the peak loads exceed the steady-state source capability.

In some cases, a variety of voltage levels for different functions may be required, but at the very least, main bus regulation will be needed. This concerns the point, discussed earlier, as to whether the power conditioning system should supply most or all specific subsystem requirements, or whether it should merely be a source of stable bus voltage, allowing individual subsystem designers to deal with their own requirements.

Any electrical noise generated by the power source or the control electronics must be isolated from the main bus. The main bus in turn must be isolated from any power source faults, such as loss of part of a solar array or voltage transients due to entry into and exit from an eclipse period.

Finally, the power system and other spacecraft subsystems must be protected from faults in any other subsystem. We ignore here the issues and design trades surrounding the use, or not, of dual-bus power and spacecraft "housekeeping" systems, and other aspects of redundancy architecture and management. Some of the issues are discussed further in Chapter 12.

Although it is beyond the scope of this text to explore the details of power control circuitry, it will be useful to discuss the basic concepts at the block-diagram level. Those needing more detail are referred to other texts.[4-6]

Power control systems for spacecraft using solar arrays are broadly categorized as dissipative and nondissipative systems. In dissipative systems, as the name implies, excess power is shed resistively, while in nondissipative systems, the solar array itself is regulated through a DC-DC converter to operate at its peak-power point, according to the load demanded from it. As loads decrease, the array output is shifted toward its open-circuit high-voltage operating point, which yields the lower current that is required. Conversely, as loads increase, operation is shifted to a lower-voltage, higher-current operating point, up to the maximum power that can be delivered by the array.

Dissipative systems are also called direct energy transfer (DET) systems, because they are not in series with the array output. For this reason, they offer excellent overall efficiency and have the additional advantage of inherently simpler design, and thus lower parts count. Nondissipative systems are commonly called peak power tracking (PPT) systems, a name aptly descriptive of their operation. PPT systems are more complex and introduce some inherent inefficiency due to the requirement for a DC-DC power converter in series between the solar array and the load. However, for LEO spacecraft encountering a wide range of operating requirements, and for spacecraft needing maximum EOL array operating efficiency, PPT systems can be appropriate.

In most DET systems, a shunt regulator will be connected across the solar array, in parallel with the battery and its charge controller, and with the spacecraft loads, as shown in Fig. 10.14. As the name implies, the shunt regulator controls spacecraft power by dissipating current in excess of that required by the instantaneous load, which consists of battery charging requirements plus spacecraft operational needs. Shunt regulators are common because they are efficient and because they are simply and reliably implemented.

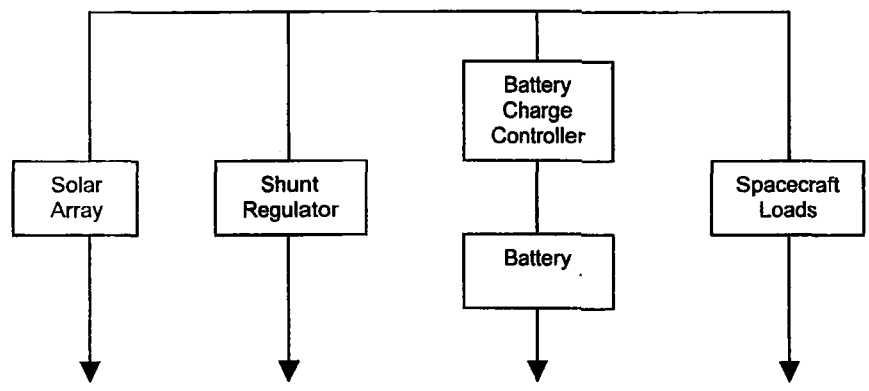


Fig. 10.14 Basic shunt regulator concept.

In the basic shunt-regulator concept, the bus voltage is unregulated and varies between the post-eclipse cold-array voltage on the charge cycle and the battery discharge voltage on the discharge cycle. As indicated earlier, this will result in considerable bus voltage variation, which may be acceptable. If unacceptable to a particular instrument or subsystem, the raw bus voltage must be further regulated within the subsystem.

Alternatively, the spacecraft bus may itself be more carefully regulated, either on the charge cycle, the discharge cycle, or both. Figure 10.15 shows a block diagram example of a shunt-regulated array with a fully regulated bus.

The shunt regulator itself can be either a simple linear controller or a switching shunt. If a switching shunt is used, the output is pulse-width modulated to produce the desired average level, a process resulting in a higher level of self-generated electromagnetic interference (EMI) than for a linear shunt. This will generally result in the requirement for additional shielding of the system and smoothing of the output power to avoid interference with other spacecraft systems.

Series regulation of solar array power to the bus is also possible. Figure 10.16 provides an example of the concept. In this case, the bus is controlled by dissipating excess power through a voltage drop in series with the load. Series regulation tends to be more complex than shunt regulation and is therefore less common, though there are advantages to providing better control of bus voltage as delivered by the solar array.

As will be obvious, more elaborate electronic control circuitry, operating at less overall efficiency, is required if the bus voltage is to be closely regulated. The engineering assessment as to whether overall mass and complexity are minimized by carefully regulating the main bus, as compared with performing the power conditioning at the subsystem level, must be made in each case. In general, larger, higher power, more complex satellites will benefit at the system level by having a central bus controller, and conversely for smaller, simpler spacecraft.

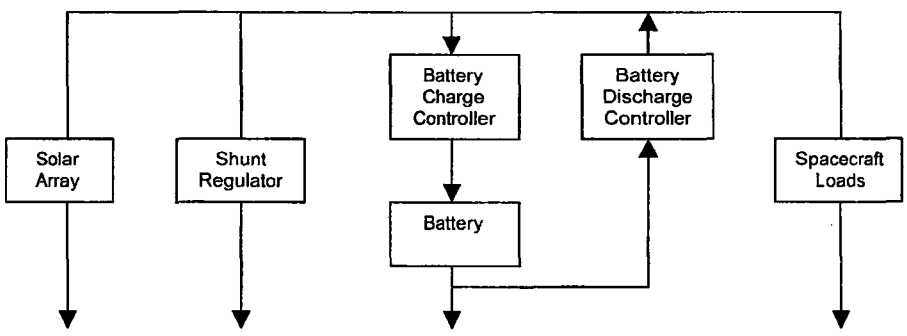


Fig. 10.15 Shunt regulator with battery charge-discharge regulation.

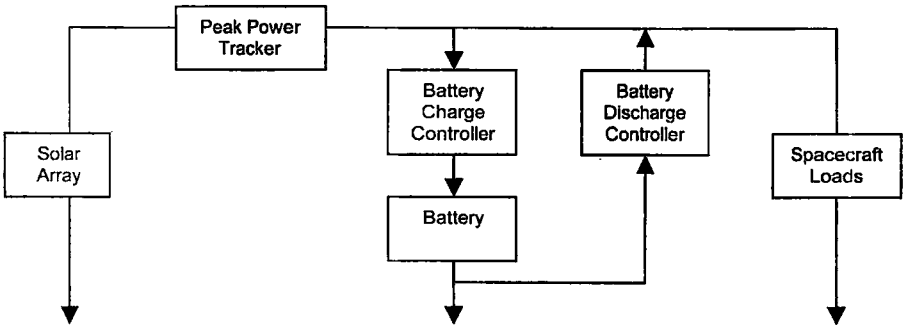


Fig. 10.16 Series regulation with peak-power tracking.

10.13 Future Concepts

10.13.1 Nuclear Reactors

Nuclear reactors offer considerable promise for the future. For very large power levels, hundreds of kilowatts to megawatts, reactors may be the only viable source in the next several decades. The nuclear reaction supplies heat, which is converted to electricity by a variety of techniques. Candidate energy conversion techniques include thermionics, thermoelectrics, and Stirling, Brayton, or Rankine cycle engines driving an alternator.

Thermoelectric energy conversion was discussed briefly in connection with RTGs. An allied concept, in the sense that it requires no moving parts, is that of thermionic energy conversion, which, however, uses a completely different concept. In thermionic conversion, heat is converted to electricity by boiling electrons from a hot emitter, or cathode, and collecting them at a cooler anode. This is exactly the mechanism at work in old-fashioned vacuum tubes, except that the heat is supplied by a nuclear reactor rather than by resistive heating of a wire filament.

Practical thermionic systems require the cathode temperature to be very hot, 1600–2000 K, while the anode must be cooler, 800–1000 K, to avoid significant back-emission of electrons. The spacing between cathode and anode must be relatively small, e.g., <1 mm. Power densities in the 100–1000 W/m^2 range can be achieved at a conversion efficiency of 10–15%. The astute reader will note that although this is higher than for thermoelectric conversion, it is only about half the intrinsic Carnot efficiency at the given cathode/anode temperature difference. A more subtle disadvantage is the tendency of the anode and cathode to expand differentially, threatening to eliminate the gap between them, which must be small but not zero. Manufacturing tolerances for thermionic converters are obviously critical. However, these disadvantages are compensated by their very high tolerance to heat and radiation, high reliability, and compactness.

All conversion concepts require radiators to reject waste heat to space. These radiators become very large for high-power units and present a major design challenge. Dynamic conversion concepts are generally much more efficient than static designs and therefore require a smaller reactor and, in some cases, a smaller radiator. Mitigating these advantages is the fact that the vibration and other disturbances typical of dynamic systems may be a problem.

Reactors have the advantage that, until they are in operation, they are not highly radioactive and can be handled with relative safety. When in operation, however, the radiation is very intense and much more damaging than that characteristic of RTGs. Heavy shielding is required even in unmanned applications, because electronic components cannot otherwise withstand the radiation from the reactor. For manned applications, the shielding and separation requirements become far more stringent. Figure 10.17 shows a typical reactor-powered spacecraft design using geometric separation to reduce shield mass. In the configuration shown, a shadow shield is used to protect only a relatively small portion of the volume of space surrounding the reactor, thus saving substantial mass. Of course, this design precludes close proximity operations outside the shield shadow. An early concept for the SP-100 nuclear reactor-based space power system was to have a mass of 3000 kg for a 100-kWe system. However, as

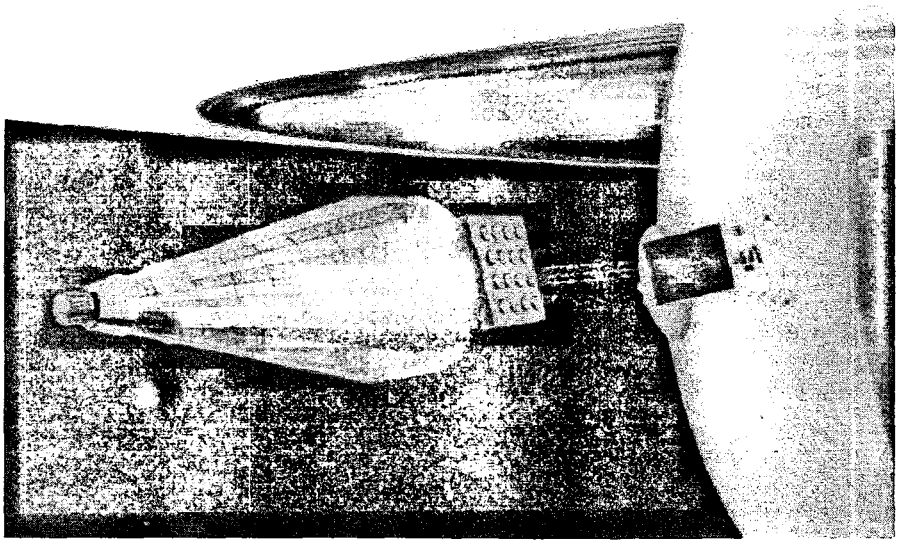


Fig. 10.17 Nuclear-electric spacecraft. (Courtesy of Jet Propulsion Laboratory.)

the SP-100 design matured, the weight essentially doubled, a factor that contributed to the program's eventual cancellation.

The mass-to-power ratio of a reactor design improves somewhat as the size increases. Essentially, reactors perform best where traditional solar power systems do poorly, such as on planetary surfaces, at great distances from the sun, and when large amounts of power are needed. Because of the severe limitations of non-nuclear alternatives, interest has once again arisen in space nuclear power options, especially in the larger power output category, but specific details are unavailable as this is written.

10.13.2 Dynamic Isotope Systems

The dynamic isotope system is a concept for obtaining more electrical power from the same isotope heat source as used for a traditional RTG. In this approach, the heat from the decaying isotope is used to heat the working fluid of a Brayton, Rankine, or Stirling cycle engine, which in turn drives an alternator. Because of the much higher conversion efficiency of these dynamic systems as compared to that of thermoelectric or thermionic systems, 500–700% more power can be obtained from a given quantity of isotope. This has advantages in reducing cost, radiation exposure, and mass. Detrimental factors include a reduction in reliability due to the added moving parts and the vibration that any dynamic system will tend to generate. Another possible disadvantage is the requirement to shed waste heat at a temperature lower than that of a typical RTG, thus requiring larger radiator area. (This comment might appear to conflict with the advantage cited earlier regarding the possibility of having smaller radiators. Indeed, either result may be true. High thermal efficiency requires a low cold-side temperature, and thus a larger radiator. However, higher intrinsic conversion efficiency by itself allows a smaller radiator. The net result depends on the particular system parameters for a given case.) Dynamic isotope conversion systems have been tested extensively but, as this is written, have not been flown.

10.13.3 AMTEC

An interesting energy conversion concept for potential future use is the alkali metal thermal-to-electric conversion (AMTEC). This device has no moving parts (if we may ignore the sodium working fluid being circulated by electromagnetic pumps) but offers potential conversion efficiencies approaching those of dynamic systems. In the AMTEC concept, sodium heated to the point of ionization by the primary energy source is applied to one side of a ceramic membrane that conducts sodium ions but not electrons. Thus, the positive sodium ions pass through, but electrons tend to accumulate. A conductive film on the membrane collects the electrons, which are then conducted through a load to the downstream side of the membrane to neutralize the sodium ions. A number of problems must be solved, including membrane life, sodium condensation management in $0g$, and

other materials concerns, before this intriguing concept can be considered for operational use.

10.13.4 Solar Dynamic Systems

Solar dynamic systems, as the name implies, feature machines such as Brayton, Rankine, or Stirling cycle engines driving an electrical generator or alternator and using the sun as the primary energy source. These units offer potential conversion efficiency five to seven times that of solar photovoltaic arrays, which becomes very attractive at high power levels, e.g., above 100 kW. At such levels, photovoltaic arrays are expensive and pose attitude control and atmospheric drag problems due to their large size. The reduction in area of collectors for the dynamic system greatly reduces drag and stability concerns and becomes cost-competitive as well at high power. However, this approach does carry the usual dynamic system problems of reduced reliability, possible vibration, and possible attitude control system interactions. Also, the size advantage may be partially offset by the requirement for waste heat radiators associated with these conversion concepts.

10.13.5 Radiators

We have referred on several occasions to the need for radiator surfaces to dispose of waste heat. As systems become larger, the significance of the radiator increases until, for very large systems, it may be the largest single item. Present radiator concepts utilize large thin skins, usually made of metal. The heat to be dissipated may be delivered by conduction, by a pumped fluid loop, or by an array of heat pipes. Conventional radiators are limited by such factors as the allowable material temperature, achievable surface-to-mass ratio, surface emissivity, and thermal conductivity. A variety of innovative concepts have been proposed to provide higher capability radiators, including droplet radiators, membrane radiators, and rotating band radiators.

The droplet radiator offers very high performance because of the large surface-to-volume ratio of the droplets and the possibility of allowing a liquid-to-solid phase change, thus greatly increasing the energy removal. However, several practical problems must be solved before this concept can be implemented, including droplet generation and collection (especially while maneuvering) and materials selection.

The membrane radiator achieves high efficiency by allowing a fluid to flow down the inside of a contoured rotating membrane. The resulting convective heat transfer, itself a high-efficiency heat transfer mechanism (see Chapter 9), may be enhanced by a gas-to-liquid phase change. Small punctures in the membrane can be tolerated, because surface tension in the fluid will prevent leakage. As always, the requirement to rotate may be a problem. Material selection and the

development of credible launch configurations and deployment scenarios may also present difficulties.

The rotating band radiator is simply a broad, thin continuous loop of high-temperature metal moving between heated rollers in the spacecraft, from which it is extruded out into space to reject heat, then back into the spacecraft through other rollers. Effective transfer of heat to the band is crucial to this concept. A similar approach using a rotating disk has also been suggested.

References

[1]Hacker, B. C., and Grimwood, J. M., *On the Shoulders of Titans*, NASA SP-4303, 1977.

[2]*Electrical Grounding Architecture for Unmanned Spacecraft*, NASA HDBK-4001, Feb. 1998.

[3]*Scientific American*, "Photovoltaic Finesse," Sept. 2003, p.33.

[4]Agrawal, B. N., *Design of Geosynchronous Spacecraft*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[5]Wertz, J. R., and Larson, W. (eds.), *Space Mission Analysis and Design*, 3rd ed., Microcosm Press, Torrance, CA, and Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[6]Hyder, A. K., Wiley, R. L., Halpert, J., Flood, D. J., and Sabripour, S., *Spacecraft Power Technologies*, Imperial College Press, London, 2000.

Problems

10.1 Size a spacecraft power system consisting of a solar array and Ni-Cd batteries to supply 7.5 kW of prime power using 12% efficient 2 × 4 cm silicon cells with sun-tracking flat panels. The orbit is 500-km circular at 28.5 deg, i.e., approximately that of the Hubble Space Telescope. Assume a five-year life, with a minimum of 28 V required during eclipse. For the battery, assume an average discharge voltage of 1.1 V/cell and a minimum allowed discharge voltage of 1.0 V/cell. Use good design practice to determine:

(a) the number of solar cells in series and parallel.
(b) the size of the array.
(c) the number of battery cells in series and parallel.

10.2 Assume the same situation as problem 10.1, but with 18% efficient 2 × 4 cm Ga-As solar cells and Ni-H_2 as the battery type. For this case, we will assume a 25% allowed DOD, a minimum allowed discharge

voltage of 1.1 V/cell, and an average discharge voltage of 1.3 V/cell. Determine, again using good design practice where parameters are not specified, the following:
(a) the number of solar cells in series and parallel.
(b) the size of the array.
(c) the number of battery cells in series and parallel.

10.3 A geostationary orbital spacecraft requires 10 kW of power for a nominal 10-year lifetime. The bus voltage is to be 42 V in sunlight at EOL. The solar cells to be used are 2 × 6 cm in size, and at 298 K have maximum-power operating characteristics of $V_{mp} = 0.45$ V and $I_{mp} = 0.40$ A. The radiation degradation factors over 10 years for V_{mp} and I_{mp} are 0.95 and 0.97, respectively. Specified solar panel temperature design points are 273, 285, and 340 K, respectively, for summer solstice, autumnal equinox, and post-eclipse. The temperature coefficients for the solar cell at end-of-life are $\gamma_I = 0.25$ mA/K and $\gamma_V = -2.2$ mV/K. Sun-tracking flat panels with a 90% packing factor are assumed.
(a) How many cells are required in series?
(b) How many cells are required in parallel?
(c) What is the required total solar panel area?
(d) What is the end-of-life post-eclipse power output?

10.4 An RTG power system is being designed for a 20-year mission to Pluto. EOL power required is 100 W. ^{238}Pu has been selected for the isotope, and initial thermoelectric conversion efficiency is 7%, degrading to 5% at EOL due to radiation damage to the thermoelectric elements.
(a) What is the approximate mass of the RTG?
(b) What is the required isotope mass if a 30% efficient dynamic energy conversion mechanism is used, assuming this unit does not degrade with radiation exposure?

11
Telecommunications

11.1 Introduction

Telecommunications in space differs from the earthbound version in two major respects: 1) its long range, which may be anything from a few hundred to several billion kilometers, and 2) the potentially large relative velocity between transmitter and receiver, so that Doppler shift becomes significant ($\pm$ 50 kHz in the *S*-band for low Earth orbit), requiring complex frequency-tracking loops in the receiver. Also, spacecraft in low orbit see very limited communications coverage from any single surface station. A station that can track to within 5° of the horizon will view a spacecraft in a 300-km orbit for only 6.5 min, even for a zenith pass. At the opposite extreme, distant spacecraft move very slowly against the background of the fixed stars, thus, the pass time is essentially governed by the rotation of the Earth. Signals from distant spacecraft, because they are very weak, require tracking by large, specialized equipment, such as NASAs Deep Space Network (DSN).

These factors complicate spacecraft design because of the mismatch between the rates of data acquisition and return. In low Earth orbit (LEO) a spacecraft may collect data throughout the orbit period of perhaps 95 min. Given only one downlink station, the spacecraft can dump data only a few times per day. Clearly, the downlink data rate must be many times that of the acquisition rate even with onboard processing and compression of the data. Power limitations and range restrict the rate at which data can be returned from a spacecraft at another planet. Data may be acquired very rapidly during an encounter and then played back at a relatively low rate over a long period.

Moreover, passage through the Earth's troposphere and ionosphere complicates signal propagation, as a result of energy absorption, rotation of polarized signals, etc. We will examine these effects in more detail later.

Spacecraft telecommunications hardware has power, mass, and volume limitations more extreme than in other applications, even aircraft avionics. Meeting these challenges, in fact, was the original spur that has led to the technology of low-power, low-mass electronics seen in today's consumer electronics market. As discussed in Chapter 3, spacecraft electronics experiences a variety of environmental stresses, such as mechanical shock and the acoustics and vibration of launch and atmospheric flight. Spacecraft are exposed to radiation that can damage electronics over a period of time. Extremes of thermal

environment normally do not unduly affect the electronics of the telecommunications system, which is usually located in the temperature-controlled interior of the spacecraft, but external equipment such as antennas may be strongly driven by thermal design considerations.

Because of its role in accepting ground commands and returning data, the telecommunications system interfaces directly or indirectly with virtually every spacecraft subsystem and experiment. The earthbound end of a link interfaces with tracking stations, and through them with operating agencies around the world.

11.2 Command Subsystem

The command subsystem allows instructions and data to be sent to the spacecraft. In some cases the command will be acted upon immediately; in others it may be stored to be acted upon when a particular clock time is reached, some event is sensed, or a particular spacecraft state is attained.

Conceptually, the two basic command types may be characterized as relay commands and data commands. The former are functionally equivalent to switch closures and may provide a simple on/off function or initiate a complex, stepwise operational sequence. Such commands may provide a pulse signal or may latch in a new state until a further command is received (in the switch analogy, a momentary contact vs a toggle).

Data commands, as the name implies, provide information upon which the spacecraft acts, such as the direction and magnitude of a translation maneuver. Later in this chapter we will discuss how such commands are structured.

A complex operation such as a thruster firing to cause a midcourse correction might involve the transmission of a substantial number of data commands involving directions and magnitude of attitude maneuvers, rocket motor burn time, or required change in velocity, and maneuvers back to cruise attitude. In addition, a number of relay commands might be required to configure the spacecraft for the maneuver (e.g., science instruments off, telecommunications from high gain to omnidirectional antenna, etc.). A final relay command to enable the sequence of actions would probably be required as well.

This example of midcourse correction illustrates the need for the two types of commands and also the need for delayed commands. The maneuver may need to take place out of sight of ground stations, or the timing may be so critical that it is not acceptable to depend on ground commands, where the communications uplink might be lost at a critical time.

The length and structure of command messages and individual words will depend on the amount of information to be sent and the capability of the equipment. In addition to the actual information, there will be address, identification, and other formatting data bits that must be transmitted and can substantially increase the overall data rate.

The component choices for the command subsystem are much the same as those for Earth applications: bipolar transistors, n-type metal-oxide semiconductors (NMOS), and complementary metal-oxide semiconductors (CMOS). Bipolar transistors are slowest and use the most power but are usually the most resistant to radiation. The higher speed and lower power consumption of the conventional metal-oxide semiconductors come at the cost of much greater sensitivity to radiation, unless special radiation-hardening measures are taken in their design. Although such issues are well understood today, it remains true that the radiation-hardening requirement for spaceborne electronic systems represents the single greatest departure from designs suitable for ground-based applications, and almost by itself accounts for the "generation gap" between the sophistication of state-of-the-art consumer electronics and that which is intended for space applications.

Radiation causes long-term degradation of components and eventual loss of function. This can occur through a variety of mechanisms, depending on the type of radiation. As a more immediate problem, energetic charged particles passing through the junctions of the components can cause "soft," or temporary, errors. Deposition of sufficient energy in a junction can cause it to "flip," leaving, for example, a logical 1 where a 0 had been. If this particular junction contains a bit that is part of a data command, erroneous data now reside in that register.

In the more common soft-error case, the error can be corrected by reloading the command. The damage is not permanent. In the case of CMOS circuitry, the energy deposition can destroy the junction, in what is called a latchup condition. Modern CMOS circuits normally have latchup protection for space applications but availability may be limited.

The smaller the junction—and small size is the means by which high speed and low power consumption are achieved—the lower is the energy required to cause the phenomena just discussed. Thus, the improvements in electronic component technology that have allowed us to design more capability into given power and volume constraints have simultaneously increased the susceptibility to radiation damage. Because the problem primarily concerns space operations, most research into radiation-hardened electronics has been done by NASA and the Departments of Defense and Energy. Production of such components is limited, which makes them expensive. One obvious solution to the radiation problem is shielding. Unfortunately, this is often of only limited practicality, as discussed in Chapter 3.

11.3 Hardware Redundancy

The use of appropriate functional redundancy is an important factor in achieving the level of system reliability required to achieve the desired design lifetime. A common and straightforward approach simply uses two completely separate parallel systems. Although this probably (but not certainly) improves

reliability, a single failure in each string will still cause loss of the command function. A more sophisticated approach employs redundancy at the subsystem level, with cross-strapping such that a given subassembly can be used in either string.

In this arrangement one or more failures can occur in each string, but as long as there are no duplicate failures in each string (i.e., at least one of each type of subassembly is working), a working command system can be assembled by selective cross-strapping between the strings.

Control and management of redundancy is a complex issue, which, when improperly done, can lead to serious pitfalls. The redundancy scheme must be examined with care to avoid inadvertent and irreversible switchovers, possible untested modes, etc. Also, care must be taken so that the system is truly redundant; for example, two fully duplicated strings operating off a single fused power cable are not redundant.

Although an exhaustive discussion of reliability and redundancy management is beyond the scope of this text, these topics are addressed in somewhat more detail in Chapter 12.

11.4 Autonomy

With the development of ever increasing computer capability and the undertaking of more complex missions at more distant targets, spacecraft have become more autonomous. Moore's Law—the empirical rule enunciated by Intel's Gordon Moore and now well into several decades of apparent applicability—implies that computational throughput doubles every 24 months. The implications for spacecraft software systems have been as profound as those for the conventional consumer electronic systems with which every reader will be familiar.

Thus, for many years the trend in spacecraft management has been away from very detailed command sequences, exhaustively vetted on the ground, and toward the use of high-level commands. As an example, a spacecraft might simply be commanded to apply a specified ΔV in a given direction. It would then autonomously compute and execute the required attitude maneuvers and the rocket motor burn. Still more advanced (and not yet practical) spacecraft would perform navigation onboard and autonomously decide that a course correction is required and perform it.

The advantages of higher levels of autonomy are obvious. The size of the ground operations crew and support team is reduced—a major savings because for long missions the cost of flight operations can easily exceed that of development and launch. Reliability can be enhanced because success is no longer as dependent on the link to Earth. Indeed, in some cases autonomous systems can prevent damage that would have occurred on distant spacecraft even

before the telltale telemetry indicative of any concern could have arrived at Earth!

However, an autonomous spacecraft for a given mission will always be more complex than a ground-controlled machine. (This overlooks very simple spacecraft such as the early Explorers and Pioneers, which had no uplink, used a single operating state, and were always on, collecting data and sending it to the Earth whether or not anyone was listening.) The increased complexity implies a greater variety of spacecraft states and operating modes to be considered in system design and testing. Because it is probably impossible to test every conceivable mode, a great deal of consideration must go into the design of a system free of traps and testable with reasonable time and effort.

Today's reality is that all spacecraft operate in a largely autonomous fashion, with periodic monitoring by mission control personnel to an extent that depends on the characteristics of the mission. Economic realities mandate this practice for near-Earth spacecraft, and round-trip communications delays impose it on planetary spacecraft. Modern computers and the software tools that they host are directly responsible for the stunning capabilities evident today in all classes of spacecraft. However, in common with other systems incorporating complex software control schemes, it is equally true that many embarrassing failures have resulted directly from the unanticipated behavior of such systems, in concert with the difficulty of comprehensively testing all possible vehicle states.

The ill-fated Mars Polar Lander (MPL) mission provides an object lesson in this regard. Intended for a 1999 landing near the south pole of Mars, the spacecraft essentially vanished following separation from its cruise stage in preparation for the landing sequence. The report[1] of the independent failure review board provides interesting reading for those hoping to draw lessons in system engineering from this event. Worthy of note is that, while the review board found several possible causes of failure, the most probable cause was found to lie within the software design of the necessarily automated landing sequence.

The MPL landing legs were equipped with sensors to detect the transient shock of the landing event, so that when surface contact occurred, the descent engine could be shut down. However, deployment of the landing gear from its stowed position would, by itself, generate essentially the same transient shocks as the landing. There was no specific software system requirement to clear the memory buffers recording this event after the landing gear deployment, and so the landing sequence logic reacted as it was designed to do, shutting down the descent engine because the presence of a "surface contact" event had been (erroneously) detected. Because this occurred at altitude rather than on the surface, the spacecraft crashed. It was noted by the review board that, while other problems *could* possibly have caused spacecraft failure prior to this point in the landing sequence, this problem *would* certainly have resulted in destruction of the spacecraft, had it reached this point in the mission.

Compounding this inherent design problem was the fact that the landing sensors were incorrectly wired during initial system testing, preventing the logic

trap from being detected. The system test was not rerun after the wiring error was corrected, because the otherwise successful outcome of the test was not viewed as being dependent on the relatively trivial sensor wiring error! Finally, the work of the failure review team was necessarily speculative, because of a design decision to curtail communications from the lander after separation from the cruise stage. Further communication was to ensue following a successful landing. Clearly, there are many lessons to be drawn from this mission by attentive space system engineers.

11.5 Command Subsystem Elements

Figure 11.1 presents a basic functional block diagram of a typical spacecraft command subsystem. The major elements that make up the subsystem are defined in the diagram and are discussed in the following paragraphs. It is worth noting that, while the functional blocks depicted in Fig. 11.1 have changed little over the years, implementation methods have changed greatly. Where once each block was a separate hardware element, today many of the functional elements are merely different subroutines in a system consisting largely of software.

11.5.1 Antennas

For LEO missions the uplink, or command antenna, will usually be omnidirectional to facilitate communication from ground stations while the aspect angle is changing during the pass. Deep space missions, in contrast, require directional high-gain antennas and thus attitude and articulation control. Except for use near the Earth, such missions also carry omnidirectional antennas to aid operations and to avoid overdriving the receiver with excessive gain at short range. Capability to send uplink commands through a low-gain omnidirectional

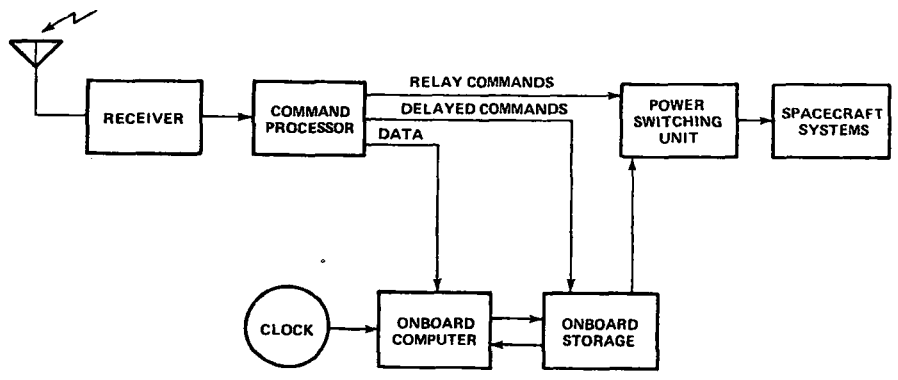


Fig. 11.1 Spacecraft command system block diagram.

antenna should be retained for emergency use even at long range. Anomalous behavior resulting in loss of high-gain antenna pointing may cause loss of the mission if no other way exists to get commands into the spacecraft. The capability to broadcast commands from the Earth at very high power to the omnidirectional antenna might save the spacecraft.

11.5.2 Receivers

The receiver may be either two types: tuned radio frequency (TRF) or, more commonly, superheterodyne. Details of the two types go beyond the scope of this text. Briefly, the TRF is a radio-frequency (RF) amplifier, tuned for a narrow bandwidth around the transmitted frequency, followed by a detector or demodulator stage and several stages of low-frequency amplification. In a superheterodyne receiver the received signal is shifted in frequency (heterodyned or mixed) to a frequency lower than the transmitted one. Two or more shifts are common, and the resulting signal is amplified and filtered at each stage to provide greater sensitivity to weak signals and better selectivity for rejection of unwanted signals.

11.5.3 Modulation

The receiver and uplink may use amplitude modulation (AM), phase modulation (PM), or frequency modulation (FM). The choice depends on several factors, including required signal-to-noise ratio (SNR), desire for graceful degradation, available RF bandwidth, required data rate, hardware complexity, and compatibility with existing ground tracking systems.

FM and PM systems can operate with lower RF signal-to-noise ratios than AM systems because FM and PM provide improved performance at the cost of greater bandwidth. FM and PM systems suffer the penalty of a threshold effect. As the SNR on the RF link decreases, the performance of the link degrades very slowly until the threshold SNR is reached. As the SNR progresses below the threshold, the performance of the link drops precipitously. AM systems do not exhibit this behavior. They require more SNR to achieve a given performance but degrade gracefully as the SNR is reduced. (The reader can experience this with an automobile radio when driving away from a station. FM will remain reasonably clear up to some distance from the station, when the signal abruptly deteriorates and is lost. AM becomes progressively weaker, probably with distortion increasing, but will remain audible through the noise for a long period.)

A high degree of frequency selectivity is essential in spacecraft receivers to enhance SNR and to reduce sensitivity to electromagnetic interference (EMI). During ground testing and launch, spacecraft operate in a very signal-rich environment and even in orbit will be illuminated by unwanted signals. Frequency selectivity is essential in such situations.

In addition to amplifying the signal and filtering out noise and EMI, the receiver demodulates the signal and provides the information-bearing portion of the received signal to the command decoder and processor.

It is a cardinal rule that the command receiver is always on. If a command exists that can turn off the receiver, and such a com-mand is inadvertently sent (erroneous commands do occur), there would be no way to undo the damage. Of course, mechanisms could be devised to turn it back on after a time, or some other recovery approach could be employed, but the straightforward and therefore preferred approach simply has the receiver permanently on.

The command decoder (not shown separately in Fig. 11.1) may be viewed as the first stage in command processing. The decoder first inspects the identifier bits that make up a part of each command word. These bits identify the word as a command and are used to synchronize the bit stream along command word boundaries. They also contain the address of the intended spacecraft. The decoder verifies that the word is a command and is intended for this spacecraft. In the case of an encrypted command string, appropriate decryption algorithms must be applied in the process. The decoder then passes the bit string to the command processor.

11.5.4 Command Processor

In early spacecraft the command processor was a simple hardwired (and therefore inflexible) circuit. The command processor in a modern spacecraft is usually just another functional block of code in a multipurpose processor. The processor interprets the command for proper destination and required action. As a precaution against erroneous action, commands are checked for validity using parity bits or more sophisticated error detection and correction (EDAC) schemes. The processor then sends the signal to the appropriate destination: elsewhere in the computer, an onboard storage memory unit, or directly to a power-switching unit.

The central processor executes the more complex operations involving data commands and delayed commands. In early spacecraft the "computer" was little more than a programmable sequencer. With increasing sophistication, spacecraft computers have become quite capable data processors. Today's microprocessor technology allows a distributed architecture in which most of the computational power resides in the subsystems, and the central computer functions primarily as a coordinator. Unless an extremely large amount of data must be stored, solid-state memory usually proves sufficient.

The power-switching elements are the interface circuits between the command subsystem and the remaining spacecraft subsystems. These elements may consist of pulsed or latching relays or solid-state switches. More complex operations may require sequences of steps to achieve the desired goal. The switching elements

may be operated immediately by relay command or by the computer, based on sequences previously loaded.

11.5.5 Telemetry Subsystem

The telemetry subsystem takes engineering or scientific data and prepares them for transmission to the ground. Figure 11.2 presents the functional block diagram of a typical spacecraft telemetry subsystem.

The data are generated by sensors or transducers responding to events in the "outside world." In this context the outside world may be other subsystems in the spacecraft that generate data concerning their status and condition, or events in the surrounding environment as sensed by the science subsystem. Parameters that are measured often include acceleration, angular rate, angular position, pressure, temperature, density, resistance, voltage, current, intensity, electric field, magnetic field, and radiant energy.

It may be said that the sensors and transducers are not truly part of the telemetry system, but rather of the subsystem in which they reside. This argument carries most weight in regard to the science instruments. In any case, these elements provide the signals upon which the telemetry system operates.

Signals from the sensors or transducers are rarely in a form immediately suitable to the data-formatting element. Generally, the signal must be "conditioned" and converted from analog to digital. Signal conditioning converts data to a form that is acceptable to the telemetry system. Weak signals may be amplified or excessively strong signals attenuated. High- or low-pass filters are used to remove bias or noise; notch filters are used to remove high-intensity signals in a specific frequency band. A signal's dynamic range may be compressed, perhaps by the use of a logarithmic amplifier. Every effort will be made to achieve isolation between signals to ensure accuracy of the data.

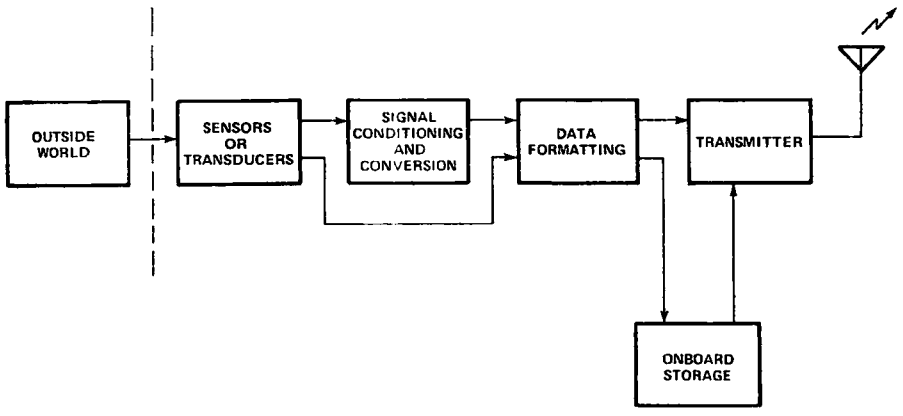


Fig. 11.2 Spacecraft telemetry system block diagram.

Analog-to-digital conversion (ADC), one of the major functions of signal conditioning, will usually be required because, generally, the real world is analog; it varies more or less continuously across the range of the phenomenon being observed. Spacecraft data, on the other hand, can be handled more efficiently in digital form, as a series of discrete steps.

Figure 11.3 shows the process schematically. The smooth curve represents the phenomenon being measured. The digital approximation is represented by the "stair steps," where the quantized level is proportional to the average value during the sampling period or the value at the sampling instant. The sampled value is restricted to one of a finite number of allowable values so that the digitized data can be represented by a finite number of bits. The analog input could be a voltage curve, between prescribed limits, which a transducer produces in response to a phenomenon. The curve may be directly proportional to the phenomenon or may have been modified, perhaps by a log amplifier, to make the range compatible with the telemetry system. The output of the digitizer is a binary word for each digitized data point. The data are quantized into one of 2^n levels, where n is the number of bits in the data word. Several types of analog-to-digital converters are available to do this. The flash ADC can operate at rates in the tens of megahertz range, but it is noisy. The successive-approximation register can handle data rates of hundreds of kilohertz. The integrating ADC can handle only tens of hertz, but produces very clean data. The details of the logic circuitry associated with different types of ADCs are beyond the scope of this text.

Several errors are inherent in the ADC process and the subsequent reconversion to analog that most data undergo on the ground. As with any telemetry system, the measured signal will be embedded in noise. A noisy analog input may fool the ADC into digitizing at a level higher than the actual level of the signal. An experienced human analyst looking at the noisy curve in raw analog form may be able to reject the noise and infer the true data. In the digitized and reconstructed data the analyst will not see the original data but rather a quantized and reconstructed version of the signal. As a result, it may be harder to estimate the correct data, or even to recognize that the data are noisy.

The quantization process can represent only a finite number of levels, and data at any point may fall between the levels. If eight bits are used to transmit each quantized level, then 256 levels can be represented; thus, the potential error is less

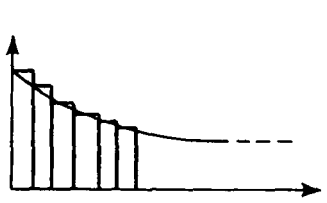


Fig. 11.3 Analog-to-digital conversion.

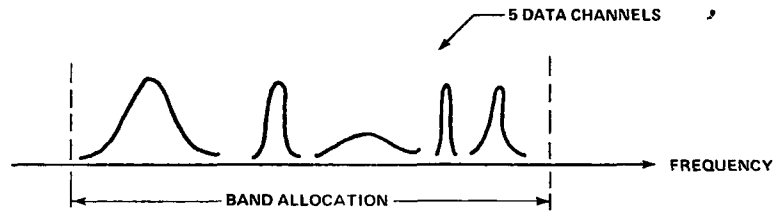


Fig. 11.4 Frequency-division multiplexing.

than 0.4%. This may or may not be sufficient; additional bits, requiring a larger ADC, may be required.

Handling more than one data type requires some form of multiplexing by the telecommunication system. This usually takes one of three forms.

Frequency-division multiplexing (FDM) subdivides the frequency bandwidth of the telemetry downlinks (Fig. 11.4) and allocates the various data streams to separate portions of the available bandwidth. In the temporal sense the data may be viewed as going out in parallel. The subdivision of the bandwidth is not necessarily equal; higher rate data streams must be allocated wider bandwidths. This approach is common when one or more channels of high rate data (e.g., video) are to be returned. It is possible to apply the other schemes discussed later to the individual channels within the FDM structure, and these channels may be analog or digital.

Time-division multiplexing (TDM) employs temporal separation to assign different sets of bits within a data frame to different users, as depicted in Fig. 11.5. The frame repeats continuously, with each user occupying the assigned bits in a cyclic fashion. Note that the cycle may be subcommutated, with more than one user sharing a particular set of bits in a subpattern within the overall sequence. By convention, frames are limited to 2048 bits with word lengths of 6–64 bits. Within these limits virtually any level of subframe definition is possible.

The final type of multiplexing, code-division multiplexing (CDM), sends the data in parallel over the same bandwidth during the same time period, but encoded using spread-spectrum techniques so that the individual streams can be separated at the receiver.[2]

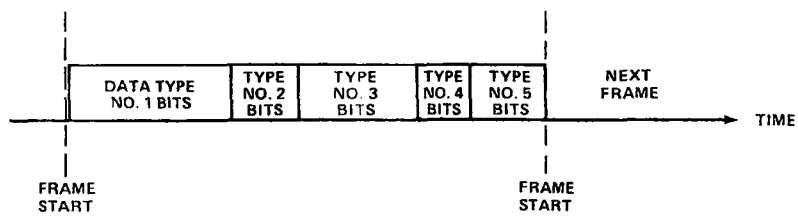


Fig. 11.5 Time-division multiplexing.

The problem of "aliasing" is a major concern in any digital communication. If a band-limited, sampled signal is to be reconstructed accurately, it must be sampled at a rate at least twice the maximum frequency contained in the signal. This minimum sampling rate is called the Nyquist rate. A simple example will illustrate the problem. Assume that a sinusoidally varying signal at 1000 Hz is being sampled at 1000 Hz (half the required Nyquist rate). In this example the digitizer would always sample the same point on the waveform, and the data appear constant—a straight line with no indication of the true signal.

As stated, the Nyquist criterion is

$$f_s \geq 2f_{\max}$$

The factor of two is, in practice, too low to achieve accurate representation of the sampled data. A more realistic sampling rate will be five or more times the maximum frequency.

It may not be necessary to use the full bandwidth of a measured signal to obtain useful mission information. An example is the use of the Earth's magnetic field for attitude control. A 20-Hz variation in the field may be of substantial scientific interest but is of no interest for attitude control, where 1-Hz resolution would normally suffice.

Sampling rates based on the Nyquist criterion are shown in Table 11.1. As noted earlier, real sample rates must be substantially higher, typically by a factor of 5–10, to represent the signal adequately. The resulting very high data rates are then usually reduced by a variety of coding and data-compression techniques. These techniques conserve bandwidth, but at the expense of greater complexity and cost of implementation.

With TDM the bit stream or data stream comprises sequentially sampled data types combined in a specified frame pattern. This process of sequential data sampling is referred to as "commutation." Figure 11.6 schematically represents a major data frame from a continuing stream of bits. The major frame is repeated continuously in the telemetry stream and is subdivided into three minor frames. In this example each minor frame contains seven data elements.

Table 11.1 Sample rates based on Nyquist criterion[a]

Signal type	Analog bandwidth, kHz	Typical resolution, bits	Nyquist digital rate, kbps
Voice	4	7	56
Music	20	10	400
BW TV	4600	4	36,800
Color TV	4600	10	92,000

[a]Binary keying is assumed. More complex coding schemes can be used to achieve lower bit rates.

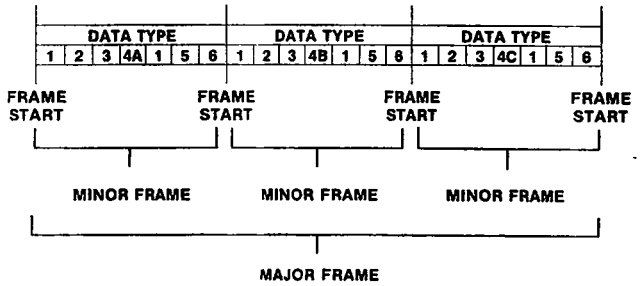


Fig. 11.6 Subcommutation and supercommutation.

In the (somewhat improbable) event that the commutation rate perfectly matches the desired sample rate for all data elements, each data element would appear once per subframe, and each subframe would be identical. In fact, however, there are always parameters that vary so slowly that sampling at a lower rate is acceptable. The same data-element position can thus carry many different low-rate parameters in each subframe, each sampled at less than the basic frame rate. This subcommutation is represented by data elements 4A, 4B, and 4C in Fig. 11.6. The majority of data elements, represented by 2, 3, 5, and 6 in the figure, will be sampled at the basic rate. Note, however, that element 1 appears twice in each minor frame, and is thus sampled at twice the basic rate. This is called supercommutation.

The engineering telemetry requirements of a spacecraft propulsion unit can illustrate this practice. The temperature of the propellant line, tank, and pressurant bottle will change slowly and might be subcommutated as element 4. Tank and bottle pressures will change more rapidly and would be sampled at the basic rate. Thrust chamber pressure—the most dynamic and crucial to performance—might be supercommutated as element 1, or at an even faster rate.

For proper telemetry interpretation some additional information must be included along with the measurement data. The major frames and often the minor frames will include additional bits for frame synchronization and identification. Extra bits will often be included for error detection and correction. The minimum would be a single bit for parity check. Incorrect parity would indicate that one bit, or some odd number of bits, had been erroneously received, but there would be no indication of where the error(s) had occurred. Note also that an error affecting an even number of bits would go undetected. More complex schemes for error correction require additional bits, and greater encoding and decoding complexity.

To allow the decoding system to synchronize with the downlinked data, frame synchronization bits are sent at the beginning of major or minor frames. The number of bits, by convention, will be 33 or less, with 24 being common for a major frame. Sending synchronization bits with each frame ensures proper synchronization, regardless of phase or frequency errors since the previous frame. Without this concern synchronization bits would be needed only at the beginning of the stream.

A final type of information required is a time tag. In most cases, for the data to be useful, it must be accurately annotated (commonly within 1 ms) as to the time of acquisition, and the time data must be included in the data frame. The time base is usually supplied by a stable crystal oscillator, often the same oscillator that comprises the clock in the onboard computer.

Telemetry formats vary from system to system, and the ratio of "overhead" bits to actual data will vary to some degree. In all cases the overhead is substantial, and the required telemetry bit rate will always be higher than the information bit rate. For many reasons, including power and antenna size, it is desirable to reduce the number of bits that must be transmitted. A variety of data-compression and data-encoding schemes have been devised and are in current use. This rather specialized field will not be covered here.

11.5.6 Onboard Processors

The onboard computer is an essential subsystem in all modern spacecraft. Categorical statements such as this are fraught with risk, but it is difficult to conceive of a modern space vehicle that does not use an onboard computer—often several computers—to control nearly all aspects of its behavior. It has not always been thus. Early onboard "computers" were little more than timers enhanced with some modest logic circuitry. However, spacecraft processing capacity has grown with the general maturation of computer and electronics technology until today most spacecraft are highly capable electronic devices.

Spacecraft computers have reflected the architectural trends in ground computers. Until at least the 1980s, a central mainframe architecture was the norm, and this approach is still sometimes used. However, the development of microprocessor technology and distributed networks has led to their adoption in spacecraft architecture, just as it has in ground applications. In this approach the central computer coordinates and directs traffic between equally powerful processors located in various subsystems. Redundancy and cross strapping can provide remarkable flexibility of operation and increase system reliability. Selection of a system architecture should be based on the type of operation anticipated, cost, and complexity.

As with any spacecraft subsystem, the onboard processor should require the minimum power, weight, and volume consistent with the required performance. The hardware must perform well under the usual environmental stresses, including temperature extremes, thermal cycling, hard vacuum, shock and vibration, and radiation. In addition, the hardware must be resistant to electromagnetic interference.

These special requirements on space hardware may, and usually do, cause a lag of several years in the technology being flown in spacecraft as compared to ground, or even aircraft, applications. In addition, special testing and flight qualification increase cost, as does the very limited production of space-qualified components and subsystems. Thus, the spacecraft designer gets a lot less

computing power at considerably higher cost that his ground-based counterpart, a situation that seems likely to continue indefinitely into the future. Some information regarding current space-qualified processors and future trends may help to put the foregoing discussion in perspective.

Clock speeds of up to 132 MHz are available, with a word length of 32 bits (although others do appear; the authors have seen 18-, 24-, and 64-bit processors in particular cases). Raw clock speed is a useful but not definitive measure of overall processor capability. The useful throughput of a computer is a complex function of architecture, word length, instruction set, and cycle time, and must usually be determined by benchmark runs of the system. The basic capability will be degraded 10% or more by error detection and correction codes (EDAC). With these qualifications, radiation-hardened processors available in the early 2000s can provide 200–300 million instructions per second (MIPS) for the typical mix of instructions encountered in spacecraft operations, with 500–700 MIPS capability expected in the near term. Random access memory (RAM) of 128–512 megabytes (MB) is available.

Processor module mass will be on the order of 0.3–0.5 kg at the board level, depending on the capacity of the machine, the technology in use, and other factors. Packaging requirements will vary with the application but will typically add several kilograms to the basic board-level mass requirement. Power requirements will be on the order of 5–10 W, depending on speed, memory size and type, and architecture.

Performance comparisons with even commercially available desktop personal computers are not impressive. Most readers will recognize that in this same early 2000s timeframe, clock speeds of several GHz are commonly available, with up to 1–2 GB of RAM. This performance gap is, as we have stated previously, primarily due to the more stringent environmental and packaging requirements placed on spaceborne electronic systems, and particularly those associated with radiation hardness requirements.

The issue of parts qualification has been discussed previously in this and other chapters. Space-qualified (class S) parts are expensive and typically have long delivery times. In fact, in some cases they may be impossible to obtain unless the project will pay for flight qualification and maintenance of a special production and test operation. This situation occurs because manufacturers show more interest in the commercial market, with production of millions of units, than in the much more limited aerospace market. One option is the use of class B parts, which are functionally equivalent to class S but lack the pedigree and the screening and testing that define class S. An in-house screening and burn-in program to provide the effective equivalent of class S parts may be a cost-effective answer. There are many subtleties in the parts selection and screening business, far more than we can treat here. Consultation with specialists in reliability, safety, and quality assurance (RS and QA) is to be recommended.

The requirement for radiation hardness presents a problem that cannot be solved through screening programs. Fundamental changes in parts design are

required, especially for spacecraft operating above low orbit and incorporating state-of-the-art electronics (e.g., 0.15 μm design rules as this is written). The soft-error rate experienced on the first Tracking and Data Relay Satellite (TDRS) was on the order of one per day, a value characteristic of unhardened processors of that (early 1980s) era. Extra shielding usually is not effective in reducing this rate because of the high energy of the offending particles. Special components and failure-tolerant architectures may be required.

Current spacecraft processor technology offers total dose hardening in the range of 0.25–1 Mrad, with a latchup-protected single-event upset rate of less than 10^{-10} errors/bit/day. Specially designed systems for military application can exceed 1 Mrad of total dose hardening and can provide a prompt dose upset tolerance on the order of 10^9 rad/s, with survival tolerance to about 10^{12} rad/s.

11.5.7 Onboard Storage

Mass storage is required when data cannot be sent over the downlink at the time they are taken or at the rate of acquisition. Storage is also required in the common case where significant amounts of reference data are required.

Ground-based computer systems have used a wide variety of storage media, including, especially in earlier decades, punch cards, paper tape, magnetic drums, magnetic tape, magnetic core, "floppy" disks, plated wire memory, and occasionally bubble memory. More recently, many of these media have been rendered essentially obsolete, with favored storage media today including floppy disks, "hard" disks, Zip™ disks, CD-ROM and CD-RW optical disks, solid state memory, and memory "sticks." Occasionally one still finds requirements for mass storage of serial data, for which magnetic tape may be appropriate, in part because tape storage can be very cheap.

Figures of merit for mass storage for ground applications today are little short of miraculous by standards applicable only a few years past. In the early 2000s, laptop and desktop computers are available with several gigabytes of solid-state random access memory (RAM), and hard disk capacity of a hundred GB or more. Solid-state memory costs substantially less than \$1/MB, and hard disk capacity retails for no more than \$1/GB. Portable Zip™ disks capable of holding 250 MB are available for a few dollars, and rewriteable compact disks (CD-RW) are even cheaper.

As noted earlier, this rosy picture—which is improving on time scales so short that no text could hope to present a current assessment of the state of the art—deteriorates rapidly when space applications are considered. Harsher environmental requirements, especially in regard to radiation hardness, the desire to avoid the use of moving parts wherever possible, and (except in manned spacecraft) the lack of human interaction, have severely restricted the types of mass storage used in space applications.

Core and plated-wire memory saw early use in manned systems, and tape recorders served faithfully for many years in both manned and unmanned

spacecraft. Magnetic bubble memory has seen application in certain high-radiation, primarily military, applications. However, as this is written, the overwhelming choice for mass memory storage on spacecraft is solid-state memory, which has evolved at a pace far outstripping other technologies. The early-2000s market offers radiation-hardened, space-qualified mass storage capability in standard packages featuring up to 2 terabits (Tb) capacity, with I/O on as many as 10 independent channels at a Gbps on each channel. More capacity can be obtained, albeit on a custom-design basis.

Conventional hard disks, familiar to any desktop personal computer user, are seeing increased utility. Currently available systems offer packaged solutions featuring, for example, up to eight "ganged" hard disks with I/O capability of about 0.5 GBps and over 500 GB total capacity. Hard disk systems, as with the tape recorders so common in earlier years, carry the liability of moving parts, always a matter of concern when extended lifetime is required.

Optical disks would seem to have tremendous potential for relatively low-cost onboard storage applications in which extreme radiation hardness is important. CD-ROMs alone can be used to make very large amounts of reference information available to the onboard computer, allowing much greater autonomy in spacecraft operations, while CD-RW devices add a reprogramming capability to the mix of design choices. However, as of this writing, the authors are unaware of any space applications of compact disk technology.

11.5.8 Modulation Methods

The modulation scheme is the method by which command and telemetry systems encode a "baseband" information-bearing signal upon an RF carrier. The carrier, $S(t)$, is an RF signal characterized as

$$S(t) = A(t)\cos[\omega(t)t + \phi(t)] \tag{11.1}$$

where A is the amplitude, ω the frequency, and ϕ the phase angle of the signal. All of these quantities can vary as a function of time. Three modulation schemes are commonly used. AM varies the amplitude of the RF carrier wave signal according to the baseband signal. FM varies the ω of the carrier according to the baseband signal. Both of these schemes are in common use in terrestrial radio systems. The third scheme, PM, varies the ϕ with the baseband signal.

All three modulation schemes offer particular advantages, some touched on briefly in the section on command. The selection of the best method will depend on the application, available bandwidth, required signal-to-noise ratio, and the capability of the ground stations. However, it is true that few modern communications systems employ amplitude modulation.

Both analog and digital input signals may be used with any of the three methods. Discrete amplitude changes, frequency shifts, or phase shifts can represent the 1s and 0s of digital data. Similarly, continuous changes in amplitude,

frequency, or phase can represent an analog signal. Modern systems use digital modulation almost exclusively. However, there may be cases in which analog modulation offers advantages (e.g., with high bandwidth video data).

Pulse-code modulation (PCM) is a technique for converting analog signals into digital form rather than a fundamental modulation method such as AM, FM, or PM. PCM samples the signal and quantizes it into one of 2^n levels at a sample rate appropriate to the application. Because the allowed 2^n levels are finite, the unique level of each sample can be represented by a digital word n bits long. These data words are then formed into a serial bit stream arranged in minor frames that in turn make up major frames, as discussed earlier. Figure 11.7 diagrams the process of converting from the original analog waveform to the final digital bit stream.

This digital bit stream varies, or "keys," the carrier signal by one of the methods discussed earlier. If each bit is encoded onto the carrier independent of the other bits in the bit stream, the digital modulation is called binary. In binary amplitude-shift keying (BASK), the amplitude of the basic signal is varied to represent a 1 or a 0. (Morse code is a simple example of BASK.) In binary frequency-shift keying (BFSK), two frequencies, close to but distinct from the carrier, represent the 1 and 0. This may be accomplished, for example, by switching between the outputs of two crystal oscillators, one just above and the other just below the basic frequency. Binary phase-shift keying (BPSK) transmits a sinusoidal carrier with one of two allowable phases that represent a 1 or a 0. Of these three techniques BPSK and BFSK are the most common in space communications.

If often seems to the casual observer that there are as many digital coding schemes as there are practitioners of the art. Figure 11.8 shows eight common approaches. The system engineer, unless specifically involved in coding

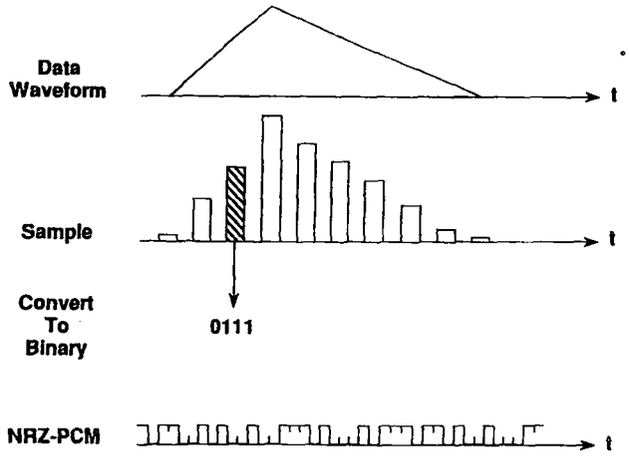


Fig. 11.7 Pulse-code modulation.

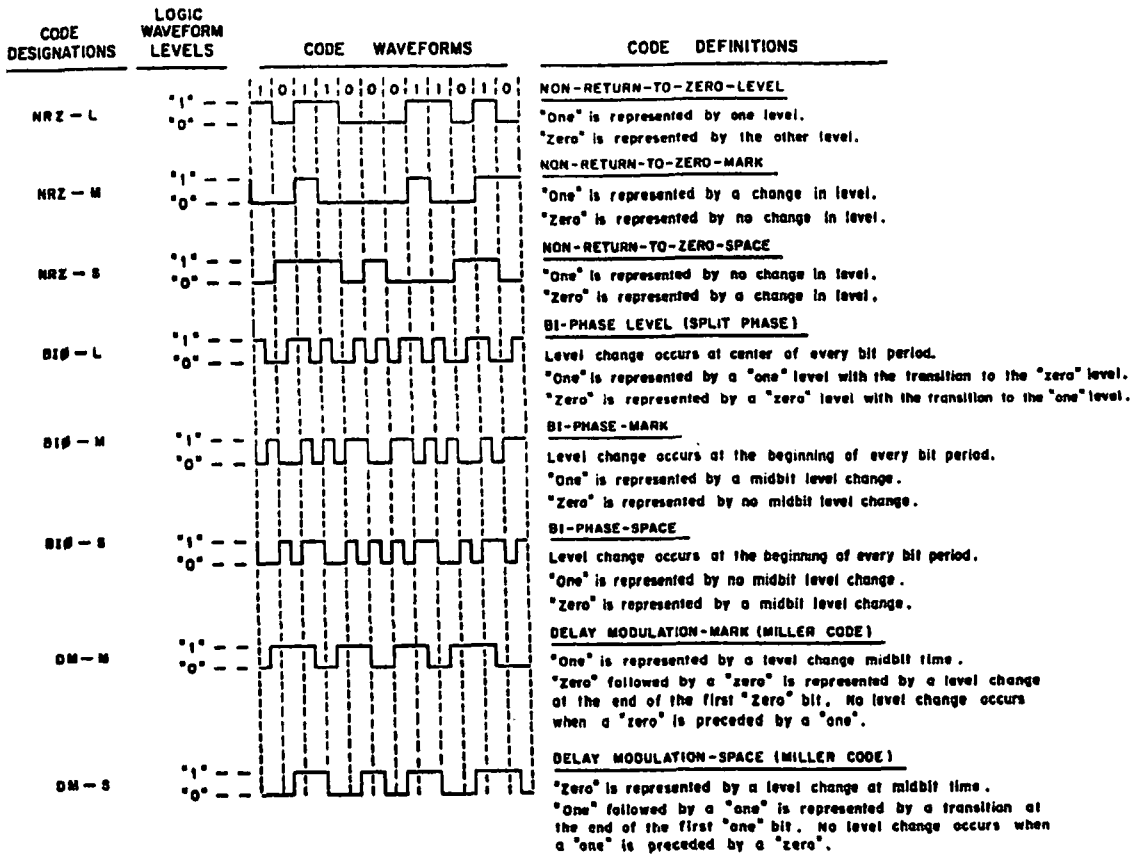


Fig. 11.8 Digital encoding schemes.

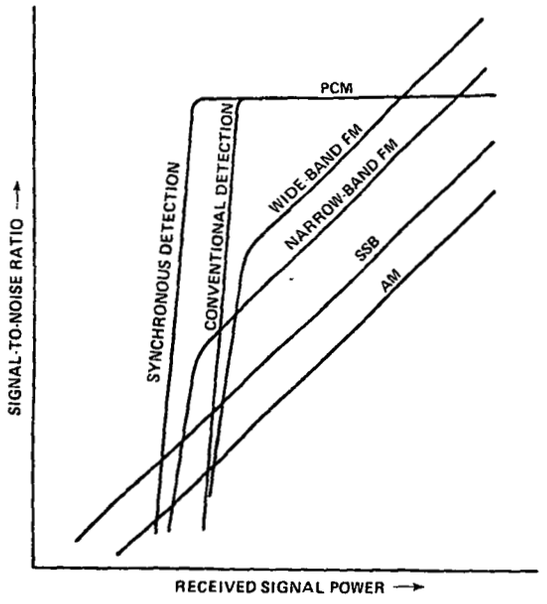


Fig. 11.9 Signal-to-noise behavior of various modulation techniques.

schemes, need not be able to interpret them. The principal differences between the schemes are the bandwidth efficiencies and the ease of determining the clock rate and bit boundaries. It is important to recognize that a variety of schemes exist and to ensure consistency across subsystem interfaces.

Figure 11.9 qualitatively compares performance of the various modulation schemes as a function of signal power and SNR. A point of interest is the very sharp threshold of the FM and PCM schemes as power and SNR decrease. This graphically displays the phenomena discussed earlier.

11.6 Radio Frequency Elements

11.6.1 Antennas and Gain

Antennas are generally categorized as omnidirectional (omni) or directional. The latter come in a variety of types, with beamwidths ranging from a few tenths to several tens of degrees. The beamwidth of a directional antenna is defined as the angle between the -3 dB (half-power) points relative to the power on the boresight axis. Figure 11.10 shows this and other beam characteristics.

The gain of a directional antenna relates the power on the boresight axis to that of an ideal isotropic radiator, a point source that radiates energy equally in all directions. The gain of an isotropic radiator is 0 dB. An antenna with gain gathers up the radiation that is distributed evenly to the celestial sphere by an isotropic

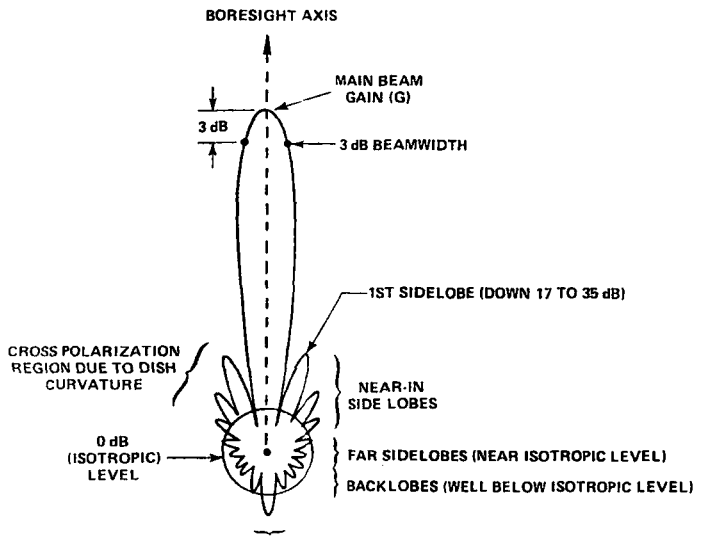


Fig. 11.10 Antenna pattern.

radiator and concentrates it into a smaller area. Thus, the gain of an ideal parabolic dish illuminating 1 deg^2 of sky is 41,253. Put another way, the product of gain G and the subtended angle ϕ (or field of view) is always 41,253$^{\circ 2}$ for an idealized antenna:

$$G\phi = 4\pi \text{ sr} = 41{,}253 \text{ deg}^2 \tag{11.2}$$

However, real antennas are not this efficient. For example, the gain-beamwidth product of a parabolic dish is approximately

$$G\phi = 2.6 \text{ sr} = 27{,}000 \text{ deg}^2 \tag{11.3}$$

This is a good rule of thumb for parabolic antennas, although coefficient values as low as 20,000 and as high as 30,000 are common in communications literature.

Because the goal is to provide a certain minimum signal strength at the receiver, a tradeoff always exists among transmitter power, antenna beamwidth, and the derived pointing accuracy. Figure 11.11 gives field-of-view (FOV) definitions and provides the equations for computing them.

Beamwidth is also related to the frequency of transmissions f, transmitted wavelength λ, and the speed of light c. The gain of an antenna as a function of the area of the aperture A, or diameter D for a circular aperture, is given by

$$G = \eta \frac{4\pi f^2 A}{c^2} = \eta \frac{\pi^2 f^2 D^2}{c^2} \tag{11.4}$$

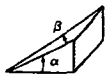
where η is the antenna efficiency. Common satellite directional antennas have η values of 0.50–0.80.

$\phi_{\text{STERADIANS}} \triangleq \dfrac{\text{SUBTENDED SPHERICAL SURFACE AREA}}{r^2}$

$\therefore$ 4π STERADIANS IN THE ENTIRE SPHERE

CONICAL FOV:

$$\phi = 2\pi\left(1-\cos\frac{\theta}{2}\right) = 4\pi\sin^2\frac{\theta}{4}$$



RECTANGULAR FOV:

$$\phi = \alpha\beta$$

Fig. 11.11 Solid angle.

If we assume that the antenna has a rectangular FOV with beamwidth in one plane of $\theta = \sqrt{\phi}$, then

$$G\theta^2 = 2.6\,\text{sr} = 27{,}000\,\text{deg}^2 \tag{11.5}$$

The beamwidth θ for a parabolic dish can be approximated by

$$\theta = \frac{164\lambda}{\pi D} \tag{11.6}$$

Figure 11.12 shows a variety of antenna types, particular characteristics of which go beyond the scope of this book. Table 11.2 presents gain and effective area for several types.

Antenna polarizations, mentioned earlier, can either be linear with vertical or horizontal orientation or circular with left or right orientation. Improper matching of transmit and receive antenna polarization will result in losses ranging from moderate to severe. Linear to circular mismatch will result in a loss on the order of 3 dB. A left/right mismatch of circular polarization will result in a loss of 25 dB or greater. Worst of all is a vertical/horizontal mismatch in linear polarization, where the losses theoretically become infinite; losses in the range of 25–35 dB are common.

Figure control—maintaining a very smooth and accurate surface shape—must be nearly exact for parabolic and phased-array antennas. It is desirable to maintain the surface contour within 1/20 of a wavelength. For high frequencies or large antennas, this becomes very demanding on structural design and fabrication. An empirical factor for loss in antenna gain due to surface roughness is

$$g = \exp(-4\pi^2 E_{\text{rms}}^2 A) \tag{11.7}$$

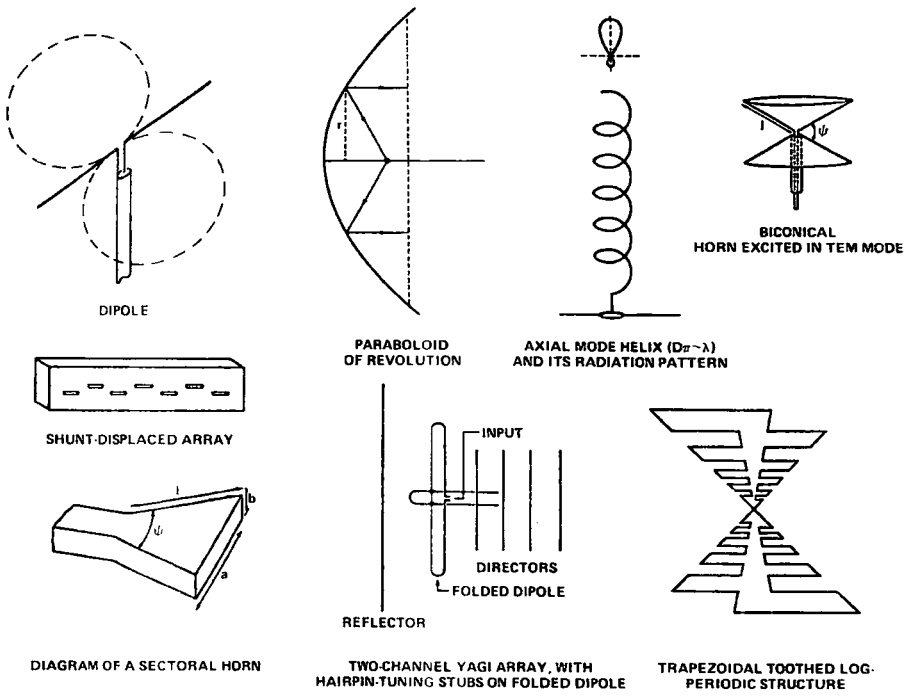


Fig. 11.12 Typical space communications antennas.

where g is the gain/loss factor, E_{rms} the average surface error in fractions of a wavelength, and

$$A = \frac{1}{1 + (D/4F)^2} \tag{11.8}$$

where D is the diameter of the antenna, and F is the focal length.

Table 11.2 Gain and effective area of several antennas

Type of antenna	Gain	Effective area
Isotropic	1	$\lambda^2/4\pi$
Elementary dipole	1.5	$1.5\ (\lambda^2/4\pi)$
Halfwave dipole	1.64	1.64
Halfwave dipole	1.64	$1.64\lambda^2/4\pi$
Horn (optimum)	$10A/\lambda^2$	$0.81A$
Parabolic reflector (or lens)	$6.2–7.5(A/\lambda^2)$	$0.5A–0.6A$
Broadside array (ideal)	$4\pi A^2/\lambda^2$	A

11.6.2 Radio-Frequency Link

The received signal power over a communication link, when combined with the total noise power, provides the fundamental measure of the quality of service available for communications.

Consider a transmission at power level P_t to a receiver system R meters distant. If the power is transmitted through an isotropic radiator, then the spherically expanding wavefront will have a flux density (power per unit area, or W/m^2) of[3]

$$F = \frac{P_t}{4\pi R^2} \tag{11.9}$$

when the wavefront arrives at the receiver. If the same power is transmitted through an antenna of gain G_t, then the flux density at the receiver will be

$$F = \frac{G_t P_t}{4\pi R^2} = \frac{EIRP}{4\pi R^2} \tag{11.10}$$

where $EIRP = P_t G_t$ is the effective isotropic radiated power. This is simply the power that would have to be transmitted through an isotropic radiator to achieve the same flux density obtained from an antenna of gain G_t.

At the receiver, an antenna with physical area A_r and effective area $A_e = \eta A$ intercepts a portion of the flux density, its total received power being

$$P_r = FA_e = \frac{P_t G_t A_e}{4\pi R^2} \tag{11.11}$$

From Eq. (11.4), the gain of the receiving antenna can be expressed in terms of its area as

$$G_r = \frac{4\pi A_e}{\lambda^2} \tag{11.12}$$

The received power, therefore, will be

$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi R}\right)^2 \tag{11.13a}$$

Equation (11.13a) gives the received power as a function of range and wavelength when both transmitter and receiver gain are assumed fixed. As a matter of engineering practice, we may customarily specify either the gain or area of either the transmitter or receiver. From Eqs. (11.13a) and (11.12) applied in various permutations, we can obtain the received power when both receiver and antenna areas are fixed:

$$P_r = P_t \left(\frac{1}{\lambda R}\right)^2 A_r A_t \tag{11.13b}$$

as well as for the cases where one gain and one antenna area are fixed:

$$P_r = P_t\left(\frac{1}{4\pi R^2}\right)A_r G_t \tag{11.13c}$$

$$P_r = P_t\left(\frac{1}{4\pi R^2}\right)G_r A_t \tag{11.13d}$$

The term $[\lambda/(4\pi R)]^2$ is known as the path loss. This is not an absorption loss, but rather a "dilution" of the transmitted energy as the wavefront expands in traveling toward the receiver. In terms of the path loss, the received power will be

$$P_r = EIRP \times G_r/\text{path loss} \tag{11.14}$$

Path losses of some 200 dB characterize links between Earth stations and geosynchronous satellites. As a result, high-gain antennas, low-noise receivers, and careful selection of modulation method and bandwidth are required to achieve acceptable signal-to-noise ratios over that long link. Large path loss is a distinctive feature of space communication systems.

Several other loss mechanisms may be present as well. These additional losses are included with the path loss in the previous equation to arrive at the overall received signal power.

Multipath loss occurs when copies of the signal arrive at the receiving antenna after being reflected to the receiver off other objects. Because these signals have traveled a greater distance, they arrive after the main signal and may interfere destructively or appear as noise. Many readers will recall the most common example of such interference in the form of television "ghosts" caused by reflection of the signal from nearby mountains, overflying aircraft, or other large objects. Such ghosts have essentially been relegated to the status of ancient memories, at least since the mid-1990s, when commercial television stations began including a ghost cancellation reference, or GCR, in analog television broadcasts, to allow the receiver to distinguish between the direct-path signal and its later-arriving doppelgangers. More modern digital television broadcasts include frame synchronization information that similarly allows accomplishment of the task of identifying the direct-path signal.

Faraday rotation is a problem for linearly polarized signals. As the linearly polarized electromagnetic field of the RF signal penetrates the Earth's magnetic field, it is rotated as shown in Fig. 11.13. Any rotation between the polarization direction of the antenna and that of the incoming signal causes a loss that can become very large as the rotation approaches 90°, as shown in Fig. 11.13. This problem can be avoided in space applications by using circular polarization. Thus, one often sees helical antennas for HF/VHF/UHF links and cruciform antennas for X-band in low- and mid-gain applications.

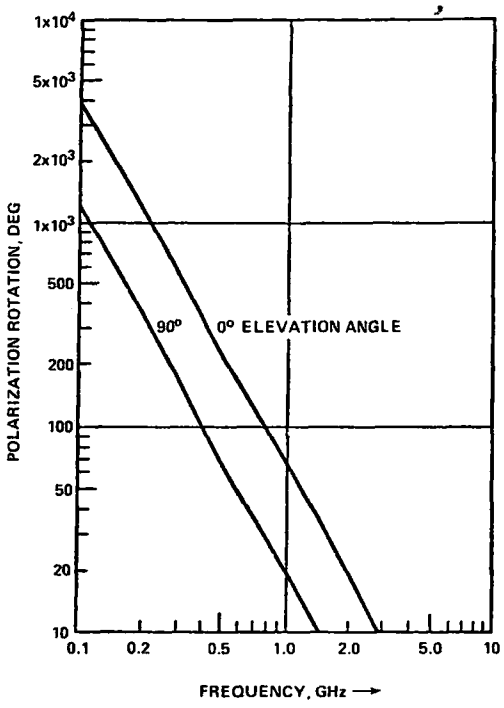


Fig. 11.13 Faraday rotation of polarized waves.

Losses related to atmospheric absorption afflict both high- and low-frequency ends of the scale (Fig. 11.14). Signal loss due to absorption by ionospheric electrons first becomes noticeable at 500–600 MHz and begins to be significant at about 100 MHz as frequency is decreased. Atmospheric absorption causes essentially no loss between about 600 MHz and 4 GHz. This convenient window explains the extensive use of S-band in the 2-GHz range for spacecraft communication.

The desire for higher data rate and narrower beamwidth pushes spacecraft designers toward higher frequencies. However, as shown in Fig. 11.13, atmospheric losses rise rapidly with increasing frequency in the higher ranges because of absorption by oxygen and by water vapor. The figure shows data for an average atmosphere. Severe clouds, fog, or rain will greatly increase the loss. The absorption region begins at approximately X-band (8 GHz), and systems in this frequency range perform quite satisfactorily except in case of significant rain at the receiving site. Systems operating at Ku-band (12–14 GHz) are further into the absorption range but still satisfactory in good weather. Note the significance of elevation angle. This simply reflects that, at lower elevation, the signal must traverse a longer path through the atmosphere. The higher frequency ranges may thus be more useful for geostationary communications satellites, which are usually well above the horizon for typical Earth station locations. Above Ku-

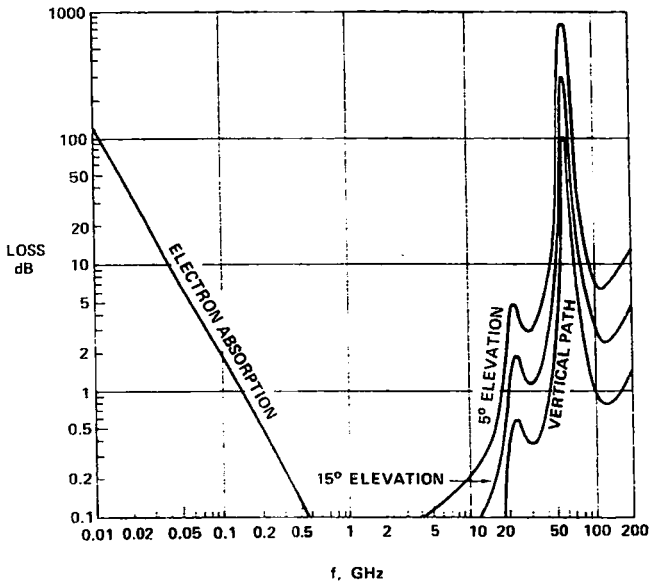


Fig. 11.14 Total absorption loss (excluding weather).

band, absorption rises rapidly. Dips do occur, however, since the offending molecules absorb at discrete frequencies. The two windows of potential future interest are in the vicinity of 35 and 94 GHz—regions referred to as Ka-band and 'millimeter wave,' respectively. For high elevation angles and good weather, these bands may have application, but there is always the risk of data loss due to inconveniently timed bad weather.

Note that most of these concerns do not apply for spacecraft-to-spacecraft communication. This, plus unstable politics on Earth, has generated interest in placing large relay spacecraft in Earth orbit to communicate with planetary spacecraft using Ku-band or higher frequency. This would reduce power requirements and antenna size on the deep space vehicle. Data from the Earth orbiter would then be relayed to the ground using frequency bands better suited to traversing the atmosphere. Cost has prevented pursuit of the idea, but it has much to recommend it.

11.6.3 Noise

The signal power from a communications link is always received in the presence of noise. Given this noise, the problem of receiving and correctly decoding the transmitted signal becomes a problem in estimation theory, a branch of mathematical statistics. The noise degrades the ability of the receiver to estimate the transmitted signal precisely, leading to incorrect bits in a digital signal or, more obviously, overt corruption of an analog signal.

Strictly speaking, noise is any received power that interferes with the desired signal, a definition that includes many things that could be "signals" to other users. Spacecraft communications engineers and radio astronomers have fundamentally differing views about which electromagnetic waves constitute signal and which constitute noise. The principal figure of merit used to specify the quality of a communications link is the ratio of signal power to noise power, the signal-to-noise ratio, or SNR. Interfering signals aside, most noise is of thermal origin, either from blackbody emission as discussed in Chapter 9, or from thermally induced motion of electrons in the receiver circuitry, referred to as Johnson noise. Noise that is not thermally generated will usually be approximated in its effects by assuming that it is thermally induced, as we will discuss later.

As an aside, it should be noted that the discussion in this section is relevant only for information transmission at conventional radio frequencies. Optical communications links are becoming quite common in space applications, both internally to the spacecraft and between vehicles. Thermal noise is not typically important in optical and electro-optical communications systems, which are dominated by shot noise, quantum noise, or the optical interference background, depending on the wavelength and system characteristics.

Electrons in receiver circuits are thermally agitated when the circuits operate above 0 K. This motion causes a random voltage or current to be induced in all portions of the circuit. Subject to the assumption that we are operating at radio frequencies with circuits of moderate temperature (at least a few tens of Kelvins), the noise power from a given device in a given bandwidth interval may be approximated by

$$P_N = kTB \tag{11.15}$$

where P_N is the noise power (in W), k is the Boltzmann constant (1.38×10^{-23} Ws/K), T is the device temperature in Kelvins, and B is the bandwidth interval.

A few comments are in order. T is referred to as the effective noise temperature of the device and is commonly used as a figure of merit in characterizing noise performance in a communications system. With the agitated electron conceptual model discussed earlier, the total noise power will be due to the collective output of a multitude of individual oscillators. The central limit theorem of statistics thus guarantees the noise to be Gaussian, i.e., the noise power per unit frequency interval at any given instant will have a level drawn from a normal distribution, characterized by a mean N_0 and a standard deviation σ. If N_0 is constant across all frequencies (always an idealization, since this implies infinite total energy content in the noise), then the noise is called white, by analogy to white light as an equal mixture of all colors. White Gaussian noise (WGN) is the standard assumption in communications link analysis.

As indicated earlier, the WGN assumption cannot ever be strictly correct. Nonetheless, receiving system bandwidths are normally quite small with respect to the frequency scale over which N_0 varies significantly. Thus, the WGN assumption with constant noise power spectral density N_0 given by

$$N_0 = kT = \frac{P_N}{B} \tag{11.16}$$

is normally quite good. The utility of this concept is so great that communications engineers commonly characterize even highly colored (e.g., nonthermal) noise by an equivalent "noise temperature." This would be selected to yield a white noise power level comparable in its effects to that of the actual colored noise.

Noise in the communications link is not limited to that generated in the receiver circuitry. As we discussed in Chapter 9, any object with a temperature above 0 K emits electromagnetic radiation distributed across all wavelengths according to Planck's law, Eq. (9.15) (modified, of course, by the fact that no surface is ideally "black"). Though blackbody noise is highly colored, it is again modeled in its effect on communications systems by assuming WGN at an equivalent temperature, usually that of the blackbody spectral peak. Thus, the sun typically acts as a noise source at approximately 5780 K. Table 11.3 shows effective noise temperatures for several potential sources. Figure 11.15 depicts the combined effect of common noise sources as a function of frequency and for several elevation angles.

The cursory discussion here can only touch on the subject of noise modeling for space communication links. Reference 4 provides an exhaustive discussion of environmental noise effects on the JPL Deep Space Network (DSN) and may also be of interest in collateral applications.

Table 11.3 Effective noise temperature of various sources

Source	Temperature
Sun	5780 K
Earth	290 K
Galaxy	Negligible above 1 GHz
Sky	30–150 K
Atmosphere	Noise due to absorption and reradiation

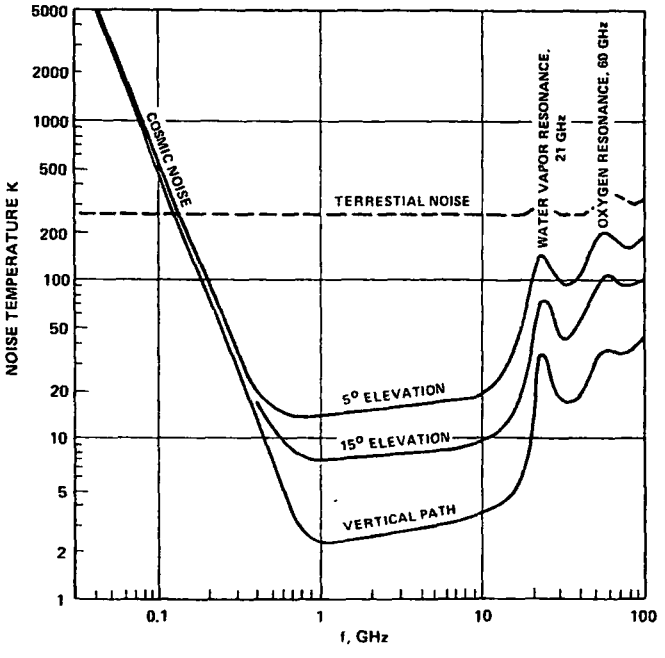


Fig. 11.15 Composite link noise plot (excluding weather).

11.6.4 Noise Figure

Another common measure of merit for the noise performance of a device is the noise figure, defined for a "two-port" device as

$$F_n = \frac{\text{Noise output of a real device}}{\text{Noise output of an ideal device with input at temperature } T_0} \tag{11.17}$$

where $T_0 = 290$ K is the standard reference temperature.

To develop a mathematical expression for noise figure, note that an ideal device adds no additional noise to the signal passing through it. The noise at its output will be the noise at the input, amplified by gain. Thus, the noise figure is the actual noise power output divided by the noise power output due only to the input noise generated by the input termination at temperature T_0. The input noise power P_{N_i} to the device due to the input at temperature T_0 will be kT_0B_n. If P_{N_o} is the actual noise output power from the real device, then the noise figure is

$$F_n = \frac{P_{N_o}}{GP_{N_i}} = \frac{P_{N_o}}{GkT_0B_n} = \frac{P_{N_i}G + \Delta P_N}{GkT_0B_n} = 1 + \frac{\Delta P_N}{GkT_0B_n} \tag{11.18}$$

where ΔP_N is the noise at the output of the device due to the device itself, and G is the two-port device gain. Note that $F_n \geq 1$; the best two-port noise figure a device can achieve is a value of one.

If we represent the noise added by the device, ΔP_N, as an additive noise source at an effective noise temperature of T_e at the input of the device, then the noise power added by this source has a value of $\Delta P_N = GkT_eB_n$, and the noise figure is

$$F_n = 1 + \frac{T_e}{T_0} \tag{11.19}$$

This gives the relationship between noise temperature and noise figure for a two-port device.

If a two-port device is purely passive, such as a transmission line or attenuator, the effective input noise temperature T_e will be a function of the device's thermodynamic temperature T_t and the available loss L,

$$T_e = T_t(L - 1) \tag{11.20}$$

The available loss is defined as $L = 1/G$. This equation gives the effective noise temperature of the device referred to the input of the passive device. The effective noise temperature of the passive device referred to the output of the device T_l is

$$T_l = T_t\left(1 - \frac{1}{L}\right) \tag{11.21}$$

This effective noise source at the output can be referred to the input by multiplying the effective output noise temperature T_l by the available loss of the device:

$$T_e = LT_l = \frac{T_l}{G} \tag{11.22}$$

The noise figure of a cascade of devices can be calculated from the noise figures and gains of the individual stages. Let F_1, F_2, and F_3 be the two-port noise figures for the first, second, and third stages in a cascade of devices; G_1, G_2, and G_3 be the power gains of the three stages; and T_1, T_2, and T_3 be the effective input noise temperatures of the three stages. It can be shown that

$$F_{123} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1G_2} \tag{11.23}$$

where F_{123} is the equivalent two-port noise figure of the cascade of the three stages. A similar result can be obtained for the equivalent input noise temperature of the cascade T_{123}:

$$T_{123} = T_1 + \frac{T_2}{G_1} + \frac{T_3}{G_1G_2} \tag{11.24}$$

These results can be extended to larger numbers of stages. The technique can be used to calculate the overall effective noise temperature of an entire receiving system.

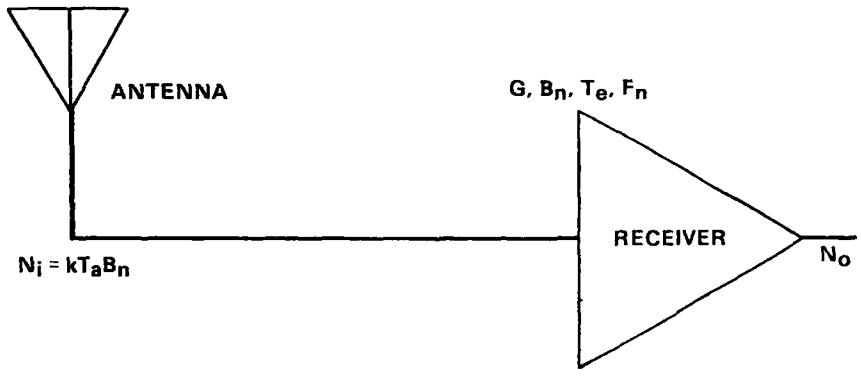


Fig. 11.16 Block diagram of receiver system.

11.6.5 Noise Figure for a Receiver System

Consider the noise performance of an entire receiver system, including the antenna, with a system noise figure, as diagrammed in Fig. 11.16. The "electronics" portion of the receiver is modeled as a two-port device with noise figure F_n, or, equivalently, noise temperature T_e. Figure 11.17 shows a block diagram of the equivalent model of the system. The receiver model implicitly includes any passive loss components between the antenna and the receiver.

All noise, other than that generated internally by the receiver, is modified by an additive noise source of value kT_aB_n. The temperature T_a is called the effective antenna temperature. Note that no additional noise enters the system through the antenna in this model. All external noise sources, such as galactic noise, atmospheric noise, and warm-body emission, are included in the appropriate selection of the value of T_a. If $T_s = T_a + T_e$ is the effective system noise temperature, then the noise power output from the system will be

$$P_{N_o} = kT_sB_nG \tag{11.25}$$

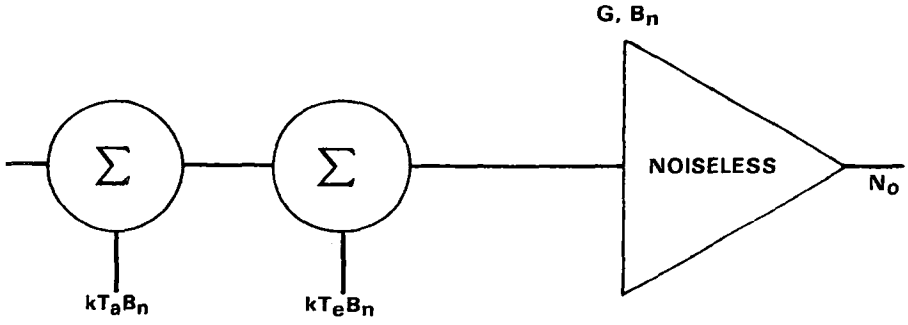


Fig. 11.17 Equivalent model of receiver system.

For the system modeled in Fig. 11.17, the noise figure can be written as

$$F_s = \frac{kB_nG(T_a + T_e)}{kB_nGT_0} \tag{11.26}$$

from which we find

$$F_sT_0 = T_a + T_e = T_s \tag{11.27}$$

$$F_s = \frac{T_a}{T_0} + \frac{T_e}{T_0}$$

$$= \frac{T_a}{T_0} + F_n - 1 \tag{11.28}$$

Note that $F_s \geq 0$. The noise performance of the entire system can be modeled in terms of either T_s, or T_e and T_a, or F_s. Because they provide equivalent results, the choice is primarily one of convenience.

Noise figures and effective noise temperatures for various types of amplifiers appear in common references. However, the reader should note that authors and designers do not always use these same (or even a consistent) set of definitions for these figures of merit. One should always determine the fundamental definitions being used before basing system performance on stated numerical noise-performance figures.

The total noise power, derived from the specified system noise temperature (or noise figure), can be combined with the received signal power to yield a communication system's SNR. For example, when both transmitter and receiver antenna gain are fixed, we have from Eqs. (11.13a) and (11.15):

$$\text{SNR} = \frac{P_r}{N} = P_tG_t\left(\frac{1}{R^2}\right)\left(\frac{\lambda}{4\pi}\right)^2\left(\frac{1}{kB}\right)\left(\frac{G_r}{T}\right) \tag{11.29}$$

From the point of view of a receiving ground station, all parameters in Eq. (11.29) are fixed by nature or by the spacecraft except for the term (G_r/T). This term, the ratio of the receiving antenna gain and system noise temperature, essentially defines the quality of the communications link with a given spacecraft. Figure 11.18 gives system G/T vs antenna aperture for several frequency bands under the assumption of a typical system noise temperature of 315 K, a very ordinary performance figure.

When the received signal power is expected to be very weak, as for interplanetary communication, special low-noise-figure receiver systems are required. Such low-noise receivers are in general composed of two parts, a preamplifier mounted on the antenna at its focal point and a receiver located away from the antenna at the signal processing center.

The preamplifier increases the power of the incoming signal without adding appreciable noise, and thus strengthens the signal before sending it to the remote receiver. This is critical because the resultant increase in signal power minimizes

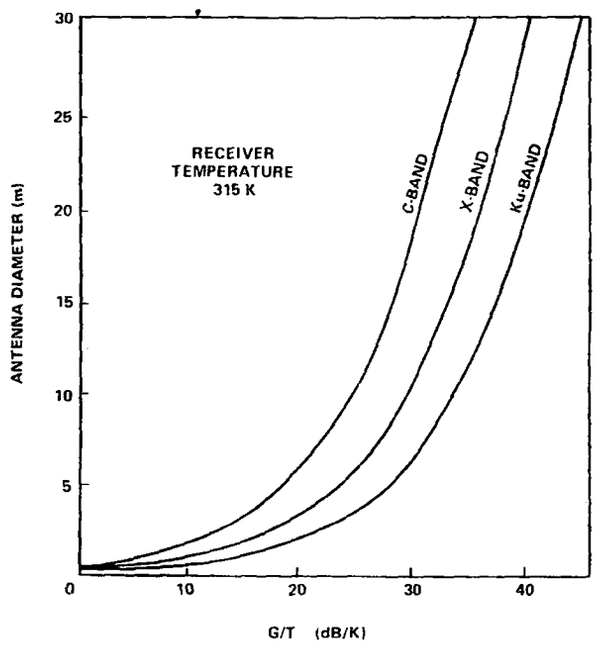


Fig. 11.18 System G/T vs antenna aperture.

the effect of noise generated by subsequent handling, as described in earlier sections.

The ability to minimize noise generation lies with the low-noise amplifier (LNA). Several varieties of LNAs are used in practice, the most sensitive being cryogenic systems cooled with liquid helium to approximately 4 K. This slows random thermal motion of the electrons in the components of the amplifier, thus resulting in a better signal-to-noise ratio for the amplified signal.

11.6.6 Communications in Noise

As we have indicated, the SNR is the fundamental quantity characterizing the quality of a communications link. This is because the data rate, or channel capacity, of the link is directly related to the SNR. The fundamental theory of communication in the presence of noise was developed by Shannon;[5] all subsequent work in the field is an extension of his efforts. According to Shannon's theorem, the error-free channel capacity of a link is given by

$$C = B \log_2 (1 + \text{SNR}) \tag{11.30}$$

where C is the channel capacity in bps, and B is the link bandwidth in Hz. (Usually the data rate in bps will be specified; this is approximately equal to the required link bandwidth in Hz.) Thus, if we consider a standard C-band

communications satellite transponder with a nominal bandwidth of 36 MHz and assume SNR = 15, the theoretical channel capacity is $C = 144$ Mbps. In actual practice, such a transponder could handle one color television signal or 1200 voice links.[6] Each voice channel is assigned a nominal 56 kbps data rate, adequate for normal speech, and so the link carries a total data rate of 67 Mbps. This is less than 50% of the channel capacity promised by Shannon's error-free limit, and the satellite link will still contain some bit errors, typically on the order of one in a million.

This simple example illustrates some important points with respect to Shannon's theorem. The theorem tells us what the limit is but not how to reach it. Today's satellite communications links employ very sophisticated coding techniques to achieve their capacity. Presumably even more complex techniques can be found to increase channel capacity, but Shannon's theorem gives no clue. Equation (11.30) gives the limit for error-free transmission, but says nothing about how system performance, i.e., bit error rate (BER), degrades as the limit is exceeded. Real systems degrade is shown in Fig. 11.19, which plots bit error rate as a function of, essentially, SNR. The analysis leading to this result is beyond the scope of this text but is found in standard references on digital communication theory.[7]

11.6.7 Link Analysis

The results presented in this section are typically distilled by the communications engineer into what is termed a link analysis for the given situation. Typical inputs to the analysis will be the intended frequency (or wavelength) band to be used, the required data rate and maximum range, noise

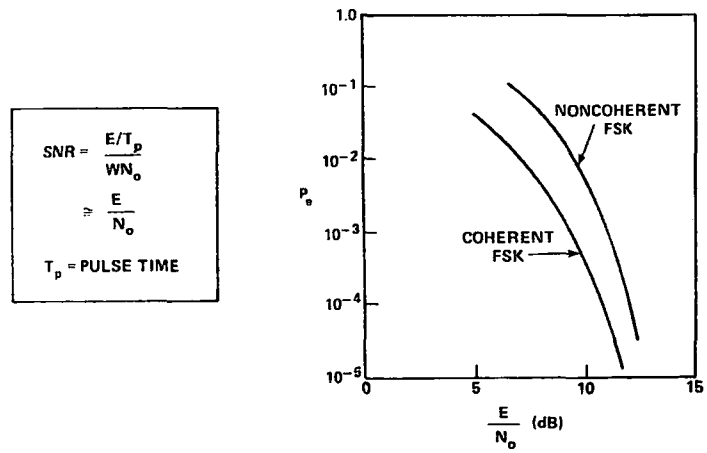


Fig. 11.19 Bit error rate for FSK.

temperature (or G/T) of the receiving system, and the minimum signal-to-noise ratio required by the receiver to achieve a specified bit error rate.

A simple example of a space-to-space communications link analysis is presented in Table 11.4. For pedagogical reasons, comments are supplied in somewhat more detail than is usual, so that the reader may apply the principles to other cases of interest. To interpret the example properly, we must first recast some of the preceding results into the conventional terminology of the communications engineer. We begin by considering again the signal-to-noise ratio for the gain–gain case of Eq. (11.29), which we rewrite slightly as

$$\text{SNR} = \frac{P_r}{N} = P_t G_t G_r \left(\frac{1^2}{R}\right)\left(\frac{\lambda}{4\pi}\right)^2 \left(\frac{1}{kTB}\right) \tag{11.31}$$

It will be more convenient to express this relationship in decibel form,

$$\text{SNR}_{\text{dB}} = P_{t\text{dB}} + G_{t\text{dB}} + G_{r\text{dB}} - 20\log_{10} R - 20\log_{10}\left(\frac{\lambda}{4\pi}\right) - 10\log_{10} kTB \tag{11.32}$$

which has the advantage of reducing all subsequent calculations to simple addition and subtraction, as well as eliminating numerous large-magnitude exponents. In what follows, we assume reference values of a single Watt, Hertz, meter, etc., where necessary, i.e., everywhere dimensioned quantities are employed.

Let us now consider a simple space-to-space communications example, such as might apply in the case of a rendezvous operation. We assume the target vehicle is equipped with a 401.5 MHz UHF data link operating at 57.6 kbps and intended to be acquired by the chase vehicle at a distance of 100 km. The chase vehicle is equipped with a receiver that offers a 10^{-6} BER, provided the input SNR is at least 11.5 dB. The target vehicle uses a 5-W transmitter and an antenna with approximately hemispherical coverage, but offering a gain of -8 dB relative to isotropic. The chase vehicle receiver is not of especially high performance, having a system noise temperature of 600 K and a boresight gain of only 2 dB. Because of the requirement to orient the chase vehicle according to the dictates of the rendezvous maneuvers, it cannot be guaranteed that the antenna will be pointed directly at the target vehicle at all times, so that even though the antenna has a fairly broad beamwidth, we assume worst-case operation at the half-power point of the pattern. As a matter of routine conservative practice, transmit and receive efficiency factors of 0.5 are assumed. In addition, line losses of 2 dB in the received signal path are assumed. The question confronting the communications engineer is, can the desired link be closed, and if so, with what margin? Table 11.4 provides the answer; even with fairly conservative assumptions such as we have employed here, a "link margin" in excess of 6 dB over the requirement is found to exist.

Table 11.4 Space-to-space link analysis example

Parameter	Units	Value	Value, dB	Comments
Requirements data				
Data rate, B	bps	5.76E+04	47.60	dB_{Hz}
Range, R	m	1.00E+05	50.00	$dB_m = 10 \log R$
Frequency, f	Hz	4.015E+08	86.04	dB_{Hz}
Signal to noise ratio, SNR_{dB}	dB	11.50	11.50	Receiver specification for 10^{-6} BER
Constants				
Pi, π		3.1416	4.97	dB
Speed of light, c	m/s	3.00E+08	84.77	$dB_{m/s}$
Boltzmann constant, k	W/Hz/K	1.38E−23	−228.60	$dB_{w/Hz/K}$
Transmitted power				
Transmitter output, P_t	W	5.00	6.99	dB_w, specification value
Antenna gain, G_t	dB_i	−8.00	−8.00	Gain relative to isotropic (=1.0)
Efficiency factor, η_t		0.50	−3.01	Miscellaneous losses
EIRP	W		**−4.02**	dB_w
Received power				
Spreading loss, $1/R^2$	m^{-2}	1.00E−10	−100.00	$SL_{dB} = -20 \log R$
Free space loss, $1/4\pi^2$	m^2	3.53E−03	−24.52	dB_{m^2}
Atmosphere loss	dB	0.00	0.00	
Polarization loss	dB	−1.00	−1.00	
Ionosphere loss	dB	0.00	0.00	
Antenna gain, G_r	dB	2.00	2.00	
Antenna pointing loss	dB	−3.00	−3.00	Half-power point of antenna pattern
Efficiency factor, η_r		0.50	−3.01	Miscellaneous losses
Implementation loss	dB	−2.00	−2.00	Cable losses, etc., for receiver system
Received power, S	W		**−135.55**	dB_w

(*continued*)

Table 11.4 Space-to-space link analysis example (continued)

Parameter	Units	Value	Value, dB	Comments
Noise power				
System noise temperature, T	K	600	27.78	dB_K, conservative assumption
Specific thermal noise, N_o	W/Hz		−200.82	$N_o = kT$
System noise	W		**−153.22**	dB_w; $N = N_oB = kTB$
System requirement				
Received $S/N = E_b/N_o$ (SNR_{dB})	dB		17.66	$SNR = S/N = S/N_oB = St_b/N_o = E_b/N_o$
Required $S/N = E_b/N_o$ (SNR_{dB})	dB	11.50	11.50	Receiver specification for 10^{-6} BER
Link margin			**6.16**	DB

The details of link margin calculations will of course vary from case to case, but in broad outline are as we have shown here. Every space system engineer quickly becomes familiar with this tool of the trade.

11.7 Spacecraft Tracking

Ground-based tracking stations were at one time the sole means of fulfilling three crucial functions: receiving spacecraft telemetry and routing it to the control center for processing and distribution; uplinking commands from the control center to the spacecraft; and obtaining Doppler range-rate data and spacecraft azimuth and elevation necessary for orbit determination. Not all stations necessarily had all these functions; some might perform tracking or data reception only. However, this was, and in many cases remains, the general case.

The technology of spacecraft tracking systems underwent a major revolution beginning in the early 1990s with the routine use of the Tracking and Data Relay Satellite System (TDRSS) instead of an extensive network of ground stations for major programs, i.e., shuttle, International Space Station, Hubble Space Telescope, etc. This trend continued beginning in the mid-1990s with the implementation of global positioning system (GPS) navigation receivers directly onboard spacecraft and launch vehicles. With these two systems, it is now at least technically possible to communicate with essentially any Earth orbiting spacecraft via TDRSS, and to allow the spacecraft or launch vehicle to

determine its own position and to relay this to ground control via TDRSS. Such an approach obviates the need for much of the preexisting ground tracking infrastructure.

However, as we will discuss, TDRSS and GPS, together or separately, are not useful for all Earth-orbiting spacecraft, and in any case there remains the problem of tracking and communicating with interplanetary missions. Also, if onboard GPS is used but TDRSS is not, then it remains necessary to be able to track the spacecraft at least well enough to establish communication, after which more accurate state vectors can be supplied by the spacecraft. For these reasons, it will perhaps be useful to consider older systems.

Spacecraft uplink and downlink operations have been supported by various networks and individual stations in S-band (2–4 GHz), C-band (4–6 GHz), X-band (7–9 GHz), Ku-band (12–14 GHz), and Ka-band (20–30 GHz). Some older systems still use UHF frequencies in the 400-MHz band. Numerous U.S. government tracking stations are spaced throughout the world as part of various networks, including the following:

1) NASA Space Network (SN, formerly Tracking and Data Relay Satellite System, or TDRSS);
2) NASA Ground Network (formerly Space Tracking and Data Network, or STDN);
3) NASA/Jet Propulsion Laboratory Deep Space Network (DSN);
4) USAF Space Ground Link System (SGLS);
5) North American Air Defense Command Space Data Acquisition and Tracking System (NORAD SPADATS).

It is worth noting that the various systems are generally not compatible with one another; for example, a C-band radar for the NASA SN will not function within the SGLS network.

Though it is true that these networks are still comprised of "numerous" stations, it is equally true that they are not as numerous as in earlier decades. The advent of TDRSS has, especially for larger spacecraft or those returning large quantities of data (e.g., Hubble Space Telescope, the space shuttle, etc.), has resulted in the diminution of importance of many individual ground stations.

11.7.1 NASA Space Network

As this is written, the NASA Space Network is comprised of the TDRSS constellation, plus associated ground facilities at White Sands, New Mexico, and on Guam, in the western Pacific. The TDRSS constellation consists of nine large satellites in geostationary orbit that relay data between low-orbit satellites and the ground, thus eliminating the problem of long communication gaps for such spacecraft. Of these, three are of the so-called enhanced variety, launched in the early 2000s, and six are of the original design, some of which have been in service for nearly 20 years.

The original six TDRSS spacecraft support single- and multiple-access S-band and Ku-band communications. One of the significant enhancements in the new series is the addition of a steerable single-access Ka-band antenna, intended to provide high data rate support, including high-definition television (HDTV), to the shuttle and International Space Station programs.

The primary ground link with TDRSS is the White Sands Complex (WSC), with both primary and backup capabilities [White Sands Ground Terminal (WSGT) and Second TDRSS Ground Terminal (STGT)]. Additional capability, including direct communication with those satellites not in view from White Sands, is provided by the Guam Remote Ground Terminal (GRGT). To use TDRSS, the host spacecraft must, in general, carry its own state vector (so that it can find a TDRSS spacecraft) and must have a steerable antenna to allow in-flight tracking of TDRSS.

As noted, this system has essentially replaced the former utility of the STDN and SGLS networks for some programs. However, TDRSS is not well suited to support smaller satellites that cannot carry the rather cumbersome equipment necessary. TDRSS may also be inappropriate even for larger, but relatively low cost, spacecraft, simply because the equipment for the TDRSS link would comprise too great a portion of the overall spacecraft cost to be justified. Finally, use of TDRSS may be technically unjustified for spacecraft returning a relatively low volume of data.

These examples aside, however, TDRSS is admirably suited to support larger, more complex spacecraft and programs. Figure 11.20 shows the TDRSS concept,

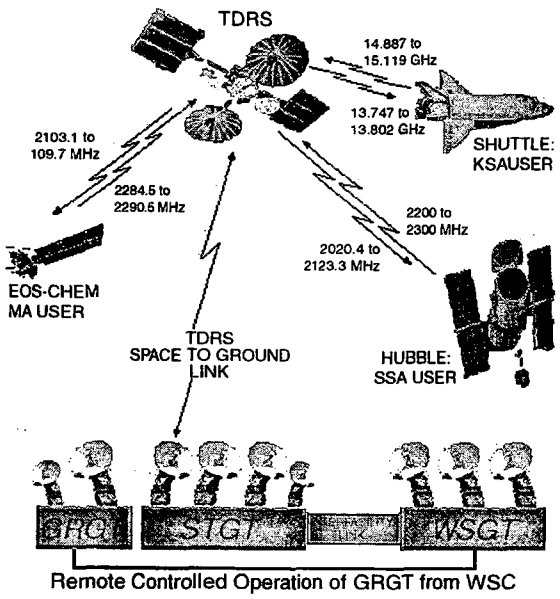


Fig. 11.20 TDRSS operations concept.

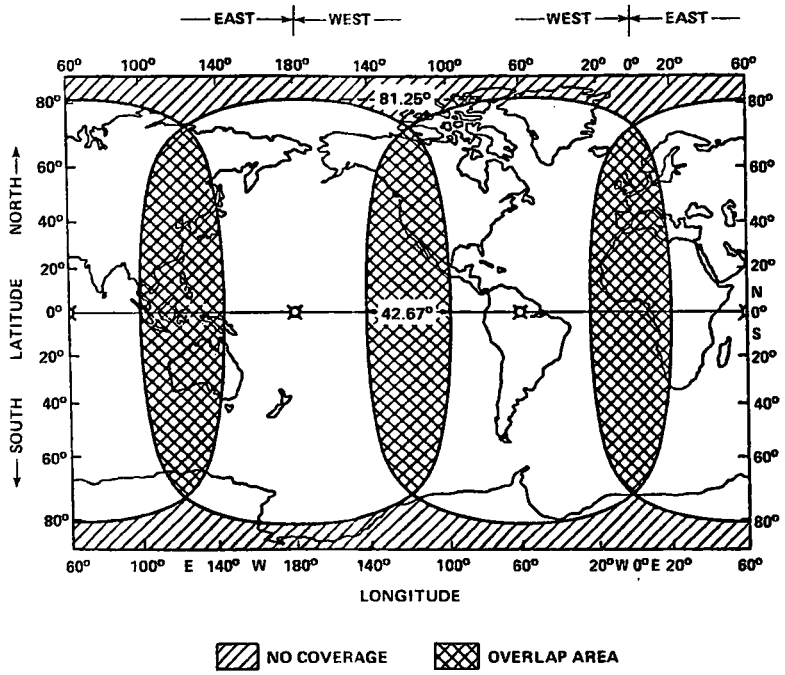


Fig. 11.21 Geosynchronous ground coverage, 0-deg elevation.

and Fig. 11.21 shows the coverage patterns characteristic of such geostationary satellites. For comparison, Figs. 11.22 and 11.23 show the coverage by the old STDN network in S-band and C-band, respectively.

A major contribution of TDRSS has been the near-elimination of the communications blackout experienced during atmospheric entry in the Mercury, Gemini, and Apollo programs, a result of the ionized flowfield surrounding, but most prominent on the windward side of, the entry vehicle. Space shuttle antennas on the leeward side of the vehicle can routinely communicate through TDRSS during reentry, though communication can still be intermittent if the vehicle executes significant roll maneuvers as a part of its trajectory-control strategy.

11.7.2 NASA Ground Network

As just indicated, the once robust Space Tracking and Data Network continues to exist, but with reduced capabilities and under a new aegis. As of this writing, operating ground stations are located in Alaska, Antarctica, Bermuda, Florida, and Norway. Several mobile installations with 2.5–7 m antenna aperture capability also exist. Overall supervision of this network is provided by NASA's

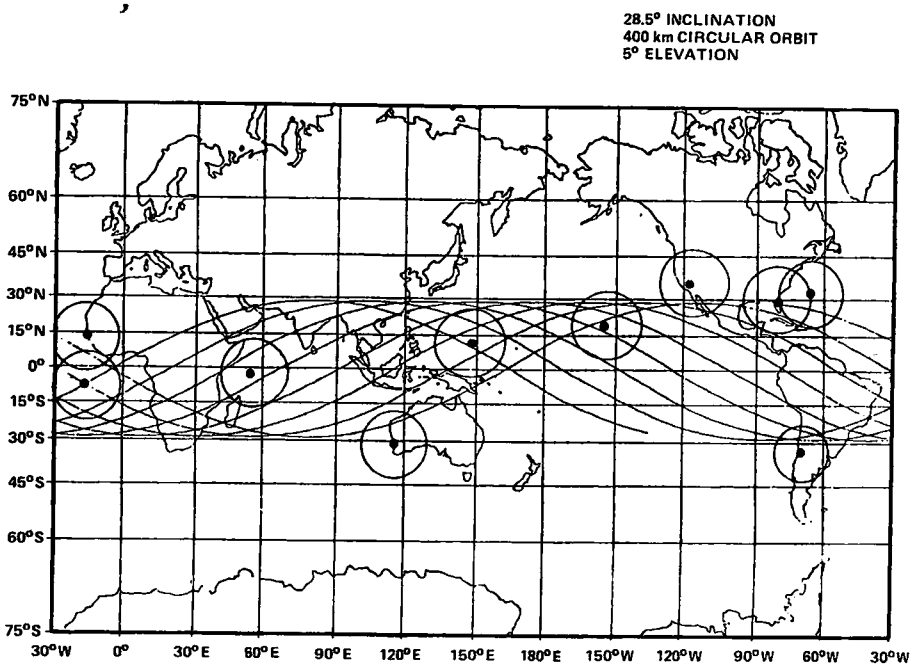


Fig. 11.22 STDN S-band coverage.

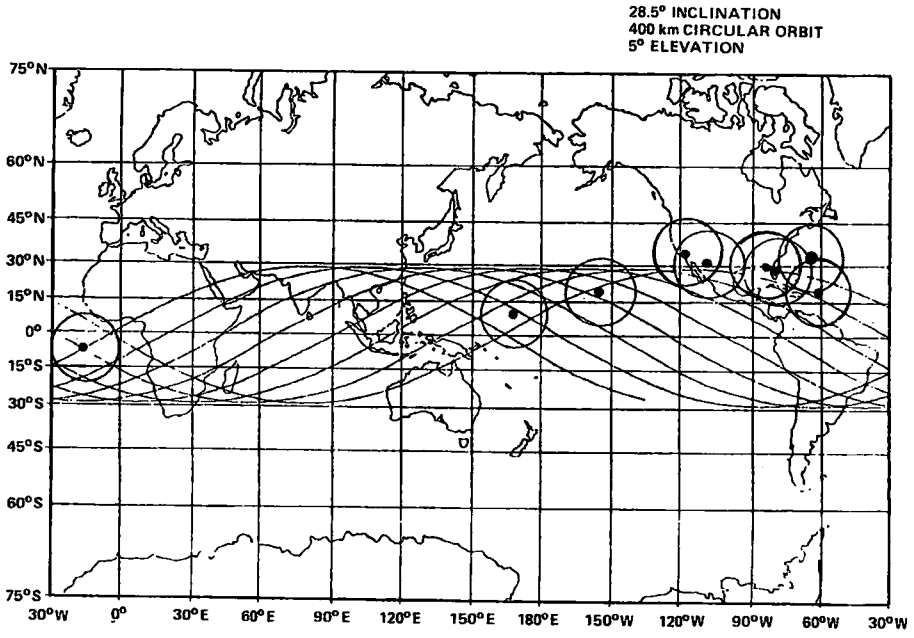


Fig. 11.23 STDN C-band coverage.

Goddard Space Flight Center, with scheduling and operations support provided by Goddard's Wallops Flight Facility at Wallops Island, Virginia.

Depending on the station, telemetry capability in UHF, L-band, S-band, and X-band is available, with the latter two being most commonly supported, along with C-band transponder tracking.[8] Both fixed and mobile tracking radars exist.

11.7.3 Deep Space Network

The JPL DSN specializes in interplanetary missions and consists of three tracking stations located near Madrid, Spain; Canberra, Australia; and Goldstone, California; thus providing nearly complete coverage of the celestial sphere. (Complete coverage of high southern declinations is provided only by the Australian installation, which in turn cannot see high northern declinations.) As might be expected from the nature of its task, the DSN features the largest antennas and lowest-noise-figure receiving systems in common use, with the possible exception of some radiotelescopes.

Each complex consists of several receiving stations equipped with large parabolic reflector antennas of 70, 34, 26, and 11 m in diameter. At the Goldstone complex, there is also a 34-m antenna used primarily for research and development.

The three complexes are controlled from JPL. The voice and data communications circuits linking the complexes to the Network Operations Control Team (NOCT) and to various U.S. and foreign flight project operations centers around the world are managed and operated by the DSNs Ground Communications and Information Service.

It is now possible to link the various antenna complexes electronically and to preserve a highly accurate time base across the link, allowing coherent reception of weak signals with a much larger effective antenna area than can be provided by any feasible single antenna. This technique has been demonstrated on many occasions, and notably as part of the recovery plan for the Galileo mission to Jupiter, follow loss of the primary spacecraft antenna as discussed in Chapter 8.

11.7.4 USAF Space Ground Link System

The SGLS network is the U.S. Air Force analog of the NASA Ground Network. The system offers S-band communications and C-band transponder tracking. SGLS provides ground station capability in Sunnyvale and Vandenberg, California; New Hampshire; Kodiak Island, Alaska; Thule, Greenland; Guam; Kaena Point, Hawaii; and on Mahe in the Seychelles.

11.7.5 Commercial Networks

It should be noted that by the late 1990s numerous commercially sponsored space programs had come into existence. For these programs, the difficulty of obtaining priority access to government tracking networks (if indeed access could be obtained at all), the associated cost of using such stations, and the often-cumbersome technical interface requirements of so doing, mitigated strongly against any involvement with the existing government networks. Accordingly, the Orbcomm, Iridium, and GlobalStar LEO communications networks, as well as the Space Imaging, Digital Globe, and OrbImage commercial imaging constellations, and others, all developed their own network of ground stations. Numerous small scientific satellite programs, even though sponsored by government organizations, have done so as well, though to a lesser degree than the commercial ventures.

Modern desktop computer workstation technology, together with the general improvement in and reduction in cost of receivers, transmitters, and antenna systems, has rendered this a practical choice in many cases. Further, as we will see in Section 11.7.7, the use of GPS to provide highly accurate onboard navigation implies that precision tracking is no longer a required ground station function, reducing historical costs even further.

The choice of a given system, and indeed the issue of whether to build or buy tracking capability, will usually be driven by the sponsor and the mission and will have strong implications for the system engineering process.

11.7.6 Tracking Accuracy

In addition to their role in receiving spacecraft telemetry and providing a channel for command and data uplinks, tracking stations can be (and formerly were required to be) used to supply position and velocity data needed for the orbit determination algorithms discussed in Chapter 4 and detailed in common references such as Bate et al.[9] or Battin.[10] The accuracy of the resulting orbit and the ability to propagate the estimated orbit forward in time are ultimately limited by the tracking precision. For this reason it is instructive to consider state-of-the-art tracking system precision with respect to position and velocity determination and prediction.

Figure 11.24 shows a typical position estimate for a low Earth orbiting satellite as a function of time as *reconstructed* from a tracking data file. The emphasis on reconstruction is deliberate; note that the actual measurements are made at sporadically indicated times. These offer the greatest accuracy, on the order of 10 m for all three velocity components (selected here to be radial, in-track, and cross-track). In between the actual measurements, the ability of interpolation algorithms to specify precisely the spacecraft location grows poorer, reaching a maximum halfway between data points. Figure 11.25 shows similar behavior for reconstructed velocity.

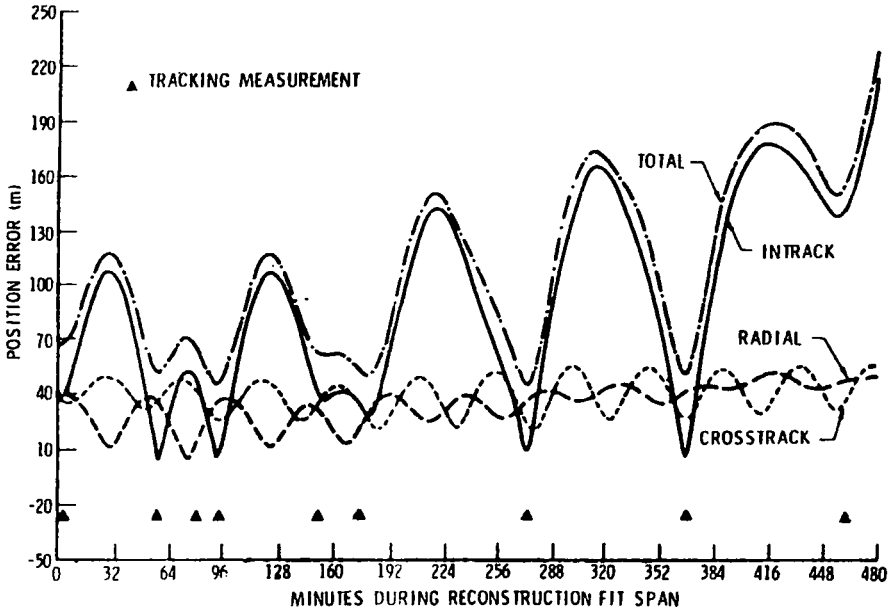


Fig. 11.24 Position reconstruction error estimate.

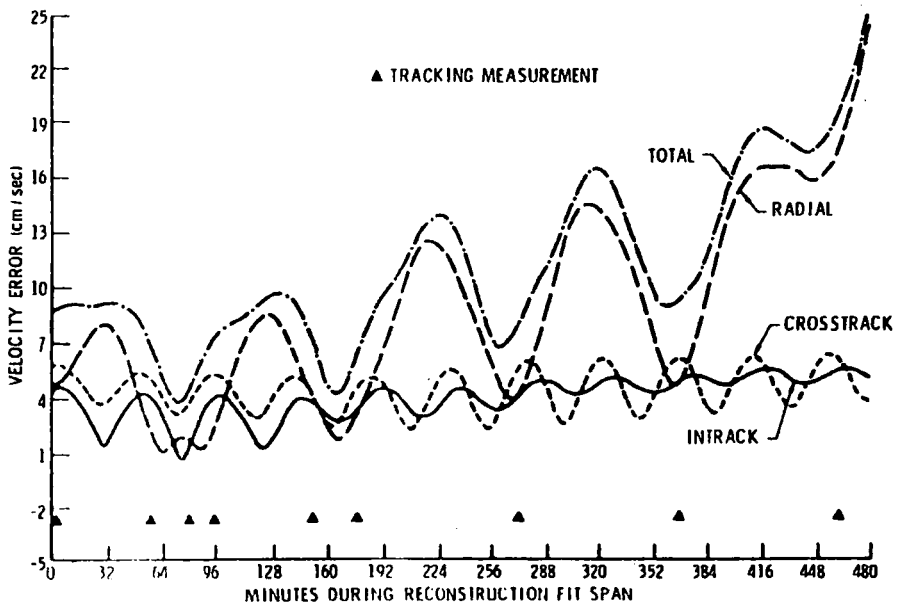


Fig. 11.25 Velocity reconstruction error estimates.

Figures 11.26 and 11.27 depict the ephemeris error growth in the hours following cessation of measurements. Note the rapid growth of in-track position error in Fig. 11.26. This is due primarily to the effect of atmospheric drag, which, as discussed in Chapter 4, is extremely difficult to quantify.

Probably the best single radar (or, for that matter, optical) tracking facility in the world is the U.S. Army's Reagan Test Site (RTS), formerly known as the Kwajalein Missile Range (KMR), located on various islands in the Kwajalein Atoll in the western Pacific Ocean.[11] Table 11.5 describes the capability of the various KMR radars for targets with and without beacons. Although these particular systems may not be available to, or even of interest in, most space operations, they are presented here to illustrate the accuracy to be expected from a good tracking radar. The optical tracking accuracy data are included for comparison purposes.

11.7.7 GPS Navigation

It is well beyond the scope of this text to explore in detail the technology of GPS navigation; indeed, we can only touch on the essentials of this increasingly ubiquitous technology. Many excellent texts are available, including the definitive work edited by Parkinson and Spilker,[12] and we refer the interested reader to these for details of interest.

As mentioned earlier, GPS navigation has rendered many earlier systems either obsolete or of marginal utility. While we must omit many details of interest

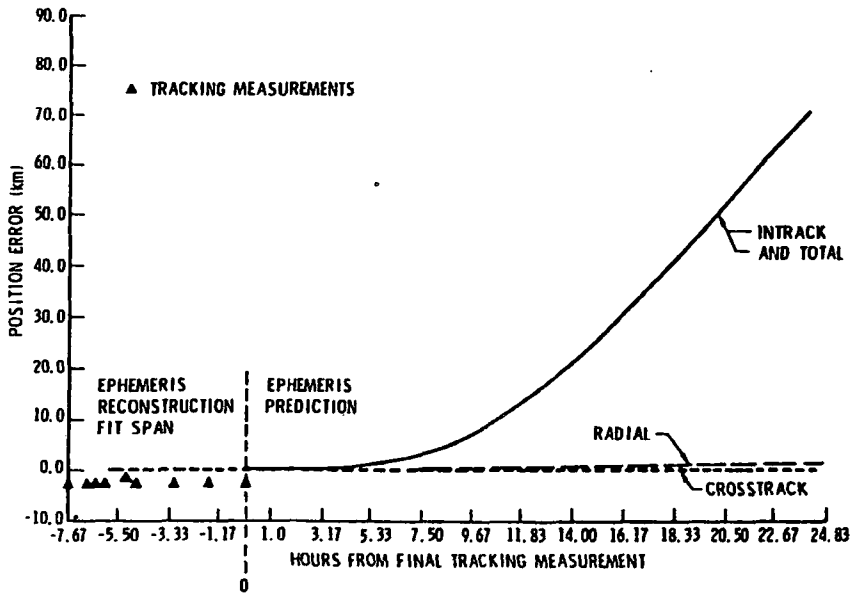


Fig. 11.26 Position error estimates reconstructed and predicted.

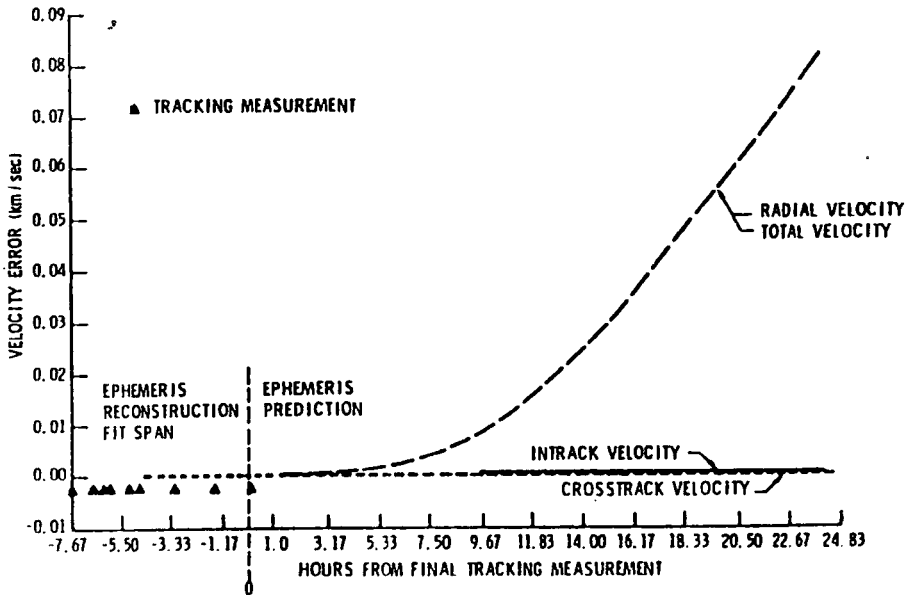


Fig. 11.27 Velocity error estimates reconstructed and predicted.

Table 11.5 Sensor metric accuracy at Kwajalein Missile Range[a]

	Accuracy, 1σ bias variability	
Sensor	Range, m	Angle, μrad
ALCOR (BCN)	1.0	100
(skin)	0.2	100
MMW (skin)	0.2	30
AN/MPS-36 (BCN)	7.0	150
TRADEX (skin)	2.0	100
ALTAIR (skin)	5.0	200
RADOT (optical)		70
SUPER RADOT (optical)		40
Ballistic Camera		60

[a]The data assume a "good target" (conical, symmetrical), with range and angle errors at pierce point. Position errors can be derived from the data but are trajectory dependent.

to the specialist in this field, a brief description of the system and the fundamentals of its operation may be of value. Strictly speaking, of course, GPS is not a "tracking" system at all. Rather, as we shall see, it is a system designed to allow a properly equipped Earth-orbiting spacecraft to determine accurately its state vector ($\boldsymbol{r}$, $\boldsymbol{V}$). This information can then be telemetered to the appropriate ground control center (possibly via TDRSS), eliminating the need for the tracking station to perform such a function. As already noted, if the spacecraft is to be tracked directly from the ground, then at least coarse state vector information must be available, for example from orbit injection conditions at launch vehicle separation, but usually this is not a difficult problem.

The GPS constellation is nominally comprised of 24 satellites, with four satellites in each of six circular Earth orbits, at altitudes of 20,335 km and inclinations of 55 deg, with the six orbital planes spaced equally in their ascending node locations. The chosen altitude provides two complete satellite revolutions per sidereal day, with between 6 and 11 spacecraft always in view above the horizon from any location on Earth. Only 21 operational spacecraft are required to guarantee full system performance, so that three may be viewed as on-orbit spares. (As this is written, the constellation actually contains 27 working spacecraft, with 10 of these having experienced degradation of some subsystems, rendering them "one failure away" in those subsystems from loss of mission.)

Each spacecraft has a very accurately known ephemeris and carries a very precise atomic clock, stable to one part in 10^{13} or better and synchronized via ground control with all other satellites, comprising what is known as "system time." The spacecraft continually broadcasts timing and ephemeris information on a set of L-band signals.

Let us assume that a satellite S_1 broadcasts from known position vector $\boldsymbol{R}_1$ in Earth-centered inertial coordinates at system time t_{s1}. A receiver with a clock synchronized to system time is positioned at unknown position vector $\boldsymbol{R}$ and receives the signal at time t_{r1}. Then the receiver-to-satellite range r_1 can be written as

$$r_1 = |\boldsymbol{R} - \boldsymbol{R}_1| = c(t_{r1} - t_{s1}) \tag{11.33a}$$

Similarly, receipt of signals from satellites S_2 and S_3 allows us to write

$$r_2 = |\boldsymbol{R} - \boldsymbol{R}_2| = c(t_{r2} - t_{s2}) \tag{11.33b}$$

and

$$r_3 = |\boldsymbol{R} - \boldsymbol{R}_3| = c(t_{r3} - t_{s3}) \tag{11.33c}$$

The system of Eqs. (11.33) is sufficient to solve for the three unknown components of receiver position vector $\boldsymbol{R}$, provided again that the receiver clock is accurately synchronized to system time and neglecting various other errors including relativistic effects, atmospheric refraction, the change in the speed of light within the atmosphere as opposed to free space, etc.

In practice, the receiver clock cannot normally be well synchronized with the satellite clocks (atomic clocks not being in common use) but will be offset by an unknown bias, which for the present discussion we assume to be constant over the course of the measurement interval. The determination of times t_{s1} through t_{s3} will be in error by the amount of this bias, δt, resulting in substantially degraded estimates of $\boldsymbol{R}$. However, use of a fourth (otherwise redundant) satellite signal allows the clock bias to be determined together with the components of the location vector, thus completing the solution.

In this discussion, we treat the navigation problem deterministically, i.e., we assume that there exist four unknown quantities, $\boldsymbol{R}$ and δt, and that we seek four equations to allow their determination. In practice, the techniques of optimal estimation theory are used to provide a "best guess" as to the unknown quantities, using as many measurements as can be obtained. Modern GPS receivers can handle essentially simultaneous receipt of signals from up to a dozen satellites (e.g., all that are in view) and can appropriately weight the additional information to obtain an optimal position estimate given all available data. Additional filtering allows estimation of velocity information as well, so that a full state vector can be computed at the receiver. The details of actual system implementation can become extremely complex, especially when great precision is desired, but this is the essence of the method.

Despite its origins within and continued maintenance by the U.S. Department of Defense, GPS has become a *de facto* global public utility. As this is written, military use in fact comprises less than 2% of all GPS use; automotive navigation by itself accounts for over a third of such use.[13] High quality portable navigation systems costing \$100–1000 are available for private and recreational uses including hunting, boating, aviation, golf, and of course automobile routing. Commercial and scientific applications include heavy equipment and shipping container tracking, surveying, geodesy, and an expanding host of other uses. Concerns over the dependence of GPS on U.S. military policy and operations have led to a European proposal for a similar but independent system, to be called *Galileo*.

The Russian Global Navigation Satellite System (GLONASS) operates in a fashion quite similar to that of the U.S. GPS; however, its performance is substantially degraded at present because of numerous satellite failures and the lack of Russian budgetary resources adequate to maintain the constellation. Still, several GPS receiver manufacturers make available surveying equipment offering dual-system capability. When high precision is required, inclusion of additional satellites in the measurement base improves the quality of the "fix," and so the ability to incorporate any functioning GLONASS satellites is helpful. The same will be true of the ability to use both GPS and *Galileo* satellites, should the latter system be developed.

The accuracy available from GPS navigation is impressive. Civilian systems (which lack access to the most accurate timing signals) provide a position fix to within a few tens of meters. Ultimate unaided system capability is on the order of

10 m. However, on a local or regional scale, very high accuracy, on the order of centimeters, is possible through the use of differential GPS (DGPS). The success of this concept relies upon the knowledge that most GPS errors, whether environmental or system-induced, are essentially the same across any area that is using basically the same group of satellites to determine a position fix. It is possible to determine these systematic errors by obtaining a GPS position measurement at a carefully surveyed known reference point in the desired area. Any deviation (which likely is time varying) of the GPS measurement from the known reference coordinates is therefore the result of the systematic errors, which now become known to the accuracy with which the reference position is known. These errors can in turn be broadcast locally and used by any suitably equipped GPS receiver in the area to correct its own measurements. As always, complications beyond the scope of this text remain to be explored, but the basic approach is easily implemented and quite useful.

As is always the case when space vehicle applications are considered, the use of GPS in space systems has lagged behind, and comes at a considerably greater cost, than its ground-based applications. Complicating factors include, of course, the higher level of environmental stress and parts qualification required. We have discussed these issues previously and need not comment further on them. With GPS, however, an additional difficulty arises in that the basic technical problem becomes substantially more difficult for space and missile guidance than for typical ground applications.

Most space vehicle applications will require determination of a full six-element state vector, i.e., position and velocity. The GPS system approach, as described, fundamentally allows a given user to determine only his position. Velocity is not directly measured, but must be inferred, basically by differentiating a series of position measurements. This is an inherently noisy process, requiring a higher level of estimator sophistication to achieve than when only position is required. Moreover, the fact of a receiver in motion at high velocity by itself impedes the convergence of the position estimator. The net result is that, in many respects, higher quality receivers are needed to perform GPS navigation for high-dynamics applications than otherwise, and space and missile systems are among the most demanding of these. In fact, for these reasons, GPS is more commonly used in space vehicle applications as an aid to, rather than as the sole means of, navigation. Often the GPS data will be used in concert with onboard inertial navigation, which is much better suited to the high-dynamics environment.

An additional complexity associated with the use of GPS on atmospheric entry vehicles such as the space shuttle is the requirement to cope with intermittent loss of signal as a result of the ionized plasma sheath that is generated during high-speed atmospheric flight. The shuttle encounters periods of degraded GPS performance for up to several minutes during entry flight, from about 100 km to 80 km altitude, with some effects persisting down to about 60 km.[14] Unaided

GPS receiver systems flown on the shuttle on an experimental basis have so far been unable to maintain continuous tracking throughout this phase of flight.[15]

Table 11.6 provides performance parameters achievable, as this is written, by spacecraft-quality GPS navigation devices. The terminology "SA off" refers to a "selective availability," a DoD-enabled feature that "dithers" the signal to prevent maximum accuracy being attained by other than military users in the event of a conflict. The "cold start" acquisition time refers to the case in which the unit is activated but has no initial position fix. The more typical in-flight situation is the "warm start," wherein the GPS must reacquire a state vector following a maneuver, a computer reset, or other event that upsets the state vector estimate, but allows use of recent values as a starting point for a new solution.

Space systems applications of GPS are only now developing beyond straightforward use of the navigation state vector ($\boldsymbol{r}$, $\boldsymbol{V}$). Even here, the full suite of possibilities is only gradually emerging. In the immediately foreseeable future, it is reasonable to suppose that the use of tracking stations for spacecraft and launch vehicle operations, at least for those in Earth orbit, and for range safety assessment will become obsolete. The spacecraft or launch vehicle itself will know its navigation state to within a few tens of meters and can telemeter this information to the ground. If range safety constraints are violated, the vehicle can be programmed to destroy itself. It will be a matter of careful engineering design to ensure adequate redundancy in such systems, but the economic benefits would seem to ensure that ultimately it will be done.

As mentioned in Chapter 7, it is also possible to use GPS for spacecraft attitude determination. Indeed, for large spacecraft (because of the lengthy baseline between separate antennas) it is almost trivial and may well represent the cheapest alternative. The method has already been demonstrated on certain experimental unmanned aerial vehicles. Basically, the approach involves fixing several antennas to widely separated portions of the spacecraft. Knowing the spacecraft ephemeris, position differences between the various antennas are immediately convertible to spacecraft attitude angles.

Table 11.6 Space-qualified GPS receiver capability

Parameters	Performance
Position accuracy (SA off, 1σ)	30 m
Velocity accuracy (SA off, 1σ)	± 1 m/s
Output data rate	1 Hz
Acquisition time (cold start)	<5 min
Acquisition time (warm start)	<3 s
Satellites tracked	1–12 (i.e., all in view)
Power	<5 W
Mass (including two antennas)	<2 kg
Volume (excluding antennas)	<1400 cm^3

11.7.8 Optical Navigation

The key advantage in using GPS navigation onboard a spacecraft or launch vehicle lies, of course, in the ability of the vehicle itself to determine its navigation state and, as required, relay it to the ground. Although GPS should prove adequate for near-Earth spacecraft in the foreseeable future, this capability is inapplicable to interplanetary missions, as we have noted on several prior occasions. Thus, the JPL DSN has remained fully as essential to the conduct of interplanetary space programs as in the earliest days of spaceflight. For purposes of communication, this is likely to remain the case until optical laser communication links become practical. However, it may well be that the development of autonomous optical spacecraft navigation can in many cases remove or reduce the burden of providing tracking data for interplanetary spacecraft.

Autonomous optical navigation (AON) has been a "holy grail" of the interplanetary mission community for decades, and rightly so. In addition to the high cost and limited availability of DSN-based navigation, the entire human-in-the-loop approach to spacecraft tracking and navigation as it is currently practiced is fraught with the potential for error. The most recent example is that of the Mars Climatology Orbiter (MCO). The primary cause of the loss of MCO was attributed[1] to a discrepancy between JPL and the spacecraft contractor in the system of measurement units employed by each, with JPL working in SI units and the contractor in English units. An erroneous conversion factor between the two systems was applied, resulting in trajectory correction maneuvers that were smaller than commanded. The resulting errors were detectable, but small, and in the event went unnoticed, leading to atmospheric entry of the spacecraft rather than injection into orbit upon arrival at Mars.

MCO is far from the only spacecraft to have been put at risk, or lost, through errors (navigational or otherwise) by ground controllers. It is a truism of the human engineering discipline that humans perform most poorly on repetitive tasks that are viewed as simple or mundane.

The underlying method in AON is essentially identical to that for orbit determination of spacecraft or natural objects from Earth-based optical measurements. Discussed briefly in Chapter 4, efficient methods due to Laplace and Gauss date back fully 200 years, and aside from the use of modern estimation algorithms which are better adapted for computer application than some of the older techniques, these methods are little changed today. (It is worth noting, and perhaps a bit humbling, to realize that the original Gauss method of orbit determination, when applied to problems of the class for which it was developed, remains among the most computationally efficient techniques yet developed.)

The technique is, then, from the vantage point of the spacecraft to observe optically the angles between the fixed stars and relatively nearby planetary bodies (e.g., Earth, Mars, the moon, or prominent asteroids). Because the ephemerides of the planets are known, the orbit of the spacecraft can be determined. The

difficulty, of course, is to attain adequate accuracy with sensors that are smaller than typical ground-based telescopes, and to do so in relatively short amounts of time.

Optical navigation has the key advantage that, as the spacecraft approaches its target, the various systematic errors grow smaller, thereby enhancing the quality of the encounter, at least in theory. Certainly this is an inherent advantage relative to tracking with Earth-based sensors, which becomes more difficult in a planetary-relative sense as the spacecraft approaches its target. For these reasons, optical navigation has been an important part of most planetary missions since the Voyager encounters with Jupiter and the other outer planets and their moons. However, interpretation of the data and closure of the navigation loop was always accomplished on the ground.

On *Deep Space 1*, it was attempted for the first time to perform AON directly onboard the spacecraft, during a July 1999 encounter with the asteroid Braille.[16] While the effort met with only partial success, it would seem that the potential rewards attendant to this method of tracking interplanetary spacecraft will continue to justify the research necessary to advance the state of the art to useful levels.

References

[1]Young, A. Thomas, "Mars Program Independent Assessment Team Report," NASA, March 2000.

[2]Fthenakis, E., *Manual of Satellite Communications*, McGraw-Hill, New York, 1984.

[3]Skolnik, M. I. (ed.), *Radar Handbook*, McGraw-Hill, New York, 1970.

[4]Slobin, S. D., "Atmospheric and Environmental Effects," DSMS Telecommunications Link Design Handbook, Jet Propulsion Lab., Doc. 810-005, Rev. E, Pasadena, CA, Jan. 2001.

[5]Shannon, C. E., "A Mathematical Theory of Communications," *Bell System Technical Journal*, Vol. 27, 1948, pp. 379–423 and 623–651.

[6]Martin, J., *Communications Satellite Systems*, McGraw-Hill, New York, 1978.

[7]Viterbi, A. J., and Omura, J. K., *Digital Communications and Coding*, McGraw-Hill, New York, 1979.

[8]"Ground Network User's Guide," NASA Doc. 452-GNUG-GN, Feb. 2001.

[9]Bate, R. R., Mueller, D. D., and White, J. E., *Fundamentals of Astrodymanics*, Dover, New York, 1971.

[10]Battin, R. H., *An Introduction to the Mathematics and Methods of Astrodynamics*, AIAA Education Series, AIAA, New York, 1987.

[11]"Kwajalein Missile Range Instrumentation and Support Facilities Manual," U.S. Army Space and Missile Defense Command, Huntsville, AL, Jan. 2000.

[12]Parkinson, B. W., and Spilker, J. J., Jr. (eds.), *Global Position System: Theory and Applications I & II*, Vols. 163 and 164, Progress in Astronautics and Aeronautics, AIAA, Reston, VA, 1996.

[13]U.S. Department of Commerce, Washington, DC, 2002.

[14]Goodman, J. L., Space Shuttle Navigation in the GPS Era," Institute of Navigation 2001 National Technical Meeting, Long Beach, CA, 22–24 Jan. 2001.

[15]Gomez, S. F., "Flying High – GPS on the International Space Station and Crew Return Vehicle," *GPS World*, Vol. 13, No. 6, 2002.

[16]Riedel, J. E., "Automated Optical Navigation (AutoNav) Technical Validation Final Report," Jet Propulsion Lab., Pasadena, CA, Feb. 2000.

[17]Meer, D. E., "Noise Figures," *IEEE Transactions on Education*, Vol. 32, May 1989, pp. 66–72.

Problems

11.1 The Apollo CSM transmitted 10 W at S-band (2 GHz) through a 2-m parabolic dish antenna from lunar orbit, a distance of 400,000 km from Earth.

(a) What was the received signal power at the Goldstone 60-m dish antenna?

(b) What was the signal power expressed in dB_W?

11.2 It is necessary to send a 1 Gbps digital video signal through a satellite-to-ground link with a SNR of at least 7. At which of the established space communications bands (L, S, C, X, Ku, Ka, etc.) would you recommend this be done? Why not lower or higher?

11.3 From geostationary orbit ($r = 42{,}164$ km) we require a SNR of 7, with 10 W being transmitted through a 2-m dish antenna at 30 GHz. What receiving system G/T is needed to obtain a usable bandwidth of 0.5 GHz? Express your final answer in units of dB/K.

11.4 An Earth ground station uses a 4.5-m parabolic dish antenna to receive a 4-pw signal from a geostationary communications satellite. Assume that the satellite and ground stations use a typical C-band transponder channel bandwidth of 36 MHz.

(a) If the composite system noise temperature for the ground station is 290 K, what is the SNR?

(b) What is the SNR if the Earth station uses a 2.25-m-diam antenna?

11.5 An existing QPSK encoded satellite-to-ground telecommunications system is operating with a BER of 10^{-6} and must be upgraded to a BER of 10^{-9}. What would be a simple way to do this? Justify your answer analytically or graphically.

11.6 An interfering signal spreads 1 pw (10^{-12} W) across a 36-MHz C-band transponder channel. What is the effective noise temperature of this interference? Could a typical receiving system be expected to tolerate it, by comparison with other naturally induced noise sources? Why or why not?

11.7 Voyager 2 encountered Uranus during 1986 at a distance from Earth of about 20 A.U. and has a transmitter power P_t of 20 W in X-band (8.4 GHz). The spacecraft has a 3.7-m parabolic dish antenna, while the receiving station at Goldstone has a 64-m aperture.

(a) What was the beamwidth, θ, of the transmitted signal?
(b) What was the received power at the Goldstone antenna?
(c) Assuming SNR = 7 at Uranus, and assuming the noise floor at the receiving station remained essentially constant over the duration of the Voyager primary mission, what was the ratio of the Voyager data rate at Uranus compared to the Saturn encounter at about 10 A.U. from Earth?

11.8 Assume a Mars lander carries an 8-GHz X-band transmitter with a nominal 10-W power output, feeding a 1.0-m-diam parabolic dish antenna, for which the pointing accuracy is sufficient to allow operation inside the half-power beamwidth. The DSN will be used for communications; for purposes of discussion we will assume a composite receiving system noise temperature of 10 K. A SNR_{dB} of 6 dB is required to achieve the desired bit error rate.

(a) Assuming a 70-m-diam antenna is used at the receiving station, and assuming reasonable conditions (i.e., minimum 5° elevation angle of the receiving antenna, no rain, the sun and moon are not in the antenna field of view, etc.), what data rate can be supported when Mars is at maximum distance from the Earth, about 2.5 A.U.?
(b) What data rate is possible with the 70-m dish when Mars is roughly at closest approach, about 0.5 A.U.?
(c) If the maximum data rate the lander telemetry system can support is 56 kbps, what is the smallest DSN antenna that can be used to obtain the data when Mars is at closest approach?

11.9 An S-band (2.2 GHz) Earth ground station has a system noise temperature of 190 K and a gain of 46 dB relative to isotropic (dB_i). A 10-dB SNR is required to obtain the bit error rate specified by the receiver manufacturer. What data rate can be supported for a spacecraft in a 1000-km altitude circular orbit, assuming that the station can track to within 5° of the local horizon?

11.10 Calculate the spacecraft effective isotropic radiated power (EIRP) required to produce a BER of 10^{-6} at 120 Mbps with an Earth station G/T of 40.7 dB/K. The downlink frequency is 4 GHZ, and the modulation is QPSK. A margin of 2.5 dB is required over the ideal theoretical performance.

12
Reliability Analysis

12.1 Introduction

An important measure of the worth of any engineering system is its reliability, which we define as the probability that the system will function as expected. Space systems tend to be quite expensive as well as unusually complex and are not repairable other than in highly unusual circumstances, and so it is both extremely important and very difficult to ensure that a spacecraft is highly reliable. Attaining the desired degree of space vehicle reliability, and trading incremental improvements in reliability against incremental costs, is the overall responsibility of the spacecraft system engineer. As with other aspects of system design, he will be aided in this effort by a knowledgeable specialist, in this case a safety, reliability, and quality assurance (SR&QA) engineer. As one of the so-called "ilities" (including such disciplines as maintainability, producibility, and operability), reliability analysis often lacks the visibility associated with seemingly more prestigious specialties. However, the cost of unreliability, both in lives and property damage, can be far greater than that of a suboptimal, but generally workable, design in almost any other subsystem.

In this chapter, we will develop an understanding of the tools and methods of reliability analysis and illustrate the use of these methods through the use of relatively simple examples. To do this, we begin with a treatment of the fundamentals of probability theory and the analysis of random variables. With these tools, the system engineer will be equipped to deal with the most common features of engineering reliability analysis. However, we offer the caution that, within the scope of this text, we can provide only the briefest outline of this very rich branch of mathematics. The interested reader is referred to more comprehensive works.[1–3]

We will also address the basic properties of random processes in Appendix A. While this latter topic is not strictly germane to the subject of reliability analysis at the level presented here, it is highly relevant to portions of Chapters 3, 7, 8, and 11, and is best introduced in the present context rather than piecemeal throughout the earlier material.

12.2 Review of Probability Theory

Most people, including most engineers, freely employ the language of probability and statistical theory in everyday conversation without considering the rigorously correct definitions of these terms. Thus, we commonly refer to "random" failures and speak casually of the "probability" of such events occurring. We possess, or at least believe that we possess, an intuitive understanding of the meaning of such terms that is sufficient for many purposes. This is indeed fortunate, because substantial mathematical sophistication is required to provide a rigorous definition of a random event or of the probability of such an event occurring. Most working definitions of the fundamental terms in probability theory are circular in the extreme, rooted in the very intuition that is sought to be avoided through the use of mathematics. In this text, where the application of knowledge is more important than its strictly rigorous development, we will appeal to these intuitively acceptable concepts to ground our discussion. Where necessary, we will try to clarify the limits of applicability of such concepts.

In this spirit, we offer as the definition of a random event an experiment for which the outcome cannot be modeled by a cause-and-effect relationship. For our purposes, it is sufficient to ignore the differences between an event that appears random due to an observer's ignorance or computational limitations, and an event whose underlying nature is intrinsically random, such as the outcome of a quantum experiment.

Continuing, we define a sample space as the set of all possible outcomes of a random event. In the simple case of a coin toss (the random event), the sample space consists of the set of possible outcomes, heads or tails, provided that we believe it is impossible for the coin ever to come to rest on its edge. In set theory notation, we might write this as $S = \{H, T\}$. For the random event of rolling a single die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. The sample space can be continuous as well as discrete. For example, if the random event is the measurement of the height of an adult human male, we might say that $S = \{h \in \Re, 2\,\text{ft} < h < 10\,\text{ft}\}$, i.e., the height h is a real number in the range between 2 and 10 ft. A given sample space may be empty, i.e., it may consist of the null set, $\emptyset$.

We define the probability of an event through a set of axioms. We employ standard set-theory notation, with $A \cup B$ denoting the union of sets A and B, and $A \cap B$ their intersection. The notation $A \cap B$ is read "A and B," while $A \cup B$ is read "A or B." The "or" in this latter case is the inclusive or, meaning A or B or both. Figure 12.1 illustrates the concepts involved.

The fundamental axioms of probability are:

1) $P(A) \equiv$ probablity of an event A; $0 \leq P(A) \leq 1$. Note that the event A may be defined as a group of several individual events, as, for example, the event of rolling either a 7 or 11 in the game of craps.

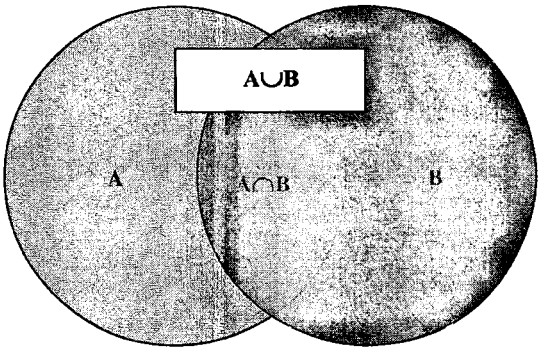


Fig. 12.1 Venn diagram for events *A* and *B*.

2) $P(A \cup B) \equiv P(A) + P(B)$ if and only if $A \cap B = \emptyset$ (A and B are mutually exclusive).

3) $P(S) \equiv 1$ (some outcome *must* occur).

From these axioms one can deduce several basic theorems. For example, suppose that events A and B are *not* mutually exclusive, i.e., $A \cap B \neq \emptyset$. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{12.1}$$

Further, if A and B are not mutually exclusive, then the knowledge that one event has occurred influences, or conditions, our expectation as to the probability of occurrence of the other event. We define

$P(A/B) \equiv$ conditional probability of event A given that event B has occurred

and we note $P(A/B) = 0$ if events A and B are mutually exclusive. Then,

$$P(A \cap B) = P(A)P(B/A) = P(B)P(A/B) \tag{12.2}$$

If A and B are independent, it is clear that $P(A/B) = P(A)$ and $P(B/A) = P(B)$, hence

$$P(A \cap B) = P(A)P(B) \qquad \text{(independent events)} \tag{12.3}$$

Now suppose there exists an event A that is dependent on several possible but mutually exclusive events B_j, $j \in [1, N]$. Then, using the preceding results on conditional probabilities, it can be shown that

$$P(A) = \sum_{j=1}^{N} P(A/B_j)P(B_j) \tag{12.4}$$

This is the Law of Conditional Probabilities. The simple example given in Example 12.1 may help to illustrate its utility in reliability analysis.

Example 12.1

A manned space launch system has a generic reliability, or probability of success, of 0.98. An abort system for the crew module is provided and has a reliability of 0.95. What is the overall probability of crew survival?

Solution: Let A = event of crew death, B_1 = event of launch vehicle success, and B_2 = event of launch vehicle failure. Note that B_1 and B_2 are mutually exclusive, and that

$P(B_1) = 0.98$	$P(A/B_1) = 0$	(abort system not needed)
$P(B_2) = 0.02$	$P(A/B_2) = 0.05$	(abort system fails)

Then from the Law of Conditional Probabilities,

$$P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) = (0.98)(0) + (0.02)(0.05) = 0.001$$

The reliability of crew survival is then

$$R_S = 1 - P(A) = 0.999$$

i.e., the crew has a 99.9% chance of survival, even though neither the launch vehicle nor the abort system is anywhere close to being 99.9% reliable.

The example provides considerable insight into the benefits of redundancy, i.e., the provision of independent, parallel systems to accomplish a given function (in Example 12.1, the required function is ensuring crew survival, not reaching orbit). Upon closer examination, it provides equal insight into one of the hazards of relying on the predictions of an analytical model, that is, the conclusions are only as valid as the assumptions underlying the model.

In Example 12.1, we explicitly assume that launch vehicle and abort system failures are independent events. Such an assumption can be extremely difficult to realize in practice and equally difficult, if not impossible, to verify. It is a simple matter to postulate launch vehicle failure modes that result in damage to the crew module abort system or, for that matter, to note that some types of abort system failures could result in destruction of the launch vehicle. The launch and abort systems would then not be independent and would then not be, from a crew survival viewpoint, fully redundant. In all likelihood, a 99.9% total system reliability for crew survival would be extremely difficult to achieve.

Two of the greatest hazards in reliability analysis are the accurate determination of the underlying failure probabilities of components, subsystems and systems, and the problem of ensuring that the failure modes of a given item and its effects on the remainder of the system are sufficiently well understood that the reliability model accurately represents the actual system. Both of these issues are addressed in later sections of this chapter.

The production volume of typical space systems greatly limits the accuracy with which most reliability analysis can be performed. For example, no launch system in existence has flown sufficiently often, in an identical configuration from launch to launch, to allow failure probabilities to be as well established as just implied. An abort system would typically be subject to much less testing than the launcher, and most such tests would, by their nature, be under conditions similar but not identical to those encountered after a true launch vehicle failure, of which there could in any case be many types. Excellent and subtle methods of statistical inference exist to establish reliability information from limited samples of data. We will touch on this subject in a later section. However, no statistical method can yield really accurate results from the sample sizes typical of space flight systems. It is crucially important that the system engineer understand these inherent limits on any reliability analysis that is performed.

This discussion illustrates a simple case in which an a priori failure analysis or prediction is required. In other cases, we may know or postulate that a failure occurs and wish to know the conditional probability that it is due to particular cause B_j. To determine this, we note from Eq. (12.2) that

$$P(A \cap B_j) = P(A)P(B_j/A) = P(B_j)P(A/B_j) \tag{12.5}$$

from which we obtain Bayes' Theorem,

$$P(B_j/A) = \frac{P(B_j)P(A/B_j)}{P(A)} \tag{12.6}$$

The various $P(B_j/A)$ are called a posteriori probabilities. As before, the events B_i must be mutually exclusive. Again, an example, given in Example 12.2, best serves to illustrate the potential uses of Bayes's Theorem.

Example 12.2

A spacecraft prime contractor obtains Ni-Cd batteries from three different sources: 50% come from source 1 and have a defective proportion of 0.3%; 30% are obtained from source 2 and are 0.2% defective; the remaining 20% are procured from source 3 and are 0.6% defective. A given battery fails during acceptance testing. What is the probability that it came from source B_3?

Solution: Let

B_1 = event that a battery is from source 1
B_2 = event that a battery is from source 2
B_3 = event that a battery is from source 3
P = event that a battery passes
F = event that a battery fails

From the statement of the problem,

$$\begin{array}{lll} P(B_1) = 0.50 & P(B_2) = 0.30 & P(B_3) = 0.20 \\ P(P/B_1) = 0.997 & P(P/B_2) = 0.998 & P(P/B_3) = 0.994 \\ P(F/B_1) = 0.003 & P(F/B_2) = 0.002 & P(F/B_3) = 0.006 \end{array}$$

Using Eq. (12.4), we find

$$\begin{aligned} P(P) &= P(B_1)P(P/B_1) + P(B_2)P(P/B_2) + P(B_3)P(P/B_3) \\ &= (0.5)(0.997) + (0.3)(0.998) + (0.2)(0.994) \\ &= 0.9967 \\ P(F) &= P(B_1)P(F/B_1) + P(B_2)P(F/B_2) + P(B_3)P(F/B_3) \\ &= (0.5)(0.003) + (0.3)(0.002) + (0.2)(0.006) \\ &= 0.0033 \\ &= 1 - P(P) \end{aligned}$$

We are given that the battery is defective, and so from Bayes's Theorem,

$$P(B_3/F) = \frac{P(B_3)P(F/B_3)}{P(F)} = \frac{(0.2)(0.006)}{0.0033} = 0.3636\ldots$$

i.e., the probability is about 36% that the defective battery is from source B_3.

12.3 Random Variables

A random variable is a real-valued function on the sample space S that assigns a numerical value according to the outcome of a random event. We denote the random variable as X, with x taken as the value of X in a specific instance. Random variables may be either discrete, with

$$P(X = x_i) \equiv f(x_i) \tag{12.7}$$

or continuous, in which case

$$P(X = x) \equiv \lim_{\Delta x \to 0} f(x)\Delta x \tag{12.8}$$

where $f(x)$ is the probability density function. In both cases $0 \leq f(x) \leq 1$. The probability density function satisfies

$$\sum_{i=1}^{N} f(x_i) = 1 \tag{12.9}$$

or in the continuous case

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \tag{12.10}$$

We define the probability distribution function as

$$F(x_n) \equiv \sum_{i=1}^{n} f(x_i) = P(X \leq x_n);\ n \in [1, N] \tag{12.11}$$

or

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = P(X \leq x) \tag{12.12}$$

Equivalently, of course, the density function may be obtained from a given distribution function as

$$f(x) = \frac{dF(x)}{dx} \tag{12.13}$$

The expected value (also first moment, mean, or average) of a random variable is

$$E(X) \equiv \sum_{i=1}^{N} x_i f(x_i) \equiv \mu \tag{12.14}$$

for a discrete random variable, and

$$E(X) \equiv \int_{-\infty}^{+\infty} xf(x)\,dx \equiv \mu \tag{12.15}$$

when X is continuous. The second moment of a random variable is defined by

$$E(X^2) \equiv \sum_{i=1}^{N} x_i^2 f(x_i) \tag{12.16}$$

or

$$E(X^2) \equiv \int_{-\infty}^{+\infty} x^2 f(x)\,dx \tag{12.17}$$

for the discrete and continuous cases, respectively. The commonly used root mean square, or rms, value of a random variable is given by the square root of Eqs. (12.16) or (12.17).

The variance of a random variable is defined as

$$\sigma^2 \equiv \sum_{i=1}^{N} [x_i - E(x)]^2 f(x_i) \tag{12.18}$$

or

$$\sigma^2 \equiv \int_{-\infty}^{+\infty} [x - E(x)]^2 f(x)\, dx \tag{12.19}$$

It is easily shown that

$$\sigma^2 = E(X^2) - E^2(X) \tag{12.20}$$

The variance (or its square root, σ, the standard deviation) of a random variable X is a measure of the deviation or spread about its mean value. Two random variables X and Y having the same mean but different standard deviations are shown in Fig. 12.2. X and Y might represent, as one of many possible examples, the measurement of the range to an orbiting satellite using two different techniques. There is a true underlying value of the range to the spacecraft, knowledge of which is corrupted in both cases by measurement errors and uncertainties, collectively considered to be "noise." If the noise itself has a mean value of 0, then a sufficiently accurate approximation to the underlying range can be obtained by averaging a large enough group of measurements. With a very large number of measurements available, either X or Y would prove to be

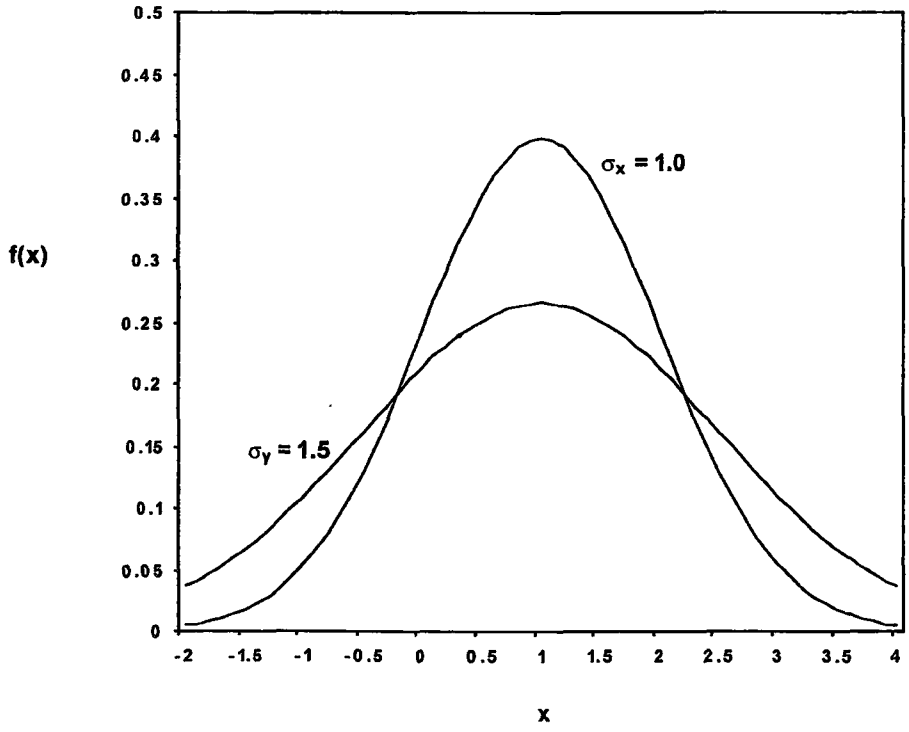


Fig. 12.2 Random variables with identical mean and different variance.

a suitable technique. However, in few cases do we have such an excess of measurement data, and so in general lower variance estimates are to be preferred. In this case, X would be considered to provide a more accurate measurement than Y, because any individual measurement of X would have a much higher probability of being near the true value (the mean) than Y.

One often hears the terms rms and standard deviation discussed as if they were the same; however, as Eq. (12.20) makes clear, the equality holds only for zero-mean random variables.

It can be of interest to examine events depending on the values of two or more random variables, a situation to which we have already alluded. The probability distribution (or density) function for such a case must then depend simultaneously upon two or more random variables. We are led to define an n-dimensional random vector $\boldsymbol{X}$ of random variables as

$$\boldsymbol{X} \equiv [X_1, X_2, \ldots, X_n]^T \tag{12.21}$$

for which there will exist joint distribution and density functions for the vector components X_1 through X_n given by

$$F(x) \equiv F(x_1, x_2, \ldots, x_n) = P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n] \tag{12.22}$$

and a joint density function defined as

$$f(\boldsymbol{x}) \equiv f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \, \partial x_2 \cdots \partial x_n} \tag{12.23}$$

The expectation operator can be applied in a straightforward manner to generate various moments of the joint density function, i.e., the mean or first moment of $\boldsymbol{X}$ is given by

$$\boldsymbol{\mu} \equiv E(\boldsymbol{X}) \equiv [E(X_1), E(X_2), \ldots, E(X_n)]^T \tag{12.24}$$

Higher order moments can be generated as well; we define the covariance of the random variable $\boldsymbol{X}$ as

$$\boldsymbol{P} \equiv E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] \tag{12.25}$$

$\boldsymbol{P}$ is an $n \times n$ matrix composed of elements σ_{ij}, where

$$\sigma_{ij} \equiv \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) f(x_1, x_2, \ldots, x_n) \, dx_i \, dx_j \tag{12.26}$$

Obviously,

$$\sigma_{ii} \equiv \sigma_i^2 = E(X_i^2) - E^2(X_i) = \text{variance of } i\text{th random variable} \tag{12.27}$$

Conceptually, the elements of the covariance matrix express the degree of mutual correlation between the ith and jth random variables. If X_i and X_j are completely uncorrelated, their product can be expected to be negative as often as it is

positive, and so the integral taken over the range $(-\infty, \infty)$ will be zero. Conversely, if X_i and X_j are correlated, they will to some extent vary in phase, their product will tend to be predominantly of one sign or the other, and their integrated product will be non-zero. The magnitude of the σ_{ij} provides a measure of this correlation, especially when appropriately normalized. To do this, we define the correlation coefficient

$$\rho_{ij} \equiv \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{12.28}$$

which varies over the range $[-1, +1]$, a result that depends on having defined $\boldsymbol{P}$ in Eq. (12.25) in terms of the zero-mean random variable $(\boldsymbol{X} - \boldsymbol{\mu})$.

We note in passing that while independent random variables X_i and X_j are clearly uncorrelated ($\rho_{ij} = 0$), the converse is not true; lack of correlation does not imply independence. A trivial example is provided by the sine and cosine functions, which are completely dependent in the statistical sense (knowledge of $\sin \theta$ allows immediate calculation of the value of $\cos \theta$), but whose integrated product over $(-\infty, \infty)$ is zero, i.e., they are uncorrelated.

If X_i and X_j are independent, then the joint probability density function is obtained very simply as

$$f(x_i, x_j) = f_i(x_i) f_j(x_j) \tag{12.29}$$

12.4 Special Probability Distributions

There are several special probability distribution (or density) functions that assume particular importance in probability and statistics generally, and with system reliability analysis and redundancy management in particular. We consider some of these special distributions in the sections that follow.

12.4.1 Gaussian Distribution

The most important probability distribution in practical applications is the Gaussian or normal distribution. The density function for the Gaussian distribution is illustrated in Fig. 12.3 and is given by

$$f(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{12.30}$$

The mean μ and standard deviation σ completely characterize the distribution. For this reason, the notation $X = N(\mu, \sigma)$ is often used to indicate that the random variable X is normally distributed with mean μ and variance σ^2. For convenience, the probability distribution function, which is of course the area under the curve of Eq. (12.30), is tabulated in common references[4] for the standard normal random variable $Z \equiv N(0, 1)$. Any Gaussian random variable X can be so

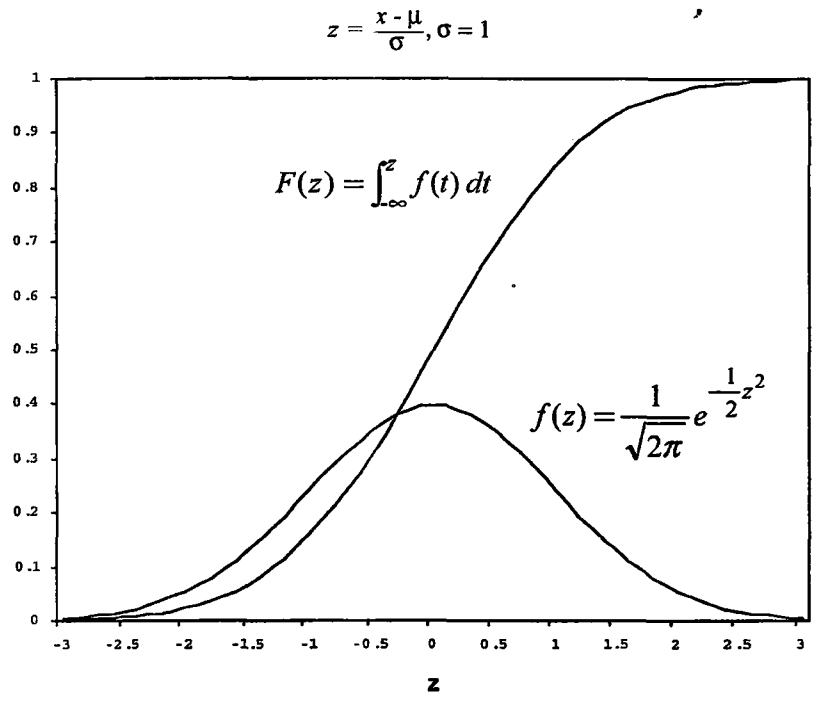


Fig. 12.3 Gaussian probability density and distribution functions.

expressed by means of the independent variable transformation $z = (x - \mu)/\sigma$. "Applets" for the computation of Gaussian probabilities are also widely available on line.

The importance of the normal distribution arises from a number of sources. The sum of independent Gaussian random variables is itself Gaussian. A linear system acting upon an input signal that is a Gaussian random variable produces as its output another Gaussian distributed signal. The Gaussian distribution is representative of an enormous number of real-world processes, a consequence of the central limit theorem of statistics. Loosely speaking, the central limit theorem demonstrates that, under certain conditions whose details are not important here, the sum of a sufficiently large number of individual probability distributions is, in the limit, a Gaussian distribution. It is emphasized that this result is true *whether or not the underlying distributions are themselves Gaussian*! Because most practical systems are susceptible to a large number of underlying noise or error sources that are individually unknown, and unknowable, the central limit theorem assures us that in such cases, noise or errors may be characterized as normally distributed. Finally, the normal distribution is among the more analytically tractable distributions, a factor that should not be underestimated when it is necessary to obtain a real conclusion as opposed to a theoretically intriguing but computationally useless result.

12.4.2 Uniform Distribution

As the name implies, the uniform distribution implies an equal probability of selecting $X = x$ across the valid domain of x. Thus,

$$f(x) = \frac{1}{(b-a)}, \qquad a \leq x \leq b \tag{12.31}$$

12.4.3 Binomial Distribution

A common problem in reliability analysis concerns questions of the following type: A process results in the production of items, a certain proportion, q, of which will be found to be defective. Suppose that a group of n such items is selected. What is the probability that $k \leq n$ will be defective? The answer to this question gives rise to the binomial distribution. We should consider first a trivial example, Example 12.3, prior to stating the more general theory.

Example 12.3

Suppose a spacecraft needs three working gyroscopes to complete its mission, and that the selected gyros have a 90% chance of surviving the required mission duration. The spacecraft is initially equipped with four such gyros. What is the probability of successfully completing the mission?

Solution: Note that we seek the probability of $k \leq 1$ defectives from a group of $n = 4$ gyros selected from a population containing a proportion $q = 0.1$ defective gyros, i.e., those that will not survive for the required mission duration. With four gyros, there are 16 possible, mutually exclusive, combinations of good (G) and bad (B) gyros:

GGGG—no defective gyros (event A)

GGGB—one defective gyro (event B)
GGBG—" " "
GBGG—" " "
BGGG—" " "

GGBB—two defective gyros (event C)
GBGB—" " "
BGGB—" " "
GBBG—" " "
BGBG—" " "
BBGG—" " "

BBBG—three defective gyros (event D)
BBGB—" " "

BGBB—" " "
GBBB—" " "

BBBB—four defective gyros (event E)

The probability of initially selecting a single good gyro is $p_G = 0.9 = 1 - q$; because the quality of each gyro is independent of all others (i.e., gyros fail from random causes after some period of use, rather than due to a systematic design flaw), each subsequent selection is an independent event, and so from Eq. (12.3) the probability of selecting four good, and no bad, gyros is

$$p(A) = p_{4,0} = (0.9)^4 = 0.6561 = \text{probability of no failed gyros}$$

Similarly, three good and one bad gyro (event B) may be selected with probability

$$p_{3,1} = (0.9)^3(0.1)^1 = 0.0729$$

Now, because no one cares *which* gyro fails, there are four possible ways to do this, and so

$$p(B) = 4p_{3,1} = 0.2916 = \text{probability of one bad gyro}$$

Similarly,

$$p(C) = 6p_{2,2} = (6)(0.9)^2(0.1)^2 = 0.0486 = \text{probability of two bad gyros}$$

$$p(D) = 4p_{1,3} = (4)(0.9)^1(0.1)^3 = 0.0036 = \text{probability of three bad gyros}$$

$$p(E) = p_{0,4} = (0.9)^0(0.1)^4 = 0.0001 = \text{probability of four bad gyros}$$

Note, as required,

$$p(A) + p(B) + p(C) + p(D) + p(E) = 1$$

Next, we note that the event that *at most* one gyro fails is the event $(A \cup B)$. Because A and B are mutually exclusive,

$$p(A \cup B) = p(A) + p(B) = 0.9477$$

e.g., there is about a 95% probability that four gyros (i.e., one spare) will allow the mission to be completed. If this probability of success is inadequate, then either more gyros or better gyros (higher p_G) are required.

Example 12.3 is both illustrative and complete, but obviously we do not wish to work through such a laborious analysis in every case, particularly as the number of individual units in question increases. However, one may proceed by induction from Example 12.3 to the general case, where

$$p(k) = B_{nk}q^k(1 - q)^{n-k} \tag{12.32}$$

with

$p(k)$ = probability of observing k "defectives" (i.e., in the set we are seeking) in a group of n objects drawn from a population having a defective proportion q.

$B_{nk} = n!/[k!\ (n - k)!]$ = binomial coefficient

Note, as in Example 12.3, that

$$P(X \leq k) = \sum_{m=1}^{k} p(m) \tag{12.33}$$

Eqs. (12.32) and (12.33) are applicable in cases where there are a finite number of discrete outcomes, with each outcome consisting of one of two (and only two) mutually exclusive states. Thus, in the example provided, gyros are either defective or not, and in either case can be explicitly enumerated. Situations that may be characterized in these terms are referred to as binomial processes.

12.4.4 Poisson Distribution

Often we must deal with situations in which the appearance of an outcome (an event) can be noted, but the *non*-appearance of the event cannot be explicitly observed, and may in fact be non-denumerable. A simple example is the receipt of a telephone call during a given time interval. All calls can be counted, but non-calls cannot be so recorded. In such cases, the either/or nature of the binomial process exists, but the finite set of non-outcomes does not. An enormous number of natural processes (e.g., charged particle radiation dosage, "shot" noise in photomultiplier tubes, micrometeoroid strikes on a spacecraft, etc.) fall into such a category, to be described next in terms of Poisson statistics, which may be derived as the limit of a binomial random process.

To determine this limit, we assume that there exists a random event that occurs at average rate λ in a given interval. We further assume that the interval is divided into n subintervals such that each subinterval is small enough to contain at most one event (e.g., the duration of a telephone ring). Each subinterval thus represents an individual experimental trial, during which we either record a success, or not. Then the probability of the event occurring in a given subinterval (i.e., on a given trial of the experiment) is

$$q = \frac{\lambda}{n} = \frac{\text{successes}}{\text{trials}} = P(\text{event occurring}) \tag{12.34}$$

while, of course, the probability of a non-event during this subinterval is $(1 - q)$.

With these assumptions we have, as required for applicability of the binomial theorem, a countable set of discrete outcomes (n intervals in which we either did or did not experience the event), as well as a probability q of experiencing the event during each experiment. To model this, we compute the probability of k

occurrences of the event during the given interval a

$$p(k) = \left[\frac{n!}{k!(n-k)!}\right]\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^k \tag{12.35}$$

To obtain the continuum result, we define the Poisson distribution as

$$p(k, \lambda) = \lim_{n\to\infty} p(k) = \lim_{n\to\infty}\left[\frac{n!}{k!(n-k)!}\right]\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^k \tag{12.36}$$

or

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{12.37}$$

where $p(k, \lambda)$ is the probability of observing k occurrences of an event that occurs with average frequency λ over the interval in question.

We note in passing that for $n > 100$ and $q < 0.01$, the binomial and Poisson distributions are essentially indistinguishable in their results, while the Poisson distribution (with $\lambda = nq$) is much easier to use.

Example 12.4

The first Tracking and Data Relay Satellite System (TDRSS) spacecraft experienced approximately one "soft error" (see Chapter 3) per day, an event requiring memory to be re-initialized from the ground. What was the probability of having a day free of such an event, and of having two such events in one day?

Solution: The nature of the soft-error process is such as to imply that it is Poisson distributed with $\lambda = 1$ per day. Then

$$p(0, 1) = \frac{1}{e} = 0.3679$$

$$p(2, 1) = \frac{(1^2)e^{-1}}{2!} = \frac{1}{2e} = 0.184$$

Example 12.5

As of October 2002, NASA's estimate of space shuttle flight risk, based on analytical models and flight history, included a loss-of-crew probability of 1/265. If this estimate was correct, and assuming all space shuttle flights to be identical (an approximation), what were the odds that two failures would occur in 113 missions?

Solution: Since $n = 113 > 100$ and $q = 1/265 = 0.00377 < 0.01$, we can use the continuous form of the Poisson distribution to find $\lambda = nq = 0.426$ and

$$p(2, 0.426) = \frac{(0.426)^2 e^{-0.426}}{2!} = 0.059$$

12.5 System Reliability

As we have discussed, space systems, more so than many other engineering systems, are expected to be reliable. It will therefore often be of interest to consider the probability of system failure during some time interval Δt. To do this, we assume the existence of a failure density function $f(t)$, which is a probability density function expressing the probability per unit time of failing. Note that this is by definition the *first* failure. From Eq. (12.8) the probability of failure during time interval Δt is then $f(t)\Delta t$, and the probability of the occurrence of failure by time t is given by

$$F(t) = \int_0^t f(t)\,dt \tag{12.38}$$

Note that $F(t)$ is a probability distribution function. The reliability of the system is then the probability that *no* failure occurs, i.e.,

$$R(t) = 1 - F(t) \tag{12.39}$$

Now, for the system to fail between time t and $t + \Delta t$, it must first survive to time t. Let S be the event of survival to time t, and F the event of failure in the time interval Δt. Then the conditional probability of failure between time t and $t + \Delta t$, given survival to time t, is from Eq. (12.2)

$$P(F/S) = \frac{P(S \cap F)}{P(S)} \tag{12.40}$$

From the preceding discussion,

$$P(S) = R(t) \tag{12.41}$$

$$P(S \cap F) = f(t)\Delta t \tag{12.42}$$

Therefore,

$$P(F/S) = \frac{f(t)\Delta t}{R(t)} \tag{12.43}$$

We define

$$Z(t) \equiv \frac{f(t)}{R(t)} \tag{12.44}$$

as the conditional failure rate function, or hazard function, or hazard rate. Again, this hazard rate is the probability of failure between time t and $t + \Delta t$, given survival to time t. Note $Z(t) > f(t)$ because $R(t) < 1$. For example, the number of spacecraft failing between, say, 10 and 11 years is quite small, $f(t)\Delta t \ll 1$, because so few last through the first 10 years. Of those that do, a relatively high proportion will fail in the 11th year, because $R(t)$ is small.

From these results, we have

$$\frac{dR}{dt} = -\frac{dF}{dt} = -f(t) = -Z(t)R(t) \tag{12.45}$$

so that

$$\frac{dR}{R} = -Z(t)\,dt \tag{12.46}$$

and

$$R(t) = e^{-\int Z(t)dt} \tag{12.47}$$

12.5.1 Constant Failure Rate Systems, Exponential Distribution

The ability to obtain a closed-form expression for $R(t)$ depends on our ability to integrate $Z(t)$. If $Z(t) = \lambda =$ a constant (i.e., age has no effect on failure rate), then we obtain the so-called exponential distribution,

$$R(t) = e^{-\lambda t} \tag{12.48}$$

Equation (12.48) gives the reliability function for the important case of a system with a constant failure rate hazard function. The probability of experiencing at least one failure by time t, the failure distribution function, is then

$$F(t) = 1 - R(t) = 1 - e^{-\lambda t} \tag{12.49}$$

and the failure density function in this case is

$$f(t) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t} \tag{12.50}$$

Example 12.6

A particular type of reaction control thruster used on a manned spacecraft has an established failure rate of approximately one failure in six months of normal usage. The attitude and translation control system for a given spacecraft consists of 16 of these thrusters, arranged in four groups of four thrusters each, all in a

plane that contains the spacecraft center of mass. What are the odds of a thruster failure during a week-long mission?

Solution: The average failure rate per thruster is

$$\lambda = \frac{1 \text{ failure}}{26 \text{ weeks}} = \frac{0.0385 \text{ failures}}{\text{week}}$$

There are 16 thrusters in the system, each independent of the others, and so the system failure rate is

$$\Lambda = 16\lambda = \frac{0.616 \text{ failures}}{\text{week}}$$

For $t = 1$ week, the chance of at least one failure is then

$$F(t) = 1 - e^{-\Lambda t} = 1 - 0.54 = 0.46$$

12.5.2 Mean Time to Failure (MTTF)

If we calculate the mean of the failure density function, we can compute the average time to the first failure or the mean time between failures (MTBF), if it is possible to repair the system and return it to service. We have

$$\text{MTTF (MTBF)} \equiv \int_0^\infty tf(t)\,dt \tag{12.51}$$

As an example, for the constant failure rate case, we have

$$\text{MTTF} = \int_0^\infty \lambda e^{-\lambda t}\,dt = \frac{1}{\lambda} \equiv \tau \tag{12.52}$$

Because constant-average-failure-rate systems are so important in reliability analysis, the preceding result is quite useful, and indeed we often express the reliability function of such a system as

$$R(t) = e^{-t/\tau} \tag{12.53}$$

How useful is the assumption that $Z(t) = \lambda =$ constant? In reality, nearly all systems (including humans!) have a failure rate that *does* depend on the age of the system, but in a rather standard fashion, as shown in Fig. 12.4 and known as the "bathtub curve." It is seen that there exist, early and late in life, two periods of significantly higher failure rates known respectively as the "infant-mortality" and "old-age" regions of the hazard function. Between these regions normally lies an extended period of approximately constant failure rate. Systems operating in this region can be adequately characterized by the simplified analysis just given.

In practical terms, one of the major goals of spacecraft subsystem and system testing is to ensure that all subsystems have operated long enough to be past their infant-mortality region. At the system level, concerns sometimes arise over

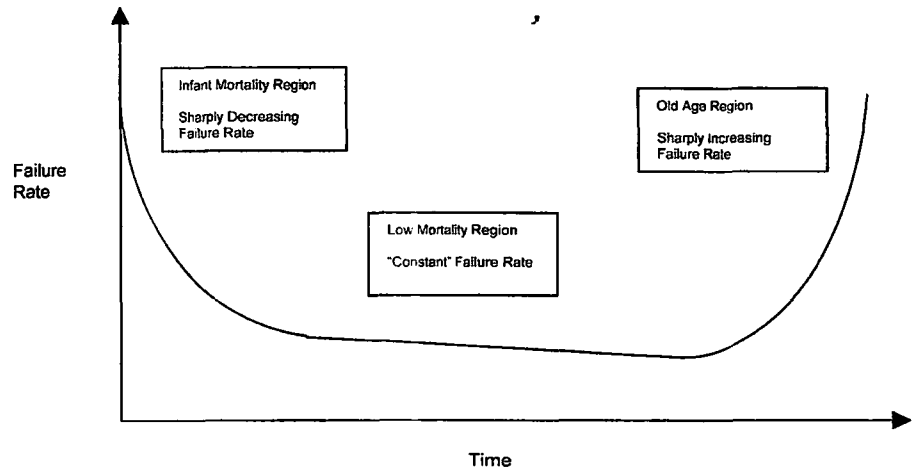


Fig. 12.4 Bathtub reliability curve.

ensuring that testing is not so protracted as to cause certain subsystems to be overused, i.e., driven into their old-age region.

12.5.3 Non-Constant Failure Rate Systems, Weibull Distribution

One implication of the preceding discussion is that a system that is either newly in service, or possibly of a relatively unproven design, or which has substantially exceeded its expected service lifetime, may not be appropriately characterized as having a constant failure rate. The most commonly assumed hazard rate in such cases follows a power-law dependence, i.e., from Eq. (12.44) we assume

$$Z(t) \equiv \frac{f(t)}{1 - F(t)} = \beta \tau^{-\beta} t^{\beta - 1} \tag{12.54}$$

The corresponding failure distribution function (again, the probability of at least one failure by time t) for this case can be shown to be

$$F(t) = 1 - \exp\left\{ -\left[\frac{t - t_0}{\tau - t_0} \right]^{\beta} \right\} \tag{12.55}$$

while the failure density function is

$$f(t) = \left\{ \left[\frac{\beta}{\tau - t_0} \right] \left[\frac{t - t_0}{\tau - t_0} \right]^{\beta - 1} \right\} \exp\left\{ -\left[\frac{t - t_0}{\tau - t_0} \right]^{\beta} \right\} \tag{12.56}$$

$F(t)$ is the three-parameter Weibull distribution, first developed in connection with the theory of failure in brittle materials and often referred to as weakest link

theory. As earlier, $\tau = 1/\lambda$ is the failure time constant (often called, for obvious reasons, the $1/e$ point). The constant t_0, often taken as 0, is the value prior to which no failure is ever observed to occur. It is seen that the hazard function depends on the constant β (called the Weibull modulus) for its character; if $\beta = 1$, we recover the constant-failure-rate law. If $\beta < 1$, the hazard rate is seen to decrease with time, i.e., the older the system, the less likely it is to fail in a given time interval, and conversely for $\beta > 1$. Thus, the Weibull distribution can be used to represent systems in either the infant-mortality region or the old-age region of their service life.

The Weibull reliability [e.g., the reliability based on Eq. (12.55) rather than on Eq. (12.49) for constant-failure-rate systems] with $\beta < 1$ is of particular interest in spacecraft design. It has been shown by Hecht and Hecht[5] that such a distribution more accurately characterizes the reliability of modern spacecraft than does the more pessimistic assumption of $Z(t) = \lambda =$ constant.

12.5.4 System Availability

Often when a subsystem or component of a system fails, circumstances are such that a repair can be effected and, after some period of time, the system returned to service. This may be true even for a space vehicle, where no physical repair is possible but redundant systems or procedures may be activated in the event of a failure in the primary system. The consideration of systems that may be repaired leads to the concepts of system availability and downtime, to be discussed next.

Let us assume that N failures occur over total time T, and that after any failure the system is not working, or "down" for some average time T_r while repairs are made. The total downtime is then

$$T_d = NT_r \tag{12.57}$$

while the system is available for a total time of

$$T_a = T - T_d = T - NT_r \tag{12.58}$$

It is more useful to define a fractional downtime D as

$$D \equiv \frac{T_d}{T} = \frac{NT_r}{T} \tag{12.59}$$

and a fractional availability A as

$$A \equiv \frac{T_a}{T} = 1 - \frac{T_d}{T} = 1 - \frac{NT_r}{T} = 1 - D \tag{12.60}$$

A and D represent, respectively, the probabilities that the system is available for use or is down. For a simple failure-and-repair model such as this, and again

assuming a constant average failure rate, we see that

$$\lambda = \frac{N}{T_a} \tag{12.61}$$

because the downtime must be removed before computing the failure rate. Then from the preceding we find

$$N = \frac{\lambda T}{1 + \lambda T_r} \tag{12.62}$$

hence

$$D = \frac{\lambda T_r}{1 + \lambda T_r} \cong \lambda T_r \quad \text{for } \lambda T_r \ll 1 \tag{12.63}$$

and

$$A = \frac{1}{1 + \lambda T_r} \tag{12.64}$$

Example 12.7

A designer plans to use control moment gyros (CMGs, see Chapter 7) to provide attitude control for a space station. A relatively inexpensive CMG being considered for use has an established MTTF of approximately three months; however, it can be removed and replaced in two hours, and spare CMG packages can be kept onboard for use by the crew. What is the availability of the station's attitude control system, assuming no other built-in redundancy?

Solution:

$$\lambda = \frac{1}{\text{MTTF}} = \frac{1}{91 \text{ days}} = \frac{1}{2184 \text{ h}} = 0 : 000458 \text{ h}^{-1}$$

$$T_r = 2 \text{ h}$$

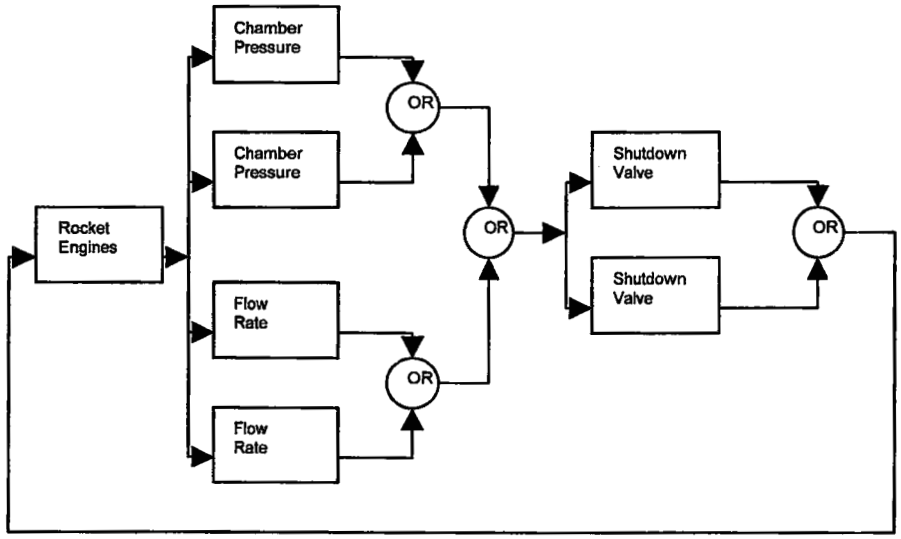
$$A = \frac{1}{1 + \lambda T_r} = 0.9991 = 99.9\%$$

. This might be a good approach, provided that the logistics train to supply spare CMGs is not a problem, and assuming that the cost of supplying cheaper replacement units is, overall, less than that for a more expensive unit that is also replaced less frequently.

Example 12.8

A flyback booster stage for a proposed partially reusable launch vehicle will have an abort sensing system that redundantly senses two pieces of information,

chamber pressure and mass flow rate. An out-of-limits value for either of these quantities is sufficient to cause an engine shutdown and abort, wherein the intact vehicle is flown back to its departure point. The engine shutdown is itself commanded by redundant valves. The logic flow for the abort system is



The functional blocks have failure rates and repair times as indicated in the following. The abort system failures may be taken to occur at the block level, are independent, and are quickly repaired; however, they are not directly detectable but can exist in the failed state for some mean time between certification checks. This required time to detect a latent failure is the effective repair time T_r:

Block	λ, failures/year	T_r, years
Pressure sensor	5	0.02
Flow sensor	6	0.02
Shutdown valve	0.4	0.10

What is the probability that the abort system will be available when needed?

Solution: The downtime on the pressure sensors is

$$D_P = \frac{\lambda T_r}{1 + \lambda T_r} = 0.0909$$

for a single sensor. Because the two pressure sensors are (it is fervently hoped!) independent, the total probability of the pressure sensing system being in a failed

state is

$$D_{PF} = D_P^2 = 0.0083$$

Similarly, each mass flow sensor has a downtime of

$$D_M = 0.1071$$

and for the system we compute the probability that the mass flow system is in a failed state as

$$D_{MF} = D_M^2 = 0.0115$$

The pressure sensor and mass flow systems are themselves independent (either alone can trigger the abort, and so both must fail for the system to fail), hence the total failure probability (downtime) of the sensor portion of the abort system is

$$D_{SF} = D_{PF}D_{MF} = 0.00009487$$

Now, the mean fractional downtime for the shutdown system behaves the same way, i.e., for one control valve,

$$D_C = 0.0385$$

and for the system,

$$D_{CF} = D_C^2 = 0.0015$$

To achieve a successful abort, both the sensing and control systems must be available when needed. We note

$$A_{SF} = 1 - D_{SF} = 0.9999$$
$$A_{CF} = 1 - D_{CF} = 0.9985$$

The availability of each of these systems is independent of the other, and so the total availability of the abort system is

$$A_A = A_{SF}A_{CF} = 0.9984$$

12.6 Statistical Inference

The topics heretofore presented share the common assumption that the underlying statistical information necessary to undertake a given calculation is available to the reliability engineer. This is sometimes true, but often it will be necessary for failure rates, defective product statistics, etc., to be derived from historical or experimental performance data for the components or systems in question. The task of deriving representative statistical properties for a population, given the observation of a limited sample of that population, is part of the process of statistical inference that is an essential element of the reliability engineer's task.

12.6.1 Sample Statistics

We begin by defining some basic quantities that will be useful. We assume there exists a population (e.g., people, launch vehicles, gyroscopes, fuses) that consists of N members and is in principle denumerable, whether or not this is practical in a given case. The population will have, with respect to its relevant characteristics (e.g., height, reliability, MTTF, etc.), certain underlying statistical properties, including a probability density function, mean, variance, etc., which are a priori unknown and which we seek to determine. If all members of the population could be sampled and categorized, the task of computing population statistics would be trivial, if possibly somewhat laborious. However, this may be impossible in principal or in practice. For example, we might wish to know the MTTF of a particular type of gyroscope; however, unless all members of the population are tested to failure, the exact answer cannot be known. If known, the result is no longer useful.

Thus, instead of a fully characterized population, we have available certain limited observations, namely a sample set consisting of $n < N$ members $\{x_i\}$ drawn as a simple random sample from the population. A simple random sample of size n is one in which all possible samples of size n have equal likelihood of being selected. We ignore in this discussion more sophisticated issues such as how the sample should be obtained, whether sampling is with or without replacement, etc., which are treated in standard texts. From Eq. (12.14), and noting that $f(x_i) = 1/n$ across the sample set, we define the sample mean as the common arithmetic average,

$$m = \left(\frac{1}{n}\right) \sum_{i=1}^{n} x_i \tag{12.65}$$

Less obviously, the sample variance is

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - m)^2}{(n-1)} \tag{12.66}$$

with s being the sample standard deviation. The form of the denominator in Eq. (12.66) is due to the fact that there are only $(n - 1)$ independent parameters available to compute the standard deviation, since the sample mean has been determined from the $\{x_i\}$.

Note carefully that m and s^2 are *not* the population mean and variance μ and σ^2, but are *estimates* of the underlying quantities based on the randomly drawn, but presumably representative, set of samples $\{x_i\}$. In the limit $n \to N$ we intuitively expect that $m \to \mu$ and $s^2 \to \sigma^2$, expectations that we will shortly justify. Depending on the particular $\{x_i\}$ that are drawn from the population, many different values of m and s^2 are possible, so that m and s^2 are random variables in their own right, with their own sample distributions, which we will now examine.

The alert reader will be unsurprised to learn that the distribution of the sample mean m approaches a Gaussian as $n \to \infty$, regardless of the underlying population distribution. In our earlier notation, $m = N(\mu_m, \sigma_m)$, with

$$\mu_m = \mu \tag{12.67}$$

$$\sigma_m^2 = \left[\frac{N-n}{N-1}\right]\left(\frac{\sigma^2}{n}\right) \tag{12.68}$$

where, again,

μ = population mean
σ^2 = population variance
n = sample size
N = population size

This is a consequence of the central limit theorem, discussed earlier. Typically, a few dozen samples are sufficient to allow the assumption that m is normally distributed, depending on the accuracy desired. (If the population distribution is itself Gaussian, then m is Gaussian regardless of sample size.) The term $[(N-n)/(N-1)]$ is called the *finite population correction factor*, used when n and N are of comparable size. When $n \ll N$, the bracketed term approaches unity, and

$$\sigma_m \approx \frac{\sigma}{\sqrt{n}}, \quad n \ll N \tag{12.69}$$

a convenient result that is often satisfied in practice. Obviously, if $N \to \infty$, Eq. (12.69) is always used. It is also appropriate when sampling with replacement is performed.

Thus, the best estimate of the population mean μ is the sample mean m, given by Eq. (12.65). The error associated with this estimate is indicated by the standard deviation σ_m, often called the *standard error of the mean*. Equation (12.68) gives the very important result that this error is proportional to $1/\sqrt{n}$.

Because the population variance σ^2 is usually unknown, we must approximate it by using Eq. (12.66); we defer for the present a discussion of the accuracy of this approximation.>

12.6.2 Estimating Population Mean

Figure 12.5 illustrates the situation thus far; for large n we have a sampling distribution for the sample mean m that is normal around the population mean μ and has variance $\sigma_m^2 = \sigma^2/n$. We have a point estimate for m in a given case, but because μ is unknown, the accuracy of the estimate for the given case is unknown. However, if we renormalize to standard form, i.e., define a random

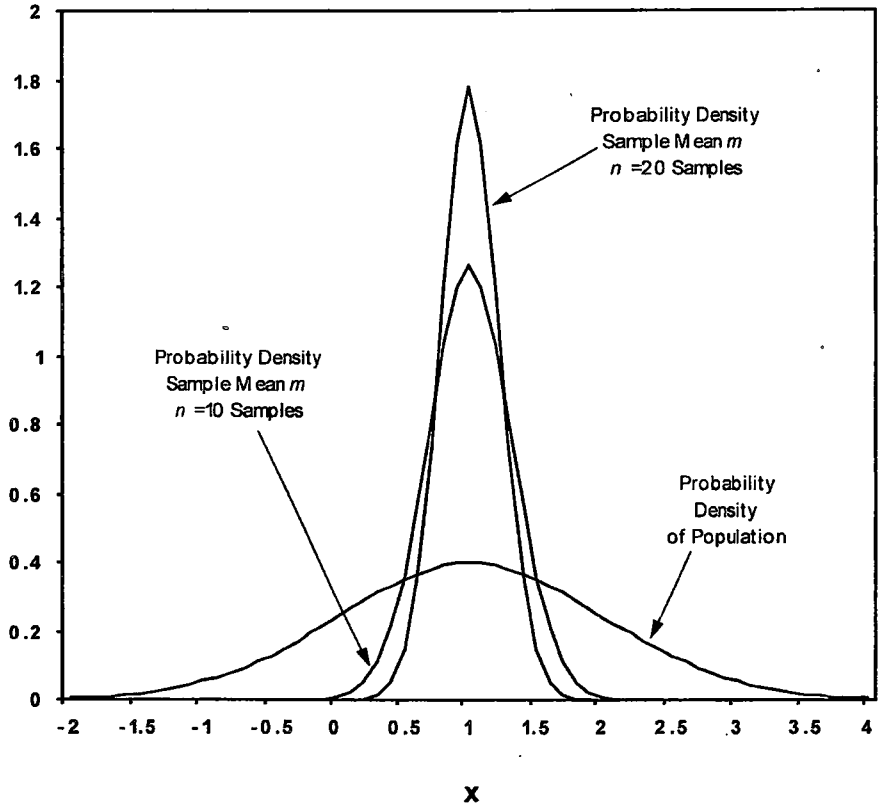


Fig. 12.5 Probability density function for sample mean m.

variable z such that

$$z = \frac{m - \mu}{\sigma_m}. \tag{12.70}$$

then $z = N(0, 1)$, and we can be confident that 68% of all sample means m will yield values of z between ± 1. More generally, for any *confidence coefficient* or *confidence level* $\beta < 1$, representing a selected value of area under the normal curve in Fig. 12.5, there is associated a confidence interval $\pm z_\beta$ such that, with probability β,

$$-z_\beta < z < z_\beta \tag{12.71}$$

From Eqs. (12.70) and (12.71), we then obtain the interval estimate for μ, at confidence level β, given by

$$m - z_\beta \sigma_m < \mu < m + z_\beta \sigma_m \tag{12.72}$$

Table 12.1 Confidence coefficients and associated z_β for normal distribution

$\beta = 1 - \alpha$	$\alpha/2$	$z_\beta = z_{\alpha/2}$
0.68	0.160	1.000
0.90	0.050	1.645
0.95	0.025	1.960
0.954	0.023	2.000
0.98	0.010	2.330
0.99	0.005	2.575
0.9973	0.0013	3.000

where σ_m is given by Eq. (12.69), unless the population variance is unknown, in which case Eq. (12.66) is used to supply an estimate of σ.

The notation $z_{\alpha/2}$ is often used instead of z_β, with $\alpha = 1 - \beta$ representing the area in the combined upper and lower tails of the normal curve, as shown in Fig. 12.5. Some commonly used confidence levels and associated values of z_β are given in Table 12.1.

12.6.3 Sampling Error

The error associated with the estimate of the population mean m in Eq. (12.72) is

$$\varepsilon = z_\beta \sigma_m = \frac{z_\beta \sigma}{\sqrt{n}} \approx \frac{z_\beta s}{\sqrt{n}} \tag{12.73}$$

with the latter equality applicable when the sample standard deviation must be used instead of the population standard deviation. If the allowable maximum error magnitude is known, the required sample size is then

$$n \geq \left(\frac{z_\beta \sigma}{\varepsilon}\right)^2 \approx \left(\frac{z_\beta s}{\varepsilon}\right)^2 \tag{12.74}$$

Example 12.9

A launch operation has a density-weighted average headwind constraint of 40 km/h, above which payload capacity suffers. The launch will be scrubbed unless there is 90% confidence that the mean headwind is below this value. Weather balloon data obtained roughly an hour before launch yielded 101 data points with a sample mean of 30 km/h and a sample standard deviation of 25 km/h, largely due to wind gusts. Should the launch be scrubbed?

Solution: From the given data, $m = 30$ km/h, $s = 25$ km/h, and $n = 101 \gg 1$. The sample is of adequate size to assume a Gaussian sampling distribution for m, irrespective of the underlying wind pattern. N is implicitly very large, and so Eq. (12.69) applies with $\sigma \approx s$, hence $\sigma_m = s/\sqrt{n} = 2.5$ km/h. For a 90% confidence level, from Table 12.1, z_β is 1.645. The interval estimate for the average headwind speed is then

$$30\,\text{km/h} - 4.12\,\text{km/h} = 25.8\,\text{km/h} < \mu_{\text{headwind}} < 30\,\text{km/h} + 4.12\,\text{km/h}$$
$$= 34\,\text{km/h}$$

Thus, the launch should not be scrubbed according to the existing headwind guidelines. Gust load constraints could well be another matter.

12.6.4 Small Sample Sets, the t Distribution

It is not uncommon to encounter sample sizes too small to justify the use of the previous results. This can be especially true in aerospace applications, where it will be appreciated that the number of spacecraft, missiles, launch vehicles, etc., for which a reliability analysis is to be performed is often quite limited. However, if the sample set is small (less than 30 samples or so), if the underlying population is normally distributed, and if the sample variance s^2 is used as an estimator of the population variance σ^2, then it is found that the interval estimate for μ satisfies

$$m - \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} < \mu < m + \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} \tag{12.75}$$

where $t_{n-1,\alpha/2}$ denotes the t distribution for $(n-1)$ degrees of freedom having area $\alpha/2$ under each tail of the distribution. The t distribution is shown in Fig. 12.6 for a few representative degrees of freedom. As previously, $(1-\alpha) = \beta$ is the confidence coefficient or confidence level, and $\pm t_{n-1,\alpha/2}s/\sqrt{n}$ is the confidence interval.

The t distribution may be viewed as a more general form of the normal distribution and is actually a separate distribution for each value of n, converging to the Gaussian for $n \to \infty$, but valid for all sample sizes as long as the underlying population is Gaussian. (This additional and possibly restrictive assumption is unnecessary for large sample sizes, as discussed earlier, allowing us to use the simpler normal distribution in such cases.) As with other special probability distributions discussed in this text, the t distribution is tabulated in standard references[4] or available online. Table 12.2 provides a few t values for representative confidence levels and degrees of freedom.

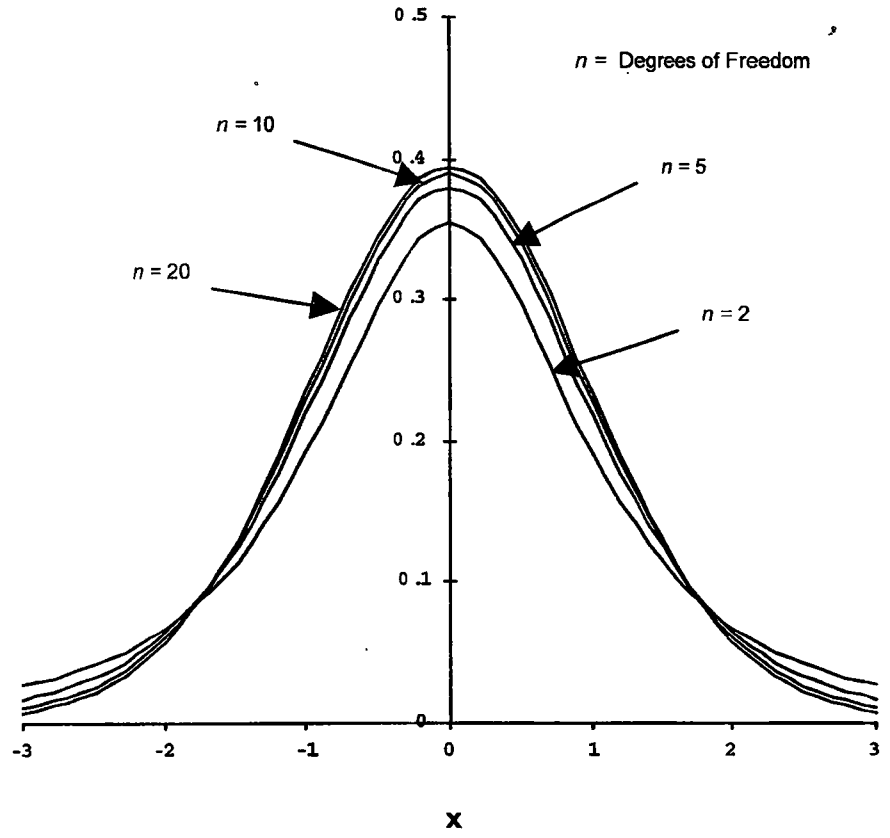


Fig. 12.6 Probability density for *t*-distribution.

12.6.5 Estimating Population Proportion

Very often in reliability analysis the quantity of interest is a proportion rather than a mean value; indeed, the reliability statistic itself is such a quantity. For example, when we speak of the reliability of a given expendable launch vehicle being 98%, we are identifying that proportion of the launch vehicle population that is successful; specifically, it is the ratio of successes to trials. This is a familiar quantity; we have encountered it earlier in connection with Poisson statistics. Accordingly, we define a sample proportion q as

$$q = \frac{\lambda}{n} \tag{12.76}$$

where

λ = number of successes
n = number of trials

Table 12.2 t Distribution for various confidence levels

DOF, $n-1$	$\beta = 0.90$, $\alpha/2 = 0.05$	$\beta = 0.95$, $\alpha/2 = 0.025$	$\beta = 0.98$, $\alpha/2 = 0.01$	$\beta = 0.99$, $\alpha/2 = 0.005$
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
14	1.761	2.145	2.624	2.977
19	1.729	2.093	2.539	2.861
24	1.711	2.064	2.492	2.797
29	1.699	2.045	2.462	2.756
∞	1.645	1.960	2.326	2.576

The sample proportion is intended to be an estimate of the underlying population proportion p, which would be obtained in the limit $n \to \infty$, but which of course is unknown.

The correspondence of the sample proportion to the sample mean m discussed earlier can be seen by defining a random variable x_i such that

$$x_i = 1 \text{ if event occurs, else } 0 \tag{12.77}$$

It is immediately seen that the sample mean of x_i is

$$m = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i = q \tag{12.78}$$

It is thus to be expected that many of the results just presented in connection with inferences about population means will be applicable to proportions as well. Specifically, q is a random variable with a normal sampling distribution satisfying $q = N(p, \sigma_q)$, where

$$\sigma_q^2 = \left[\frac{N-n}{N-1}\right]\frac{p(1-p)}{n-1} \approx \frac{p(1-p)}{n}, \quad 1 \ll n \ll N \tag{12.79}$$

In the common case where p is unknown, we can use the estimate q as a substitute, precisely as was done earlier. However, when working with proportions, we may alternatively note that the expression $p(1-p)$ is maximized

when $p = 1/2$. Thus, we have the inequality

$$\sigma_q^2 \leq \frac{(N-n)/(N-1)}{4(n-1)} \approx \frac{1}{4n}, \quad 1 \ll n \ll N \tag{12.80}$$

which yields a useful upper bound for σ_q^2.

Interval estimates of p are obtained exactly as before. Defining a standard normal random variable z as

$$z = \frac{q-p}{\sigma_q} \tag{12.81}$$

and choosing a confidence coefficient $\beta < 1$, we know that z satisfies the inequality

$$-z_\beta < z < z_\beta \tag{12.82}$$

with probability β, from which we obtain

$$q - z_\beta \sigma_q < p < q + z_\beta \sigma_q \tag{12.83}$$

Finally, assuming $1 \ll n \ll N$, and substituting q as an estimate for the unknown p in the computation of σ_q, we have for the interval estimate of p

$$q - z_\beta \left[\frac{q(1-q)}{n}\right]\frac{1}{2} < p < q + z_\beta \left[\frac{q(1-q)}{n}\right]\frac{1}{2} \tag{12.84}$$

The required sample size for a given level of error control is, by analogy to Eq. (12.74),

$$n \geq \left(\frac{z_\beta \sigma_p}{\varepsilon}\right)^2 = \left(\frac{z_\beta}{\varepsilon}\right)^2 p(1-p) \approx \left(\frac{z_\beta}{\varepsilon}\right)^2 q(1-q) \tag{12.85}$$

If we seek a conservative choice for n, we can set $q = 1/2$, maximizing the right-hand side, in which case the sample size requirement becomes

$$n \geq \frac{(z_\beta/\varepsilon)^2}{4} \tag{12.86}$$

Example 12.10

Following Example 12.5, the space shuttle has experienced two fatal accidents in 113 flights. Using again the approximation that all space shuttle flights are identical and independent, give an interval estimate for generic space shuttle system safety (the probability of not having a fatal accident) at the 95% confidence level.

Solution: The sample proportion of shuttle flight successes is $q = 0.982 = 111/113$. Since $n = 113 \gg 1$, and $z_\beta = z_{0.95} = 1.96$, we have

$$z_\beta\left[\frac{q(1-q)}{n}\right]\frac{1}{2} = (1.96)(0.0124) = 0.0243 = \varepsilon$$

and from Eq. (12.84),

$$q - \varepsilon = 0.956 < p < q + \varepsilon = 1.007$$

Because the overall reliability cannot be greater than unity, the interval estimate of reliability based on the sample proportion q is

$$0.96 < p < 1.0$$

with the midrange value 0.98 taken as a reasonable single-point estimate at the 95% confidence level.

12.6.6 Estimating Population Variance

Earlier we identified the sample variance, given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - m)^2}{n-1} \tag{12.87}$$

as an estimator for the often unknown population variance σ. It is of interest to understand the accuracy to be expected of this estimate for a given sample size. As before, the sampling distribution can be used to derive an interval estimate for the population variance when the population is normally distributed. However, unlike the sample mean, the distribution of the sample variance is not itself Gaussian. Rather, the parameter $(n-1)s^2/\sigma^2$ follows the more complex $\chi^2_{n-1,\xi}$ distribution. Several normalized (having unit total area under the curve) χ^2 probability density functions for representative values of $(n-1)$ are plotted in Fig. 12.7.

As with the t distribution, the parameter $(n-1)$ indicates the number of degrees of freedom in the distribution. The parameter ξ denotes the probability, or area under the curve of the density function, for values of $\chi^2 > \chi^2_\xi$, i.e., to the right of the value χ^2_ξ. Note that the χ^2 distribution is not symmetrical about its peak.

The interval estimate for the population variance σ is obtained in the same fashion as in our earlier discussion, yielding

$$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \tag{12.88}$$

As before, the confidence level of the estimate is the probability $1 - \alpha = \beta < 1$, with the preceding notation indicating that an area, or probability, of $\alpha/2$ remains

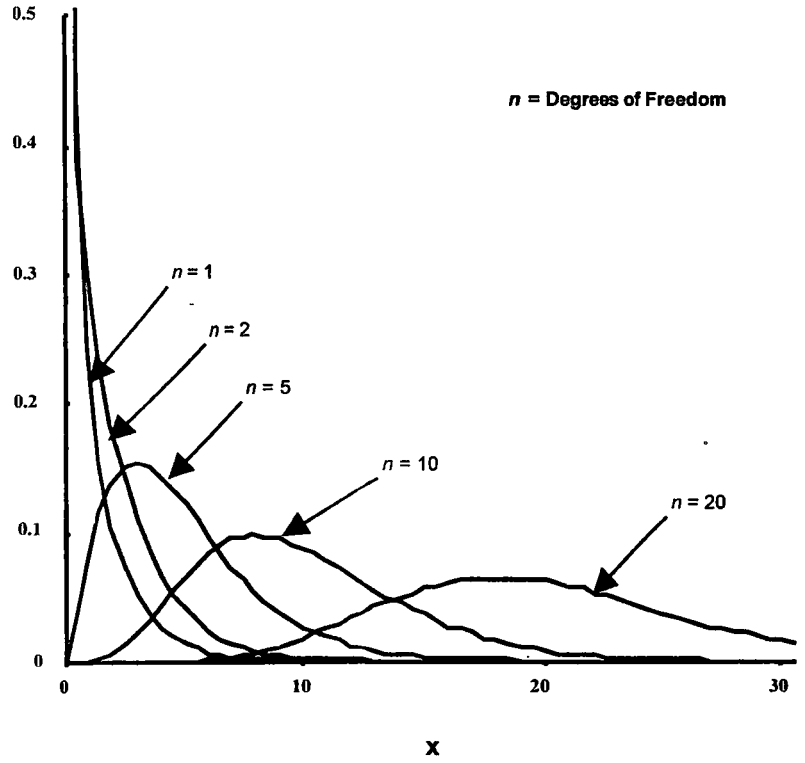


Fig. 12.7 Probability density function for χ^2 distribution.

in the upper and lower tails of the distribution. Table 12.3 gives values of χ^2 for several values of ξ and numerous degrees of freedom; more extensive tables are available in standard references.[4]

Example 12.11

Following Example 12.9, the launch vehicle also has a wind gust constraint of 30 km/h. The launch will be scrubbed unless there is 95% confidence that the gusts will be below this value. Weather balloon data obtained roughly an hour before launch yielded 101 data points with a sample standard deviation of 25 km/h that is ascribed to wind gusts. Should the launch be scrubbed?

Solution: The 95% confidence interval requirement implies $\alpha = 5\%$, hence $\alpha/2 = 0.025$. Thus, we want the area under the χ^2 curve between $\xi = 0.975$ and $\xi = 0.025$, i.e., between $\chi^2_{100,0.975} = 74.2219$ and $\chi^2_{100,0.025} = 129.561$. We then

Table 12.3 Values of χ^2

DOF, $n-1$	$\xi = 0.99$	$\xi = 0.975$	$\xi = 0.95$	$\xi = 0.05$	$\xi = 0.025$	$\xi = 0.01$
1	1.57088E−04	9.82069E−04	3.93214E−03	3.84146	5.02389	6.63490
2	0.0201007	0.0506356	0.102587	5.99147	7.37776	9.21034
3	0.114832	0.215795	0.351846	7.81473	9.34840	11.3449
4	0.297110	0.484419	0.710721	9.48773	11.1433	13.2767
5	0.554300	0.831211	1.145476	11.0705	12.8325	15.0863
6	0.872085	1.237347	1.63539	12.5916	14.4494	16.8119
7	1.239043	1.68987	2.16735	14.0671	16.0128	18.4753
8	1.646482	2.17973	2.73264	15.5073	17.5346	20.0902
9	2.087912	2.70039	3.32511	16.9190	19.0228	21.6660
14	4.66043	5.62872	6.57063	23.6848	26.1190	29.1413
19	7.63273	8.90655	10.1170	30.1435	32.8523	36.1908
24	10.8564	12.4011	13.8484	36.4151	39.3641	42.9798
29	14.2565	16.0471	17.7083	42.5569	45.7222	49.5879
40	22.1643	24.4331	26.5093	55.7585	59.3417	63.6907
50	29.7067	32.3574	34.7642	67.5048	71.4202	76.1539
100	70.0648	74.2219	77.9295	124.342	129.561	135.807

have from Eq. (12.88)

$$\frac{(100)(25\,\text{km/h})^2}{(129.561)} < \sigma^2 < \frac{(100)(25\,\text{km/h})^2}{(74.2219)}$$

hence

$$482.4\,\text{km}^2/\text{h}^2 < \sigma^2 < 842.1\,\text{km}^2/\text{h}^2$$

or, at the 95% confidence level,

$$22.0\,\text{km/h} < \sigma < 29.0\,\text{km/h} < 30\,\text{km/h}$$

Because the maximum wind gust at the 95% confidence level is below the constraint, the launch should proceed.

12.7 Design Considerations

The preceding text and examples enable the reader to analyze and assess the reliability of a given system in many simple but nonetheless realistic and interesting cases. It is hoped that this has also fostered some insight into how systems must be designed to attain desired levels of reliability. In this section, we explore these design techniques in more detail.

Because physical repair is normally not an option, two basic approaches are used to achieve a reliable spacecraft design. These are fault avoidance and fault tolerance, and they may be used separately or in combination in any given system.

The goal of fault avoidance, as the name implies, is simply to ensure that a part, subsystem, or complete system does not fail. This is normally accomplished through the provision of ample environmental and performance margins in the basic design, the use of carefully selected, screened, parts, rigorously controlled assembly procedures conducted in very clean environments, extensive subsystem and system-level testing, and extensive review and documentation of all steps in the process. This documentation will include all design drawings and analysis, assembly history, test results, and historical information concerning the parts and components used in the spacecraft, quite possibly down to the materials from which the parts were fabricated. Such documentation allows the most rigorous possible understanding of the systemic causes of mistakes, design flaws, component failures, and test anomalies when and as they are discovered. These and other procedures are provided in excruciating detail in applicable military standards (and therefore the de facto government and industry standard as well) governing this subject.[6,7]

With enough care, it is indeed possible to develop almost fault-free systems. However, it will be apparent to the reader that "enough care" can be exceedingly expensive and time consuming, and equally apparent that complex systems (e.g., those with many components) will always be vulnerable to random failure of isolated components. As an elementary example, consider a large system with one million individual parts, each of which has a failure probability of 10^{-6} over the duration of a given mission. From our earlier discussion of Poisson statistics, it is seen that such a system has a substantial risk of failure, approximately 63%. It would in practice be exceedingly difficult to achieve a mission failure rate as low as 10^{-6} for each of a million parts used in a spacecraft. Indeed, while specific details vary, it may be stated as a rule of thumb that the reliability of the best screened class S parts is only about 10 times that of good commercial parts. If achieved, such performance levels are often even more difficult to verify. Thus, even with the best materials and procedures, it may be impossible to know whether the desired level of reliability has been reached.

For these reasons, fault avoidance is rarely if ever employed as the sole means of attaining a particular level of system reliability. It is of greatest value in simple systems, in systems such as launch vehicles whose operating lifetime is relatively short, or when no other approach is physically possible. (One cannot, for example, have redundant airplane wings.) In other cases, some of the techniques of fault avoidance are normally combined with those of fault tolerance, to which many of our previous examples have alluded.

As the name implies, fault tolerance means that the system or subsystem is designed to operate after one or more random failures. For example, a common design criterion for manned or very expensive unmanned space systems is two-

fault tolerance, sometimes referred to as "fail-op, fail-op, fail-safe" design. The idea is that the spacecraft should continue to function after any two random failures, and should remain at least safely non-operational after a third failure.

Any fault-tolerant design requires the incorporation of redundancy, i.e., the provision of extra components or systems by means of which the desired task can be accomplished despite the failure of the first component or system. Usually some means of detecting the initial fault and switching the old and new systems is also required. Redundancy can be provided within components, among components, across subsystems, and at the whole-system level. As an example, at the component level a spacecraft power system might feature a main bus consisting of several wires (each oversized) in case one wire or connector pin fails. Multiple main buses, each capable of carrying the entire load, might be used to guard against a damaged harness. At the subsystem level, redundant power supplies could be provided. Additionally, if the mission is very important, more than one spacecraft could be launched to improve the probability of success. (Early planetary missions routinely featured the launch of two identical spacecraft during a given mission opportunity for just this reason.)

The incorporation of redundant systems, and the resultant effect on system reliability, is easily analyzed with the tools we have developed, subject as always to our assumption of the independence of subsystem-level failures. Indeed, many of the principles of design redundancy have been illustrated in the examples in this chapter. Figure 12.8 depicts the use of redundant blocks to achieve a given system function.

When, for whatever reason, such design redundancy fails, mission controllers may on occasion employ functional redundancy to achieve their goals. Functional redundancy refers to the use of physically different systems to accomplish the originally planned task. Too often, this occurs according to the rule that "necessity is the mother of invention" rather than as a planned strategy. A classic example is provided by the Mariner 10 mission to Mercury and Venus, wherein the attitude control system failed to provide roll stability because of an unanticipated flaw in the original design. Roll stability was provided throughout the mission through the use of differential solar radiation pressure torque (see Chapter 7) caused by individually tilting the spacecraft solar arrays.

A more dramatic, indeed spellbinding, example of the use of functional redundancy occurred during the Apollo 13 lunar mission. When an oxygen tank in the command and service module (CSM) exploded, all CSM power and oxygen was quickly lost. The lunar module (LM) was used to supply power, propulsion, attitude control, water, and oxygen for the crew until shortly before separation and reentry. Numerous accounts of this mission are available[8,9] and should be required reading for every space systems engineer.

It is especially worth noting that the Apollo CSM designers *believed* they had provided subsystem-level redundancy through the provision of redundant oxygen tanks and fuel cells, either of which could provide sufficient oxygen and power to return to Earth. That an explosion of one tank could occur, and by so doing

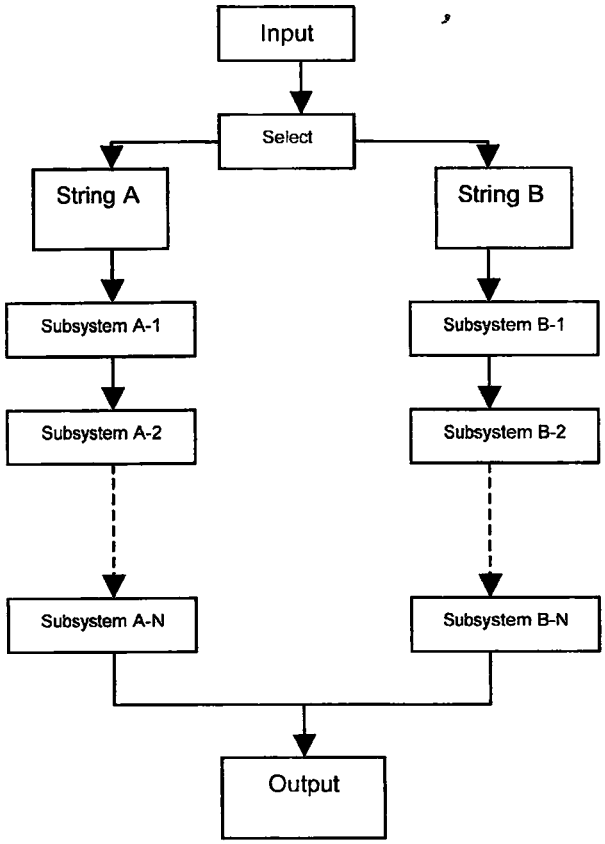


Fig. 12.8 Block-level functional redundancy.

remove both systems from service, was not anticipated. This highlights again a crucial problem in reliability analysis, wherein our calculations frequently depend on the assumption that all of the possible failure modes are known, and that failures of individual subsystems are independent. Nature frequently disobeys the rules set down by design engineers in this matter.

As a practical matter, no single level of redundancy can typically be implemented uniformly throughout a spacecraft. For example, it may be quite effective to employ parallel redundant plumbing lines to convey propellant from a tank to a thruster. However, it would normally be considered much more practical to carry two command receivers rather than to design a single radio receiver with every internal circuit redundantly wired. The choice of redundancy partitioning or cross-strapping thus varies from system to system, but can be illustrated conceptually as shown in Fig. 12.9. The dual-string system is single-failure tolerant, whereas the cross-strapped system is single-failure tolerant for given subsystems, and multiply-failure-tolerant for nonidentical subsystems.

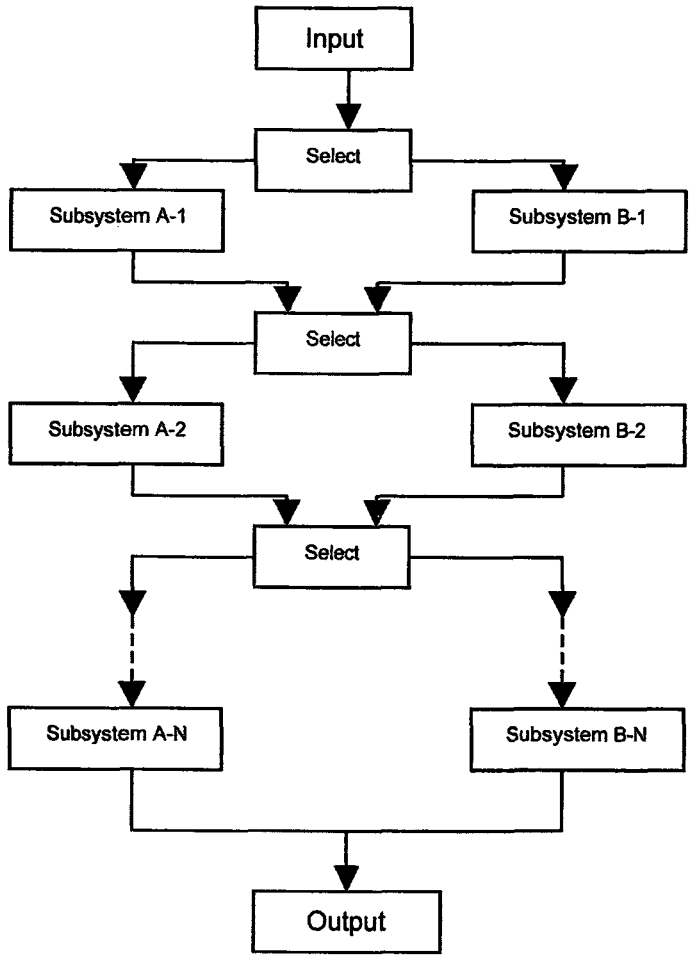


Fig. 12.9 Cross-strapped block redundancy.

Clearly, the reliability of the cross-strapped system is higher than that for the simple dual-string system, as long as the failure rates of the failure detection and switching mechanisms are negligible in comparison with the block failure rates. However, the complexity of the cross-strapped system is also much greater, a factor that normally results in higher cost, longer development time for the system, as well as a substantially more involved testing regimen to ensure that the cross-strapping works as intended.

Obviously, as with fault avoidance, designing for fault tolerance carries its own set of penalties. The redundant hardware requires additional design complexity, cost, mass, volume, power, and time to integrate and test. A redundant system may be more reliable once launched, but it offers twice as many (or more) opportunities for failure and delay while still on the ground as a

nonredundant system, because there are more components. Of course, any broken component must be repaired before launch or the desired redundancy will not exist. Moreover, a first-order analysis of the reliability offered through the use of redundant systems will often neglect the failure modes introduced by the detection and switching systems. In the final analysis, and in the real world, these cannot be ignored. (Is it the oil, or is it the warning light?) Indeed, net system safety can be reduced, if one is not careful, by the additional failure modes introduced by very complex systems. Furthermore, as we see from the Apollo 13 example, it sometimes occurs that the catastrophic failure of one redundant system component can destroy other perfectly functioning systems. It may well be true in particular cases that the theoretical gains from system redundancy are offset by the practical difficulties of implementation, and that the resources available to the project are best invested in making, and thoroughly testing, a simpler and more robust system. The challenge for the system engineer is to know when this is so.

Testing of highly redundant systems is a particular challenge. To have full confidence in the system, all logic paths must of course be tested. In Fig. 12.9, the dual-string system has only two paths, whereas the cross-strapped system has many. In a modern complex system, where redundancy management is often implemented in a powerful onboard computer, it may easily result that more logic paths exist than could ever be tested in the time available. In such cases the spacecraft will be launched with incomplete, and often very incomplete, knowledge of all the states into which it could theoretically be commanded. The potential for difficulty is obvious.

None of this is to say that redundancy in a spacecraft is bad. Indeed, it will be employed at some level in nearly all spacecraft, certainly those within the present authors' experience. However, as with many other tools in the system engineer's repertoire, it must be employed with discretion and engineering judgment.

References

[1]Anderson, D. R., Sweeney, D. J., and Williams, T. A., *Statistics: Concepts and Applications*, West Publishing, St. Paul, MN, 1986.

[2]Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.

[3]Rheinfurth, M. H., and Howell, L. W., "Probability and Statistics in Aerospace Engineering," NASA TP-1998-207194, March 1998.

[4]Selby, S. M., *Standard Mathematical Tables*, 22nd ed., CRC Press, Cleveland, OH, 1974.

[5]Hecht, H., and Hecht, M., "Reliability Predictions for Spacecraft," USAF Rome Air Development Center, Technical Rept. RADC-TR-85-229, Rome, NY, 1985.

[6]"Reliability Program Requirements for Space and Missile Systems," MIL-HDBK-1543, Department of Defense, 1988.

[7]"Procedures for Performing a Failure Mode Effects and Criticality Analysis (FMECA)," MIL-STD-1629, Department of Defense, 1980.

[8]Murray, C., and Cox, C. B., *Apollo: The Race to the Moon*, Simon and Schuster, New York, 1989, pp. 387–446.

[9]Chaikin, A., *A Man on the Moon: The Voyages of the Apollo Astronauts*, Penguin Books, New York, 1994, pp. 285–336.

Problems

12.1 A manned space launch system has an overall reliability of 98%, or one "failure" in 50 launches. There are three categories of failure, i.e., those that lead to in-flight destruction of the vehicle, those that lead to a safe return-to-launch-site or downrange abort, and those that lead to an abort to a stable but degraded orbit allowing primary mission completion. These failures occur with relative probabilities of 5, 75, and 20%. Of the abort-to-orbit cases, 40% allow the primary mission to be completed, while 60% lead to loss of mission because of the degraded orbit. What is the overall probability of loss of mission for a given launch?

(a) Given that a failure occurs, what is the probability of primary mission completion?

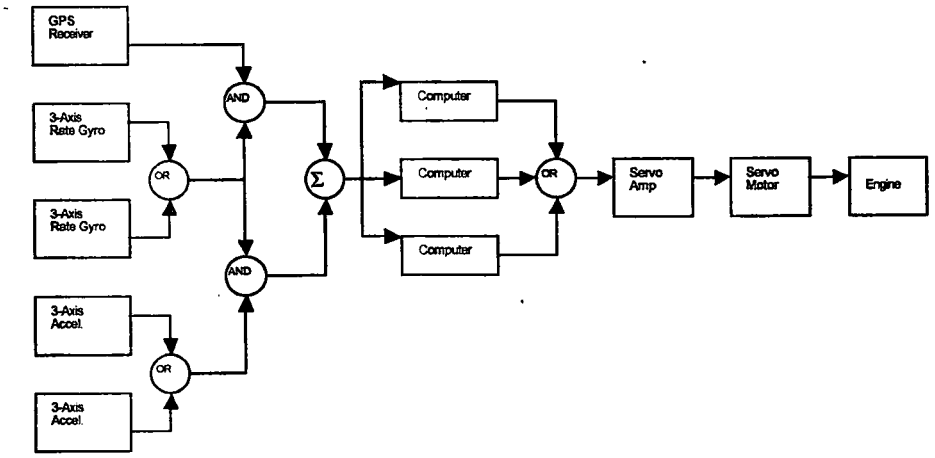
(b) Given that the vehicle reached orbit, what is the probability that an abort-to-orbit occured?

(c) Given that loss of mission has occurred, what is the probability of crew survival?

12.2 Integrated circuits (ICs) are supplied to a flight project from three sources, H, M, and L: 75% come from source H, which has a proportion of 0.1% defectives; 20% are from source M, with a 0.5% defective rate; and 5% come from source L, with a 1% population of defectives. The SR and QA department screens incoming parts and rates them "P" or "F" for pass/fail according to established criteria. A given IC is tested and found to be defective. What is the probability that it came from source H?

12.3 The navigation/autopilot system shown in the following is planned for a proposed new launch vehicle. Primary guidance is via GPS; however, with somewhat degraded accuracy, the system can function with conventional inertial navigation using strapdown gyros and accelerometers. The failure rates (assumed constant) are included in the table for each component. What is the probability of a system-level failure during the half-hour period necessary for ascent and orbital injection?

Item number	Item	λ_i, failures/hour
1	GPS	5×10^{-4}
2	Rate gyro	3×10^{-3}
3	Accelerometers	1×10^{-3}
4	Computer	2×10^{-3}
5	Servo amplifier	2×10^{-5}
6	Servomotor	1×10^{-4}



12.4 A geostationary communications satellite is placed in orbit with, unfortunately, inadequate protection against soft errors due to heavy-ion cosmic rays, which strike on a random basis having a long-term average of about once per day. It takes about an hour to do a new memory upload when this happens. What is the availability of the system?

12.5 A space launch is scheduled for a given day, but historical data show that due to various exigencies (weather, winds aloft, vehicle subsystem failures, conflicts over tracking range priorities with other launches, etc.), the launch occurs on the planned day only 50% of the time. Assuming an average delay of two days to recycle the launch operation following a scrub, what is the availability of the system?

12.6 A new rocket engine is being designed and tested; the specification requires a vacuum $I_{sp} \geq 450$ s. A heavy test engine, faithful to the planned production geometry but unsuited for flight, is constructed and used to generate the following 20 data points for specific impulse in

seconds (corrected to vacuum conditions from test-stand conditions):

452	449	447	453	448
449	451	453	452	449
453	450	450	449	452
448	452	451	452	449

Engine tests are expensive and time consuming; however, it will be vastly more expensive and time consuming to put the wrong design into production. It is desired to be 95% confident that the engine design will meet the specific impulse requirement before commencing production.

(a) What is the sampling error associated with the data?

(b) What is the 95% confidence interval estimate for the average specific impulse?

12.7 Consistency of performance is also important for the engine in problem 12.6, with the variance of I_{sp} required to be less than 1 s^2 at the 95% confidence level. Given the preceding data, is this requirement being met?

12.8 A kinetic energy penetrator (i.e., no explosive is carried) is dropped from a high-altitude airplane and is used as a bunker-busting weapon to destroy buried targets without causing substantial above-ground damage. The guidance system has a demonstrated circular error probable (CEP) of 10 m. (This is the radius of the circle around the designated target within which 50% of the penetrators will hit.) To be effective, such a weapon must effectively score a direct hit on the buried target. Therefore, the targeting criterion is that two penetrators must be delivered to within the CEP. How many penetrators must be dropped to achieve a 90% probability of meeting this criterion?

Appendix A
Random Processes

A.1 Introduction

As noted in the introduction to Chapter 12, the material in this appendix is not required for a discussion of system reliability at the level presented in this text. However, some discussion of random processes is useful in connection with the material covered elsewhere in this text, and its treatment logically follows from that already presented. We therefore include the required discussion in this appendix to avoid interrupting the continuity of the material on reliability analysis. As always, we omit derivations that can be found in standard texts, seeking instead to provide the reader with an understanding of the key ideas and results.

A.2 Concept of a Random Process

If a random variable X is a function of time, i.e., $X = X(t)$, then $X(t)$ is said to be a *random process* or *stochastic process*. Unlike simple random variables, random processes are characterized both by their properties at a given time and by their behavior as it evolves across time.

The value of $X(t)$ at any particular time, for example $X(t_0) = x_0$, is a random variable characterized by a probability density function $f(x, t_0)$ and having a mean, variance, etc., just as for any random variable. For example, if the density function is Gaussian, we have by analogy to Eq. (12.30),

$$f(x, t) = \left[\frac{1}{2}\pi\sigma^2(t)\right]^{\frac{1}{2}} \exp\left\{-\frac{[x-\mu(t)]^2}{2\sigma^2(t)}\right\} \tag{A.1}$$

then the process is said to be a Gaussian random process.

A random process governed by a density function that is constant in time is called a stationary process. (Technically, such a process would be strictly stationary, to distinguish it from those that are stationary only through one or more moments of the distribution.[1] This distinction and its implications are well beyond the scope of this text, as is the discussion of nonstationary processes in general.) Note that a stationary process does not imply that any given outcome $X(t_0)$ must be the same as another outcome $X(t_1)$ at a different time. However, the

density function that determines the range and frequency of values for $X(t)$ is not a function of time and can therefore be written as $f(x, t) = f(x)$. The moments of a stationary random process, $E[X(t)]$, $E[X^2(t)]$, etc., are of course also constant; thus, if the Gaussian process of Eq. (A.1) were stationary, μ and σ would be constant.

A given random function $X(t)$ is considered to be a representative sample, or sample function, taken from an ensemble of such functions, denoted by $\{X(t)\}$ and shown graphically in Fig. A.1. A given sample function $X(t)$ may be viewed as the result of a particular trial run of an experiment; the ensemble $\{X(t)\}$ is the set of all trials that could occur. As an example, any sample function in the ensemble shown in Fig. A.1 might represent the attitude history (e.g., pointing angle vs time) of a given spacecraft axis. The entire ensemble might represent the set of all possible attitude histories that could be produced by the given spacecraft operating in its environment.

At any point in time, $\{X(t)\}$ represents all possible values of x that the random process $X(t)$ can produce. This range of values is governed by the density function $f(x, t)$. The expected value of the random process $X(t)$ is then calculated by averaging across the ensemble $\{X(t)\}$ in the usual fashion,

$$E[X(t)] \equiv \int_{-\infty}^{+\infty} xf(x, t)\,dx \equiv \mu(t) \tag{A.2}$$

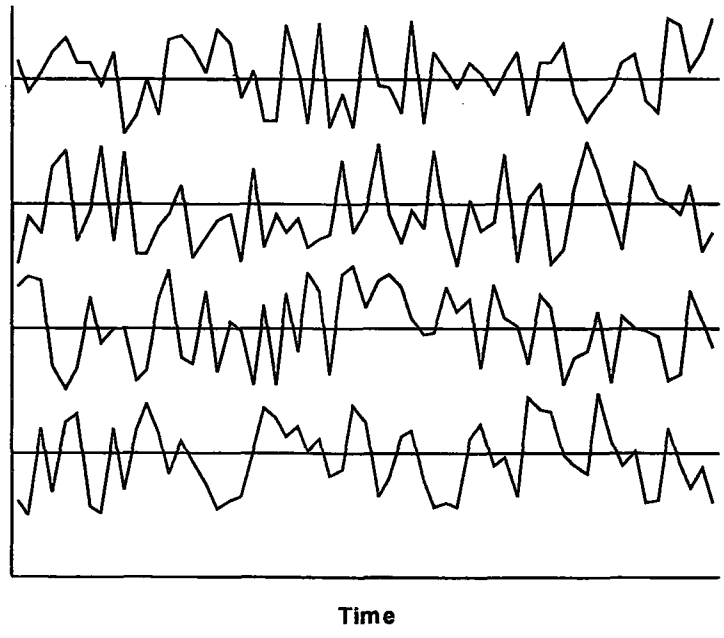


Fig. A.1 Ensemble of sample functions of random process $\{X(t)\}$.

and similarly for the higher order moments, precisely as we have seen earlier for random variables. The mean or expected value of $X(t)$, $E[X(t)]$, is thus seen to be the ensemble average across all possible sample functions $\{X(t)\}$ at time t_0.

Because the random process $X(t)$ evolves in time, it is of course also possible to define and compute moments based on time-averaging the data from a given sample function, i.e.,

$$E[X] = \lim_{T\to\infty} (1/T) \int_{-T/2}^{T/2} X(t)\,dt \tag{A.3}$$

$$E[X^2] = \lim_{T\to\infty} (1/T) \int_{-T/2}^{T/2} X^2(t)\,dt \tag{A.4}$$

An ergodic random process is one for which, loosely speaking, time averages and the ensemble averages are identical. That is, any process statistic (e.g., mean, variance, etc.) is the same regardless of whether the calculation is performed across the members of the ensemble or by averaging the behavior of one sample function from the ensemble over a sufficiently long period of time. Obviously, any ergodic process is stationary; however, the converse is not true.

The ergodic hypothesis, when employed in engineering practice with respect to a given random process, is usually unverifiable but is nonetheless crucial to the practical application of stochastic theory. The difficulty of verification follows from the fact that, as observers, we usually see only one or a few members of the total ensemble of sample functions $\{X(t)\}$. Theoretical work can always proceed under the assumption that a particular probability distribution is of interest, whether this is justified in practice or not. However, in a given application, the time-averaged statistics of a given sample function are typically all with which we have to work. In the preceding attitude control example, the reader will note that we can observe only one attitude history, not the entire set that might have been possible had different trial runs been performed. Thus, since in engineering practice ensemble averages cannot usually be found, time averages of a given sample function will be used to obtain estimates of the mean and variance for the process (only rarely are higher order moments used). The assumption of ergodicity will be applied, and the moments obtained via Eqs. (A.3) and (A.4) will be taken as equivalent to ensemble averages.

A.3 Autocorrelation and Cross-Correlation Functions

As a random process evolves in time, the sample function $X(t)$ is generated and, by analogy with the question of correlation among joint random variables, the issue arises as to the relationship, if any, between $X(t_1)$ and $X(t_2)$. The implications of such a relationship will shortly be seen to have a profound influence on the response of systems to random inputs. This goes to the heart of

several topics discussed earlier in this text, e.g., the response of a spacecraft structure to the random vibration generated by its launch vehicle, the boresight disturbance of a given sensor in response to jitter from a source elsewhere on the spacecraft, etc. In keeping with the assumptions made elsewhere in this text and in this section, we will shortly specialize our discussion to the response of linear time-invariant (LTI), single-input single-output (SISO) systems to ergodic (hence stationary) random processes. For the moment, we begin with the more general case.

Retreating to first principles, and as noted earlier in connection with random variables, there will exist a joint probability distribution function (even if unknown and unknowable) analogous to Eq. (12.22),

$$F(x_1, t_1, x_2, t_2) \equiv \int_{-\infty}^{x_1} dx_1 \int_{-\infty}^{x_2} f(x_1, t_1, x_2, t_2)\mathrm{d}x_2$$
$$= P[X(t_1) \leq x_1 \cap X(t_2) \leq x_2] \tag{A.5}$$

which gives the probability of the joint event that $X(t_1) \leq x_1$ and $X(t_2) \leq x_2$. The joint probability density function $f(x_1, t_1, x_2, t_2)$ is defined by Eq. (A.5), or equivalently as

$$f(x_1, t_1, x_2, t_2) = \frac{\partial^2 F(x_1, t_1, x_2, t_2)}{\partial x_1 \partial x_2} \tag{A.6}$$

Reflecting common engineering practice, the need for analytic tractability, and the oft-stated limitations on the scope of this text, we restrict ourselves to the second-order statistical treatment just implied by considering values of the process at only t_1 and t_2.

As an example of a particularly useful random process, we offer the bivariate Gaussian distribution with $\boldsymbol{x} = (x_1, x_2)$,

$$f(x_1, t_1, x_2, t_2) = f(\boldsymbol{x}, t_1, t_2)$$
$$= \left(\frac{1}{2\pi}\right)|\boldsymbol{P}(t_1, t_2)|^{-1/2}$$
$$\times \exp\left\{-\frac{[\boldsymbol{x} - \boldsymbol{\mu}(t_1, t_2)]^T \boldsymbol{P}^{-1}(t_1, t_2)[\boldsymbol{x} - \boldsymbol{\mu}(t_1, t_2)]}{2}\right\} \tag{A.7}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are given by Eqs. (12.24) and (12.25).

If the joint density function is known, then various moments can be computed in the usual way. The most important of these is the autocorrelation function analogous to Eq. (12.25), and given by the ensemble average

$$E[X(t_1), X(t_2)] = \int_{-\infty}^{x_1} x_1 \, \mathrm{d}x_1 \int_{-\infty}^{x_2} x_2 f(x_1, t_1, x_2, t_2)\mathrm{d}x_2 \equiv R_{XX}(t_1, t_2) \tag{A.8}$$

$E[X(t_1), X(t_2)]$ is seen to be the average correlation, across all sample functions, of the values of the random process $X(t)$ obtained at times t_1 and t_2.

In what follows we will consider the effect of a linear time-invariant system on a random input $X(t)$, thus producing a random output $Y(t)$. A key result of this section will be to describe the statistical properties of $Y(t)$ in terms of both the system parameters and the properties of $X(t)$.

We will therefore be interested the cross-correlation function,

$$E[X(t_1), Y(t_2)] = \int_{-\infty}^{x} x\,dx \int_{-\infty}^{y} yf(x, t_1, y, t_2)dy \equiv R_{XY}(t_1, t_2) \tag{A.9}$$

If we invoke the usual assumption that $X(t)$ and $Y(t)$ are zero-mean processes, then R_{XX} and R_{XY} are covariance functions, exactly as noted earlier in connection with joint random variables. Note that R_{XX}, R_{XY}, R_{YX}, and R_{YY} are identical to the σ_{ij} of Eq. (12.26), defined earlier in connection with joint random variables.

At this point we specialize our discussion to the case in which $X(t)$ and $Y(t)$ are at least stationary processes. It is then clear that R_{XX} and R_{XY}, being results of an expectation operation, cannot depend on t_1 for their computation, but only on the *difference* $\tau = t_2 - t_1$. Equations (A.8) and (A.9) then become

$$E[X(t), X(t+\tau)] = \int_{-\infty}^{x_1} x_1\,dx_1 \int_{-\infty}^{x_2} x_2 f(x_1, t, x_2, t+\tau)\,dx_2$$

$$\equiv R_{XX}(\tau) \tag{A.10}$$

$$E[X(t), Y(t+\tau)] = \int_{-\infty}^{x} x\,dx \int_{-\infty}^{y} yf(x, t, y, t+\tau)\,dy$$

$$\equiv R_{XY}(\tau) \tag{A.11}$$

where, again, the choice of absolute time t is irrelevant.

As before, we note that while the analyst may postulate any desired probability density function and so compute R_{XX} and R_{XY} as ensemble averages, the engineer whose goal is to interpret test or telemetry data has no such luxury. He must work with the single, or very few, sample functions that can be obtained. As discussed earlier, we can integrate over a representative (theoretically infinite) segment of a given sample function to obtain the time-averaged correlation between $X(t)$ and $X(t+\tau)$ to yield

$$R_{XX}(\tau) = \lim_{T\to\infty} (1/T) \int_{-T/2}^{T/2} X(t)\,X(t+\tau)\,dt \tag{A.12}$$

$$R_{XY}(\tau) = \lim_{T\to\infty} (1/T) \int_{-T/2}^{T/2} X(t)\,Y(t+\tau)\,dt \tag{A.13}$$

Under the additional assumption of ergodicity, Eqs. (A.12) and (A.13) are taken equal to the ensemble averages of Eqs. (A1.10) and (A1.11).

The auto- and cross-correlation functions of stationary random processes have several easily derived but interesting properties:

$$R_{XX}(\tau) = R_{XX}(-\tau) \tag{A.14}$$

$$R_{XY}(\tau) = R_{YX}(-\tau) \tag{A.15}$$

$$E[X^2(t)] = R_{XX}(0) \geq |R_{XX}(\tau)| \tag{A.16}$$

The latter property is worth emphasizing; a valid autocorrelation function is symmetric about and attains its maximum value at the origin, reflecting that fact that the highest possible correlation must occur when $X(t)$ is correlated with itself. It is common to normalize the autocorrelation function by the factor $1/E[X^2(t)]$, thus guaranteeing $R_{XX}(0) = 1$.

Finally, we note before leaving this section that a Gaussian random process has the unique analytical advantage that because the distribution is fully characterized if $E[X(t)]$ and $E[X^2(t)]$ are known, knowledge of the autocorrelation function is sufficient to describe the process completely.

It remains to provide an interpretation of the autocorrelation function. Recall that $R_{XX}(\tau)$ is a measure of the degree, on average, to which a given value x_1 of a sample function $X(t)$ at time t_1 is correlated with another given value x_2 from the same sample function at a later time, $t_2 = t_1 + \tau$. Thus, if $R_{XX}(\tau)$ is sharply peaked, later values of $X(t)$ are only poorly correlated with earlier values; knowledge of $X(t)$ at time t_1 will be of little help in predicting $X(t)$ at $t_1 + \tau$. In that way, the stochastic process $X(t)$ is more 'random', in the colloquial sense, than another process with a more broadly peaked autocorrelation function.

The extreme case of a broadly peaked autocorrelation function would be

$$R_{XX}(\tau) = R_{XX}(0) = R_0 \tag{A.17}$$

i.e., the average correlation between $X(t)$ and $X(t + \tau)$ is a constant. While $X(t)$ and $X(t + \tau)$ are separate random variables drawn from the same probability distribution, on average they are correlated to an extent given by the magnitude of R_0. If the process $X(t)$ is viewed as 'noise' that is corrupting an underlying 'signal' of interest, then $X(t)$ may be visualized as a random bias.[2]

At the opposite extreme would be the case in which $R_{XX}(\tau)$ is given by the Dirac delta function,

$$R_{XX}(\tau) = R_0\delta(\tau) \tag{A.18}$$

where $\delta(\tau)$ is defined by the properties

$$\delta(\tau) = \infty, \tau = 0 \tag{A.19a}$$

$$\delta(\tau) = 0, \tau \neq 0 \tag{A19b}$$

and

$$\int_{-\infty}^{\infty} \delta(\tau)\, d\tau = 1 \tag{A.19c}$$

The delta function is the idealized mathematical representation of the unit impulse function first mentioned in Chapter 7 and discussed next; it is a peak of infinitesimal width and infinite height, with unitary area under the curve.

Clearly, when $R_{XX}(\tau) = R_0\delta(\tau)$, the process $X(t)$ has, on average, no correlation at all with $X(t + \tau)$ for any non-zero value of τ; knowledge of $X(t)$ is useless as a predictor of the future behavior of the given sample function. This is the well-known and often-utilized white noise process, to be discussed further in the following section.

A.4 Linear System Response to Random Processes

Our primary interest in the subject of random processes lies in the response of spacecraft systems to random inputs; such inputs are usually considered to be noise, and thus as disturbances to the intended operation of the system. It is therefore desired to characterize the statistical properties of the output, given the input and the system parameters. We consider the elementary case of a linear, single-input, single-output system, for which the response at time t to an input $x(\tau)$ is

$$y(t) = \int_{-\infty}^{t} h(t, \tau)x(\tau)\,d\tau \tag{A.20}$$

The function $h(t, \tau)$ is the impulse response of the system, i.e., the response at time t to the unit impulse $\delta(\tau)$ applied at time τ. For our purposes, it might be the response of one part of a spacecraft structure given an excitation at another point, or it might be an attitude control command issued in response to a unit disturbance. Because linear systems have the property of superposition (the total output of a sum of input signals is the sum of the individual outputs), we can omit any discussion of the desired signal and consider only the behavior of the system in response to the additive noise taken here as the random process $x(t)$.

We note that causal systems require $h(t, \tau) = 0$ for $t < \tau$; there can be no system response prior to the input. Also, if the system is time invariant, the origin in time is irrelevant, and $h(t, \tau) = h(t - \tau)$. Then we can write

$$y(t) = \int_{-\infty}^{t} h(t - \tau)x(\tau)\,d\tau = \int_{0}^{\infty} h(\tau)x(t - \tau)\,d\tau \tag{A.21}$$

The simplification to a LTI SISO system allows us to convert the preceding convolution integral to an algebraic expression, i.e.,

$$Y(\omega) = H(\omega)X(\omega) \tag{A.22}$$

where $H(\omega)$, $X(\omega)$, and $Y(\omega)$ are the Fourier transforms of $h(t)$, $x(t)$, and $y(t)$, with $\omega = 2\pi f$ being the natural frequency. Thus,

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{i\omega t}\,dt \tag{A.23}$$

and has the inverse transform

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{i\omega t}\,d\omega \tag{A.24}$$

and similarly for $y(t)$ and $h(t)$. The Fourier transform may also be obtained analytically from the Laplace transform, with $s \to i\omega$. Most readers will be aware that both Fourier and Laplace transforms are extensively tabulated because of their utility in theoretical work, while practical applications are greatly facilitated by the routine availability of fast Fourier transform (FFT) processors designed for precisely the sorts of tasks indicated here.

If the input function $x(t)$ is a deterministic waveform, such as a step, ramp, or sinusoidal function, then Eqs. (A.22)–(A.24) provide the tools to evaluate $y(t)$ given $x(t)$. However, when $x(t)$ is a random process, analytic evaluation of the Fourier or Laplace transforms is not possible because the sample functions lack recognizable functional form. At best we can seek the statistics of the process $y(t)$, and especially the mean and variance. These provide an indication of the behavior to be expected on average and the deviations that can be expected about that average.

The relationship between the mean values of the input and output is easily obtained by taking the expected value of Eq. (A.21). Since $h(\tau)$ is a deterministic function and $E[x(t)]$ is a constant for an ergodic process, we can exchange the order of time integration and expectation and obtain

$$E[y(t)] = E\left\{\int_0^{\infty} h(\tau)x(t-\tau)\,d\tau\right\} = E[x(t)]\int_0^{\infty} h(\tau)\,d\tau \tag{A.25}$$

With a bit more work it is found that the auto- and cross-correlation functions are related by

$$R_{YY}(\tau) = \int_0^{\infty} h(\tau_2)\,d\tau_2 \int_0^{\infty} h(\tau_1)R_{XX}(\tau+\tau_1-\tau_2)\,d\tau_1 \tag{A.26}$$

and

$$R_{XY}(\tau) = \int_0^{\infty} h(\tau_1)R_{XX}(\tau-\tau_1)\,d\tau_1 \tag{A.27}$$

A.5 Power Spectral Density

In the Fourier transform domain, these convolution integrals yield the much simpler algebraic relationships

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega) \tag{A.28}$$

and

$$S_{XY}(\omega) = |H(\omega)|S_{XX}(\omega) \tag{A.29}$$

where $S_{XX}(\omega)$, $S_{XY}(\omega)$, and $S_{YY}(\omega)$ are the Fourier transforms of $R_{XX}(\tau)$, $R_{XY}(\tau)$ and $R_{YY}(\tau)$, respectively, defined via Eq. (A.23). Once Eqs. (A.28) and (A.29) have been used to obtain $S_{YY}(\omega)$ and $S_{XY}(\omega)$, $R_{YY}(\tau)$ and $R_{XY}(\tau)$ can be obtained using the inverse Fourier transform, Eq. (A.24). From Eq. (A.16), we note that $E[y^2(t)] = R_{YY}(0)$ then gives the variance of the output random process $y(t)$, which is the result we have sought.

The terms $S_{XX}(\omega)$ and $S_{YY}(\omega)$ are known as the *power spectral density* of the random processes $x(t)$ and $y(t)$, respectively, while $S_{XY}(\omega)$ is the *cross-power spectral density* of $x(t)$ and $y(t)$. These terms arise from the general usage of "power" to indicate a squared signal amplitude, while the magnitude of S_{XX}, S_{XY}, and S_{YY} at a specific value of ω gives the power density at that frequency. Indeed, integration of $S_{XX}(\omega)$ from $(-\infty, \infty)$ gives the total power in the signal, while integration between $[\omega_1, \omega_2]$ gives the power in that frequency band. The utility of this approach is obvious, to the point where it is probably more common to characterize joint random processes in terms of their power spectral density, or cross-power spectral density, than otherwise.

When we speak of "white noise," here and in Chapters 7 and 11, the reference is to a process with a constant noise power spectral density across all frequencies, i.e., $S(\omega) = 2\pi S_0 \equiv R_0$. We had earlier referred to the special autocorrelation function, $R_{XX}(\tau) = R_0\,\delta(\tau)$, as the white noise process. Recalling that the Fourier transform of $\delta(\tau)$ is a constant, i.e.,

$$S(\omega) = \int_{-\infty}^{\infty} R_0 \delta(\tau) e^{i\omega\tau}\,d\tau = R_0 e^{i\omega(0)} = R_0 \equiv 2\pi S_0 \qquad \text{(A.30)}$$

establishes the connection between the time- and frequency-domain representations of the white noise process.

As noted earlier, white noise is an idealization; such a signal would have infinite total power and thus cannot actually exist. However, the idealization is quite useful when confined to a specific frequency band; also, many "colored" noise processes can be derived theoretically by passing white noise through a shaping filter defined by $H(\omega)$, exactly as in Eq. (A.28). The use of white Gaussian noise (WGN) is without doubt the single most common assumption in the application of stochastic process theory to real system.[3]

References

[1]Lutes, L. D., and Sarkani, S., *Stochastic Analysis of Structural and Mechanical Vibrations*, Prentice-Hall, Upper Saddle River, NJ, 1997.

[2]Gelb, A. (ed.), *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

[3]Wozencraft, J. M., and Jacobs, I. M., *Principles of Communications Engineering*, Waveland Press, Prospect Heights, IL, 1965.

Appendix B
Tables

Table B.1 SI fundamental units

Quantity	Name
Mass[a]	kilogram, kg
Length[b]	meter, m
Time[c]	second, s
Thermodynamic temperature[d]	Kelvin, K
Electric current[e]	ampere, A
Amount of substance (atoms, molecules, ions)[f]	mole, mol
Luminous intensity[g]	candela, cd

[a]The meter is the distance traveled by light in vacuum during 1/299,792,458 s.

[b]The kilogram is the unit of mass equal to that of the international prototype maintained at the International Bureau of Weights and Measures in Sevres, France.

[c]The second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

[d]The ampere is the current that, if maintained constant in two straight parallel conductors of infinite length and negligible circular cross section, placed 1 m apart in vacuum, would produce between these conductors a force of 2×10^{-7} N/m of length.

[e]The Kelvin is the unit of thermodynamic temperature equal to 1/273.16 of the thermodynamic temperature of the triple point of water.

[f]The mole is the amount of substance of a system that contains as many elementary entities as there are atoms in 0.012 kg of carbon 12, where such atoms are unbound and at rest in their ground state. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles.

[g]The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian.

Table B.2 SI derived units

Quantity	Name	Formulation in terms of SI derived units	Formulation in terms of SI Units
Plane angle[a]	radian, rad	—	$m \cdot m^{-1} = 1$
Solid angle[b]	steradian, sr	—	$m^2 \cdot m^{-2} = 1$
Frequency	hertz, Hz	—	s^{-1}
Force	newton, N	—	$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal, Pa	N/m^2	$kg \cdot m^{-1} \cdot s^{-2}$
Energy, work, quantity of heat	joule, j	$N \cdot m$	$m^2 \cdot kg \cdot s^{-2}$
Power, radiant flux	watt, W	J/s	$m^2 \cdot kg \cdot s^{-3}$
Quantity of electric charge	coulomb, C	—	$s \cdot A$
Electric potential or potential difference; electromotive force	volt, V	W/A	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Capacitance of electric charge	farad, F	C/V	$m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Electrical resistance	ohm, Ω	V/A	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Magnetic flux	weber, Wb	$V \cdot s$	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux density	tesla, T	Wb/m^2	$kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry, H	Wb/A	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Luminous flux	lumen, lm	$cd \cdot sr$	$cd \cdot sr$
Illuminance	lux, lx	lm/m^2	$cd \cdot sr \cdot m^{-2}$
Radioactivity	becquerel, Bq	—	s^{-1}
Absorbed dose	gray, Gy	J/kg	$m^2 \cdot s^{-2}$
Personal dose equivalent	sieyert, Sv	J/kg	$m^2 \cdot s^{-2}$

[a]The radian is the plane angle between two radii with a vertex at the center of a circle of radius r, and which subtend a circumferential arc length r.

[b]The steradian is the solid angle, having its vertex at the center of a sphere of radius r, which subtends an area on the surface of the sphere equal to r^2.

Table B.3 Commonly used quantities in SI derived units

Quantity	SI derived units	SI fundamental units
Angular velocity	rad/s	$m \cdot m^{-1} \cdot s^{-1} = s^{-1}$
Angular acceleration	rad/s^2	$m \cdot m^{-1} \cdot s^{-2} = s^{-2}$
Dynamic viscosity	$Pa \cdot s$	$kg \cdot m^{-1} \cdot s^{-1}$
Moment of force	$N \cdot m$	$m^2 \cdot kg \cdot s^{-2}$
Surface tension	N/m	$kg \cdot s^{-2}$
Heat flux density	W/m^2	$kg \cdot s^{-3}$
Irradiance	W/m^2	$kg \cdot s^{-3}$
Radiant intensity	W/sr	$m^2 \cdot kg \cdot s^{-3} \cdot sr^{-1}$
Radiance	$W/(m^2 \cdot sr)$	$kg \cdot s^{-3} \cdot sr^{-1}$
Heat capacity	J/K	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}$
Entropy	J/K	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}$
Specific heat capacity	$J/(kg \cdot K)$	$m^2 \cdot s^{-2} \cdot K^{-1}$
Specific entropy	$J/(kg \cdot K)$	$m^2 \cdot s^{-2} \cdot K^{-1}$
Specific energy	J/kg	$m^2 \cdot s^{-2}$
Thermal conductivity	$W/(m \cdot K)$	$m \cdot kg \cdot s^{-3} \cdot K^{-1}$
Energy density	J/m^3	$kg \cdot m^{-1} \cdot s^{-2}$
Electric field strength	V/m	$m \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric charge density	C/m^3	$m^{-3} \cdot s \cdot A$
Electric flux density	C/m^2	$m^{-2} \cdot s \cdot A$
Permittivity	F/m	$m^{-3} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Permeability	H/m	$m \cdot kg \cdot s^{-2} \cdot A^{-2}$
Molar energy	J/mol	$m^2 \cdot kg \cdot s^{-2} \cdot mol^{-1}$
Molar entropy	$J/(mol \cdot K)$	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1} \cdot mol^{-1}$
Molar heat capacity	$J/(mol \cdot K)$	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1} \cdot mol^{-1}$
Radiation exposure	C/kg	$kg^{-1} \cdot s \cdot A$
Absorbed dose rate	Gy/s	$m^2 \cdot s^{-3}$

Table B.4 Selected conversion factors

Category	From	To	Multiply by
Length	inch, in.	meter	39.37
	inch	centimeter	2.54
	foot, ft	meter	3.281
	statute mile, mile	kilometer	1.609
	statute mile	foot	5280
	nautical mile, n mile	kilometer	1.852
	nautical mile	foot	6076.1
	ångström, Å	meter	1×10^{-10}
	astronomical unit, AU	meter	1.495979×10^{11}
	light year	kilometer	9.46073×10^{12}
Area	hectare	m^2	1×10^4
	ft^2	m^2	0.0928940
	square yard, yd^2	m^2	0.8361274
	acre	m^2	4046.873
Volume	gallon (U.S.), gal	m^3	3.785412×10^{-3}
	ft^3	m^3	2.83127×10^{-2}
Angle	degree, deg	radian	0.01745329
	minute, '	radian	2.908882×10^{-3}
	second, "	radian	4.848137×10^{-6}
	revolution	radian	6.283185
Angular velocity	revolution per minute, rpm	rad/s	0.1047198
	revolution per minute, rpm	deg/s	6
Mass	ounce, oz	kilogram	0.02835
	pound, lbm	kilogram	0.4536
	slug	kilogram	14.59
	slug	pound	32.17
	ton	kilogram	907.2
Density	$slug/ft^3$	kg/m^3	515.3788
	lbm/ft^3	kg/m^3	16.01846
	lbm/gal (U.S.)	kg/m^3	119.8264
Force	pound-free (lbf)	newton	4.448
	kilogram-force (kgf)	newton	9.807
Pressure	$lbf/in.^2$	lbf/ft^2	144
	$lbf/in.^2$	N/m^2	6895
	lbf/ft^2	N/m^2	47.88
	bar, bar	pascal	1×10^5
	millimeter of mercury, 0°C	pascal	133.3224
	torr	pascal	133.3224
	atmosphere, atm	N/m^2	101,325
	atmosphere	lbf/in^2	14.70
	atmosphere	lbf/ft^2	2116
Energy[a]	British thermal unit, BTU	joule	1055
	ft-lbf	joule	1.356
	BTU	ft-lbf	777.9
	calorie, cal	joule	4.1868

(*continued*)

Table B.4 Selected conversion factors (continued)

Category	From	To	Multiply by
	electron-volt, eV	joule	1.602×10^{-19}
	ton of TNT, explosive energy	joule	4.184×10^{9}
Power[b]	ft-lbf/h	watt	3.766×10^{-4}
	horsepower, hp	watt	745.7
	BTU/h	watt	0.2931
Intensity	$BTU/ft^2 \cdot s$	W/m^2	11,358
	$BTU/ft^2 \cdot hr$	W/m^2	4.089×10^{7}
Temperature[c]	degrees Rankine, °R	Kelvin	1.8
Heat capacity	BTU/lb · °R	J/kg · K	4187
Thermal conductivity	BTU/h-ft · °R	W/m · K	1.731
	BTU/s · ft · °R	W/m · K	6230.6
	$(BTU/ft^2 \cdot s)/(°R/in)$	W/m · K	519.2
	BTU/s · in · °R	W/m · K	7.4768×10^{4}
Magnetic moment	pole-cm	$A \cdot m^2 \cdot$ turns	1000
Magnetic flux	unit pole	weber	1.25664×10^{-7}
Magnetic flux density	gauss, G	tesla	1.0×10^{-4}
Illuminance	Footcandle	lux	10.76391
Luminous flux	Footlambert	cd/m^2	3.426259
Radiation[d]	rad(·)	gray	0.02
	Roentgen, R	C/kg	2.58×10^{-4}
	Roentgen-equivalent-man, rem	sievert	0.02
	Curie, Ci	becquerel	3.7×10^{10}

[a]BTU = BTU (International Table) = 1055.056 J, from the Fifth International Conference on the Properties of Steam (1956). The exact conversion factor is 1055.05585262 J/BTU. The earlier thermochemical quantity $BTU_{th} = 1.054350$ J is based on the thermochemical calorie cal_{th}, where $cal_{th} = 4.184$ J exactly. The BTU is the amount of heat required to raise the temperature of one pound of pure liquid water by 1°F at a temperature of 39°F. Water has its maximum density at 14.5°C = 39°F.

[b]The modern calorie, (International Table) = 4.1868 J exactly, is the amount of heat required to raise the temperature of one gram of pure liquid water from 14.5°C to 15.5°C. The diet Calorie is 1000 calories, archaically denoted the "kilocalorie" or "kcal."

[c]The Celsius, Farenheit, and Rankine temperature scales find common use in engineering. The Rankine scale is a thermodynamic temperature scale (i.e., 0°R = absolute zero) with 1 K = 1.8°R. The degree Celsius, or °C, is equal in magnitude to the Kelvin, and the °F is equal to the °R. The thermodynamic temperature $T_0 = 273.15$ Kelvin = 491.67°R is exactly 0.01 K below the thermodynamic temperature of the triple point of water:

$$T_C = 1.8T_C + 32°F \qquad T_R = T_F + 459.67°R \qquad T_K = T_C + T_0$$

[d]The rad(·), often the rad(Si) or rad(Al) in spacecraft applications, denotes the energy deposition in a given material, and depends both on the nature of the radiation and the nature of material. Thus, care should be taken to include the material specification when using rad(·) to specify radiation dosage; this practice also obviates confusion with the SI rad, the unit of planar angle. Though not an SI unit, usage of the rad(·) is common in engineering. The Roentgen-equivalent-man (rem) may be viewed as a rad(human).

Table B.5 Physical and mathematical constants (Courtesy of NIST)

Constant	Symbol	Value	Units
Circle circumference-to-diameter ratio	π	3.141592654	
Base of natural logarithms	e	2.718281828	
Speed of light in vacuum	c	2.997925×10^{8}	m/s
Planck's constant	h	6.626069×10^{-34}	Js
Boltzmann's constant	k	1.380650×10^{-23}	J/K
Stefan-Boltzmann constant	σ	5.670400×10^{-8}	$W/m^2 \cdot K^4$
Gravitational constant	G	6.67259×10^{-11}	$m^3/kg \cdot s^2$
Avogadro's Number[a]	n_A	6.022142×10^{23}	mol^{-1}
Molar (universal) gas constant[a]	R	8.314472	$J/mol \cdot K$
Volume of ideal gas at 1 atm, °C[a]	V_0	$22.413996 \times 10^{-23}$	m^3/mol
Electron charge	e	1.602176×10^{-19}	C
Electron mass	m_e	9.109382×10^{-31}	kg
Proton mass	m_p	1.672622×10^{-27}	kg

[a]The fundamental SI unit for the amount of a substance is the mole (mol), which contains, by definition, n_A atoms, molecules, or ions of the substance. For an ideal gas (within which intermolecular forces are negligible), Boltzmann's constant represents the energy content of each gas particle per unit temperature change, i.e., $k = 1.380650 \times 10^{-23}$ J/K. The molar energy content per unit temperature change of an ideal gas is $R = kn_A = 8.314472$ J/mol · K. R is thus the ideal gas constant per mole, while k is the gas constant per molecule. $R = R/M$ is the *specific* gas constant, the gas constant per unit mass, where M is the mole weight of the gas. An ideal gas occupies the standard molar volume V_0 under conditions of standard temperature and pressure (273.15 K and 101,325 N/m^2). In engineering it is more common to work with kilomole (kmol) = 1000 mol, for which $N_A = 1000\, n_A$ and $R = 8.314472 \times 10^3$ J/kmol · K.

Table B.6 Physical and astronomical properties of sun, Earth, and moon

Constant	Symbol	Value	Units
Astronomical unit	AU	1.49597871×10^8	km
Earth radius (equatorial)	R_E	6378.136	km
Polar flattening factor	f	0.00335281	
Mass of Earth	M_E	5.9736×10^{24}	kg
	M_E	332,946	M_S
	M_E	81.30059	M_m
Sidereal year	yr	365.25636	days
	yr	3.155815×10^7	s
Sidereal day	d	86164.09	s
Mean solar day (24-h day)	day	86400.0	s
Inertial rotation rate	ω_E	7.292116×10^{-5}	rad/s
Earth gravitational constant	$\mu_E = GM_E$	3.9860×10^5	km^3/s^2
Earth surface acceleration	g	9.80665	m/s^2
Obliquity of ecliptic, J2000		23.43928	deg
Lunar mean radius	R_m	1738	km
Mass of moon	M_m	7.349×10^{22}	kg
Orbital period	τ_m	27.3216	days
Lunar gravitational constant	$\mu_m = GM_m$	4902.801	km^3/s^2
Solar radius, visible	R_S	696,000	km
Solar mass	M_S	1.9891×10^{30}	kg
Solar gravitational constant	$\mu_S = GM_S$	1.32713×10^{11}	km^3/s^2
Solar constant, at 1 AU	I_S	1358	W/m^2
Blackbody temperature	T_S	5780	K

Table B.7 Selected physical properties of the planets (Courtesy NASA/JPL)

Name	Mass, ($\times 10^{23}$kg)	Mean radius, (km)	Sidereal rotation period, (h)	Sidereal orbital period, (yr)	Geometric albedo	Equatorial gravitation, (m/s^2)
Mercury	3.302	2440 ± 1	1407.509	0.2408467	0.106	3.701
Venus	48.685	6051.84 ± 0.01	−5832.444	0.6151973	0.65	8.87
Earth	59.736	6371.01 ± 0.02	23.93419	1.0000174	0.367	9.780327
Mars	6.4185	3389.92 ± 0.04	24.622962	1.8808476	0.15	3.69
Jupiter	18,986	69911 ± 6	9.92425	11.862615	0.52	23.12 ± 0.01
Saturn	5684.6	58232 ± 6	10.65622	29.447498	0.47	8.96 ± 0.01
Uranus	868.32	25362 ± 12	17.24 ± 0.01	84.016846	0.51	8.69 ± 0.01
Neptune	1024.3	24624 ± 21	16.11 ± 0.01	164.79132	0.41	11.00 ± 0.05
Pluto	0.1314 ± 0.0018	1151	153.28	247.92065	0.3	0.655

Table B.8 Physical properties of selected moons (Courtesy NASA/JPL)

Name	μ, (km^3/sec^2)	Mean radius, (km)	Geometric albedo
Earth			
Moon	4902.801 ± 0.001	1737.5 ± 0.1	0.12
Mars			
Phobos	0.0007138 ± 0.0000019	11.1 ± 0.15	0.071 ± 0.012
Deimos	0.0001497 ± 0.0000105	6.2 ± 0.18	0.068 ± 0.007
Jupiter[a]			
Io	5959.91 ± 0.02	1821.6 ± 0.5	0.62
Europa	3202.74 ± 0.02	1560.8 ± 0.5	0.68
Ganymede	9887.83 ± 0.03	2631.2 ± 1.7	0.44
Callisto	7179.29 ± 0.02	2410.3 ± 1.5	0.19
Saturn[b]			
Titan	8978.2 ± 1.0	2575.0 ± 2.0	0.2
Uranus[c]			
Titania	235.3 ± 6.0	788.9 ± 1.8	0.27 ± 0.03
Oberon	201.1 ± 5.0	761.4 ± 2.6	0.23 ± 0.03
Neptune[d]			
Triton	1427.9 ± 3.5	1353.4 ± 0.9	0.756 ± 0.041
Pluto			
Charon	108.0 ± 6.0	593.0 ± 13.0	0.372 ± 0.012

[a]Galilean moons only; a total of 52 are known to exist.
[b]Largest moon only; 30 are known to exist.
[c]Major moons only; 21 are known to exist.
[d]Largest moon only; 11 are known to exist.

Table B.9 Mean planetary orbital elements (Courtesy NASA/JPL) (Epoch J2000 = 1.5 Jan. 2000 = 2451545.0 JD = t_0)

Name	a, AU	e	i, deg	Ω, deg	ω, deg	L,[a] deg
Mercury	0.38709893	0.20563069	7.00487	48.33167	77.45645	252.25084
Venus	0.72333199	0.00677323	3.39471	76.68069	131.53298	181.97973
Earth	1.00000011	0.01671022	0.00005	−11.26064	102.94719	100.46435
Mars	1.52366231	0.09341233	1.85061	49.57854	336.04084	355.45332
Jupiter	5.20336301	0.04839266	1.30530	100.55615	14.75385	34.40438
Saturn	9.53707032	0.05415060	2.48446	113.71504	92.43194	49.94432
Uranus	19.19126393	0.04716771	0.76986	74.22988	170.96424	313.23218
Neptune	30.06896348	0.00858587	1.76917	131.72169	44.97135	304.88003
Pluto	39.48168677	0.24880766	17.14175	110.30347	224.06676	238.92881

[a] $L = M_0 + \Omega + \omega$, where $M_0 = (\mu_S/a^3)^{1/2}(t_0 - t_p)$ = mean anomaly at epoch; t_p = time of perihelion.

Table B.10 Mean orbital elements of selected solar system moons (Courtesy NASA/JPL)

Name	a, km	e	i, deg	Ω, deg	ω, deg	M_0, deg	$2\pi\dot{\Omega}^{-1}$, yr	$2\pi\dot{\omega}^{-1}$, yr
Earth,[a] 2000 Jan. 1.5								
Moon	384,400	0.0554	5.16	125.08	318.15	135.27	18.600	5.997
Mars,[b] 1950 Jan. 1.0								
Phobos	9380	0.0151	1.075	164.931	150.247	92.474	2.262	1.131
Deimos	23,460	0.0002	1.793	339.600	290.496	296.230	54.536	26.892
Jupiter,[b] 1997 Jan. 16								
Io	421,800	0.0041	0.036	44.208	83.898	342.021	1.624	7.431
Europa	671,100	0.0094	0.469	219.383	88.684	171.016	1.394	30.230
Ganymede	1,070,400	0.0011	0.170	63.692	203.214	306.589	59.435	131.32
Callisto	1,882,700	0.0074	0.187	294.195	57.714	180.997	190.36	301.97
Saturn,[b] 1999 Jan. 1.0								
Titan	1,221,900	0.0288	1.634	44.046	172.749	192.132	581.68	3485.64
Uranus,[a] 1980 Jan. 1.0								
Titania	436,300	0.0011	0.079	99.771	284.400	24.614	161.525	195.369
Oberon	583,500	0.0014	0.068	279.771	104.400	283.088	161.52	195.37

(*continued*)

Table B.10 Mean orbital elements of selected solar system moons (Courtesy NASA/JPL) (continued)

Name	a, km	e	i, deg	Ω, deg	ω, deg	M_0, deg	$2\pi\dot{\Omega}^{-1}$, yr	$2\pi\dot{\omega}^{-1}$, yr
Neptune,[b] 1989 Aug. 25.0								
Triton	354,800	0.0000	156.834	172.431	344.046	264.775	397.516	688.126
Pluto,[c] 1986 June 19.0								
Charon	19,410	0.0002	99.089	223.015	209.793	259.960		

[a]Inclination referenced to primary body equatorial plane.

[b]Inclination referenced to Laplace plane. This is the plane containing the satellite's average nodal regression track, also equivalent to the plane normal to the satellite's orbital precession pole. For a satellite sufficiently close to its primary to be perturbed only by the sun and the primary body's gravitational harmonics, the Laplace plane must lie between the primary body's equatorial and orbital planes. Additional perturbations, such as those due to other planets, can produce a Laplace plane outside these limits, as for example with distant satellites of Jupiter or Saturn. For all satellites listed, the Laplace plane is very close to the planet's equatorial plane, with the worst case being Deimos, with a Laplace plane tilt of 0.9 degs.

[c]Inclination referenced to International Celestial Reference Frame (ICRF) equatorial plane. The ICRF is a quasi-inertial frame defined by the radiometric positions of 212 extragalactic sources over the celestial sphere. The position accuracy of these sources has been established to within 1 milli-arcsecond in both right ascension and declination using all applicable very long baseline interferometry (VLBI) data through 1995, a total of 1.6 million observations. The orientation of the ICRF is consistent with the FK5 J2000.0 optical system within the limits of accuracy of the latter frame. The International Celestial Reference Frame has been adopted by the International Astronomical Union as the fundamental celestial reference frame, replacing the FK5 optical frame as of 1 January 1998.

Table B.11 Structural properties of common spacecraft materials[a]

Material	Density, ρ, 10^3 kg/m^2	Ultimate tensile strength, 10^6 N/m^2	Yield tensile strength, 10^6 N/m^2	Young's modulus, E 10^9 N/m^2	Shear modulus, G, 10^9 N/m^2	Poisson's ratio, ν
Aluminum	2.70			68	25.0	
2024-T6	2.80	470	415	72.4	28	0.33
2090-T83, Al-Li	2.59	550	520	76	28	0.34
2219-T62	2.84	415	290	72	27	0.33
6061-O	2.70	125	55	69	26.0	0.33
6061-T6	2.70	310	275	69	26.0	0.33
7075-T6	2.81	570	505	72	26.9	0.33
Beryllium[b]	1.844	370	240	303	135	0.07–0.18
AlBeMet™ 162, HIP	2.07	307	226	193		0.17
AlBeMet™ 162, annealed	2.07	435	321	193		0.17
AlBeMet™ IC910	2.17	207	158	193	84	0.154
Beryllium 170H, optical grade	1.844	448	303	303	135	0.01–0.18
Copper	8.90	221–455[c]	69–365[c]	115	44	0.31
Bronze, Herculoy™	8.52	386–1000[c]	145–483[c]	105	39	
Beryllium-copper 172, annealed	8.25	1207	1034	131	50	0.3
hard	8.25	1379	1241	117	45	0.31
Glass, Pyrex™ 7740 optical	2.23			62.75	26.1	0.2
Vycor™ UV transparent	2.18			66.2	28	0.19
Inconel 600, hot rolled, annealed	8.42	621	248	214	76	0.41
X750, hot rolled	8.25	1296	951	214	76	0.41
Invar, annealed	8.1	490	260	145	56	0.30

(*continued*)

Table B.11 Structural properties of common spacecraft materials[a] (continued)

Material	Density, ρ, 10^3 kg/m^2	Ultimate tensile strength, 10^6 N/m^2	Yield tensile strength, 10^6 N/m^2	Young's modulus, E 10^9 N/m^2	Shear modulus, G, 10^9 N/m^2	Poisson's ratio, ν
Magnesium, pure; sand cast	1.74	90	21	44	19	0.35
pure; extruded	1.74	169–205	69–105	44	19	0.35
annealed sheet	1.74	160–196	90–105	44	19	0.35
hard rolled sheet	1.74	180–220	115–140	44	19	0.35
Monel 400, annealed	8.83	545	207	179	66	0.37
405, cold drawn	8.83	689	517	179	66	0.37
Nickel, annealed	8.88	45		207	76	0.31
Silicon	2.33			112.4	49	0.28
Silicone RTV, adhesive/sealant	1.16	5	5	0.62		
encapsulating	1.2	4	4			
Stainless Steel 304, annealed	8.03	586	241	193	75	0.28
Titanium, pure	4.35/4.51[d]			105	45	0.34
Grade 12, annealed	4.5	620	480	103	41	0.26
Ti-6Al-4V	4.43	900	830	114	44	0.33
Ti-3Al-13V-11Cr	4.84	1241	1172	103	43	0.21
Tungsten	19.3	980	750	400	175	0.28

[a]Material property data vary considerably depending on the source and manufacturing details; definitive values should be obtained from the manufacturer for critical design. Particular attention should be paid to possible deviations about the representative values provided here. Materials selected for inclusion in this table are for preliminary design purposes only and do not represent a comprehensive list of the many choices available and in common use.
[b]Caution: Health hazard due to airborne particles generated during processing. Variations in Poisson's ratio reflect variations in measured data.
[c]Depending on temper, cold or hot rolling process, or other heat treatment during fabrication.
[d]Densities for alpha/beta titanium.

Table B.12 Thermal properties of common spacecraft materials[a]

Material	Density, ρ, 10^3 kg/m^2	Thermal conductivity, κ, W/m · K	Heat capacity, c_p, J/kg · K	Melting point, minimum, K	Thermal expansion coefficient[b], 10^{-6}/K
Aluminum	2.70	210	900	933	24
2024-T6	2.80	155	880	880	23
2090-T83, Al-Li	2.59	88	1203	833	23.6
2219-T62	2.84	120	864	816	12.4
6061-O	2.70	180	896	855	23.6
6061-T6	2.70	167	896	855	23.6
7075-T6	2.81	130	960	805	23.6
Beryllium	1.844	216	1925	1546	12
AlBeMet™ 162, HIP	2.07	210	1560	1355	13.9
AlBeMet™ 162, annealed	2.07	210	1560	917	13.9
AlBeCast™ IC910	2.17	110	1560		14.6
Beryllium 170H, optical grade	1.844	216	1925	1546	11.4
Copper	8.90	385	385	1356	17
Bronze, Herculoy™	8.52	36		1243	18
Beryllium-copper 172, annealed	8.25	115	419	1144	16.7
hard	8.25	115	419	1144	16.7
Glass, Pyrex™ 7740 optical	2.23			1094	3.25
Vycor™ UV transparent	2.18	1.38	750	1803	0.75
Inconel 600, hot rolled, annealed	8.42	15	444	1700	13.3
X750, hot rolled	8.25	12	431	1700	12.6

(*continued*)

Table B.12 Thermal properties of common spacecraft materials[a] (continued)

Material	Density, ρ, 10^3 kg/m^2	Thermal conductivity, κ, W/m · K	Heat capacity, c_p, J/kg · K	Melting point, minimum, K	Thermal expansion coefficient[b], 10^{-6}/K
Invar, annealed	8.1	13.5	514	1700	1.25
Magnesium, pure; sand cast	1.74	159	1025	921	26.1
pure; extruded	1.74	159	1025	921	26.1
annealed sheet	1.74	159	1025	921	26.1
hard rolled sheet	1.74	159	1025	921	26.1
Monel 400, annealed	8.83	22	427	1622	13.9
405, cold drawn	8.83	22	427	1622	13.9
Nickel, annealed	8.88	60.7	460	1738	13.1
Silicon	2.33	124	702	1685	2.49
Silicone RTV, adhesive	1.16	0.18		213–477[c]	350
encapsulating	1.2	0.17		204–493[c]	200
Stainless steel 304	8.03	16.2	500	1727	17.3
Titanium,	4.35/4.51[d]	17	523	1943	8.41
Grade 12, annealed	4.5	19	555	1933	9.9
Ti-6A1-4V	4.43	6.7	526	1877	8.6
Ti-3Al-13V-11Cr	4.84	7.0	502	1922	9.7
Tungsten	19.3	163.3	134	3643	4.4

[a]Material property data vary considerably depending on the source and manufacturing details; definitive values should be obtained from the manufacturer for critical design. Particular attention should be paid to possible deviations about the mean or representative values provided here. Materials selected for inclusion in this table are for preliminary design purposes only and do not represent a comprehensive list of the many choices available and in common use.
[b]At 293 K.
[c]Service temperature range, minimum to maximum.
[d]Densities for alpha/beta titanium.

Table B.13 Absorptivity and emissivity of selected materials[a]

Material/coating	Solar, α	Infrared, ε
Aluminum, polished	0.2	0.1
highly polished	0.1	0.05
black anodized	0.6	0.85
Beryllium, polished	0.4	0.05
Beta cloth	0.4	0.85
Black paint, Martin Black	0.94	0.94
polyurethane	0.95	0.85
Copper, polished	0.3	0.05
highly polished	0.2	0.02
black oxidized	0.7	0.8
FEP (silver) Teflon™, 5 mil	0.11	0.8
2 mil	0.08	0.62
Gold, on aluminum foil	0.26	0.03
Kapton™, 1 mil aluminized, BOL	0.35	0.6
1 mil aluminized, EOL	0.65	0.6
Magnesium, polished	0.2	0.1
Mylar™, 3–5 mil Al-backed	0.18	0.76
Nickel, pure, polished	0.35	0.08
electroplated	0.4	0.05
Quartz, polished	0.06	0.8
Silver, polished	0.02	0.15
Silicon solar cell, with cover	0.8	0.8
Stainless steel, polished	0.4	0.15
Tungsten, highly polished	0.4	0.05
White paint, silicone base	0.25	0.9
TiO_2 base	0.2	0.9

[a]Data are provided for convenience in preliminary design only. Radiative properties data are notoriously sensitive to surface finish, temperature, coating thickness, aging, contamination, solar UV and atomic oxygen exposure, etc. Data in table are for normal total absorptivity and emissivity.

Table B.14 Properties of common gases

Gas	Density,[a] ρ, kg/m^3	Mole weight,[b] M, kg/kmol	Specific gas constant,[b] R, J/kg	Heat capacity,[b] C_p, J/kg · K	Ratio of specific heats,[b] $k = C_p/C_v$	Boiling point,[c] K
Air	1.296	29.0	287	1009	1.40	78.7
Carbon dioxide, CO_2	1.972	44.0	189	858	1.30	194.8
Carbon monoxide, CO	1.255	28.0	297	1017	1.40	81.5
Hydrogen, H_2	0.179	2.0	4157	5234	1.41	4.3
Helium, He	0.090	4.0	2079	14,320	1.67	20.4
Methane, CH_4	0.716	16.0	520	2483	1.32	112.0
Nitrogen, N_2	1.259	28.0	297	1034	1.41	77.6
Oxygen, O_2	1.433	32.0	260	909	1.40	90.4
Propane, C_3H_8	1.972	44.1	189	1645	1.15	228.7

[a]Density at standard conditions, i.e., $p = 1$ atm $= 101{,}325$ N/m^2, $T = 273.15$ K.
[b]For calorically perfect gas at standard temperature, $T = 273.15$ K.
[c]At $p = 1$ atm.

Table B.15 Variation of specific heat ratio of air with temperature[2]

T, °R	$k = C_p/C_v$	T, °R	$k = C_p/C_v$	T, °R	$k = C_p/C_v$
500	1.400	1400	1.355	2600	1.313
600	1.399	1500	1.349	2800	1.309
700	1.396	1600	1.344	3000	1.306
800	1.392	1700	1.339	3500	1.301
900	1.387	1800	1.335	4000	1.298
1000	1.381	1900	1.331	4500	1.296
1100	1.375	2000	1.328	5000	1.294
1200	1.368	2200	1.322		
1300	1.361	2400	1.317		

Table B.16 Standard normal probability density and distribution[a]

z	$f(z)^b$	$F(z)^c$	z	$f(z)^b$	$F(z)^c$
0.0000	0.3989	0.50000	2.1000	0.0440	0.98214
0.1000	0.3970	0.53983	2.2000	0.0355	0.98610
0.2000	0.3910	0.57926	2.3000	0.0283	0.98928
0.3000	0.3814	0.61791	2.4000	0.0224	0.99180
0.4000	0.3683	0.65542	2.5000	0.0175	0.99379
0.5000	0.3521	0.69146	2.6000	0.0136	0.99534
0.6000	0.3332	0.72575	2.7000	0.0104	0.99653
0.7000	0.3123	0.75804	2.8000	0.0079	0.99744
0.8000	0.2897	0.78814	2.9000	0.0060	0.99813
0.9000	0.2661	0.81594	3.0000	0.0044	0.99865
1.0000	0.2420	0.84134	3.1000	0.0033	0.99903
1.1000	0.2179	0.86433	3.2000	0.0024	0.99931
1.2000	0.1942	0.88493	3.3000	0.0017	0.99952
1.3000	0.1714	0.90320	3.4000	0.0012	0.99966
1.4000	0.1497	0.91924	3.5000	0.0009	0.99977
1.5000	0.1295	0.93319	3.6000	0.0006	0.99984
1.6000	0.1109	0.94520	3.7000	0.0004	0.99989
1.7000	0.0940	0.95543	3.8000	0.0003	0.99993
1.8000	0.0790	0.96407	3.9000	0.0002	0.99995
1.9000	0.0656	0.97128	4.0000	0.0001	0.99997
2.0000	0.0540	0.97725	∞	0	1

[a]The general normal probability density function is

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \equiv N(\mu, \sigma)$$

with the cumulative probability distribution function

$$F(x; \mu, \sigma) = \int_{-\infty}^{x} f(x; \mu, \sigma)\, dx$$

Tabulated here is the standard normal density function,

$$f(z; 0, 1) = f(z) = N(0, 1)$$

and its corresponding cumulative distribution $F(z)$, which may be related to the general normal distribution via the variable transformation

$$z = \frac{x-\mu}{\sigma}, \quad \sigma = 1$$

[b]Since $f(z)$ is an even function, $f(-z) = f(z)$.
[c]$F(-z) = 1 - F(z)$, $z \geq 0$.

Table B.17 1976 U.S. Standard Atmosphere

Altitude, km	Temperature, K	Pressure, N/m^2	Density, kg/m^3
0	288.150	1.01325 E+5	1.2250 E+0
1	281.651	8.9876 E+4	1.1117 E+0
2	275.154	7.9501 E+4	1.0066 E+0
3	268.659	7.0121 E+4	9.0925 E−1
4	262.166	6.1660 E+4	8.1935 E−1
5	255.676	5.4048 E+4	7.3643 E−1
6	249.187	4.7217 E+4	6.6011 E−1
7	242.700	4.1105 E+4	5.9002 E−1
8	236.215	3.5651 E+4	5.2579 E−1
9	229.733	3.0800 E+4	4.6706 E−1
10	223.252	2.6449 E+4	4.1351 E−1
11	216.774	2.2699 E+4	3.6480 E−1
12	216.650	1.9399 E+4	3.1194 E−1
13	216.650	1.6579 E+4	2.6660 E−1
14	216.650	1.4170 E+4	2.2786 E−1
15	216.650	1.2111 E+4	1.9476 E−1
16	216.650	1.0352 E+4	1.6647 E−1
17	216.650	8.8497 E+3	1.4230 E−1
18	216.650	7.5652 E+3	1.2165 E−1
19	216.650	6.4674 E+3	1.0400 E−1
20	216.650	5.5293 E+3	8.8910 E−2
21	217.581	4.7289 E+3	7.5715 E−2
22	218.574	4.0475 E+3	6.4510 E−2
23	219.567	3.4668 E+3	5.5006 E−2
24	220.560	2.9717 E+3	4.6938 E−2
25	221.552	2.5492 E+3	4.0084 E−2
26	222.544	2.1883 E+3	3.4257 E−2
27	223.536	1.8799 E+3	2.9298 E−2
28	224.527	1.6161 E+3	2.5076 E−2
29	225.518	1.3904 E+3	2.1478 E−2
30	226.509	1.1970 E+3	1.8410 E−2
31	227.500	1.0312 E+3	1.5792 E−2
32	228.490	8.8906 E+2	1.3555 E−2
33	230.973	7.6730 E+2	1.1573 E−2
34	233.743	6.6341 E+2	9.8874 E−3
35	236.513	5.7459 E+2	8.4634 E−3
36	239.282	4.9852 E+2	7.2579 E−3
37	242.050	4.3324 E+2	6.2355 E−3
38	244.818	3.7713 E+2	5.3666 E−3
39	247.584	3.2882 E+2	4.6268 E−3
40	250.350	2.8714 E+2	3.9957 E−3
41	253.114	2.5113 E+2	3.4564 E−3

(continued)

Table B.17 1976 U.S. Standard Atmosphere (continued)

Altitude, km	Temperature, K	Pressure, N/m^2	Density, kg/m^3
42	255.878	2.1996 E+2	2.9948 E−3
43	258.641	1.9295 E+2	2.5989 E−3
44	261.403	1.6949 E+2	2.2589 E−3
45	264.164	1.4910 E+2	1.9663 E−3
46	266.925	1.3134 E+2	1.7142 E−3
47	269.684	1.1585 E+2	1.4965 E−3
48	270.650	1.0229 E+2	1.3167 E−3
49	270.650	9.0336 E+1	1.1628 E−3
50	270.650	7.9779 E+1	1.0269 E−3
51	270.650	7.0458 E+1	9.0690 E−4
52	269.031	6.2214 E+1	8.0562 E−4
53	266.277	5.4873 E+1	7.1791 E−4
54	263.524	4.8337 E+1	6.3901 E−4
55	260.771	4.2525 E+1	5.6810 E−4
56	258.019	3.7362 E+1	5.0445 E−4
57	255.268	3.2782 E+1	4.4738 E−4
58	252.518	2.8723 E+1	3.9627 E−4
59	249.769	2.5132 E+1	3.5054 E−4
60	247.021	2.1958 E+1	3.0968 E−4
61	244.274	1.9157 E+1	2.7321 E−4
62	241.527	1.6688 E+1	2.4071 E−4
63	238.781	1.4515 E+1	2.1178 E−4
64	236.036	1.2605 E+1	1.8605 E−4
65	233.292	1.0929 E+1	1.6321 E−4
66	230.549	9.4609 E+0	1.4296 E−4
67	227.807	8.1757 E+0	1.2503 E−4
68	225.065	7.0529 E+0	1.0917 E−4
69	222.325	6.0736 E+0	9.5171 E−5
70	219.585	5.2209 E+0	8.2829 E−5
71	216.846	4.4795 E+0	7.1966 E−5
72	214.263	3.8362 E+0	6.2374 E−5
73	212.308	3.2802 E+0	5.3824 E−5
74	210.353	2.8008 E+0	4.6386 E−5
75	208.399	2.3881 E+0	3.9921 E−5
76	206.446	2.0333 E+0	3.4311 E−5
77	204.493	1.7286 E+0	2.9448 E−5
78	202.541	1.4673 E+0	2.5239 E−5
79	200.590	1.2437 E+0	2.1600 E−5
80	198.639	1.0524 E+0	1.8458 E−5
81	196.688	8.8923 E−1	1.5750 E−5
82	194.739	7.5009 E−1	1.3418 E−5
83	192.790	6.3167 E−1	1.1414 E−5

(continued)

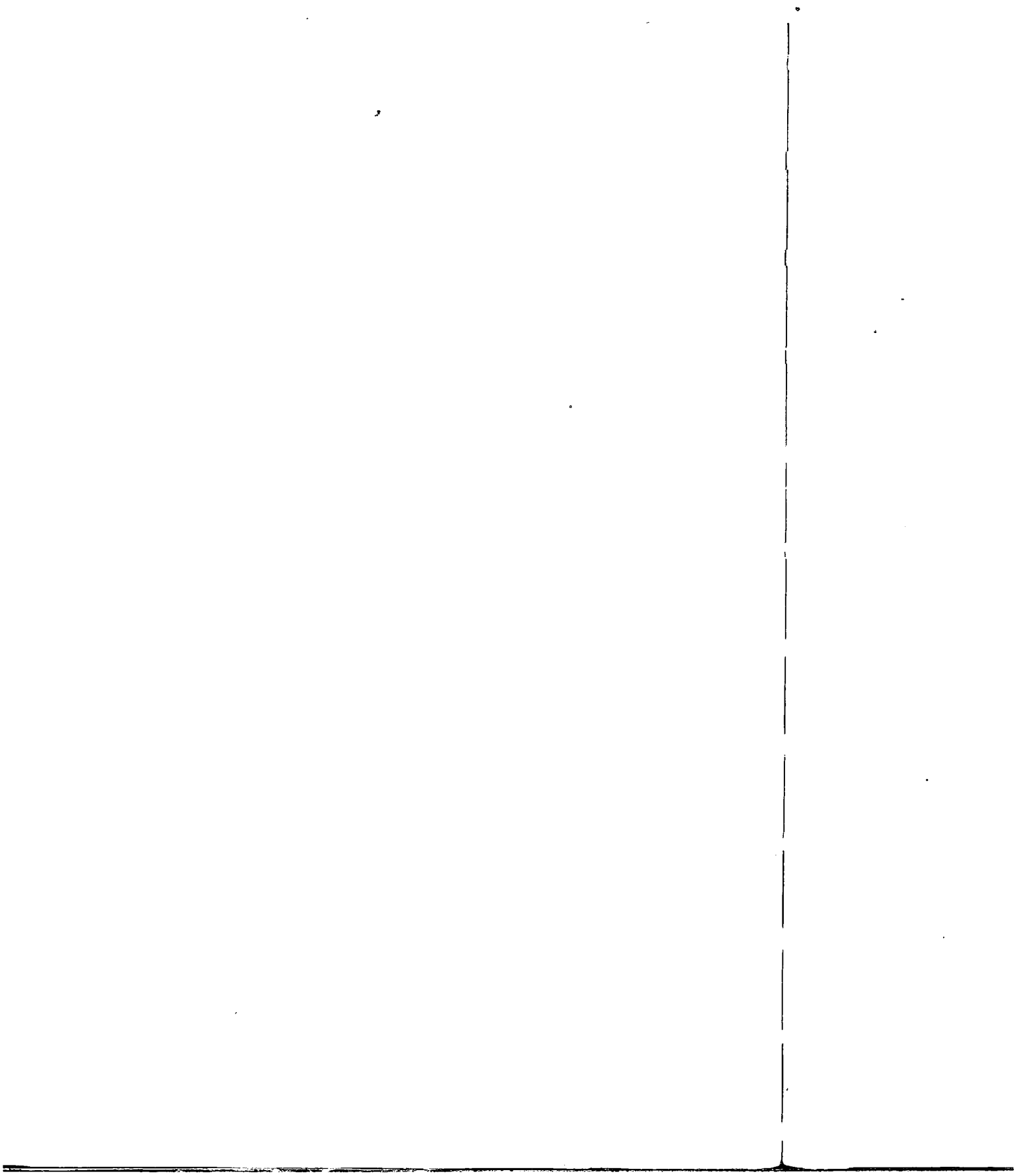
Table B.17 1976 U.S. Standard Atmosphere (continued)

Altitude, km	Temperature, K	Pressure, N/m^2	Density, kg/m^3
84	190.841	5.3105 E−1	9.6940 E−6
85	188.893	4.4568 E−1	8.2196 E−6
86	186.87	3.7338 E−1	6.958 E−6
87	186.87	3.1259 E−1	5.824 E−6
88	186.87	2.6173 E−1	4.875 E−6
89	186.87	2.1919 E−1	4.081 E−6
90	186.87	1.8359 E−1	3.416 E−6
91	186.87	1.5381 E−1	2.860 E−6
92	186.96	1.2887 E−1	2.393 E−6
93	187.25	1.0801 E−1	2.000 E−6
94	187.74	9.0560 E−2	1.670 E−6
95	188.42	7.5966 E−2	1.393 E−6
96	189.31	6.3765 E−2	1.162 E−6
97	190.40	5.3571 E−2	9.685 E−7
98	191.72	4.5057 E−2	8.071 E−7
99	193.28	3.7948 E−2	6.725 E−7
100	195.08	3.2011 E−2	5.604 E−7
110	240.00	7.1042 E−3	9.708 E−8
120	360.00	2.5382 E−3	2.222 E−8
130	469.27	1.2505 E−3	8.152 E−9
140	559.63	7.2028 E−4	3.831 E−9
150	634.39	4.5422 E−4	2.076 E−9
160	696.29	3.0395 E−4	1.233 E−9
170	747.57	2.1210 E−4	7.815 E−10
180	790.07	1.5271 E−4	5.194 E−10
190	825.31	1.1266 E−4	3.581 E−10
200	854.56	8.4736 E−5	2.541 E−10
210	878.84	6.4756 E−5	1.846 E−10
220	899.01	5.0149 E−5	1.367 E−10
230	915.78	3.9276 E−5	1.029 E−10
240	929.73	3.1059 E−5	7.858 E−11
250	941.33	2.4767 E−5	6.073 E−11
260	950.99	1.9894 E−5	4.742 E−11
270	959.04	1.6083 E−5	3.738 E−11
280	965.75	1.3076 E−5	2.971 E−11
290	971.34	1.0685 E−5	2.378 E−11
300	976.01	8.7704 E−6	1.916 E−11
310	979.90	7.2285 E−6	1.552 E−11
320	983.16	5.9796 E−6	1.264 E−11
330	985.88	4.9630 E−6	1.035 E−11
340	988.15	4.1320 E−6	8.503 E−12
350	990.06	3.4498 E−6	7.014 E−12

(*continued*)

Table B.17 1976 U.S. Standard Atmosphere (continued)

Altitude, km	Temperature, K	Pressure, N/m^2	Density, kg/m^3
360	991.65	2.8878 E−6	5.805 E−12
370	992.98	2.4234 E−6	4.820 E−12
380	994.10	2.0384 E−6	4.013 E−12
390	995.04	1.7184 E−6	3.350 E−12
400	995.83	1.4518 E−6	2.803 E−12



Bibliography

AIAA Aerospace Design Engineers Guide, 5th ed., AIAA, Reston, VA, 2003.

"Equations, Tables, and Charts for Compressible Flow," Ames Aeronautical Laboratory, National Advisory Committee for Aeronautics, TR-1135, U.S. Government Printing Office, Washington, DC, 1953.

Baumeister, T. F., Avallone, E. A., and Baumeister, T. F. III (eds.), *Marks Standard Handbook for Mechanical Engineers*, 8th ed., McGraw-Hill, New York, 1978.

Gray, D. E. (ed.), *American Institute of Physics Handbook*, 3rd ed., McGraw-Hill, New York, 1972.

Ma, C., Arias, E. F., Eubankds, T. M., Fey, A. L., Gontier, A.-M., Jacobs, C. S., Sovers, O. J., Archinal, B. A., and Charlot, P., "The International Celestial Reference Frame as Realized by Very Long Baseline Interferometry," *Astronomical Journal*, Vol. 116, July 1998, pp. 516–546.

Mechtly, E. A., "The International System of Units," 2nd Rev., NASA SP-7012, 1973.

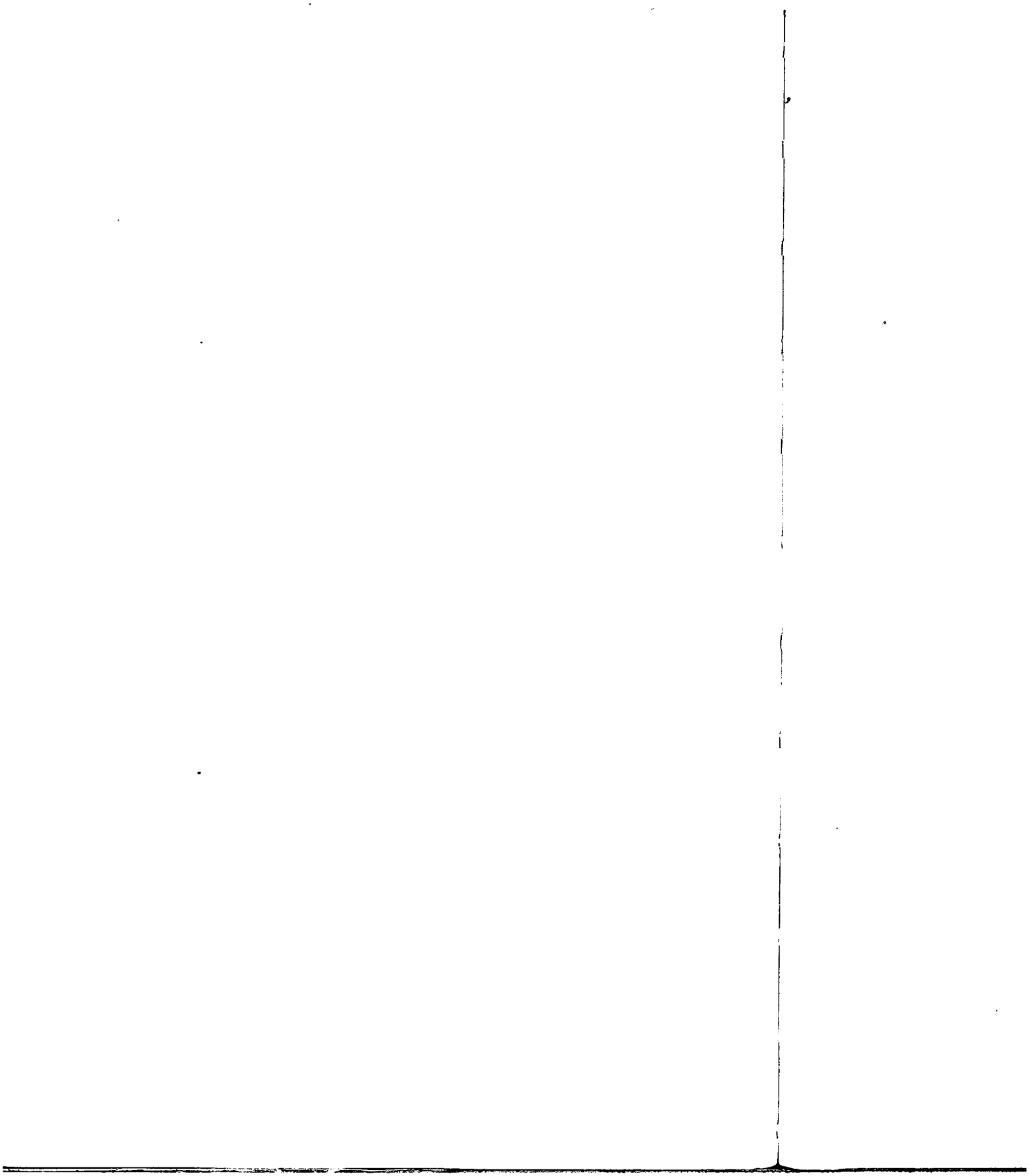
Taylor, B. N. (ed.), "The International System of Units (SI)," National Institute of Standards and Technology, NIST-SP-330, U.S. Government Printing Office, Washington, DC, 1991.

Taylor, B. N., "Guide for the Use of the International System of Units (SI)," National Institute of Standards and Technology, NIST-SP-811, U.S. Government Printing Office, Washington, DC, 1995.

U.S. Standard Atmosphere, National Oceanic and Atmospheric Administration, NOAA S/T 76-1562, U.S. Government Printing Office, Washington, DC, 1976.

Weissman, P. R., McFadden, L.-A., and Johnson, T. V. (eds.), *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999.

Yoder, C. F., "Astrometric and Geodetic Properties of Earth and the Solar System," *Global Earth Physics, A Handbook of Physical Constants, AGU Reference Shelf 1*, American Geophysical Union, Tables 6, 7, 10, 1995.



Index

TEXTS PUBLISHED IN THE AIAA EDUCATION SERIES

Space Vehicle Design, Second Edition
Michael D. Griffin and James R. French 2004
ISBN 1-56347-539-1

Performance, Stability, Dynamics, and Control of Airplanes, Second Edition
Bandu N. Pamadi 2004
ISBN 1-56347-583-9

Applied Mathematics in Integrated Navigation Systems, Second Edition
Robert M. Rogers 2003
ISBN 1-56347-656-8

Finite Element Multidisciplinary Analysis, Second Edition
K. K. Gupta and J. L. Meek 2003
ISBN 1-56347-580-4

Flight Testing of Fixed-Wing Aircraft
Ralph D. Kimberlin 2003
ISBN 1-56347-564-2

The Fundamentals of Aircraft Combat Survivability Analysis and Design, Second Edition
Robert E. Ball 2003
ISBN 1-56347-582-0

Analytical Mechanics of Space Systems
Hanspeter Schaub and John L. Junkins 2003
ISBN 1-56347-563-4

Introduction to Aircraft Flight Mechanics
Thomas R. Yechout with Steven L. Morris, David E. Bossert, and Wayne F. Hallgren 2003
ISBN 1-56347-577-4

Aircraft Design Projects for Engineering Students
Lloyd Jenkinson and James Marchman 2003
ISBN 1-56347-619-3

Elements of Spacecraft Design
Charles D. Brown 2002
ISBN 1-56347-524-3

Civil Avionics Systems
Ian Moir and Allan Seabridge 2002
ISBN 1-56347-589-8

Helicopter Test and Evaluation
Alastair K. Cooke and Eric W. H. Fitzpatrick 2002
ISBN 1-56347-578-2

Aircraft Engine Design, Second Edition
Jack D. Mattingly, William H. Heiser, and David T. Pratt 2002
ISBN 1-56347-538-3

Dynamics, Control, and Flying Qualities of V/STOL Aircraft
James A. Franklin 2002
ISBN 1-56347-575-8

Orbital Mechanics, Third Edition
Vladimir A. Chobotov, Editor 2002
ISBN 1-56347-537-5

Basic Helicopter Aerodynamics, Second Edition
John Seddon and Simon Newman 2001
ISBN 1-56347-510-3

Aircraft Systems: Mechanical, Electrical, and Avionics Subsystems Integration
Ian Moir and Allan Seabridge 2001
ISBN 1-56347-506-5

Design Methodologies for Space Transportation Systems
Walter E. Hammond 2001
ISBN 1-56347-472-7

Tactical Missile Design
Eugene L. Fleeman 2001
ISBN 1-56347-494-8

Flight Vehicle Performance and Aerodynamic Control
Frederick O. Smetana *2001*
ISBN 1-56347-463-8

Modeling and Simulation of Aerospace Vehicle Dynamics
Peter H. Zipfel *2000*
ISBN 1-56347-456-6

Applied Mathematics in Integrated Navigation Systems
Robert M. Rogers *2000*
ISBN 1-56347-445-X

Mathematical Methods in Defense Analyses, Third Edition
J. S. Przemieniecki *2000*
ISBN 1-56347-396-6

Finite Element Multidisciplinary Analysis
Kajal K. Gupta and John L. Meek *2000*
ISBN 1-56347-393-3

Aircraft Performance: Theory and Practice
M. E. Eshelby *1999*
ISBN 1-56347-398-4

Space Transportation: A Systems Approach to Analysis and Design
Walter E. Hammond *1999*
ISBN 1-56347-032-2

Civil Jet Aircraft Design
Lloyd R. Jenkinson, Paul Simpkin, and Darren Rhodes *1999*
ISBN 1-56347-350-X

Structural Dynamics in Aeronautical Engineering
Maher N. Bismarck-Nasr *1999*
ISBN 1-56347-323-2

Intake Aerodynamics, Second Edition
E. L. Goldsmith and J. Seddon *1999*
ISBN 1-56347-361-5

Integrated Navigation and Guidance Systems
Daniel J. Biezad *1999*
ISBN 1-56347-291-0

Aircraft Handling Qualities
John Hodgkinson *1999*
ISBN 1-56347-331-3

Performance, Stability, Dynamics, and Control of Airplanes
Bandu N. Pamadi *1998*
ISBN 1-56347-222-8

Spacecraft Mission Design, Second Edition
Charles D. Brown *1998*
ISBN 1-56347-262-7

Computational Flight Dynamics
Malcolm J. Abzug *1998*
ISBN 1-56347-259-7

Space Vehicle Dynamics and Control
Bong Wie *1998*
ISBN 1-56347-261-9

Introduction to Aircraft Flight Dynamics
Louis V. Schmidt *1998*
ISBN 1-56347-226-0

Aerothermodynamics of Gas Turbine and Rocket Propulsion, Third Edition
Gordon C. Oates *1997*
ISBN 1-56347-241-4

Advanced Dynamics
Shuh-Jing Ying *1997*
ISBN 1-56347-224-4

Introduction to Aeronautics: A Design Perspective
Steven A. Brandt, Randall J. Stiles, John J. Bertin, and Ray Whitford *1997*
ISBN 1-56347-250-3

Introductory Aerodynamics and Hydrodynamics of Wings and Bodies: A Software-Based Approach
Frederick O. Smetana *1997*
ISBN 1-56347-242-2

An Introduction to Aircraft Performance
Mario Asselin *1997*
ISBN 1-56347-221-X

Orbital Mechanics, Second Edition
Vladimir A. Chobotov, Editor *1996*
ISBN 1-56347-179-5

Thermal Structures for Aerospace Applications
Earl A. Thornton *1996*
ISBN 1-56347-190-6

Structural Loads Analysis for Commercial Transport Aircraft: Theory and Practice
Ted L. Lomax *1996*
ISBN 1-56347-114-0

Spacecraft Propulsion
Charles D. Brown *1996*
ISBN 1-56347-128-0

Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modeling
Gareth D. Padfield *1996*
ISBN 1-56347-205-8

Flying Qualities and Flight Testing of the Airplane
Darrol Stinton *1996*
ISBN 1-56347-117-5

Flight Performance of Aircraft
S. K. Ojha *1995*
ISBN 1-56347-113-2

Operations Research Analysis in Test and Evaluation
Donald L. Giadrosich *1995*
ISBN 1-56347-112-4

Radar and Laser Cross Section Engineering
David C. Jenn *1995*
ISBN 1-56347-105-1

Introduction to the Control of Dynamic Systems
Frederick O. Smetana *1994*
ISBN 1-56347-083-7

Tailless Aircraft in Theory and Practice
Karl Nickel and Michael Wohlfahrt *1994*
ISBN 1-56347-094-2

Mathematical Methods in Defense Analyses, Second Edition
J. S. Przemieniecki *1994*
ISBN 1-56347-092-6

Hypersonic Aerothermodynamics
John J. Bertin *1994*
ISBN 1-56347-036-5

Hypersonic Airbreathing Propulsion
William H. Heiser and David T. Pratt *1994*
ISBN 1-56347-035-7

Practical Intake Aerodynamic Design
E. L. Goldsmith and J. Seddon *1993*
ISBN 1-56347-064-0

Acquisition of Defense Systems
J. S. Przemieniecki, Editor *1993*
ISBN 1-56347-069-1

Dynamics of Atmospheric Re-Entry
Frank J. Regan and Satya M. Anandakrishnan *1993*
ISBN 1-56347-048-9

Introduction to Dynamics and Control of Flexible Structures
John L. Junkins and Youdan Kim *1993*
ISBN 1-56347-054-3

Spacecraft Mission Design
Charles D. Brown *1992*
ISBN 1-56347-041-1

Rotary Wing Structural Dynamics and Aeroelasticity
Richard L. Bielawa *1992*
ISBN 1-56347-031-4

Aircraft Design: A Conceptual Approach, Second Edition
Daniel P. Raymer *1992*
ISBN 0-930403-51-7

Optimization of Observation and Control Processes
Veniamin V. Malyshev, Mihkail N. Krasilshikov, and Valeri I. Karlov 1992
ISBN 1-56347-040-3

Nonlinear Analysis of Shell Structures
Anthony N. Palazotto and Scott T. Dennis 1992
ISBN 1-56347-033-0

Orbital Mechanics
Vladimir A. Chobotov, Editor 1991
ISBN 1-56347-007-1

Critical Technologies for National Defense
Air Force Institute of Technology 1991
ISBN 1-56347-009-8

Space Vehicle Design
Michael D. Griffin and James R. French 1991
ISBN 0-930403-90-8

Defense Analyses Software
J. S. Przemieniecki 1990
ISBN 0-930403-91-6

Inlets for Supersonic Missiles
John J. Mahoney 1990
ISBN 0-930403-79-7

Introduction to Mathematical Methods in Defense Analyses
J. S. Przemieniecki 1990
ISBN 0-930403-71-1

Basic Helicopter Aerodynamics
J. Seddon 1990
ISBN 0-930403-67-3

Aircraft Propulsion Systems Technology and Design
Gordon C. Oates, Editor 1989
ISBN 0-930403-24-X

Boundary Layers
A. D. Young 1989
ISBN 0-930403-57-6

Aircraft Design: A Conceptual Approach
Daniel P. Raymer 1989
ISBN 0-930403-51-7

Gust Loads on Aircraft: Concepts and Applications
Frederic M. Hoblit 1988
ISBN 0-930403-45-2

Aircraft Landing Gear Design: Principles and Practices
Norman S. Currey 1988
ISBN 0-930403-41-X

Mechanical Reliability: Theory, Models and Applications
B. S. Dhillon 1988
ISBN 0-930403-38-X

Re-Entry Aerodynamics
Wilbur L. Hankey 1988
ISBN 0-930403-33-9

Aerothermodynamics of Gas Turbine and Rocket Propulsion, Revised and Enlarged
Gordon C. Oates 1988
ISBN 0-930403-34-7

Advanced Classical Thermodynamics
George Emanuel 1987
ISBN 0-930403-28-2

Radar Electronic Warfare
August Golden Jr. 1987
ISBN 0-930403-22-3

An Introduction to the Mathematics and Methods of Astrodynamics
Richard H. Battin 1987
ISBN 0-930403-25-8

Aircraft Engine Design
Jack D. Mattingly, William H. Heiser, and Daniel H. Daley 1987
ISBN 0-930403-23-1

Gasdynamics: Theory and Applications
George Emanuel *1986*
ISBN 0-930403-12-6

Composite Materials for Aircraft Structures
Brian C. Hoskin and Alan A. Baker, Editors *1986*
ISBN 0-930403-11-8

Intake Aerodynamics
J. Seddon and E. L. Goldsmith *1985*
ISBN 0-930403-03-7

The Fundamentals of Aircraft Combat Survivability Analysis and Design
Robert E. Ball *1985*
ISBN 0-930403-02-9

Aerothermodynamics of Aircraft Engine Components
Gordon C. Oates, Editor *1985*
ISBN 0-915928-97-3

Aerothermodynamics of Gas Turbine and Rocket Propulsion
Gordon C. Oates *1984*
ISBN 0-915928-87-6

Re-Entry Vehicle Dynamics
Frank J. Regan *1984*
ISBN 0-915928-78-7