IRIS FailSafe™ 2.0 Programmer's Guide

Document Number 007-3900-001

CONTRIBUTORS

Written by Lori Johnson
Illustrated by Dany Galgani
Edited by Rick Thompson
Production by Linda Rae Sande
Engineering contributions by Michael Nishimoto, Bill Sparks, Paddy Sreenivasan, Dan Stekloff, Rebecca Underwood, and Manish Verma
St. Peter's Basilica image courtesy of ENEL SpA and InfoByte SpA. Disk Thrower image courtesy of Xavier Berenguer, Animatica.

IRIS FailSafe™ 2.0 Programmer's Guide
Document Number 007-3900-001

Contents

List of Figures

List of Tables

About This Guide

This guide explains how to write the set of scripts that are required to turn an application into a high-availability service in conjunction with IRIS FailSafe 2.0 software. It also tells you how to create a new resource type and provides instructions for migrating script information from Release 1.2 to Release 2.0.

This guide assumes that the IRIS FailSafe system has been configured as described in the *IRIS FailSafe 2.0 Administrator's Guide.*

This guide was prepared in conjunction with Release 2.0 of IRIS FailSafe.

Audience

This guide is written for system programmers who are developing scripts for the IRIS FailSafe system. These scripts allow the failover of applications that are not handled by the base and optional IRIS FailSafe products. These programmers must be familiar with the operation and administration of nodes running IRIS FailSafe, with the applications that are to be failed over, and with the *IRIS FailSafe 2.0 Administrator's Guide*.

Structure of This Document

This guide contains the following chapters:

- Chapter 1, "Introduction to IRIS FailSafe Programming"
- Chapter 2, "Using the Script Library"
- Chapter 3, "Writing the Action Scripts and Adding Monitoring Agents"
- Chapter 4, "Creating a Failover Policy"
- Chapter 5, "Defining a New Resource Type"

- Chapter 6, “Testing Scripts”
- Appendix A, “Migrating to IRIS FailSafe 2.0”

A glossary is also provided.

Related Documentation

Besides this guide, other documentation for the IRIS FailSafe system includes

- *IRIS FailSafe 2.0 Administrator’s Guide*
- *IRIS FailSafe 2.0 INFORMIX Administrator’s Guide* (IRIS FailSafe INFORMIX option)
- *IRIS FailSafe 2.0 NFS Administrator’s Guide* (IRIS FailSafe NFS option)
- *IRIS FailSafe 2.0 Oracle Administrator’s Guide* (IRIS FailSafe Oracle option)
- *IRIS FailSafe 2.0 Netscape Server Administrator’s Guide* (IRIS FailSafe Netscape Web option)

The IRIS FailSafe reference pages are as follows:

- cbeutil(1M)
- cdbBackup(1M)
- cdbRestore(1M)
- cdbutil(1M)
- cluster_mgr(1M)
- failsafe(7M)
- fs2d(1M)
- ha_cilog(1M)
- ha_cmsd(1M)
- ha_exec2(1M)
- ha_fsd(1M)
- ha_gcd(1M)
- ha_ifd(1M)

- ha_ifdadmin(1M)
- ha_macconfig2(1M)
- ha_srmd(1M)
- ha_statd2(1M)
- haStatus(1M)

Release notes are included with each IRIS FailSafe product. The names of the release notes are as follows:

Table i IRIS FailSafe Release Notes

Release Note	Product
failsafe2	IRIS 2.0 FailSafe
cluster_ha	Cluster high availability services
cluster_admin	Cluster administration services
cluster_control	Cluster node control services

Conventions Used in This Guide

These type conventions and symbols are used in this guide:

Bold Function names literal command-line arguments (options/flags)

`Bold fixed-width type`
Commands and text that you are to type literally in response to shell and command prompts

Italics New terms, manual/book titles, commands, variable command-line arguments, filenames, and variables to be supplied by the user in examples, code, and syntax statements

`Fixed-width type`
Code examples, error messages, prompts, and screen text

`#` IRIX shell prompt for the superuser (*root*)

Chapter 1

Introduction to IRIS FailSafe Programming

IRIS FailSafe 2.0 provides high-availability services for a cluster that contains up to 8 nodes. High-availability services are monitored by the IRIS FailSafe software. You can create additional high-availability services by using the instructions in this guide.

This chapter provides an introduction to IRIS FailSafe programming. The sections are as follows:

IRIS FailSafe Concepts

In order to use IRIS FailSafe, you must understand the following concepts.

Cluster Node (or Node)

A *cluster node* is a single IRIX image. Usually, a cluster node is an individual computer. The term *node* is also used in this guide for brevity; this use of node does not have the same meaning as a node in an Origin system.

Pool

A *pool* is the entire set of nodes involved with a group of clusters. The group of clusters are usually close together and should always serve a common purpose. A replicated database is stored on each node in the pool.

Cluster

A *cluster* is a collection of one or more nodes coupled to each other by networks or other similar interconnections. A cluster is identified by a simple name; this name must be unique within the pool. A particular node may be a member of only one cluster. All nodes in a cluster are also in the pool; however, all nodes in the pool are not necessarily in the cluster.

Node Membership

A *node membership* is the list of nodes in a cluster on which IRIS FailSafe can allocate resource groups.

Process Membership

A *process membership* is the list of process instances in a cluster that form a process group. There can be multiple process groups per node.

Resource

A *resource* is a single physical or logical entity that provides a service to clients or other resources. For example, a resource can be a single disk volume, a particular network address, or an application such as a web server. A resource is generally available for use over time on two or more nodes in a cluster, although it can be allocated to only one node at any given time.

Resources are identified by a *resource name* and a *resource type*. One resource can be dependent on one or more other resources; if so, it will not be able to start (that is, be made available for use) unless the dependent resources are also started. Dependent resources must be part of the same *resource group* and are identified in a *resource dependency list*.

Resource Name

A *resource name* identifies a specific instance of a *resource type*. A resource name must be unique for a given resource type.

Resource Type

A *resource type* is a particular class of resource. All of the resources in a particular resource type can be handled in the same way for the purposes of *failover*. Every resource is an instance of exactly one resource type.

A resource type is identified by a simple name; this name must be unique within the cluster. A resource type can be defined for a specific node, or it can be defined for an entire cluster. A resource type that is defined for a specific node overrides a clusterwide resource type definition with the same name; this allows an individual node to override global settings from a clusterwide resource type definition.

Like resources, a resource type can be dependent on one or more other resource types. If such a dependency exists, at least one instance of each of the dependent resource types must be defined. For example, a resource type named *Netscape_web* might have resource type dependencies on resource types named *IP_address* and *volume*. If a resource named *web1* is defined with the *Netscape_web* resource type, then the resource group containing *web1* must also contain at least one resource of the type *IP_address* and one resource of the type *volume*.

The IRIS FailSafe software includes many predefined resource types. If these types fit the application you want to make into a high-availability service, you can reuse them. If none fit, you can create additional resource types by using the instructions in this guide.

Resource Group

A *resource group* is a collection of interdependent resources. A resource group is identified by a simple name; this name must be unique within a cluster. Table 1-1 shows an example of the resources for a resource group named *WebGroup.*

Table 1-1 Example Resource Group

Resource	Resource Type
vol1	*volume*
/fs1	*filesystem*
199.10.48.22	*IP_address*
web1	*Netscape_web*

If any individual resource in a resource group becomes unavailable for its intended use, then the entire resource group is considered unavailable. Therefore, a resource group is the unit of failover for IRIS FailSafe.

Resource groups cannot overlap; that is, two resource groups cannot contain the same resource.

For information about configuring resource groups, see the *IRIS FailSafe 2.0 Administrator's Guide*.

Resource Dependency List

A *resource dependency list* is a list of resources upon which a resource depends. Each resource instance must have resource dependencies that satisfy its resource type dependencies before it can be added to a resource group.

Resource Type Dependency List

A *resource type dependency list* is a list of resource types upon which a resource type depends. For example, the *filesystem* resource type depends upon the *volume* resource type, and the *Netscape_web* resource type depends upon the *filesystem* and *IP_address* resource types.

For example, suppose a file system instance */fs1* is mounted on volume */vol1*. Before */fs1* can be added to a resource group, */fs1* must be defined to depend on */vol1*. IRIS FailSafe only knows that a file system instance must have one volume instance in its dependency list. This requirement is inferred from the resource type dependency list.

Failover

A *failover* is the process of allocating a resource group (or application) to another node, according to a *failover policy.* A failover may be triggered by the failure of a resource, a change in the node membership (such as when a node fails or starts), or a manual request by the administrator.

Failover Policy

A *failover policy* is the method used by IRIS FailSafe to determine the destination node of a failover. A failover policy consists of the following:

- *Failover domain*
- *Failover attributes*
- *Failover script*

IRIS FailSafe uses the failover domain output from a failover script along with failover attributes to determine on which node a resource group should reside.

The administrator must configure a failover policy for each resource group. A failover policy name must be unique within the *pool*.

Failover Domain

A *failover domain* is the **ordered** list of nodes on which a given resource group can be allocated. The nodes listed in the failover domain must be within the same cluster; however, the failover domain does not have to include every node in the cluster.

The administrator defines the *initial failover domain* when creating a failover policy. This list is transformed into a *run-time failover domain* by the *failover script*; IRIS FailSafe uses the run-time failover domain along with failover attributes and the node membership to determine the node on which a resource group should reside. IRIS FailSafe stores the run-time failover domain and uses it as input to the next failover script invocation. Depending on the run-time conditions and contents of the failover script, the initial and run-time failover domains may be identical.

In general, IRIS FailSafe allocates a given resource group to the first node listed in the run-time failover domain that is also in the node membership; the point at which this allocation takes place is affected by the failover attributes.

Failover Attribute

A *failover attribute* is a string that affects the allocation of a resource group in a cluster. The administrator must specify system attributes (such as Auto_Failback or Controlled_Failback), and can optionally supply site-specific attributes.

Failover Script

A *failover script* is a shell script that generates a *run-time failover domain* and returns it to the IRIS FailSafe process. The IRIS FailSafe process applies the failover attributes and then selects the first node in the returned failover domain that is also in the current node membership.

The *ordered* failover script is provided with the IRIS FailSafe release. This script does not change the order of the *initial failover domain*. If this script does not meet your needs, you can create a new failover script using the information in this guide.

Action Scripts

The *action scripts* are the set of scripts that determine how a resource is started, monitored, and stopped. There must be a set of action scripts specified for each resource type.

The following is the complete set of action scripts that can be specified for each resource:

- *probe*, which verifies that the resource is configured on a server
- *exclusive*, which verifies that the resource is not already running
- *start*, which starts the resource
- *stop*, which stops the resource
- *monitor*, which monitors the resource
- *restart*, which restarts the resource on the same server after a monitoring failure occurs

The IRIS FailSafe software includes action scripts for predefined resource types. If these scripts fit the resource type that you want to make into a high-availability service, you can reuse them by copying them and modifying them as needed. If none fits, you can create additional action scripts by using the instructions in this guide.

High-Availability Services that are Available for IRIS FailSafe

The base IRIS FailSafe release includes the software required to make the following high-availability services:

- IP addresses (the *IP_address* resource type)
- XLV logical volumes (the *volume* resource type)
- XFS file systems (the *filesystem* resource type)
- MAC addresses (the *MAC_address* resource type)

Optional software packages, known as *plug-ins*, are available to make additional applications into high-availability services. For example:

- IRIS FailSafe Oracle
- IRIS FailSafe INFORMIX
- IRIS FailSafe Netscape Web
- IRIS FailSafe Mediabase

Note: IRIS FailSafe NFS is not part of the core IRIS FailSafe software, but it is documented with the base release.

If you want to create new high-availability services, or change the functionality of the provided failover scripts and action scripts by writing new scripts, you will use the instructions in this guide. However, not all resources can be made into high-availability services; see "Appropriate Applications for High Availability."

Appropriate Applications for High Availability

The characteristics of an application that can be made into high-availability service are as follows:

- The application can be easily restarted and monitored.

  It should be able to recover from failures as does most client/server software. The failure could be a hardware failure, an operating system failure, or an application failure. If a node crashed and reboots, client/server software should be able to attach again automatically.

- The application must have a start and stop procedure.

  When the resource group fails over, the resources that constitute the resource group are stopped on one node and started on another node, according to the failover script and action scripts.

- The application can be moved from one node to another after failures.

  If the resource has failed, it must still be possible to run the resource stop procedure. In addition, the resource must recover from the failed state when the resource start procedure is executed in another node.

- The application does not depend on knowing the host name; that is, those resources that can be configured to work with an IP address.

- Other resources on which the application depends can be made into high-availability services. If they are not provided by IRIS FailSafe and its optional products (see "High-Availability Services that are Available for IRIS FailSafe" on page 7), you must make these resources highly available, using the information in this guide.

Note: An application itself is not modified to make it into a high-availability service.

Overview of the Programming Steps

Note: If you do not want to write the scripts yourself, you can establish a contract with the Silicon Graphics Professional Services group to create customized scripts. See: http://www.sgi.com/services/index.html.

To turn an application into a high-availability service, follow these steps:

1. Configure and test the base IRIS FailSafe system as described in the *IRIS FailSafe Administrator's Guide*.
2. Understand the application and determine:
   - The configuration required for the application, such as user names, permissions, volumes, and so on. For more information about configuration, see the *IRIS FailSafe Administrator's Guide*.
   - The other resources on which the resource depends. All interdependent resources must be part of the same resource group. Additional resources may also be included in the resource group.
   - The number of instances of the resource type. Each instance of a given resource type is a separate *resource*.
   - The commands and arguments required to start, stop, and monitor the resources.
   - The relationships between this resource and other high-availability services; specifically, the order in which all high-availability services need to be started and stopped.
3. Determine whether existing action scripts can be reused. If they cannot, write a new set of action scripts, using existing scripts and the templates in */var/cluster/ha/resource_types/template* as a guide. See Chapter 3, "Writing the Action Scripts and Adding Monitoring Agents."
4. Determine whether the existing *ordered* failover script can be reused for the resource group. If it cannot, write a new failover script. See Chapter 4, "Creating a Failover Policy."
5. Determine whether an existing resource type can be reused. If none applies, create a new resource type or clone and modify an existing resource. See Chapter 5, "Defining a New Resource Type."

6. Configure the following in the IRIS FailSafe database (for more information, see the *IRIS FailSafe Administrator's Guid*e):
   - Resource group
   - Resource type
   - Failover policy
7. Test the action scripts and failover script. See Chapter 6, "Testing Scripts."

Note: Do not modify the scripts included with the IRIS FailSafe product. New or customized scripts must have different names from the files included with IRIS FailSafe.

IRIS FailSafe System Software

This section describes the software layers, communication paths, and database.

Layers

An IRIS FailSafe system has the following software layers:

- IRIS FailSafe plug-ins, which create high-availability services. Some plug-ins are included with the IRIS FailSafe release, others are available for separate purchase. If the application you want is not available, you can hire the Silicon Graphics Professional Services group to develop the required software, or you can use this guide to write the software yourself.
- IRIS FailSafe base, which includes the ability to define resource groups and failover policies
- High-availability infrastructure, which lets you define clusters, resources, and resource types (this consists of the *cluster_ha* installation package)
- Cluster software infrastructure, which lets you do the following:
  - Perform node logging
  - Administer the cluster
  - Define nodes

  The cluster software infrastructure consists of the *cluster_admin* and *cluster_control* subsystems).

Figure 1-1 shows a graphic representation of these layers.

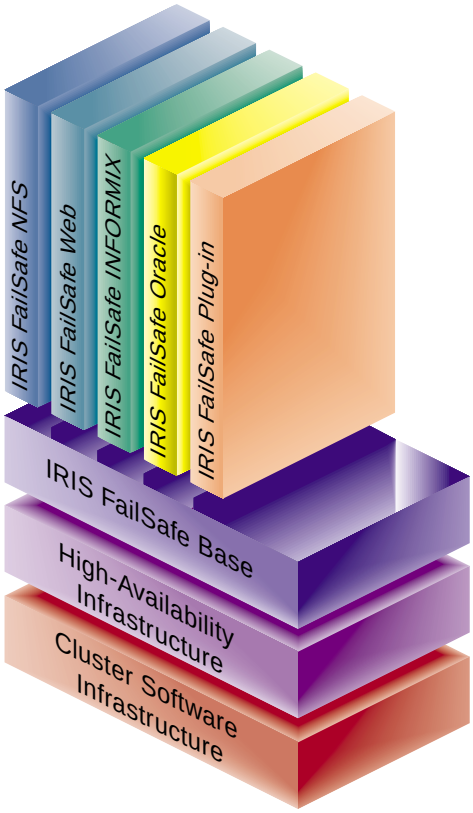

Figure 1-1 Software Layers

Table 1-2 describes the layers, which are located in the */usr/cluster/bin* directory.

Table 1-2 Contents of */usr/cluster/bin*

Layer	Subsystem	Process	Description
Plug-ins	*informix_rdbms* *oracle_rdbms*	*ha_ifmx2*	IRIS FailSafe database agents. Each database agent monitors all instances of one type of database.
IRIS FailSafe Base	*failsafe2*	*ha_fsd*	IRIS FailSafe daemon. Provides basic component of the IRIS FailSafe software.
High-availability infrastructure	*cluster_ha*	*ha_cmsd*	Cluster membership daemon. Provides the list of nodes, called *node membership*, available to the cluster.
		ha_gcd	Group membership daemon. Provides group membership and reliable communication services in the presence of failures to IRIS FailSafe processes.
		cmond	Start daemon. Starts all IRIS FailSafe daemons, and restarts them on failures.
		ha_srmd	System resource manager daemon. Manages resources, resource groups, and resource types. Executes action scripts for resources.
		ha_ifd	Interface agent daemon. Monitors the local node's network interfaces.
Cluster software infrastructure	*cluster_admin*	*cad*	Cluster administration daemon. Provides administration services and manages the configuration database.
	cluster_control	*crsd*	Node control daemon. Monitors the serial connection to other nodes. Has the ability to reset other nodes.

Communication Paths

The following figures show communication paths in IRIS FailSafe.

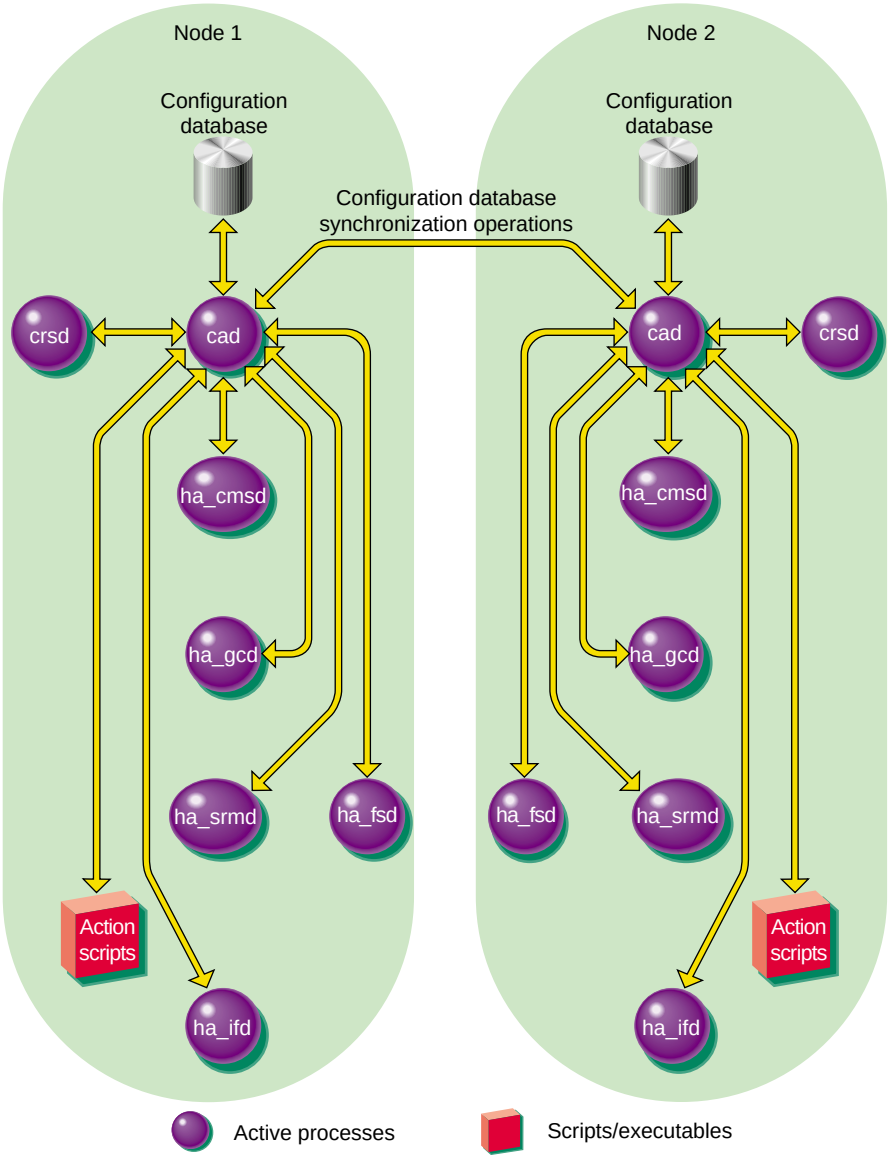

Figure 1-2 Read/Write Actions to the Configuration Database

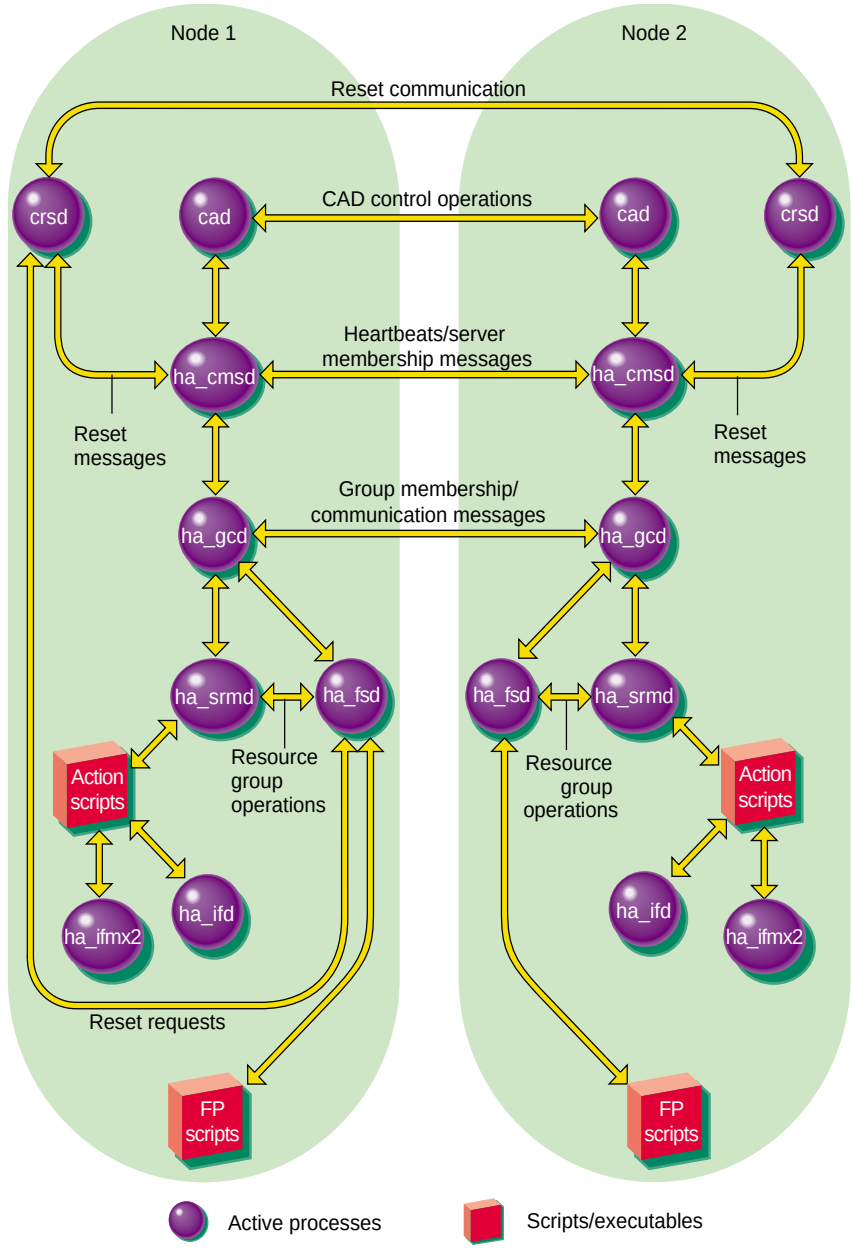

Figure 1-3 Other Message Paths

Figure 1-4 shows the communication path for a node that is in the pool but not in a cluster.

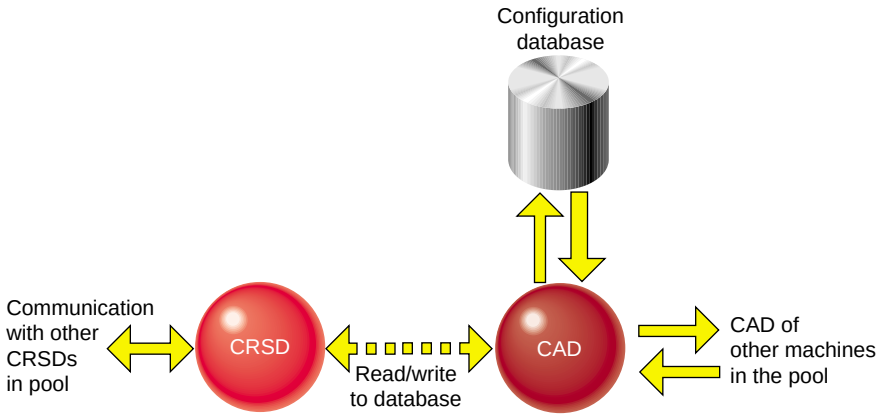

Figure 1-4 Communication Path for a Node that is Not in a Cluster

Components

The IRIS FailSafe database is a key component of IRIS FailSafe software. It contains all information about the following:

- Resources
- Resource types
- Resource groups
- Failover policies
- Nodes
- Clusters

The cluster administration daemon (*cad*) maintains identical databases on each node in the cluster.

Table 1-3 shows the contents of the */var/cluster/ha* directory.

Table 1-3 Contents of */var/cluster/ha* directory

Directory or File	Contents
comm/	Contains files that communicate between various daemons.
common_scripts/	Contains the script library (the common functions that may be used in action scripts).
log/	Contains the logs of all scripts and daemons executed by IRIS FailSafe. The outputs and errors from the commands within the scripts are logged in the *script*_**nodename** file.
policies/	Contains the failover scripts used for resource groups.
resource_types/template	Contains the template action scripts.
resource_types/rt	Contains the action scripts for the *rt* resource type.
resource_types/rt/exclusive	Verifies that the resource type is not already running.
resource_types/rt/monitor	Monitors the resource type.
resource_types/rt/probe	Verifies that the resource type is configured on the node.
resource_types/rt/restart	Restarts the resource type on the same node on a monitoring failure.
resource_types/rt/start	Starts the resource type.
resource_types/rt/stop	Stops the resource type.

Table 1-4 shows the administrative commands available with IRIS FailSafe.

Table 1-4 IRIS FailSafe Administrative Commands

Command	Purpose
ha_cilog	Logs messages to the *script*_**nodename** log files
ha_exec2	Monitors a process (similar to the IRIX *ps*(1) command)
ha_filelock	Locks a file
ha_fileunlock	Unlocks a file
ha_ifdadmin	Communicates with the *ha_ifd* network interface agent daemon
ha_http_ping2	Checks if a Netscape node is still running
ha_macconfig2	Displays or modifies MAC addresses

Chapter 2

Using the Script Library

The */var/cluster/ha/common_scripts/scriptlib* file contains the library of environment variables (beginning with uppercase *HA_*) and functions (beginning with lowercase *ha_*) available for use in your action scripts.

Note: Do not change the contents of the *scriptlib* file.

This chapter describes functions that perform the following tasks, using samples from the file:

- Set global definitions
- Check arguments
- Read an input file
- Execute a command
- Write status for a resource
- Get the value for a field
- Get resource information
- Print exclusivity check messages

Set Global Definitions

The **ha_set_global_defs()** function sets the global definitions for the environment variables shown in Table 2-1.

Table 2-1 Global Environment Variables

Variable Type	Variable Name	Default	Description
Global variable	HA_HOSTNAME	*`uname -n`*	The output of the *uname* command with the **-n** option, which is the host name or nodename. The nodename is the name by which the system is known to communications networks.
Command location	HA_CMDSPATH	*/usr/cluster/bin*	Path to user commands.
	HA_PRIVCMDSPATH	*/usr/sysadm/privbin*	Path to privileged commands (those that can only be run by root).
	HA_LOGCMD	*ha_cilog*	Command used to log into the IRIS FailSafe logs.
	HA_RESOURCEQUERYCMD	*resourceQuery*	Resource query command. This is an internal command that is not meant for direct use in scripts; use the **ha_get_info()** function of *scriptlib* instead.
	HA_SCRIPTTMPDIR	*/tmp*	Location of the script temporary directory.
Database location	HA_CDB	*/var/cluster/cdb/cdb.db*	Location of the IRIS FailSafe database.
Script log variables	HA_SCRIPTGROUP	*script*	Log for the script group.
	HA_SCRIPTSUBSYS	*script*	Log for the script subsystem.
Script log levels	HA_NORMLVL	0	Normal level of script logs.
	HA_DBGLVL	10	Debug level of script logs.

Table 2-1 (continued) Global Environment Variables

Variable Type	Variable Name	Default	Description
Script logging commands	HA_LOGQUERY_OUTPUT	*`${HA_PRIVCMDSPATH}/loggroupQuery _NUM_LOG_GROUPS=1 \ _LOG_GROUP_0=ha_script`*	Determine the current logging level for scripts.
	HA_DBGLOG	*ha_dbglog*	Command used to log debug messages from the scripts.
	HA_CURRENT_LOGLEVEL	*`echo ${HA_LOGQUERY_OUTPUT} | /usr/bin/awk '{print $2}'`*	Display the current log level. The default will be 0 (no script logging) if the *loggroupQuery* command fails or does not find configuration information.
	HA_LOG	*ha_log*	Command used to log the scripts.
Script error values	HA_SUCCESS	0	Successful execution of the script. This variable is used by the *start*, *stop*, *restart*, *monitor*, and *probe* scripts.
	HA_NOT_RUNNING	0	The script is not running. This variable is used by *exclusive* scripts.
	HA_INVAL_ARGS	1	An invalid argument was entered. This is used by all scripts.
	HA_CMD_FAILED	2	A command called by the script has failed. his variable is used by the *start*, *stop*, *restart*, *monitor*, and *probe* scripts.
	HA_RUNNING	2	The script is running. This variable is used by *exclusive* scripts.
	HA_NOTSUPPORTED	3	The specific action is not supported for this resource type. This is used by all scripts.
	HA_NOCFGINFO	4	No configuration information was found. This is used by all scripts.

Check Arguments

The **ha_check_args()** function checks the arguments specified for the script and sets the $HA_INFILE and $HA_OUTFILE variables, which specify the input and output files, respectively.

```
ha_check_args()
{

    ${HA_DBGLOG} "$HA_SCRIPTNAME called with $1 and $2"

    if [ $# -ne 2 ]; then
        ${HA_LOG} "Incorrect number of arguments"
        return 1;
    fi

    if [ ! -r $1 ]; then
        ${HA_LOG} "file $1 is not readable or does not exist"
        return 1;
    fi

    if [ ! -s $1 ]; then
        ${HA_LOG} "file $1 is empty"
        return 1;
    fi

    HA_INFILE=$1
    HA_OUTFILE=$2

    return 0;
}
```

Read an Input File

The **ha_read_infile()** function reads the $HA_INFILE input file into the $HA_RES_NAMES variable, which specifies the list of resource names.

```
ha_read_infile()
{
    HA_RES_NAMES="";

    for HA_RESOURCE in `cat ${HA_INFILE}`
    do
        HA_TMP="${HA_RES_NAMES} ${HA_RESOURCE}";
        HA_RES_NAMES=${HA_TMP};
    done
}
```

Execute a Command

The **ha_execute_cmd()** function executes the command specified by $HA_CMD. $1 is the string to be logged. The function returns 1 on error and 0 on success. On errors, the standard output and standard error of the command is redirected to the log file.

```
ha_execute_cmd()
{
    OUTFILE=${HA_SCRIPTTMPDIR}/script.$$

    ${HA_DBGLOG} $1

    eval ${HA_CMD} > ${OUTFILE} 2>&1;

    ha_exit_code=$?;

    if [ $ha_exit_code -ne 0 ]; then
        ${HA_DBGLOG} `cat ${HA_SCRIPTTMPDIR}/script.$$`
    fi

    ${HA_DBGLOG} "${HA_CMD} exited with status $ha_exit_code";

    /sbin/rm ${OUTFILE}

    return $ha_exit_code;
}
```

The **ha_execute_cmd_ret()** function is similar to **ha_execute_cmd**, except that it places the command output in the location specified by $HA_CMD_OUTPUT.

```
ha_execute_cmd_ret()
{
    ${HA_DBGLOG} $1

    # REVISIT: Is it possible to redirect the output to a log
    HA_CMD_OUTPUT=`${HA_CMD}`;

    ha_exit_code=$?;

${HA_DBGLOG} "${HA_CMD} exited with status $ha_exit_code";

    return $ha_exit_code;
}
```

Write Status for a Resource

The **ha_write_status_for_resource()** function writes the status for a resource to the $HA_OUTFILE output file. $1 is the resource name, and $2 is the resource status.

```
ha_write_status_for_resource()
{
    echo $1 $2 >> $HA_OUTFILE;
}
```

Similarly, the **ha_write_status_for_all_resources()** function writes the status for all resources. $HA_RES_NAMES is the list of resource names.

```
ha_write_status_for_all_resources()
{
    for HA_RES in $HA_RES_NAMES
    do
        echo $HA_RES $1 >> $HA_OUTFILE;
    done
}
```

Get the Value for a Field

The **ha_get_field()** function obtains the field value from a string, where $1 is the string and $2 is the field name. The string format is as follows:

name value name value . . .

```
ha_get_field()
{
    HA_STR=$1
    HA_FIELD_NAME=$2
    ha_found=0;

    for ha_i in $HA_STR
    do
        if [ $ha_i = $HA_FIELD_NAME ]; then
            ha_found=1;
            continue;
        fi
        if [ $ha_found -eq 1 ]; then
            HA_FIELD_VALUE=$ha_i
            return 0;
        fi
    done

    return 1;
}
```

Get Resource Information

The **ha_get_info()** function reads resource information. $1 is the resource type and $2 is the resource name. Resource information is stored in the HA_STRING variable. All query errors are ignored; the return value is always 0. If the *resourceQuery* command fails, the HA_STRING is set to an invalid string, and future calls to **ha_get_info()** return errors.

```
ha_get_info()
{
    HA_STRING=`${HA_PRIVCMDSPATH}/${HA_RESOURCEQUERYCMD}
_CDB_DB=$HA_CDB \
        _RESOURCE=$2 _RESOURCE_TYPE=$1`
    if [ $? -ne 0 ]; then
        ${HA_DBGLOG} "${HA_RESOURCEQUERYCMD} resource name $2 resource
type $1"
        ${HA_DBGLOG} "Failed with error ${HA_STRING}";
    fi

    return 0;
}
```

Print Exclusivity Check Messages

The **ha_print_exclusive_status()** function prints exclusivity check messages to the log file. $1 is the resource name and $2 is the exit status.

```
ha_print_exclusive_status()
{
    if [ $? -eq $HA_NOT_RUNNING ]; then
        ${HA_LOG} "resource $1 exclusive status: NOT RUNNING"
    else
        ${HA_LOG} "resource $1 exclusive status: RUNNING"
    fi
}
```

The **ha_print_exclusive_status_all_resources()** function is similar, but it prints exclusivity check messages for all resources. $HA_RES_NAMES is the list of resource names.

```
ha_print_exclusive_status_all_resources()
{
    for HA_RES in $HA_RES_NAMES
    do
        ha_print_exclusive_status ${HA_RES} $1
    done
}
```

Chapter 3

Writing the Action Scripts and Adding Monitoring Agents

This chapter provides information about writing the action scripts required to make an application a high-availability service and how to add monitoring agents. It discusses the following topics:

- Set of action scripts
- Preparation
- Script format
- Steps in writing a script
- Examples of action scripts
- Monitoring agents

Set of Action Scripts

The following set of action scripts can be provided for each resource:

- *probe*, which verifies that the resource is configured on a node
- *exclusive*, which verifies that the resource is not already running
- *start*, which starts the resource
- *stop*, which stops the resource
- *monitor*, which monitors the resource
- *restart*, which restarts the resource on the same node when a monitoring failure occurs

Note: The *start*, *stop*, and *exclusive* scripts are required for every resource type. A *monitor* script is also required, but if you wish it may contain only a return-success function. A *restart* script is required if the restart mode is set to 1; however, this script may contain only a return-success function. The *probe* script is optional.

Caution: Multiple instances of scripts may be executed at the same time.

Table 3-1 shows the state of a resource group after the successful execution of an action script for every resource within a resource group. To view the state of a resource group, use the IRIS FailSafe Cluster Manager Graphical User Interface or the *cluster_mgr* command.

Table 3-1 Successful Action Script Results

Event	Action Script Execute	Resource Group State
Resource group is made online on a node	*start*	*online*
Resource group is made offline on a node	*stop*	*offline*
Online status of the resource group	*exclusive*	No effect
Normal monitoring of online resource group	*monitor*	*online*
Resource group monitoring failure	*restart*	*online*
Configuration verification	*probe*	No effect

Table 3-2 shows the state of the resource group and the error state when an action script fails.

Table 3-2 Failure of an Action Script

Failing Action Script	Resource Group State	Error State
exclusive	No effect	No effect
failover	*online*	*monitoring failure*
monitor	*online*	*monitoring failure*
probe	No effect	No effect
start	*online*	*executable error*
stop	*online*	*executable error*

Preparation

Before you can write the action scripts, you must do the following:

- Understand the *scriptlib* functions described in Chapter 2, "Using the Script Library."
- Familiarize yourself with the script templates provided in the following directory:
  /var/cluster/ha/resource_types/template
- Familiarize yourself with the action scripts for other high-availability services in */var/cluster/ha/resource_types* that are similar to the scripts you wish to create.
- Understand how to do the following actions for your application:
  - Verify that the resource is running
  - Verify that the resource can be run
  - Start the resource
  - Stop the resource
  - Check for the process
  - Do a query and understand the expected response
  - Check for file and directory existence (as needed)

- Determine whether or not a monitoring script is required. See "Is Monitoring Necessary?" If it is not, a *monitor* script is still required, but it can contain only a return-success function.
- Determine if a resource type must be added to the IRIS FailSafe database.
- Understand the vendor-supplied startup and shutdown procedures.
- Be aware of the configuration parameters for the application.

Is Monitoring Necessary?

In the following situations, you may not need to perform monitoring:

- Heartbeat monitoring is sufficient; that is, simply verifying that the node is alive (provided automatically by IRIS FailSafe software) determines the health of the highly available service.
- There is no process or resource that can be monitored. For example, the Silicon Graphics Gauntlet Internet Firewall software performs IP filtering on firewall nodes. Because the filtering is done in the kernel, there is no process or resource to monitor.
- The resource on which the resource depends is already monitored. For example, monitoring some client-node resources might best be done by monitoring the file systems, volumes, and network interfaces they use. Because this is already done by the IRIS FailSafe base software, additional monitoring is not required.

Caution: Beware that monitoring may be so expensive that it affects system performance. In this case, monitoring should not be performed. Also, security issues may make monitoring difficult.

Types of Monitoring

There are two types of monitoring that may be accomplished in a *monitor* script:

- Is the resource present?
- Is the resource responding?

You can define multiple levels of monitoring within the monitor script, and the administrator can choose the desired level by configuring the IRIS FailSafe database.

Ensure that the monitoring level chosen does not affect system performance. For more information, see the *IRIS FailSafe 2.0 Administrator's Guide*.

What are the Symptoms of Monitoring Failure?

Possible symptoms of failure include the following:

- The resource returns an error code.
- The resource returns the wrong result.
- The resource does not return quickly enough.

How Often Should Monitoring Occur?

You must determine the probe time and time-out values for the monitor script. The time-out must be long enough to guarantee that occasional anomalies do not cause false failovers.

You must also determine if the monitor test should execute multiple times so that a node is not declared dead after a single failure. In general, testing more than once before declaring failure is a good idea.

Examples of Testing for Monitoring Failure

The test should be simple and complete quickly, whether it succeeds or fails. Some examples of tests are as follows:

- For a client-node resource that follows a protocol, the monitoring script can make a simple request and verify that the proper response is received.
- For a web node, the monitoring script can request a home page, verify that the connection was made, and ignore the resulting home page.
- For a database, a simple request such as querying a table can be made.
- For NFS, more complicated end-to-end monitoring is required. The test might consist of mounting an exported file system, checking access to the file system with a *stat()* system call to the root of the file system, and undoing the mount.

- For a resource that writes to a log file, check that the size of the log file is increasing or use the *grep* command to check for a particular message.
- The following command can be used to determine quickly whether a process exists:

  `/sbin/killall -0` *process_name*

  You can also use the *ha_exec2* command to check if a process is running. The *ha_exec2* command differs from *killall* in that it performs a more exhaustive check on the process name. (The *ha_exec2* command is used in the Web, Oracle, and INFORMIX scripts.) The command line is as follows:

  `/usr/cluster/bin/ha_exec2 -s 0 -t` *process_name*

  Note: Do not use the *ps* command to check on a particular process because its execution can be too slow.

Script Format

Templates for the action scripts are provided in the following directory:

`/var/cluster/ha/resource_types/template`

The scripts have the same general format:

- Header information
- Set local variables
- Read resource information
- Exit status
- Perform the basic action of the script, which is the customized area you must provide
- Set global variables
- Verify arguments
- Read input file

Note: Action "scripts" can be of any form -- such as Bourne shell script, perl script, or C language program.

The following sections show an example from the NFS *start* script. The contents of these examples may not match the released system.

Header Information

The header information contains comments about the resource type, script type, and resource configuration format. You must modify the code as needed.

Following is the header for the NFS *start* script:

```
#!/sbin/ksh

# ****************************************************************************
# *                                                                          *
# *                  Copyright (C) 1998 Silicon Graphics, Inc.               *
# *                                                                          *
# *  These coded instructions, statements, and computer programs  contain  *
# *  unpublished  proprietary  information of Silicon Graphics, Inc., and  *
# *  are protected by Federal copyright law.  They  may  not be disclosed  *
# *  to  third  parties  or copied or duplicated in any form, in whole or  *
# *  in part, without the prior written consent of Silicon Graphics, Inc.  *
# *                                                                          *
# ****************************************************************************

#ident "$Revision: 1.11 $"

# Resource type: NFS
# Start script NFS

#
# Test resource configuration information is present in the database in
# the following format
#
# resource-type.NFS
```

Set Local Variables

The **set_local_variables()** section of the script defines all of the variables that are local to the script, such as temporary file names or database keys. All local variables should use the *LOCAL_* prefix. You must modify the code as needed.

Following is the **set_local_variables()** section from the NFS *start* script:

```
set_local_variables()
{
    LOCAL_TEST_KEY=NFS
}

#
```

Read Resource Information

The **get**_*xxx*_**info()** function, such as **get_nfs_info()**, reads the resource information from the database. $1 is the test resource name. If the operation is successful, a value of 0 is returned; if the operation fails, 1 is returned.

The information is returned in the HA_STRING variable. For more information about HA_STRING, see Chapter 2, "Using the Script Library."

Following is the **get_nfs_info()** section from the NFS *start* script

```
get_nfs_info ()
{
    ha_get_info ${LOCAL_TEST_KEY} $1
    if [ $? -ne 0 ]; then
        return 1;
    else
        return 0;
    fi
}
```

Exit Status

In the **exit_script()** function, $1 contains the **exit_status** value. If cleanup actions are required, such as the removal of temporary files that were created as part of the process, place them before the *exit* line.

Following is the **exit_script()** section from the NFS *start* script

```
exit_script()
{
    exit $1;
}
```

Note: If you call the **exit_script** function prior to normal termination, it should be preceded by the **ha_write_status_for_resource** function and you should use the same return code that is logged to the output file.

Basic Action

This area of the script is the portion you must customize. The templates provide a minimal framework.

Following is the framework for the basic action from the *start* template:

```
start_test()
{
    # for all test resources passed as parameter
    for TEST in $HA_RES_NAMES
    do
        # HA_CMD="<command to start $TEST resource on the local machine>";
        # ha_execute_cmd "<string to describe the command being executed>";

        ha_write_status_for_resource $TEST $HA_SUCCESS;
    done
}
```

Note: When testing the script, you will add the **set -x** line to this area to obtain debugging information.

For examples of this area, see "Examples of Action Scripts" on page 40.

Set Global Variables

The following lines set all of the global and local variables and store the resource names in $HA_RES_NAMES.

Following is the **set_global_variables()** function from the NFS *start* script:

```
set_global_variables()
{
    HA_DIR=/var/cluster/ha
    COMMON_LIB=${HA_DIR}/common_scripts/scriptlib

    # Execute the common library file
    . $COMMON_LIB

    ha_set_global_defs;
}
```

Verify Arguments

The **ha_check_arg()** function verifies the arguments and stores them in the $HA_INFILE and $HA_OUTFILE variables. It returns 1 on error and 0 on success.

Following is the **ha_check_arg ()** function from the NFS *start* script:

```
ha_check_args $*;

if [ $? -ne 0 ]; then
    exit $HA_INVAL_ARGS;
fi
```

Read Input File

The **ha_read_infile()** function reads the input file and stores the resource names in the $HA_RES_NAMES variable.

Following is the **ha_read_infile()** function from the NFS *start* script:

```
ha_read_infile;
```

Complete the Action

Each action script ends with the following, which performs the action and writes the output status to the $HA_OUTFILE:

action_resourcetype;

```
exit_script $HA_SUCCESS
```

Following is the completion from the NFS *start* script:

```
start_nfs;

exit_script $HA_SUCCESS;
```

Steps in Writing a Script

Caution: Multiple copies of actions scripts can execute at the same time. Therefore, all temporary file names used by the scripts can be suffixed by PIDscript.$$ in order to make them unique, or you can use the resource name because it must be unique to the cluster.

For each script, you must do the following:

- Get the required variables
- Check the variables
- Perform the action
- Check the action

  Note: The *start* and *stop* scripts are required to be *idempotent*; that is, they have the appearance of being run once but can in fact be run multiple times. For example, if the *start* script is run for a resource that is already started, the script must not return an error.

  All action scripts must return the status to the */var/cluster/ha/log/script*_**nodename** file.

Examples of Action Scripts

The following sections use portions of the NFS scripts as examples.

Note: The examples in this guide may not exactly match the released system.

start Script

The NFS *start* script does the following:

1. Creates a resource-specific NFS status directory.
2. Exports the specified export-point with the specified export-options.

Following is a section from the NFS *start* script:

```
# Start the resource on the local machine.
# Return HA_SUCCESS if the resource has been successfully started on the local
# machine and HA_CMD_FAILED otherwise.
#
start_nfs()
{
    # for all nfs resources passed as parameter
    for resource in ${HA_RES_NAMES}
    do
        NFSFILEDIR=${HA_SCRIPTTMPDIR}$resource
        HA_CMD="/sbin/mkdir -p $NFSFILEDIR";
        ha_execute_cmd "creating nfs status file directory";

        get_nfs_info $resource
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        ha_get_field "${HA_STRING}" export-point
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        export_pt="$HA_FIELD_VALUE"
        ha_get_field "${HA_STRING}" export-info
```

```
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        export_opts="$HA_FIELD_VALUE"

        # Make the script idempotent, check to see if the filesystem
        # is already exported, if so return success. Remember that we
        # might not have any export options.
        retstat=0;
        # Check to see if the filesystem is already exported
        # (without options)
        /usr/etc/exportfs | grep "$export_pt$" >/dev/null 2>&1
        retstat=$?
        if [ $retstat -eq 1 ]; then
            # Check to see if the filesystem is already exported
            # with options.
            /usr/etc/exportfs | grep "$export_pt " | grep "$export_opts$"
>/dev/null
2>&1
            retstat=$?
        fi
        if [ $retstat -eq 1 ]; then
            # Before we try and export the file system, make sure
            # it exists.
            HA_CMD="/sbin/grep $export_pt /etc/mtab > /dev/null 2>&1";
            ha_execute_cmd "check if the export-point exists";
            if [ $? -eq 0 ]; then
                HA_CMD="/usr/etc/exportfs -i -o $export_opts $export_pt";
                ha_execute_cmd "export $export_pt directories to NFS clients";
                if [ $? -ne 0 ]; then
                    ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
                else
                    ha_write_status_for_resource ${resource} ${HA_SUCCESS};
                fi
            else
                ${HA_LOG} "Failed to find filesystem $export_pt"
                ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
            fi
        else
            ha_write_status_for_resource ${resource} ${HA_SUCCESS};
        fi
    done
}
```

stop Script

The NFS *stop* script does the following:

1. Unexports the specified export-point.
2. Removes the NFS status directory.

Following is an example from the NFS *stop* script:

```
# Stop the nfs resource on the local machine.
# Return HA_SUCCESS if the resource has been successfully stopped on the local
# machine and HA_CMD_FAILED otherwise.
#
stop_nfs()
{
    # for all nfs resources passed as parameter
    for resource in ${HA_RES_NAMES}
    do
        get_nfs_info $resource
        if [ $? -ne 0 ]; then
            # NFS resource information not available.
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        ha_get_field "${HA_STRING}" export-point
        if [ $? -ne 0 ]; then
            # NFS export-point not available.
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        export_pt="$HA_FIELD_VALUE"

        # Make the script idempotent, check to see if the filesystem
        # is already exported, if so return success. Remember that we
        # might not have any export options.
        retstat=0;
        # Check to see if the filesystem is already exported
        # (without options)
        /usr/etc/exportfs | grep "$export_pt$" >/dev/null 2>&1
        retstat=$?
        if [ $retstat -eq 1 ]; then
            # Check to see if the filesystem is already exported
            # with options.
```

```
            /usr/etc/exportfs | grep "$export_pt " | grep "$export_opts$"
>/dev/null
2>&1
            retstat=$?
        fi
        if [ $retstat -eq 0 ]; then
            # Before we unexport the filesystem, check that it exists
            HA_CMD="/sbin/grep $export_pt /etc/mtab > /dev/null 2>&1";
            ha_execute_cmd "check if the export-point exists";
            if [ $? -eq 0 ]; then
                HA_CMD="/usr/etc/exportfs -u $export_pt";
                ha_execute_cmd "unexport $export_pt directories to NFS clients";
                if [ $? -ne 0 ]; then
                    ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
                else
                    ha_write_status_for_resource ${resource} ${HA_SUCCESS};
                fi
            else
                ${HA_LOG} "filesystem $export_pt not found in export filesystem
list, unexporting anyway";
                HA_CMD="/usr/etc/exportfs -u $export_pt";
                ha_execute_cmd "unexport $export_pt directories to NFS clients";
                ha_write_status_for_resource ${resource} ${HA_SUCCESS};
            fi
        else
            ha_write_status_for_resource ${resource} ${HA_SUCCESS};
        fi
        # remove the monitor nfs status file
        NFSFILEDIR=${HA_SCRIPTTMPDIR}$resource
        HA_CMD="/sbin/rm -rf $NFSFILEDIR";
        ha_execute_cmd "removing nfs status file directory";
    done
}
```

probe Script

The NFS *probe* script does the following:

1. Verifies that the NFS daemons are running.
2. Verifies that the file system is present.

Following is an example from the NFS *probe* script:

```
# Check if the nfs resource can be online on this node. Verify if the
# database configuration is correct for the resource.
# Return HA_SUCCESS if the resource can be online on the local node
# and HA_CMD_FAILED otherwise.
#
probe_nfs()
{

    # for all nfs resources passed as parameter
    for resource in ${HA_RES_NAMES}
    do
        HA_CMD="/sbin/killall -0 nfsd"
        ha_execute_cmd "checking for nsfd processes"
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
            exit_script $HA_CMD_FAILED;
        fi
        get_nfs_info $resource
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        ha_get_field "${HA_STRING}" filesystem
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        fs="$HA_FIELD_VALUE"
        # Check if the file system is present
        if [ -b $fs ]; then
            ha_write_status_for_resource ${resource} ${HA_SUCCESS};
        else
            ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
        fi
    done
}
```

monitor Script

The NFS *monitor* script does the following:

1. Verifies that the file system is mounted at the correct mount point.
2. Requests the status of the exported file system.
3. Checks the export-point.
4. Requests NFS statistics and (based on the results) make a Remote Procedure Call (RPC) to NFS as needed.

Following is an example from the NFS *monitor* script:

```
# Check if the nfs resource is allocated in the local node
# This check must be light weight and less intrusive compared to
# exclusive check. This check is done when the resource has been
# allocated in the local node.
# Return HA_SUCCESS if the resource is running in the local node
# and HA_CMD_FAILED if the resource is not running in the local node
# The list of the resources passed as input is in variable
# $HA_RES_NAMES
#
monitor_nfs()
{
    for resource in ${HA_RES_NAMES}
    do
        get_nfs_info $resource
        if [ $? -ne 0 ]; then
            # No resource information available.
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        ha_get_field "${HA_STRING}" export-point
        if [ $? -ne 0 ]; then
            # NFS export-point not available available.
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        export_pt="$HA_FIELD_VALUE";
        ha_get_field "${HA_STRING}" filesystem
        if [ $? -ne 0 ]; then
            # filesystem not available available.
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
```

```
fi
fs="$HA_FIELD_VALUE";
# Check to see if the filesystem is mounted
HA_CMD="/sbin/mount | grep $fs | grep $export_pt >> /dev/null 2>&1"
ha_execute_cmd "check to see if $export_pt is mounted on $fs"
if [ $? -ne 0 ]; then
    ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
    exit_script $HA_CMD_FAILED;
fi
# stat the filesystem
HA_CMD="/sbin/stat $export_pt";
ha_execute_cmd "stat mount point $export_pt"
if [ $? -ne 0 ]; then
    ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
    exit_script $HA_CMD_FAILED;
fi

# check the filesystem is exported
EXPORTFS="${HA_SCRIPTTMPDIR}/exportfs.$$"
/usr/etc/exportfs > $EXPORTFS 2>&1
HA_CMD="awk '{print \$1}' $EXPORTFS | grep $export_pt"
ha_execute_cmd " check the filesystem $export_pt is exported"
if [ $? -ne 0 ]; then
    ${HA_LOG} "failed to find $export_pt in exported filesystem list:-"
    ${HA_LOG} "`/sbin/cat ${EXPORTFS}`"
    ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
    rm -f $EXPORTFS
    exit_script $HA_CMD_FAILED;
fi
rm -f $EXPORTFS
# create a file to hold the nfs stats. This will will be
# deleted in the stop script.
NFSFILE=${HA_SCRIPTTMPDIR}$resource/.nfsstat
NFS_STAT=`nfsstat -rs | tail -1 | awk '{print $1}'`
if [ ! -f $NFSFILE ]; then
   echo $NFS_STAT > $NFSFILE;
   if [ $NFS_STAT -eq 0 ];then
      # do some rpcinfo's
      exec_rpcinfo;
      if [ $? -ne 0 ]; then
         ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
         exit_script $HA_CMD_FAILED;
      fi
   fi
else
```

```
            OLD_STAT=`/sbin/cat $NFSFILE`
            if [ $NFS_STAT -gt $OLD_STAT ]; then
                 echo $NFS_STAT > $NFSFILE;
            else
                 echo $NFS_STAT > $NFSFILE;
                 exec_rpcinfo;
                 if [ $? -ne 0 ]; then
                    ha_write_status_for_resource $resource ${HA_CMD_FAILED};
                    exit_script $HA_CMD_FAILED;
                 fi
            fi
        fi
        ha_write_status_for_resource $resource $HA_SUCCESS;
    done
}
```

exclusive Script

The NFS *exclusive* script determines whether the file system is already exported. The check made by an exclusive script can be more expensive than a monitor check. IRIS FailSafe uses this script to determine if resources are running on a node in the cluster, and to thereby prevent starting resources on multiple nodes in the cluster.

Following is an example from the NFS *exclusive* script:

```
# Check if the nfs resource is running in the local node. This check can
# more intrusive than the monitor check. This check is used to determine
# if the resource has to be started on a machine in the cluster.
# Return HA_NOT_RUNNING if the resource is not running in the local node
# and HA_RUNNING if the  resource is running in the local node
# The list of nfs resources passed as input is in variable
# $HA_RES_NAMES
#
exclusive_nfs()
{
    # for all resources passed as parameter
    for resource in ${HA_RES_NAMES}
    do
        get_nfs_info $resource
        if [ $? -ne 0 ]; then
            # No resource information available
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
```

```
        fi
        ha_get_field "${HA_STRING}" export-point
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED;
        fi
        export_pt="$HA_FIELD_VALUE";
        SMFILE=${HA_SCRIPTTMPDIR}/showmount.$$
        /etc/showmount -x >> ${SMFILE};
        HA_CMD="/sbin/grep $export_pt ${SMFILE} >> /dev/null 2>&1"
        ha_execute_cmd "checking for $export_pt exported directory"
        if [ $? -eq 0 ];then
            ha_write_status_for_resource ${resource} ${HA_RUNNING};
            ha_print_exclusive_status ${resource} ${HA_RUNNING};
        else
            ha_write_status_for_resource ${resource} ${HA_NOT_RUNNING};
            ha_print_exclusive_status ${resource} ${HA_NOT_RUNNING};
        fi
        rm -f ${SMFILE}
    done
}
```

restart Script

The NFS *restart* script exports the specified export-point with the specified export-options.

Following is an example from the *restart* script for NFS:

```
# Restart nfs resource
# Return HA_SUCCESS if nfs resource failed over successfully or
# return HA_CMD_FAILED if nfs resource could not be failed over locally.
# Return HA_NOT_SUPPORTED if local restart is not supported for nfs
# resource type.
# The list of nfs resources passed as input is in variable
# $HA_RES_NAMES
#
restart_nfs()
{
    # for all nfs resources passed as parameter
    for resource in ${HA_RES_NAMES}
    do
        get_nfs_info $resource
```

```
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED
        fi
        ha_get_field "${HA_STRING}" export-point
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED
        fi
        export_pt="$HA_FIELD_VALUE"
        ha_get_field "${HA_STRING}" export-info
        if [ $? -ne 0 ]; then
            ha_write_status_for_resource ${resource} ${HA_NOCFGINFO};
            exit_script $HA_CMD_FAILED
        fi
        export_opts="$HA_FIELD_VALUE"

        HA_CMD="/usr/etc/exportfs -i -o $export_opts $export_pt";
        ha_execute_cmd "export $export_pt directories to NFS clients";
        if [ $? -ne 0 ]; then
           ha_write_status_for_resource ${resource} ${HA_CMD_FAILED};
        else
           ha_write_status_for_resource ${resource} ${HA_SUCCESS};
        fi
    done
}
```

Monitoring Agents

If any additional resource monitoring agent must be started/stopped when activating/deactivating high-availability services on a node, information about that agent should be added to the *cmond* configuration on that node. The *cmond* configuration is located in the */var/cluster/cmon/process_groups* directory. Information about every agent should go into a different file. The name of the file is not relevant to the activate/deactivate procedure.

For example, the */var/cluster/cmon/process_groups/ip_addresses* file contains information about the *ha_ifd* process that monitors network interfaces. It contains the following:

```
TYPE = cluster_agent
PROCS = ha_ifd
ACTIONS = start stop restart attach detach
AUTOACTION = attach
```

If you create a new monitoring agent, you must also a create a corresponding file in the */var/cluster/cmon/process_groups* directory that contains similar information about the new agent. To do this, you can copy the *ip_addresses* file and modify the `PROCS` line so that it lists the procss(es) that constitute your new agent. These processes must be located in the */usr/cluster/bin* directory. You should not modify the other configuration lines (`TYPE`, `ACTIONS`, and `AUTOACTION`).

Suppose you need to add a new agent called *newagent* that consists of processes *ha_x* and *ha_y*. The configuration information for this agent will be located in the */var/cluster/cmon/process_groups/newagent* file, which will contain the following:

```
TYPE = cluster_agent
PROCS = ha_x ha_y
ACTIONS = start stop restart attach detach
AUTOACTION = attach
```

In this case, the high-availability software will expect two executables (*/usr/cluster/bin/ha_x* and */usr/cluster/bin/ha_y*) to be present.

Chapter 4

Creating a Failover Policy

This chapter tells you how to create a failover policy.

Contents of a Failover Policy

A *failover policy* is the method by which a resource group is failed over from one node to another. A failover policy consists of the following:

- *Failover domain*
- *Failover attributes*
- *Failover script*

IRIS FailSafe uses the failover domain output from a failover script along with failover attributes to determine on which node a resource group should reside.

The administrator must configure a failover policy for each resource group. The name of the failover policy must be unique within the *pool*.

Failover Domain

A *failover domain* is the **ordered** list of nodes on which a given *resource group* can be allocated. The nodes listed in the failover domain **must** be within the same cluster; however, the failover domain does not have to include every node in the cluster. For example, if you have a cluster of nodes named venus, mercury, and pluto, you could configure the following initial failover domains for resource groups RG1 and RG2:

- mercury, venus, pluto for RG1
- pluto, mercury for RG2

The administrator defines the *initial failover domain* when configuring a failover policy. The initial failover domain is used when a cluster is first booted. The ordered list specified by the initial failover domain is transformed into a *run-time failover domain* by the *failover script*. With each failure, the failover script takes the current run-time failover domain and potentially modifies it; the initial failover domain is never used again. Depending on the run-time conditions and contents of the failover script, the initial and run-time failover domains may be identical.

IRIS FailSafe stores the run-time failover domain and uses it as input to the next failover script invocation.

Failover Attributes

A *failover attribute* is a value that is passed to the *failover script* and used by IRIS FailSafe for the purpose of modifying the *run-time failover domain* used for a specific resource group. There are required and optional failover attributes, and you can also specify your own strings as attributes.

Table 4-1 shows the required failover attributes.

Table 4-1 Required Failover Attributes (mutually exclusive)

Name	Description
Auto_Failback	Specifies that the resource group will be run on the first available node in the run-time failover domain. If the first node fails, the next available node will be used; when the first node reboots, the resource group will return to it. This attribute is best used when some type of load balancing is required. You must specify either this attribute or the Controlled_Failback attribute.
Controlled_Failback	Specifies that the resource group will be run on the first available node in the run-time failover domain, and will remain running on that node until it fails. If the first node fails, the next available node will be used; the resource group will remain on this new node even after the first node reboots. This attribute is best used when client/server applications have expensive recovery mechanisms, such as databases or any application that uses *tcp* to communicate. You must specify either this attribute or the Auto_Failback attribute.

When defining a failover policy, you can choose one of the recovery attributes shown in Table 4-2. The recovery attribute determines the node on which a resource group will be allocated when its state changes to online and a member of the group is already allocated (such as when volumes are present).

Table 4-2 Optional Failover Attributes (mutually exclusive)

Name	Description
Auto_Recovery	Specifies that the failover policy will be used to allocate the resource group. This attribute is optional and is mutually exclusive with the InPlace_Recovery attribute. If you specify neither of these attributes, IRIS FailSafe will use this attribute by default if you have specified the Auto_Failback attribute.
InPlace_Recovery	Specifies that the resource group will be allocated on the node that already contains part of the resource group. This attribute is optional and is mutually exclusive with the Auto_Recovery attribute. If you specify neither of these attributes, IRIS FailSafe will use this attribute by default if you have specified the Controlled_Failback attribute.

Failover Script

A *failover script* generates the run-time failover domain and returns it to the IRIS FailSafe process. The IRIS FailSafe process applies the failover attributes and then selects the first node in the returned failover domain that is also in the current node membership.

Note: The run-time of the failover script must be capped to a system-definable maximum. Hence, any external calls must be guaranteed to return quickly. If the failover script takes too long to return, IRIS FailSafe will kill the script process and use the previous run-time failover domain.

Failover scripts are stored in the */var/clusters/ha/policies* directory.

The *ordered* failover script is provided with the IRIS FailSafe release. The *ordere*d script never changes the initial domain; when using this script, the initial and run-time domains are equivalent. The script reads six lines from the input file and in case of errors logs the input parameters and/or the error to the script log.

The following example shows the contents of the *ordered* failover script.

```
#!/sbin/ksh
#
# $1 - input file
# $2 - output file
#
# line 1 input file - version
# line 2 input file - name
# line 3 input file - owner field
# line 4 input file - attributes
# line 5 input file - node membership
# line 6 input file - application failure domain


DIR=/usr/cluster/bin
LOG=${DIR}/ha_cilog
FILE=/var/cluster/ha/policies/ordered

input=$1
output=$2
cat ${input} | read version
head -2 ${input} | tail -1 | read name
head -3 ${input} | tail -1 | read attr
head -3 ${input} | tail -1 | read owner
head -4 ${input} | tail -1 | read attr
head -5 ${input} | tail -1 | read mem1 mem2 mem3 mem4 mem5 mem6 mem7 mem8
head -6 ${input} | tail -1 | read afd1 afd2 afd3 afd4 afd5 afd6 afd7 afd8


mem="("$mem1")"" ""("$mem2")"" ""("$mem3")"" ""("$mem4")"" ""("$mem5")""
""("$mem6")"" ""("$mem7")"" ""("$mem8")"

afd="("$afd1")"" ""("$afd2")"" ""("$afd3")"" ""("$afd4")"" ""("$afd5")""
""("$afd6")"" ""("$afd7")"" ""("$afd8")"


${LOG} -l 11 "${FILE}:" ‘/bin/cat ${input}‘

if [ "${version}" -ne 1 ] ; then
    echo ERROR: ${FILE}: Different version no.;
    exit 1;
elif [ -z "${name}" ]; then
    echo "ERROR: ${FILE}: Failover script not defined";
    exit 1;
```

```
elif [ -z "${attr}" ]; then
    echo "ERROR: ${FILE}: Failback attribute not defined";
    exit 1;
elif [ -z "${mem1}" ]; then
    echo "ERROR: ${FILE}: No node memberships defined";
    exit 1;
elif [ -z "${afd1}" ]; then
    echo "ERROR: ${FILE}: No failover domains defined";
    exit 1;
fi


found=0
for i in $afd1 $afd2 $afd3 $afd4 $afd5 $afd6 $afd7 $afd8; do
    for j in $mem1 $mem2 $mem3 $mem4 $mem5 $mem6 $mem7 $mem8; do
        if [ "X${j}" = "X${i}" ]; then
            found=1;
            break;
        fi
    done
done

if [ ${found} -eq 0 ]; then
    ${LOG} -l 1 "ERROR: ${FILE}: Policy script failed"
    ${LOG} -l 1 "ERROR: ${FILE}: " `/bin/cat ${input}`
    ${LOG} -l 1 "ERROR: ${FILE}: Nodes defined in membership donot match the
ones in failure domain"
    ${LOG} -l 1 "ERROR: ${FILE}: Parameters read from input file: version =
$version, name = $name, owner = $owner,  attribute = $attr, nodes = $mem, afd
=$afd"
    exit 1;
fi


if [ ${found} -eq 1 ]; then
    rm -f ${output}
    echo $afd1 $afd2 $afd3 $afd4 $afd5 $afd6 $afd7 $afd8 > ${output}
    exit 0
fi
exit 1
```

If the *ordered* script does not meet your needs, you can create a new failover script and place it in the */var/clusters/ha/policies* directory. You can then configure the IRIS FailSafe database to use your new failover script for the required resource groups.

Note: The IRIS FailSafe release also includes the *round_robin* failover script. However, this script is for demonstration purposes only and is unlikely to be useful in a production environment.

Failover Script Interface

The following is passed to the failover script:

`function(`*version*`, `*name*`, `*owner*`, `*attributes*`, `*membership*`, `*domain*`)`

version	IRIS FailSafe version. The IRIS FailSafe 2.0 release uses version number 1.
name	Name of the failover script (used for error validations and logging purposes).
owner	Logical name of the node that has the resource group allocated.
attributes	Failover attributes (Auto_Failback or Controlled_Failback must be included)
membership	List of the nodes that are currently available.
domain	Ordered list of nodes used at the last failover. (At the first failover, the initial failover domain is used.)

The failover script returns the newly generated run-time failover domain to IRIS FailSafe, which then chooses the node on which the resource group should be allocated by applying the failover attributes and node membership to the run-time failover domain.

Example Failover Policies

There are two general types of configuration:

- *N* nodes that can potentially failover their applications to any of the other nodes in the cluster. *N* is an integer from 1 through 8.
- *N* primary nodes that can failover to *M* backup nodes. The sum of *N* and *M* is an integer from 2 through 8. For example, you could have 3 primary nodes and 1 backup node.

This section shows examples of failover policies for the following types of configuration, where *N* can be an integer from 2 through 8:

- *N* primary nodes and one backup node (*N*+1)
- *N* primary nodes and two backup nodes (*N*+2)
- *N* primary nodes and *M* backup nodes (*N*+*M*)

Note: The diagrams in the following sections illustrate the configuration concepts discussed here, but they do not address all required or supported elements, such as reset hubs. For configuration details, see the *IRIS FailSafe 2.0 Installation and Maintenance Instructions*.

N+1 Configuration

Figure 4-1 shows a specific instance of an *N*+1 configuration in which there are three primary nodes and one backup node. (This is also known as a *star configuration*.) The disks shown could each be disk farms.

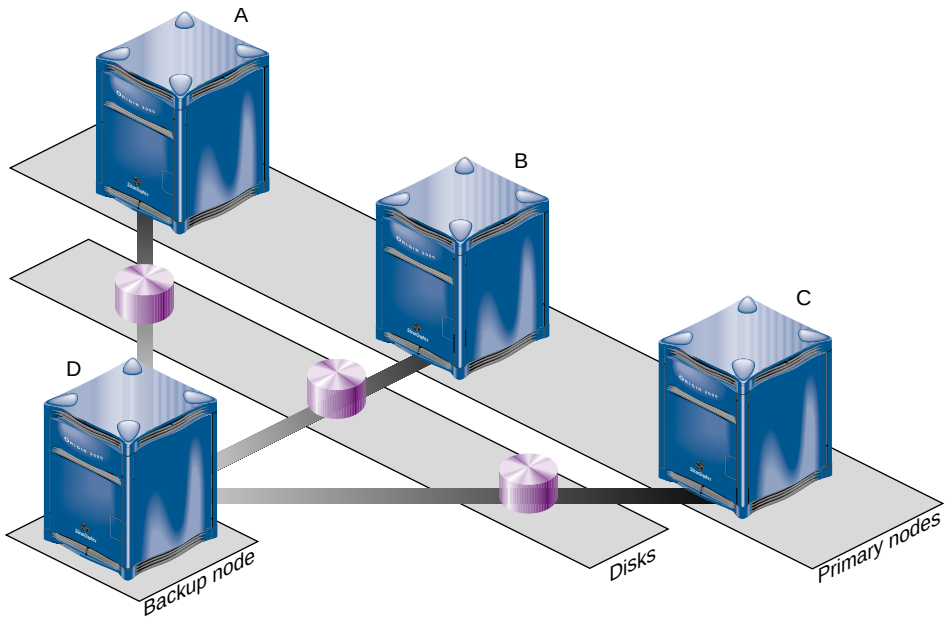

Figure 4-1 *N*+1 Configuration Concept

You could configure the following failover policies for load balancing:

- Failover policy for RG1:
  - Initial failover domain = A, D
  - Failover attribute = AutoFailback
  - Failover script = ordered
- Failover policy for RG2:
  - Initial failover domain = B, D
  - Failover attribute = AutoFailback
  - Failover script = ordered
- Failover policy for RG3:
  - Initial failover domain = C, D
  - Failover attribute = AutoFailback
  - Failover script = ordered

If node A fails, RG1 will fail over to node D. As soon as node A reboots, RG1 will be moved back to node A.

If you change the failover attribute to ControlledFailback for RG1 and node A fails, RG1 will fail over to node D and will remain running on node D even if node A reboots.

N+2 Configuration

Figure 4-2 shows a specific instance of an *N*+2 configuration in which there are four primary nodes and two backup nodes. The disks shown could each be disk farms.

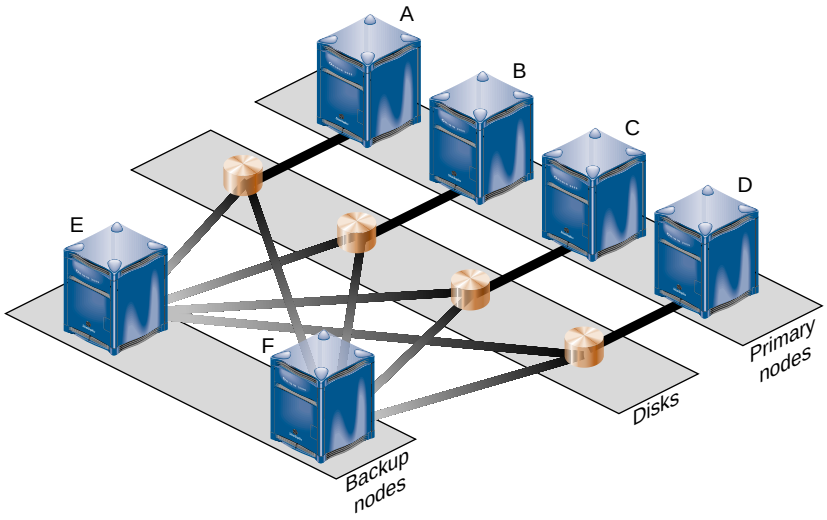

Figure 4-2 *N*+2 Configuration Concept

You could configure the following failover policy for a database resource groups RG7 and RG8:

- Failover policy for RG7:
  - Initial failover domain = A, E, F
  - Failover attribute = ControlledFailback
  - Failover script = ordered

- Failover policy for RG8:
  - Initial failover domain = B, F, E
  - Failover attribute = AutoFailback
  - Failover script = ordered

If node A fails, RG7 will fail over to node E. If node E also fails, RG7 will fail over to node F. If A is rebooted, RG7 will remain on node F.

If node B fails, RG8 will fail over to node F. If B is rebooted, RG8 will return to node B.

N+M Configuration

Figure 4-3 shows a specific instance of an *N+M* configuration in which there are four primary nodes and each can serve as a backup node. The disk shown could be a disk farm.

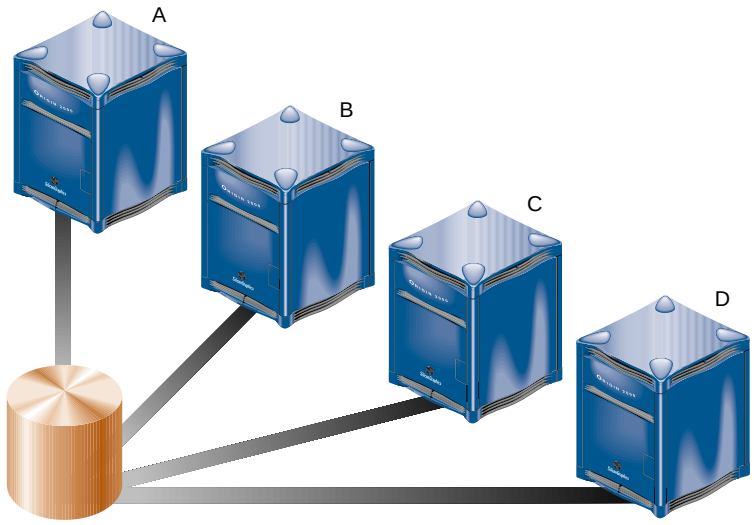

Figure 4-3 *N+M* Configuration Concept

You could configure the following failover policy for a database resource groups RG5 and RG6:

- Failover policy for RG5:
  - Initial failover domain = A, B, C, D
  - Failover attribute = ControlledFailback
  - Failover script = ordered
- Failover policy for RG6:
  - Initial failover domain = C, A, D
  - Failover attribute = ControlledFailback
  - Failover script = ordered

If node C fails, RG6 will fail over to node A. When node C reboots, RG6 will remain running on node A. If node A then fails, RG6 will return to node C and RG5 will move to node B. If node B then fails, RG5 moves to node C.

Chapter 5

Defining a New Resource Type

This section tells you how to define a new resource type.

Information You Must Gather

To define a new resource type, you must have the following information:

- Name of the resource type.
- Name of the cluster to which the resource type will apply.
- If the resource type is to be restricted to a specific node, you must know the node name.
- Order of performing the action scripts for resources of this type in relation to resources of other types:
  - Resources are started in the increasing order of this value
  - Resources are stopped in the decreasing order of this value

  Ensure that the number you choose for a new resource type permits the resource types on which it depends to be started before it is started, or stopped before it is stopped, as appropriate.

Table 5-1 shows the conventions used for order ranges. The values available for customer use are 201-400 and 701-999.

Table 5-1 Order Ranges

Range	Reservation
1-100	SGI-provided basic system resource types, such as *MAC_address*
101-200	SGI-provided system plug-ins that can be started before *IP_address*
201-400	User-defined resource types that can be started before *IP_address*
401-500	SGI-provided basic system resource types, such as *IP_address*
501-700	SGI-provided system plug-ins that must be started after *IP_address*
701-999	User-defined resource types that must be started after *IP_address*

Table 5-2 shows the order numbers of the resource types provided with the release.

Table 5-2 Resource Type Order Numbers

Order Number	Resource Type
10	*MAC_address*
20	*volume*
30	*filesystem*
201	*NFS*
401	*IP_address*
411	*statd*
501	*Netscape_web*
511	*Oracle_DB*
521	*INFORMIX_DB*

- Restart mode, which can be one of the following values:
  - 0 = Do not restart on monitoring failures
  - 1 = Restart a fixed number of times
- Number of local restarts (when restart mode is 1).
- Location of the executable script. This is always */var/cluster/ha/resource_types/rtname*, where *rtname* is the resource type name.
- Monitoring interval, which is the time period (in milliseconds) between successive executions of the *monitor* action script; this is only valid for the *monitor* action script.
- Starting time for monitoring. When the resource group is made online in a cluster node, IRIS FailSafe will start monitoring the resources after the specified time period (in milliseconds).
- Action scripts to be defined for this resource type. You must specify scripts for *start*, *stop*, *exclusive*, and *monitor*, although the *monitor* script may contain only a return-success function if you wish. If you specify 1 for the restart mode, you must specify a *restart* script. The *probe* script is optional.
- Type-specific attributes to be defined for this resource type. The action scripts use this information to start, stop, and monitor a resource of this resource type. For example, NFS requires the following resource keys:
  - *export-point*, which takes a value that defines the export disk name. This name is used as input to the *exportfs* (1M) command. For example:

    ```
    export-point = /this_disk
    ```

  - *export-info*, which takes a value that defines the export options for the file system. These options are used in the *exportfs*(1M) command. For example:

    ```
    export-info = rw,wsync,anon=root
    ```

  - *filesystem*, which takes a value that defines the raw file system. This name is used as input to the *mount*(1M) command. For example:

    ```
    filesystem = /dev/xlv/xlv_object
    ```

Using the GUI

To define a new resource type using the FailSafe Manager graphical user interface (GUI), use the following task:

```
Resources & Resource Types => Define a Resource Type
```

The GUI will prompt you for required and optional information. Online help is provided for each item. You will need to know the following information:

- Name of the new resource type
- Start/stop order for resources of this type
- Maximum duration for the execution of the *start*, *stop*, *monitor*, *exclusive*, *restart* (optional), and *probe* (optional) scripts.

  You must also provide the time interval between the start of a resource and the start of monitor checking, and the interval between subsequent monitor checks.

  If you choose to use the *restart* script, you must also decide how many restart attempts will be allowed.
- Type-specific attributes.

To define the dependencies for a given type use the following task:

```
Add/Remove Dependencies for a Resource Type
```

Using cluster_mgr Interactively

The following steps show the use of *cluster_mgr* interactively to define a resource type called **`test_rt`**.

1. Log in as **root**.
2. Execute the *cluster_mgr* command using the **-p** option to prompt you for information (the command name can be abbreviated to *cmgr*):

   ```
   # /usr/cluster/bin/cluster_mgr -p
   Welcome to IRIS FailSafe Cluster Manager Command-Line Interface

   cmgr>
   ```

3. Use the *set* subcommand to specify the default cluster used for *cluster_mgr* operations. In this example, we use a cluster named **test**:

   ```
   cmgr> set cluster test
   ```

   Note: If you prefer, you can specify the cluster name as needed with each subcommand.

4. Use the *define resource_type* subcommand. By default, the resource type will apply across the cluster; if you wish to limit the resource type to a specific node, enter the node name when prompted. If you wish to enable restart mode, enter 1 when prompted.

   Note: The following example only shows the prompts and answers for two action scripts (*start* and *stop*) for a new resource type named *test_rt*.

   ```
   cmgr> define resource_type test_rt

   (Enter "cancel" at any time to abort)

   Node[optional]?
   Order ? 300
   Restart Mode ? (0)

   DEFINE RESOURCE TYPE OPTIONS

           1) Add Action Script.
           2) Remove Action Script.
           3) Add Type Specific Attribute.
           4) Remove Type Specific Attribute.
           5) Add Dependency.
           6) Remove Dependency.
           7) Show Current Information.
           8) Cancel. (Aborts command)
           9) Done. (Exits and runs command)

   Enter option:1

   No current resource type actions

   Action name ? start
   Executable Time? 40000
   Monitoring Interval? 0
   ```

```
Start Monitoring Time? 0

        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)

Enter option:1

Current resource type actions:
        Action - 1: start

Action name stop
Executable Time? 40000
Monitoring Interval? 0
Start Monitoring Time? 0

        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)

Enter option:3

No current type specific attributes

Type Specific Attribute ? integer-att
Datatype ? integer
Default value[optional] ? 33
```

```
        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)

Enter option:3

Current type specific attributes:
        Type Specific Attribute - 1: export-point

Type Specific Attribute ? string-att
Datatype ? string
Default value[optional] ? rw

        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)

Enter option:5

No current resource type dependencies

Dependency name ? filesystem

        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)
```

```
Enter option:7

Current resource type actions:
        Action - 1: start
        Action - 2: stop

Current type specific attributes:
        Type Specific Attribute - 1: integer-att
        Type Specific Attribute - 2: string-att

No current resource type dependencies

Resource dependencies to be added:
        Resource dependency - 1: filesystem

        1) Add Action Script.
        2) Remove Action Script.
        3) Add Type Specific Attribute.
        4) Remove Type Specific Attribute.
        5) Add Dependency.
        6) Remove Dependency.
        7) Show Current Information.
        8) Cancel. (Aborts command)
        9) Done. (Exits and runs command)

Enter option:9
Successfully defined resource_type test_rt

cmgr> show resource_types

NFS
template
Netscape_web
test_rt
statd
Oracle_DB
MAC_address
IP_address
INFORMIX_DB
filesystem
volume

cmgr> exit
#
```

Using cluster_mgr With a Script

You can write a script that contains all of the information required to define a resource type and supply it to *cluster_mgr* by using the **-f** option:

`cluster_mgr -f` *scriptname*

Or, you could include the following as the first line of the script and then execute the script itself:

```
#!/usr/cluster/bin/cluster_mgr -f
```

A template script for creating a new resource type is located in */var/cluster/cmgr-templates/cmgr-create-resource_type*.

Each line of the script must be a valid *cluster_mgr* line, similar to a *here* document. Because *cluster_mgr* will run through commands as if entered interactively, you must include *done* and *quit* lines to finish a multi-level command and exit out of *cluster_mgr*.

Note: If you concatenate information from multiple template scripts to prepare your cluster configuration, you must remove the *quit* at the end of each template script, except for the final *quit*. A *cluster_mgr* script must have only one *quit* line.

For example, you could use the following script to define the same *test_rt* resource type defined interactively in the previous section:

```
# test_rt.script: Script to define the "test_rt" resource type

set cluster test
define resource_type test_rt
set order to 300
set restart_mode to 0
add action start
set exec_time to 40000
set monitor_interval to 0
set monitor_time to 0
done
add action stop
set exec_time to 40000
set monitor_interval to 0
set monitor_time to 0
done
add type_attribute integer-att
set data_type to integer
set default_value to 33
```

```
done
add type_attribute string-att
set data_type to string
set default_value to rw
done
add dependency filesystem
done
quit
```

When you execute the *cluster_mgr -f* command line with this script, you will see the following output:

```
# /usr/cluster/bin/cluster_mgr -f test_rt.script
Successfully defined resource_type test_rt

#
```

To verify that the resource type was defined, enter the following:

```
# /usr/cluster/bin/cluster_mgr -c "show resource_types in cluster test"

NFS
template
Netscape_web
test_rt
statd
Oracle_DB
MAC_address
IP_address
INFORMIX_DB
filesystem
volume
```

Chapter 6

Testing Scripts

This chapter describes how to test action scripts without running IRIS FailSafe. It also provides tips on how to debug problems that you may encounter.

Note: Parameters are passed to the action scripts as both input files and output files. Each line of the input file contains the resource name; the output file contains the resource name and the script exit status.

General Testing and Debugging Techniques

Some general testing and debugging techniques you can use during testing are as follows:

- To get debugging information, adding the following line to each of your scripts in the main function of the script:

  ```
  set -x
  ```

- To check that an application is running on a node, you may be able to use a command provided by the application. For example, the IRIS FailSafe INFORMIX option uses the INFORMIX command *onstat*.
- Another way to check that an application is running on a node, is to enter this command on that node:

  # **ps -ef | grep** *application*

  application is the name (or a portion of the name) of the executable for the application.

- To show the status of a resource, use the following *cluster_mgr* command:

  ```
  cmgr> set cluster cname
  cmgr> show status of resource rname of resource_type rtype
  ```

- To show the status of a node, use the following *cluster_mgr* command:

  `cmgr> show status of node` *nodename*

- To show the status of a resource group, use the following *cluster_mgr* command:

  `cmgr> show status of resource_group` *rgname* `in cluster` *cname*

Testing an Action Script

To test an action script, do the following:

1. Create an input file, such as */tmp/input*, that contains expected resource names. For example, to create a file that contains the resource named *disk1*, do the following:

   ```
   echo "/disk1" > /tmp/input
   ```

2. Execute the action script using the input file, as follows:

   ```
   start /tmp/input /tmp/output
   ```

The output file will contain one of the following return values for the *start*, *stop*, *probe*, *monitor*, and *restart* scripts:

```
HA_SUCCESS=0
HA_INVAL_ARGS=1
HA_CMD_FAILED=2
HA_NOTSUPPORTED=3
HA_NOCFGINFO=4
```

The output file will contain one of the following return values for the *exclusive* script:

```
HA_NOT_RUNNING=0
HA_RUNNING=2
```

Note: If you call the **exit_script** function prior to normal termination, it should be preceded by the **ha_write_status_for_resource** function and you should use the same return code that is logged to the output file.

Suppose you have a resource named /disk1 and the following files:

- The syntax for the input file is: <*resourcename*>
- The syntax for the output file is: <r*esourcename*> <*status*>

The following example shows:

- The exit status of the action script is 2
- The exit status of the resource is 2

Note: The use of *anonymous* indicates that the script was run manually. When the script is run by IRIS FailSafe, the full path to the script name is displayed.

```
# echo "/disk1" > /tmp/ipfile
# ./monitor  /tmp/ipfile /tmp/opfile
# echo $?
2
# cat /tmp/opfile
/disk1 2
# tail /var/cluster/ha/log/script_heb1
Tue Aug 25 11:32:57.437 <anonymous script 23787:0 Unknown:0> ./monitor:
./monitor called with /tmp/ipfile and /tmp/opfile
Tue Aug 25 11:32:58.118 <anonymous script 24556:0 Unknown:0> ./monitor:
check to see if /disk1 is mounted on /disk1
Tue Aug 25 11:32:58.433 <anonymous script 23811:0 Unknown:0> ./monitor:
/sbin/mount | grep /disk1 | grep /disk1 >> /dev/null 2>&1 exited with
status 0
Tue Aug 25 11:32:58.665 <anonymous script 24124:0 Unknown:0> ./monitor:
stat mount point /disk1
Tue Aug 25 11:32:58.969 <anonymous script 23525:0 Unknown:0> ./monitor:
/sbin/stat /disk1 exited with status 0
Tue Aug 25 11:32:59.258 <anonymous script 24431:0 Unknown:0> ./monitor:
check the filesystem /disk1 is exported
Tue Aug 25 11:32:59.610 <anonymous script 6982:0 Unknown:0> ./monitor:
Tue Aug 25 11:32:59.917 <anonymous script 24040:0 Unknown:0> ./monitor:
awk '{print \$1}' /var/cluster/ha/tmp/exportfs.23762 | grep /disk1 exited
with status 1
Tue Aug 25 11:33:00.131 <anonymous script 24418:0 Unknown:0> ./monitor:
echo failed to find /disk1 in exported filesystem list:-
Tue Aug 25 11:33:00.340 <anonymous script 24236:0 Unknown:0> ./monitor:
echo /disk2
```

For additional information about a script's processing, see the */var/cluster/ha/log/script_***nodename**, where **nodename** is the name of the node.

Special Testing Considerations for the monitoring Script

The *monitor* script tests the liveliness of applications and resources. The best way to test it is to induce a failure, run the script, and check if this failure is detected by the script; then repeat the process for another failure.

Use this checklist for testing a *monitor* script:

- Verify that the script detects failure of the application successfully.
- Verify that the script always exits with a return value.
- Verify that the script does not contain commands that can hang (such as using DNS for name resolution) or those that continue forever, such as *ping*.
- Verify that the script completes before the time-out value specified in the configuration file.
- Verify that the script's return codes are correct.

During testing, measure the time it takes for a script to complete and adjust the monitoring times in your script accordingly. To get a good estimate of the time required for the script to execute, run it under different system load conditions.

Appendix A

Migrating to IRIS FailSafe 2.0

This chapter provides guidelines for migrating your IRIS FailSafe 1.2 resources and monitor script information to IRIS FailSafe 2.0 action scripts. It assumes you are already familiar with the migration information provided in the *IRIS FailSafe 2.0 Administrator's Guide*.

Cautions

Multiple instances of action scripts may be executed at the same time.

The software for IRIS FailSafe 2.0 and IRIS FailSafe 1.2 can coexist in the same node. However, IRIS FailSafe 2.0 and IRIS FailSafe 1.2 cannot run at the same time.

There are no configuration checksum verification in scripts.

Resource Types

In 2.0, the *ha.conf* configuration file has been replaced by the configuration database. The configuration database is automatically copied to all nodes in the pool. See the *IRIS FailSafe 2.0 Administrator's Guide* for information about configuring a 2.0 system.

If you require new resource types, you will create them using either the IRIS FailSafe Cluster Manager GUI (graphical user interface) or the *cluster_mgr* command. See Chapter 5, "Defining a New Resource Type."

You may be able to reuse the following monitoring information from the 1.2 *ha.conf* file with regard to 2.0 resource types:

- `start-monitor-time`
- `lmon-probe-time`
- `lmon-timeout`

Note: All IRIS FailSafe 2.0 time-outs are in milliseconds.

The following examples show information (in bold) that is used in the 1.2 *ha.conf* file and reused when creating a new resource type in 2.0.

Suppose a portion of the 1.2 *ha.conf* file had this:

```
action apache
{
        local-monitor = /var/ha/actions/ha_apache_lmon
}

action-timer apache
{
        start-monitor-time = 120
        lmon-probe-time = 120
        lmon-timeout = 60
}
```

You would reuse the information when creating a resource type in 2.0, as follows:

```
cmgr> create resource_type apache in cluster apache-cluster
Enter commands, when finished enter either "done", "cancel", "check"
Resource Type Name [apache]? apache
Cluster? apache-cluster
Node? node1
Order [0]? 500
Restart Mode [0]?0
Restart Count [0]?0
Number of Actions [0]? 4
Action? start
Executable? /var/cluster/ha/resource_types/apache/start
Executable Time? 20000
Monitoring Interval? 0
Start Monitoring Time? 0
Action? stop
Executable? /var/cluster/ha/resource_types/apache/stop
Executable Time? 20000
Monitoring Interval? 0
Start Monitoring Time? 0
Action? monitor
Executable?  /var/cluster/ha/resource_types/apache/monitor
Executable Time? 60000
Monitoring Interval? 120000
Start Monitoring Time? 120000
```

```
Action? exclusive
Executable? /var/cluster/ha/resource_types/apache/exclusive
Executable Time? 60000
Monitoring Interval? 0
Start Monitoring Time? 0?0
Number of Resource Keys [0]? 1
Name of resource key? search-string
Datatype? string
Default Key? httpd
Enter dependency commands,when finished enter either "done" or "cancel"

resource_type apache? add type IP_address
resource_type apache? done
```

Reading Information

In 2.0, configuration information is read using the **ha_get_info()** and **ha_get_field()** shell functions. These functions are equivalent to the 1.2 *ha_cfginfo* command.

In 2.0, all common functions and variables are kept in */var/cluster/ha/common_scripts/scriptlib* file. This file is equivalent to the 1.2 */var/ha/actions/common.vars* file.

For more information, see Chapter 2, "Using the Script Library."

Parameter Parsing

In 2.0, action script parameters are passed in a file and information is also returned in a file. The script takes a list of resource names as parameters.

Action Scripts

Table A-1 summarizes the differences in scripts between the releases.

Table A-1 Differences between IRIS FailSafe 1.2 and 2.0 Scripts

IRIS FailSafe 1.2	IRIS FailSafe 2.0
giveaway, giveback	*stop*
takeover, takeback	*start*
check	*monitor*
(no equivalent)	*exclusive*, *probe*, *restart*

In 2.0, the action scripts are installed as */var/cluster/ha/Resource_Type_Name/Action_Name* directory, where *Resource_Type_Name* is the name of the resource type (such as *NFS*) and *Action_Name* is the name of the action script (such as *start*).

Templates of the action scripts (*start*, *stop*, *probe*, *monitor*, *exclusive*, *restart*) are provided in the */var/cluster/ha/resource_types/template* directory. For more information about action scripts, see Chapter 3, "Writing the Action Scripts and Adding Monitoring Agents."

The following sections provide example portions of 1.2 scripts and their 2.0 equivalents:

- *giveback* and *stop*
- *takeover* and *start*
- *monitor* and *monitor*

Note: There are no 1.2 equivalents for the 2.0 *probe*, *exclusive*, and *restart* scripts.

In the following examples, only the relevant portions of the scripts are shown. Areas in common between 1.2 and 2.0 are in bold.

1.2 giveback / 2.0 stop

For example, suppose you had the following in the *giveback* script in 1.2:

```
giveback()
{
    for i in `$CFG_INFO ${T_APACHE}`
    do
        SEARCH="$CFG_INFO ${T_APACHE}${CFG_SEP}${i}${CFG_SEP}${T_BACKUP}"
        BACKUP=`$SEARCH`
        if [ $? -eq 1 ]; then
            ${LOGGER} "$0: Trouble finding backup-node for apache ($SEARCH)"
            exit $INCORRECT_CONF_FILE;
        fi
        # If I am the backup
        if [ ${BACKUP} = ${HOST} ]; then
            ${LOGGER} "$0: Stopping apache for backup server."
            killall -9 /apache-fs/usr/local/apache_1.2.0/src/httpd
            if [ $? -ne "0" ]; then
                ${LOGGER} "$0: halt of apache on backup server failed."
            fi
        fi

        exit $SUCCESS
    done
}
```

In 2.0, you would have the following in the *stop* script:

```
stop_apache()
{
    for server in $HA_RES_NAMES
    do
        ${HA_DBGLOG} "Stopping apache server $server"
        killall -9 /apache-fs/usr/local/apache_1.2.0/src/httpd
        if [ $? -ne "0" ]; then
            ${HA_LOG} "halt of apache server $server failed."
            ha_write_status_for_resource $server $HA_CMD_FAILED;
        else
            ${HA_DBGLOG} "halt of apache server $server successful"
            ha_write_status_for_resource $server $HA_SUCCESS;
        fi
    done
}
```

1.2 takeover / 2.0 start

For example, suppose you had the following in the *takeover* script in 1.2:

```
takeover()
{
    for i in `$CFG_INFO ${T_APACHE}`
    do
        SEARCH="$CFG_INFO ${T_APACHE}${CFG_SEP}${i}${CFG_SEP}${T_BACKUP}"
        BACKUP=`$SEARCH`
        if [ $? -eq 1 ]; then
            ${LOGGER} "$0: Trouble finding backup-node for apache ($SEARCH)"
            exit $INCORRECT_CONF_FILE;
        fi
        # If I am the backup
        if [ ${BACKUP} = ${HOST} ]; then
            ${LOGGER} "$0: Starting apache for backup server."
            /apache-fs/usr/local/apache_1.2.0/src/httpd -d \
/apache-fs/usr/local/apache_1.2.0
          if [ $? -ne "0" ]; then
               ${LOGGER} "$0: start of apache on backup server failed."
               exit $FAILED
            fi
        fi
        exit $SUCCESS
    done

}
```

In 2.0, you would have the following in the *start* script:

```
start_apache()
{
    for server in $HA_RES_NAMES
    do
        ${HA_DBGLOG} "Starting apache server $server"
        /apache-fs/usr/local/apache_1.2.0/src/httpd -d \
/apache-fs/usr/local/apache_1.2.0
        if [ $? -ne "0" ]; then
            ${HA_LOG} "start of apache server $server failed."
            ha_write_status_for_resource $server $HA_CMD_FAILED;
        else
            ${HA_DBGLOG} "start of apache server $server successful"
            ha_write_status_for_resource $server $HA_SUCCESS;
        fi
    done
}
```

1.2 monitor/ 2.0 monitor

For example, suppose you had the following in the *monitor* script in 1.2:

```
monitor()
{

    # Read the search string entry
    for i in `$CFG_INFO ${T_APACHE}`
    do
        SEARCH="$CFG_INFO ${T_APACHE}${CFG_SEP}${i}${CFG_SEP}${T_SEARCH_STR}"
        SEARCH_STR=`$SEARCH`
        ${SEARCH_STR:=httpd};
    done

    EXEC="${KILLALL} -0 ${SEARCH_STR}";
     execute_cmd "check if apache server processes are running"

}
```

In 2.0, you would have the following in the *monitor* script:

```
monitor_apache()
{
    for server in $HA_RES_NAMES
    do
        get_apache_info $server
        if [ $? -eq 0 ]; then
            APACHE_FIELDS=${HA_STRING
            ha_get_field "${APACHE_FIELDS}" search-string;
            if [ $? -eq 0 ]; then
                SEARCH_STR=${HA_FIELD_VALUE};
            fi
        fi
        ${SEARCH_STR:=httpd};
        HA_CMD=${KILLALL} -0 ${SEARCH_STR}";
         ha_execute_cmd "check if server $server processes are running"
        if [ $? -ne 0 ]; then
            ${HA_LOG} "monitor of apache server $server failed."
            ha_write_status_for_resource $server $HA_CMD_FAILED;
        else
            ${HA_DBGLOG} "monitor of apache server $server successful"
            ha_write_status_for_resource $server $HA_SUCCESS;
        fi
    done
}
```

Ordering Script Actions

In 2.0, each resource type has a start/stop order, which is a nonnegative integer. In a resource group, the start/stop orders of the component resource types determine the order in which the resources will be started when IRIS FailSafe brings the group online and will be stopped when IRIS FailSafe takes the group offline. The group's resources are started in increasing order, and stopped in decreasing order.

Note: Resources of the same type are started and stopped in indeterminate order.

For example, if resource type *volume* has order 10 and resource type *filesystem* has order 20, then when IRIS FailSafe brings a resource group online, all volume resources in the group will be started before all file system resources in the group.

There is no need to create software links similar to those used in 1.2.

Glossary

action scripts

The set of scripts that determine how a resource is started, monitored, and stopped. There must be a set of action scripts specified for each resource type. The possible set of action scripts is: *probe*, *exclusive*, *start*, *stop*, *monitor*, and *restart*.

cluster

A collection of one or more *cluster nodes* coupled to each other by networks or other similar interconnections. A cluster is identified by a simple name; this name must be unique within the *pool*. A particular node may be a member of only one cluster.

cluster administrator

The person responsible for managing and maintaining an IRIS FailSafe cluster.

cluster configuration database

Contains configuration information about all resources, resource types, resource groups, failover policies, nodes, and clusters.

cluster node

A single IRIX image. Usually, a cluster node is an individual computer. The term *node* is also used in this guide for brevity; this use of node does not have the same meaning as a node in an Origin system.

control messages

Messages that cluster software sends between the cluster nodes to request operations on or distribute information about cluster nodes and resource groups. IRIS FailSafe sends control messages for the purpose of ensuring nodes and groups remain highly available. Control messages and heartbeat messages are sent through a node's network interfaces that have been attached to a control network. A node can be attached to multiple control networks.

A node's control networks should not be set to accept control messages if the node is not a dedicated IRIS FailSafe node. Otherwise, end users who run non-IRIS FailSafe jobs on the machine can have their jobs killed unexpectedly when IRIS FailSafe resets the node.

control network

The network that connects nodes through their network interfaces (typically Ethernet) such that IRIS FailSafe can maintain a cluster's high availability by sending heartbeat messages and control messages through the network to the attached nodes. IRIS FailSafe uses the highest priority network interface on the control network; it uses a network interface with lower priority when all higher-priority network interfaces on the control network fail.

A node must have at least one control network interface for heartbeat messages and one for control messages (both heartbeat and control messages can be configured to use the same interface). A node can have no more than eight control network interfaces.

dependency list

See *resource dependency* or *resource type dependency.*

failover

The process of allocating a *resource group* to another *node*, according to a *failover policy.* A failover may be triggered by the failure of a resource, a change in the node membership (such as when a node fails or starts), or a manual request by the administrator.

failover attribute

A string that affects the allocation of a resource group in a cluster. The administrator must specify system-defined attributes (such as Auto_Failback or Controlled_Failback), and can optionally supply site-specific attributes.

failover domain

The ordered list of nodes on which a particular *resource group* can be allocated. The nodes listed in the failover domain must be within the same cluster; however, the failover domain does not have to include every node in the cluster.The administrator defines the *initial failover domain* when creating a failover policy. This list is transformed into the *run-time failover domain* by the *failover script*; the run-time failover domain is what is actually used to select the failover node. IRIS FailSafe stores the run-time failover domain and uses it as input to the next failover script invocation. The initial and run-time failover domains may be identical, depending upon the contents of the failover script. In general, IRIS FailSafe allocates a given resource group to the first node listed in the run-time failover domain that is also in the node membership; the point at which this allocation takes place is affected by the *failover attributes*.

failover policy

The method used by IRIS FailSafe to determine the destination node of a failover. A failover policy consists of a *failover domain*, *failover attributes*, and a *failover script*. A failover policy name must be unique within the *pool*.

failover script

A failover policy component that generates a *run-time failover domain* and returns it to the IRIS FailSafe process. The IRIS FailSafe process applies the failover attributes and then selects the first node in the returned failover domain that is also in the current node membership.

heartbeat messages

Messages that cluster software sends between the nodes that indicate a node is up and running. Heartbeat messages and *control messages* are sent through a node's network interfaces that have been attached to a control network. A node can be attached to multiple control networks.

heartbeat interval

Interval between heartbeat messages. The node timeout value must be at least 10 times the heartbeat interval for proper IRIS FailSafe operation (otherwise false failovers may be triggered). The higher the number of heartbeats (smaller heartbeat interval), the greater the potential for slowing down the network. Conversely, the fewer the number of heartbeats (larger heartbeat interval), the greater the potential for reducing availability of resources.

initial failover domain

The ordered list of nodes, defined by the administrator when a failover policy is first created, that is used the first time a cluster is booted.The ordered list specified by the initial failover domain is transformed into a *run-time failover domain* by the *failover script*; the run-time failover domain is used along with failover attributes to determine the node on which a resource group should reside. With each failure, the failover script takes the current run-time failover domain and potentially modifies it; the initial failover domain is never used again. Depending on the run-time conditions and contents of the failover script, the initial and run-time failover domains may be identical. See also *run-time failover domain*.

key/value attribute

A set of information that must be defined for a particular resource type. For example, for the resource type *filesystem*, one key/value pair might be *mount_point=/fs1* where *mount_point* is the key and *fs1* is the value specific to the particular resource being defined. Depending on the value, you specify either a *string* or *integer* data type. In the previous example, you would specify *string* as the data type for the value *fs1*.

log configuration

A log configuration has two parts: a *log level* and a *log file*, both associated with a *log group*. The cluster administrator can customize the location and amount of log output, and can specify a log configuration for all nodes or for only one node. For example, the *crsd* log group can be configured to log detailed level-10 messages to the */var/cluster/ha/log/crsd-foo* log only on the node *foo*, and to write only minimal level-1 messages to the *crsd* log on all other nodes.

log file

A file containing IRIS FailSafe notifications for a particular *log group*. A log file is part of the *log configuration* for a log group. By default, log files reside in the */var/cluster/ha/log* directory, but the cluster administrator can customize this. Note: IRIS FailSafe logs both normal operations and critical errors to */var/adm/SYSLOG*, as well as to individual logs for specific log groups.

log group

A set of one or more IRIS FailSafe processes that use the same log configuration. A log group usually corresponds to one IRIS FailSafe daemon, such as *gcd*.

log level

A number controlling the number of log messages that IRIS FailSafe will write into an associated log group's log file. A log level is part of the log configuration for a log group.

node

See *cluster node*

node ID

A 16-bit positive integer that uniquely defines a cluster node. During node definition, IRIS FailSafe will assign a node ID if one has not been assigned by the cluster administrator. Once assigned, the node ID cannot be modified.

node membership

The list of nodes in a cluster on which IRIS FailSafe can allocate resource groups.

node timeout

If no heartbeat is received from a node in this period of time, the node is considered to be dead. The node timeout value must be at least 10 times the heartbeat interval for proper IRIS FailSafe operation (otherwise false failovers may be triggered).

notification command

The command used to notify the cluster administrator of changes or failures in the cluster, nodes, and resource groups. The command must exist on every node in the cluster.

offline resource group

A resource group that is not highly available in the cluster. To put a resource group in offline state, IRIS FailSafe stops the group (if needed) and stops monitoring the group. An offline resource group can be running on a node, yet not under IRIS FailSafe control. If the cluster administrator specifies the *detach only* option while taking the group offline, then IRIS FailSafe will not stop the group but will stop monitoring the group.

online resource group

A resource group that is highly available in the cluster. When IRIS FailSafe detects a failure that degrades the resource group availability, it moves the resource group to another node in the cluster. To put a resource group in online state, IRIS FailSafe starts the group (if needed) and begins monitoring the group. If the cluster administrator specifies the *attach only* option while bringing the group online, then IRIS FailSafe will not start the group but will begin monitoring the group.

owner host

A system that can control an IRIS FailSafe node remotely, such as power-cycling the node. Serial cables must physically connect the two systems through the node's system controller port. At run time, the owner host must be defined as a node in the IRIS FailSafe pool.

owner TTY name

The device file name of the terminal port (TTY) on the *owner host* to which the system controller serial cable is connected. The other end of the cable connects to the IRIS FailSafe node with the system controller port, so the node can be controlled remotely by the owner host.

pool

The entire set of nodes involved with a group of clusters. The group of clusters are usually close together and should always serve a common purpose. A replicated database is stored on each node in the pool.

port password

The password for the system controller port, usually set once in firmware or by setting jumper wires. (This is not the same as the node's root password.)

powerfail mode

When powerfail mode is turned *on*, IRIS FailSafe tracks the response from a node's system controller as it makes reset requests to a cluster node. When these requests fail to reset the node successfully, IRIS FailSafe uses heuristics to try to estimate whether the machine has been powered down. If the heuristic algorithm returns with success, IRIS FailSafe assumes the remote machine has been reset successfully. When powerfail mode is turned *off*, the heuristics are not used and IRIS FailSafe may not be able to detect node power failures.

process membership

A list of process instances in a cluster that form a process group. There can multiple process groups per node.

resource

A single physical or logical entity that provides a service to clients or other resources. For example, a resource can be a single disk volume, a particular network address, or an application such as a web server. A resource is generally available for use over time on two or more nodes in a *cluster*, although it can be allocated to only one node at any given time. Resources are identified by a *resource name* and a *resource type*. Dependent resources must be part of the same *resource group* and are identified in a *resource dependency list*.

resource dependency

The condition in which a resource requires the existence of other resources.

resource group

A collection of resources. A resource group is identified by a simple name; this name must be unique within a cluster. Resource groups cannot overlap; that is, two resource groups cannot contain the same resource. All interdependent resources must be part of the same resource group. If any individual resource in a resource group becomes unavailable for its intended use, then the entire resource group is considered unavailable. Therefore, a resource group is the unit of failover for IRIS FailSafe.

resource keys

Variables that define a resource of a given resource type. The action scripts use this information to start, stop, and monitor a resource of this resource type.

resource name

The simple name that identifies a specific instance of a *resource type*. A resource name must be unique within a given resource type.

resource type

A particular class of *resource*. All of the resources in a particular resource type can be handled in the same way for the purposes of *failover*. Every resource is an instance of exactly one resource type. A resource type is identified by a simple name; this name must be unique within a cluster. A resource type can be defined for a specific node or for an entire cluster. A resource type that is defined for a node overrides a cluster-wide resource type definition with the same name; this allows an individual node to override global settings from a cluster-wide resource type definition.

resource type dependency

A set of resource types upon which a resource type depends. For example, the *filesystem* resource type depends upon the *volume* resource type, and the *Netscape_web* resource type depends upon the *filesystem* and *IP_address* resource types.

run-time failover domain

The ordered set of nodes on which the resource group can execute upon failures, as modified by the *failover script*. The run-time failover domain is used along with failover attributes to determine the node on which a resource group should reside. See also *initial failover domain*.

start/stop order

Each resource type has a start/stop order, which is a nonnegative integer. In a resource group, the start/stop orders of the resource types determine the order in which the resources will be started when IRIS FailSafe brings the group online and will be stopped when IRIS FailSafe takes the group offline. The group's resources are started in increasing order, and stopped in decreasing order; resources of the same type are started and stopped in indeterminate order. For example, if resource type *volume* has order 10 and resource type *filesystem* has order 20, then when IRIS FailSafe brings a resource group online, all volume resources in the group will be started before all file system resources in the group.

system controller port

A port sitting on a node that provides a way to power-cycle the node remotely. Enabling or disabling a system controller port in the cluster configuration database (CDB) tells IRIS FailSafe whether it can perform operations on the system controller port. (When the port is enabled, serial cables must attach the port to another node, the owner host.) System controller port information is optional for a node in the pool, but is required if the node will be added to a cluster; otherwise resources running on that node never will be highly available.

tie-breaker node

A node identified as a tie-breaker for IRIS FailSafe to use in the process of computing node membership for the cluster, when exactly half the nodes in the cluster are up and can communicate with each other. If a tie-breaker node is not specified, IRIS FailSafe will use the node with the lowest node ID in the cluster as the tie-breaker node.

type-specific attribute

Required information used to define a resource of a particular resource type. For example, for a resource of type *filesystem*, you must enter attributes for the resource's volume name (where the file system is located) and specify options for how to mount the file system (for example, as readable and writable).

Index

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V

W

X

Tell Us About This Manual

As a user of Silicon Graphics products, you can help us to better understand your needs and to improve the quality of our documentation.

Any information that you provide will be useful. Here is a list of suggested topics:

- General impression of the document
- Omission of material that you expected to find
- Technical errors
- Relevance of the material to the job you had to do
- Quality of the printing and binding

Please send the title and part number of the document with your comments. The part number for this document is 007-3900-001.

Thank you!

Three Ways to Reach Us

- To send your comments by **electronic mail**, use either of these addresses:
  - On the Internet: techpubs@sgi.com
  - For UUCP mail (through any backbone site): *[your_site]*!sgi!techpubs
- To **fax** your comments (or annotated copies of manual pages), use this fax number: 415-965-0964
- To send your comments by **traditional mail**, use this address:

  Technical Publications
  Silicon Graphics, Inc.
  2011 North Shoreline Boulevard, M/S 535
  Mountain View, California 94043-1389