White Paper

# FailSafe™: High Availability at Low Cost

## 1.0 Introduction

In today's global business environment, the availability of information and computing resources is extremely important. Downtime in any system component means lost productivity, lost revenue, and reduced competitiveness. Minimizing the impact of computer failures is crucial but complicated.

FailSafe is an innovative solution that provides high availability. FailSafe is a lightweight, low-cost solution that significantly reduces downtime without affecting system performance or deployment time. The actions FailSafe automatically takes can be tailored by your system administrator to fit your organization's specific business requirements.

This paper discusses how FailSafe can provide solutions to common downtime problems. It gives you an overview of the hardware and software components in a FailSafe cluster and discusses how you can customize FailSafe to meet your specific needs.

**Note:** This paper discusses FailSafe for the IRIX® OS. However, the Linux FailSafe™ 1.0 product also provides support for Linux using open-source software. The FailSafe for Linux product provides much of the same functionality as the IRIX product. For more information, see the Linux FailSafe Administrator's Guide, Linux FailSafe Programmer's Guide, and Linux FailSafe Web site at *http://oss.sgi.com/projects/failsafe/*.

## 2.0 FailSafe Solutions to Downtime Problems

As shown in figure 1, a typical client/server-based business environment contains:

- End users, such as salespeople in an online transaction-processing environment, people watching set-top boxes in a video-on-demand business, users accessing their mail on an Internet service provider's server, worldwide users accessing a corporate Web site, or students using a departmental file server in a university
- Clients connected to a network, on which the end users do their work
- One or more servers connected to the network
- Data storage devices that maintain the current state of the business data
- Applications running on the servers, such as databases, video servers, NFS, or custom-made applications for a particular business

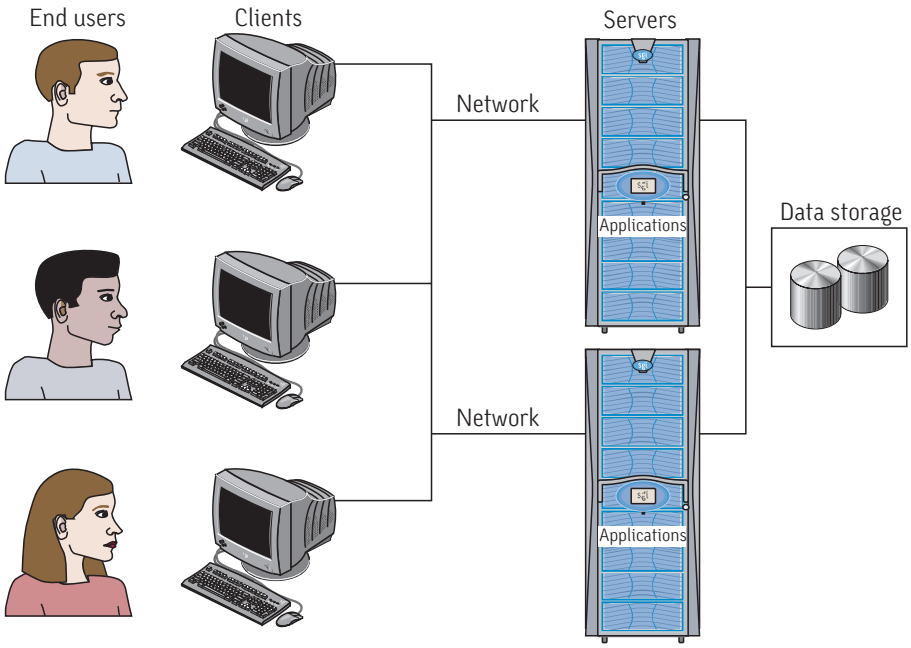


Fig. 1. Typical client/server business environment

A failure in any system component means that the end users cannot do their work. The negative impact of failures on the entire business is progressive, with server resources being the most critical, and client desktops being the least critical. The following sections analyze the kinds of failures that can occur in these components and discuss ways to minimize downtime by using FailSafe.

### 2.1 Cost/Benefit of FailSafe

FailSafe uses a clustered approach to achieve highly available applications in a cost-effective way, as compared with proprietary fault tolerance techniques.

Different degrees of availability are provided by different types of systems:

- Continuously available, or fault-tolerant, systems use redundant proprietary components and specialized logic to ensure continuous operation and to provide complete data integrity. These systems will tolerate outages due to hardware or software upgrades. This solution provides an extremely high degree of availability but is very expensive. The cost makes this solution prohibitive for most business purposes. Also, it does not guard against all failures, such as a bug in an application or an operator error.
- Highly available systems survive single points of failure by using redundant off-the-shelf components and specialized software. Typically these systems provide high availability only for client/server applications and base their redundancy on cluster architectures with shared resources. They provide a lower degree of availability than the continuously available systems but at much lower cost. It is important to understand the cost/benefit tradeoff to achieve the most appropriate level of availability for your particular business needs. The cost of deploying any system varies depending on the techniques used to increase tolerance against failures. FailSafe uses standard off-the-shelf components to provide a low-cost, lightweight solution that greatly improves the availability of applications. When compared with continuously available systems, FailSafe represents an excellent value. Figure 2 shows this concept.

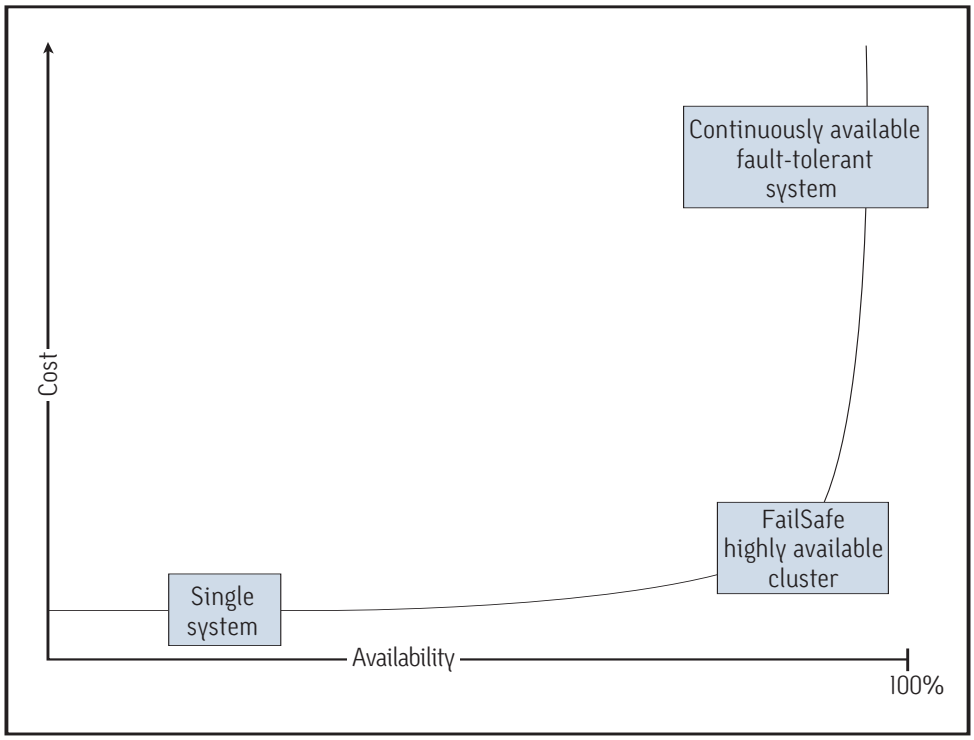


Fig. 2. Cost/benefit value of FailSafe

## 2.2 Server Downtime

Server downtime has the most impact on a client/server business application. Server downtime occurs when a server is inaccessible for useful work for a period of time. Server downtime could happen due to any of the following reasons:

- Power failure
- Critical server component failure
- Application failure
- Administrative error
- Planned outage for system maintenance or upgrades

The probability of server downtime can be reduced by using various techniques, such as the use of redundant power supplies and careful hardware and software design. However, the possibility cannot be eliminated.

Some of the reasons described above can cause lengthy downtimes. For example, critical hardware component failure will require the availability of spares and the appropriate personnel to get the system running again. Other failures, such as system crashes due to software bugs, may be fixed by a system reboot. These fixes, however, are initiated only after the failure has been detected. It is very likely that failure would be detected only after many end users discover that they cannot execute their business transactions. Therefore, significant damage would already have been done to the business before the failure was even detected and isolated. This implies two requirements:

- Efficient detection mechanisms that detect server failures and initiate recovery steps before clients are affected
- A backup server that can quickly take over all the business-critical applications after the failure of a server is detected (sometimes described as a hot standby state); the backup server, while acting as insurance against server failures, should also do useful work such as running other business applications or acting as a test bed for future deployment

A FailSafe environment consists of a cluster of two or more SGI® servers running business-critical applications and a tightly interwoven mechanism to detect failures on any of the servers. Two or more servers working independently while still picking up services from each other (in case of failure) provide high availability with low overhead cost.

The detection mechanism has negligible performance overhead. If one server fails, another server in the cluster automatically assumes the failed server's functions. FailSafe servers can run unattended while being monitored by failure detection software in the FailSafe framework solution.

FailSafe detects server failures using a heartbeat mechanism. There is a private network that connects all the servers in the FailSafe cluster. FailSafe processes use this network to exchange heartbeat messages. The heartbeat messages enable the servers in the cluster to determine each other's state.

## 2.3 I/O Subsystem Downtime

If a critical piece of data is not accessible, it will cause downtime, even if the rest of the system is up and running. This data could be a database record, a movie being requested by a set-top box, the mailbox of a user checking e-mail on an Internet service provider's server, or a critical Web page requested by a user across the continent. Data could be inaccessible because of following reasons:

- Disk error
- Power supply failure within the disk unit
- Error in the path connecting the server to the disk unit, such as a bent pin in the SCSI bus
- I/O-related system software bug
- Accidental data erasure due to operator error
- Data corruption due to application or system bug

All of the above reasons could be eliminated by careful I/O system layout and well-thought-out system maintenance techniques. All solutions employ some kind of data redundancy where multiple copies of the same data are maintained. It is quite possible that the same error could either destroy all copies of the same data or make them all inaccessible. To prevent

the last two errors, you must establish routine system backups and a lost-data recovery process. All the critical data in a FailSafe cluster is maintained using one of the following:

- SGI redundant arrays of independent disks (RAIDs) using CXFS™ clustered filesystems and the XVM volume manager
- SGI RAID using XFS™ filesystems and either the XVM or the XLV volume manager
- Just a bunch of disks (JBOD) using XFS filesystems and XLV or (when using direct attach) XVM

Each method is designed to protect data and provide continuous access to information by surviving any single disk failure.

SGI RAID consists of redundant power supplies to protect against a single power supply failure. When using XVM mirroring, the data mirrors are placed on JBODs on different Fibre Channel enclosures or SCSI vaults so that if there is a power failure at one location, another copy of the data is still available on the other location. SGI® Total Performance RAID supports two storage processors, which provide two different paths to each of the logical disk units in the RAID box. The server is connected to the two storage processors using independent cables. In case of an error anywhere along the primary path to a disk, the XVM or XLV software, in coordination with the RAID firmware, switches to the alternative path.

In the case of mirroring, XVM or XLV software automatically starts using only the accessible mirrors when one of the mirrors is rendered inaccessible due to an error in the path from the server to the disks.

**Note:** All of the above recovery actions (resulting from disk faults, power supply failures, or an error in the path to the disks) are application transparent. Applications do not get any I/O errors and do not have to execute any I/O retries. The disks containing the critical data in a FailSafe cluster are physically connected to at least two servers in the cluster. When directly connected to two servers in the cluster, they are said to be dual-hosted; they may also be SAN-attached via switches to multiple systems.

An XFS filesystem can be mounted on only one server at a time. A CXFS filesystem can be mounted on multiple servers at the same time.

FailSafe constantly monitors these filesystems and volumes and therefore detects any I/O fault that cannot be recovered by any of the above mechanisms. Examples of such faults could be system software errors or system administrator errors such as the unmounting of a critical filesystem. After detecting a fault, FailSafe stops all applications accessing these filesystems and unmounts the filesystems from their primary servers. These filesystems are then mounted on another server in the cluster and the applications using the filesystems are also restarted on the new server. Figure 3 shows an example of disk storage failover in a two-server cluster that uses direct attach rather than a SAN configuration.

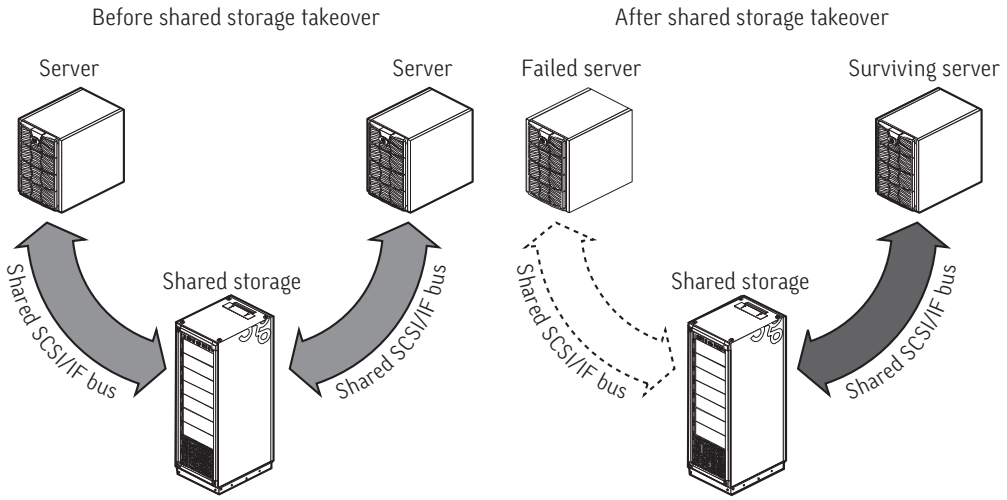


Fig. 3. Disk storage failover in a two-server cluster using direct attach

### 2.4 Applications Downtime

Applications themselves may crash or become inaccessible for various reasons:

- Application software bug
- Unavailability of a critical system resource
- Human error, whereby an application is accidentally aborted
- Application software upgrade

FailSafe provides application monitoring and recovery software, known as plug-ins, for key applications such as Samba. There are plug-ins provided with the base FailSafe release, optional plug-ins available for purchase from SGI, and customized plug-ins. It is possible to integrate most other business applications with FailSafe and make them highly available. Customized plug-ins written by SGI Professional Services or written by the customer using the script library of shell functions provided with FailSafe and the instructions in the IRIS FailSafe™ Version 2 Programmer's Guide. Each of these plug-ins monitors its respective applications and initiates a failover immediately after detecting (and confirming) application failure.

### 2.5 Network Downtime

A business server is useless if it cannot be accessed on the network, even though the server, business applications, and other system resources may be up and running. Lost network connectivity can be attributed to the following causes:

- Physical disconnection, such as a broken network cable
- Network card failure
- Network-related system software bug

All servers in a FailSafe cluster are connected to the same local area network. Clients connect to the cluster using a logical IP address. FailSafe monitors each of these networks. FailSafe initiates a logical network address failover (IP failover) in case failure on any of these networks is detected. All the highly available resources and applications are also failed over along with the network addresses. Because the logical network address itself is failed over, the clients need not be aware of two different network addresses. There is no change that must be made to the application. Also, there are no software processes related to FailSafe that must run on the client systems, which significantly saves administration costs, especially on sites with large numbers of client systems.

**Note:** Clients that connect to servers by using a connection-oriented (or session-based) protocol will lose their sessions as a result of the failover. When they try to reconnect to the same network address, they will be automatically connected to the server that has taken over the business application.

## 3.0 How FailSafe Works

FailSafe consists of user-level software running in a clustered environment that provides:

- Multiple points of failure recovery
- A single point of administration
- An economic and customizable way to make applications highly available

This section discusses the following:

- The failover process
- Servers
- Disks
- Administration tools
- Monitoring with Performance Co-Pilot™

### 3.1 The Failover Process

FailSafe monitors all critical system resources such as I/O, network, servers, and business applications. If any of these components fails, FailSafe detects the error and initiates a pre-configured recovery action procedure.

Highly available services are monitored by the FailSafe software. If a failure is detected on any of these components, a failover process is initiated. Using FailSafe, you can define a failover policy to establish which server will take over the services and under what conditions. This process consists of resetting the failed server (to ensure data consistency), performing recovery procedures required by the failed-over services, and quickly restarting the services on the server that will take them over. FailSafe supports selective failover, in which individual, highly available applications can be failed over to a backup server independent of any other highly available applications.

FailSafe also supports cascading failover, in which FailSafe allows users to specify multiple backup servers for the applications. When there are failures, the application is moved to a backup server based on user-defined failover policies. If there is failure in the backup server, the application is moved to other backup servers or to the originally active server.

Failover can happen within the same server for some failures, such as network or application failures. Servers such as SGI® Origin® 3800 servers can be partitioned into multiple servers.

In a FailSafe environment, servers can act as backup systems for other servers. Unlike the backup resources in a fault-tolerant system, which act purely as redundant hardware for backup in case of failure, the resources of each server in a highly available system can be used during normal operation to run other applications that are not necessarily highly available services. All highly available services are owned by one server in the cluster at a time.

### 3.2 Servers

FailSafe uses multiple SGI servers in a cluster configuration. More than one server has access to the disks containing business-critical data. The data is contained in a fault-tolerant disk subsystem that protects against any single disk failure, power supply failure, or I/O path error. In normal operation, all systems in a FailSafe cluster can be active, working as if they were independent servers. Figure 4 shows an example configuration for a two-server system.

FailSafe supports the following server types:

- SGI® Origin® 300 series
- SGI® Origin® 3000
- SGI® Onyx® 300 series
- SGI® Onyx® 3000 series
- SGI® Origin® 200
- Silicon Graphics® Onyx2®
- SGI® Origin® 2000 series

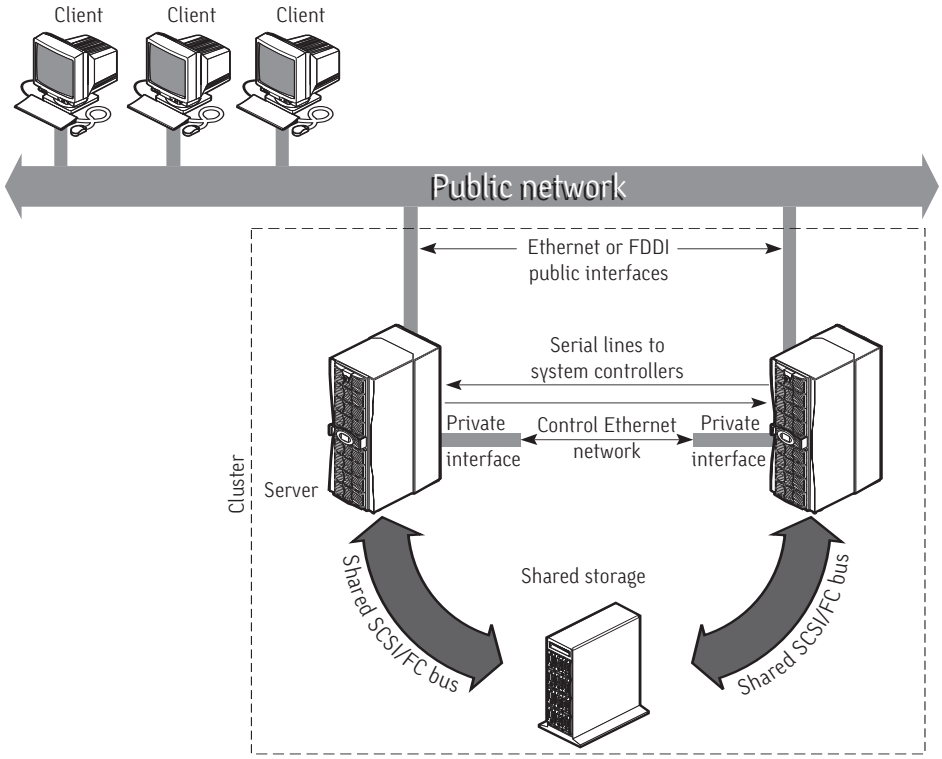


Fig. 4. Sample system components

### 3.3 Disks

FailSafe supports storage based on SCSI or Fibre Channel. Plexing must be used to mirror disks in a JBOD configuration. If highly available applications use filesystems, then XFS filesystems or CXFS filesystems must be used. When CXFS filesystems are used, they must be on XVM volumes. XVM is also available for use in local mode; XVM in local mode cannot be used with CXFS.

**Note:** No SCSI storage or Fibre Channel JBOD is supported in a SAN configuration and therefore cannot be used with CXFS in coexecution with FailSafe. The storage components should not have a single point of failure. All data should be in a RAID or mirrored. It is recommended that there are at least two paths from storage to the servers for redundancy.

For Fibre Channel RAID storage systems, if a disk or disk controller fails, the RAID storage system is equipped to keep services available through its own capabilities. For all the above storage systems, if a disk or disk controller fails, XVM or XLV will keep the service available through a redundant path as appropriate. If no alternative paths are available to the storage subsystems, then FailSafe will initiate a failover process.

### 3.4 Administration Tools

You can perform all FailSafe administrative tasks by means of the FailSafe Manager graphical user interface (GUI), which is based on Java™. You can also perform administrative tasks directly by using the cmgr command, which provides a command-line interface for the administration tasks.

The GUI provides a guided interface to configure, administer, and monitor a FailSafe software-controlled highly available cluster. The GUI also provides screen-by-screen help text. For example, Set Up a New Cluster steps you through the process for creating a new cluster and allows you to launch the necessary individual tasks by simply clicking their titles.

Figure 5 shows a sample GUI window.

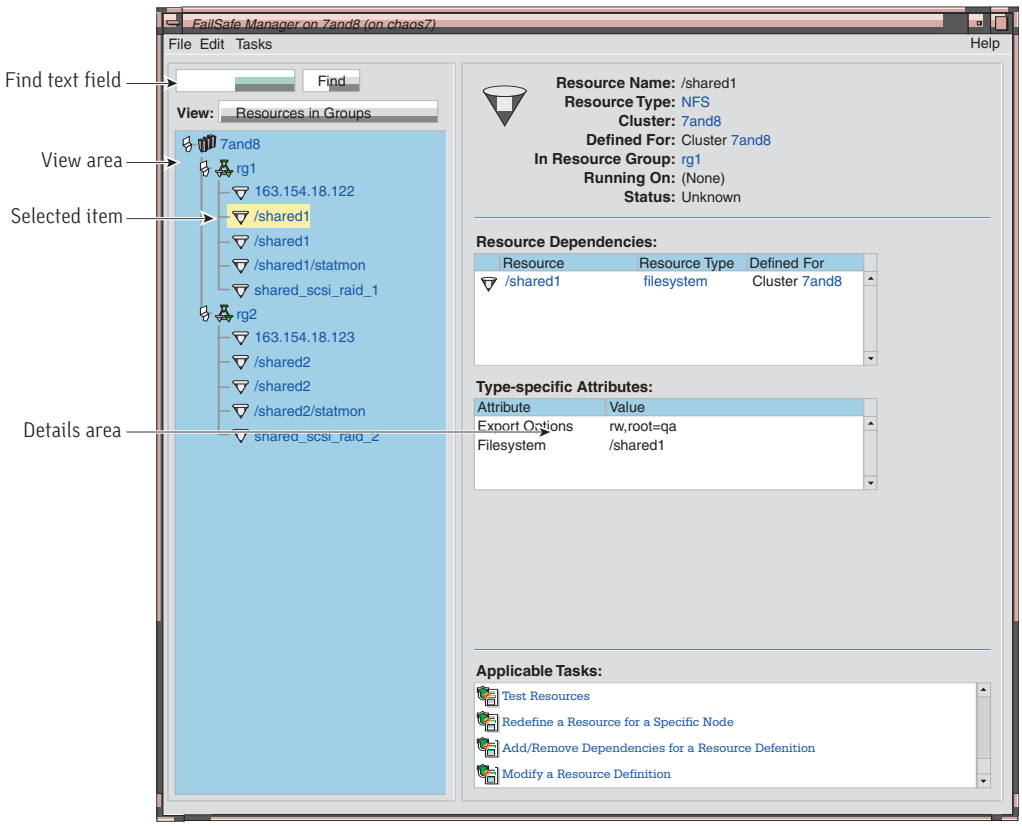


Fig. 5. GUI showing details for a resource

### 3.5 Monitoring with Performance Co-Pilot

FailSafe provides access to Performance Co-Pilot, which provides the following:

- An agent for exporting FailSafe heartbeat and resource monitoring statistics to the Performance Co-Pilot framework
- 3D visualization tools for displaying these statistics in an intuitive presentation

The visualization of statistics provides valuable information about the availability of servers and resources monitored by FailSafe. For example, it can highlight a reduction in monitoring response times that may indicate problems in availability of services provided by the cluster.

Because Performance Co-Pilot for FailSafe is an extension to the Performance Co-Pilot framework, you can use other Performance Co-Pilot tools to analyze or present FailSafe monitoring statistics and record Performance Co-Pilot for FailSafe metrics as archives for deferred analysis. You can also use Performance Co-Pilot to gather statistics about CPU and memory utilization, network and disk activity, and other performance metrics for each server in the cluster.

The Performance Co-Pilot hbvis command constructs a display showing the distribution of heartbeat response times for every server in the cluster. Figure 6 shows an example display.

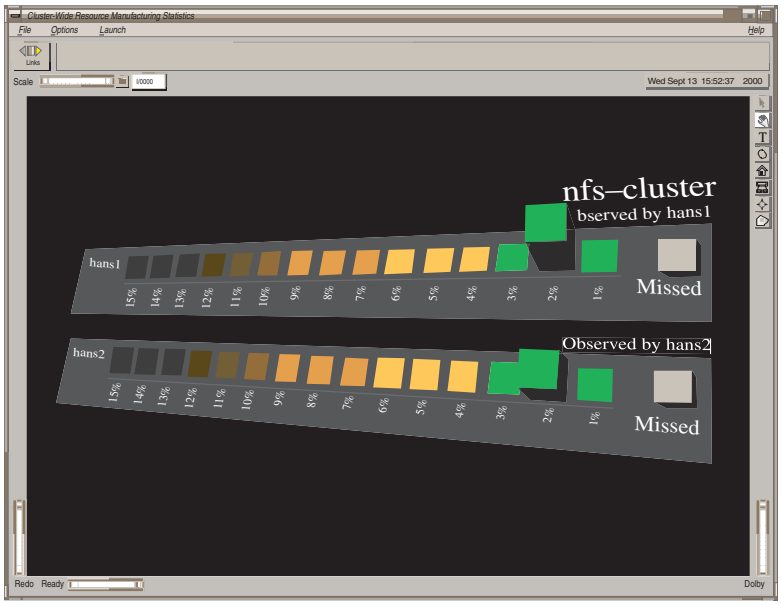


Fig. 6. Heartbeat response statistics

Key features of the display include the frequency of heartbeat responses that arrive at particular intervals within the timeout period and the frequency of heartbeat responses that have been missed (determined not to have arrived). The bar representing the frequency of missed heartbeat responses changes color to indicate the urgency of problems with the availability of a server.

The rmvis command constructs a display of the resource monitoring response times for resources monitored on every server of the cluster. Figure 7 shows an example display.

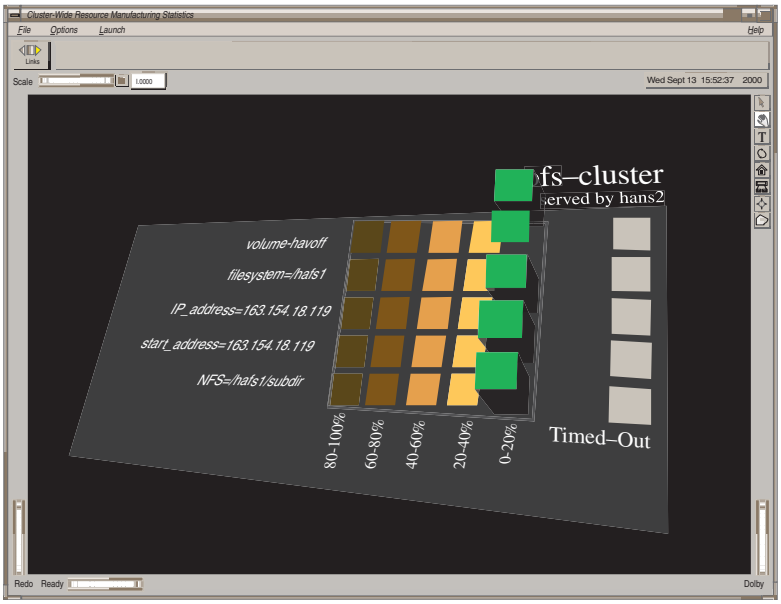


Fig. 7. Resource monitoring statistics

The display is similar in concept to that of hbvis, showing the frequency of resource monitoring responses that arrive within the timeout period and the frequency of responses that have timed out. The bar representing the frequency of resource responses that have timed out also changes color to indicate the urgency of problems with the availability of particular resources.

## 4.0 Complete Storage Solution: CXFS, DMF, TMF, and FailSafe

The FailSafe, CXFS, DMF, and TMF products are integrated to provide a complete storage solution.

### 4.1 CXFS

CXFS, the clustered XFS filesystem, allows groups of computers to coherently share large amounts of data while maintaining high performance. You can use FailSafe to provide highly available services (such as NFS or Web) running on a CXFS filesystem. This combination provides high-performance shared data access for highly available applications.

Figure 8 shows an example configuration.

For more information, see the IRIS FailSafe Version 2 Administrator's Guide.

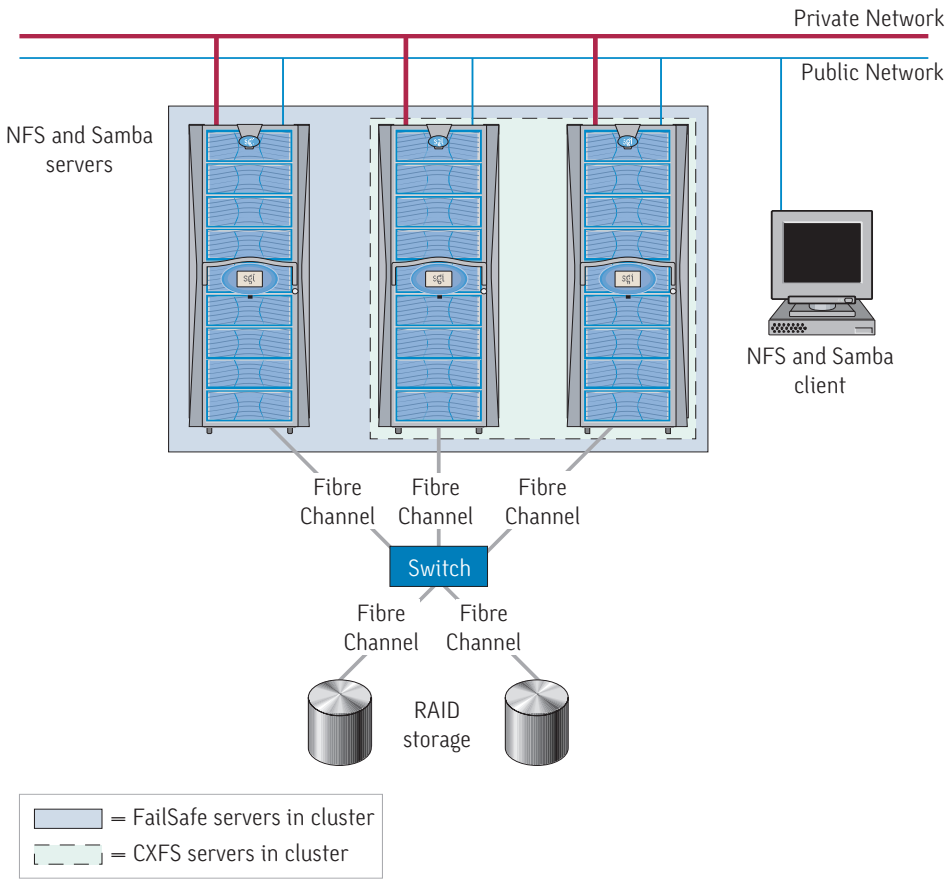


Fig. 8. An example CXFS and FailSafe configuration

### 4.2 DMF

The Data Migration Facility (DMF) is a hierarchical storage management system for SGI environments. Its primary purpose is to preserve the economic value of storage media and stored data. The high I/O bandwidth of these environments is sufficient to overrun online disk resources. Consequently, capacity scheduling, in the form of native filesystem migration, has become an integral part of

many computing environments and is a requirement for effective use of SGI systems. The FailSafe DMF plug-in enables DMF and its resources to be moved from one server to another when a FailSafe failover occurs. If the server that is running FailSafe DMF crashes, DMF fails over to another server, along with its filesystems. For more information, see the IRIS FailSafe 2.0 DMF Administrator's Guide.

### 4.3 TMF

The Tape Management Facility (TMF) is an IRIX subsystem that supports processing of labeled tapes, including multifile volumes and multivolume sets. These capabilities are most important to customers who run production tape operations where tape label recognition and tape security are requirements. The FailSafe TMF plug-in enables TMF and its resources to be failed over from one server to another when a failure occurs. For more information, see the IRIS FailSafe Version 2 TMF Administrator's Guide.

## 5.0 For More Information

For more information about FailSafe, see the following manuals:

- IRIS FailSafe Version 2 Programmer's Guide
- IRIS FailSafe Version 2 Administrator's Guide

You can access these and other SGI manuals at the Tech Pubs Library: *http://docs.sgi.com*. Various formats are available. This library contains the most recent and most comprehensive set of online books, release notes, man pages, and other information.

**Corporate Office**
1600 Amphitheatre Pkwy.
Mountain View, CA 94043
(650) 960-1980
**www.sgi.com**

North America 1(800) 800-7441
Latin America (52) 5267-1387
Europe (44) 118.925.75.00
Japan (81) 3.5488.1811
Asia Pacific (65) 6771.0290



3473 04/24/2003]

JI4245