5000 Series
Best Practices Guide

P/N 83-00004295-10
Revision A
October 2007

The material in this document is for information only and is subject to change without notice. While reasonable efforts have been made in the preparation of this document to assure its accuracy, changes in the product design can be made without reservation and without notification to its users.

Contents

Preface

This guide provides recommendations for optimizing and maximizing the reliability, accessibility, and serviceability (RAS) and for avoiding single points of failure (SPOF) of a R/Evolution™ 5000 Series storage system, and applies to the following enclosures:

- 5730 FC Controller Enclosure
- SAS Expansion Enclosure

This guide is written for system architects, system administrators, and information management professionals contemplating the purchase of storage products, as well as for those who need more information on how to optimize their R/Evolution storage system. This guide assumes a familiarity with Fibre Channel (FC), and Serial Attached SCSI (SAS) configurations, network administration, and RAID technology.

Related Documentation

Application	Title	Part Number
Site planning information	*R/Evolution Storage System Site Planning Guide*	83-00004283
Late-breaking information not included in the documentation set	*R/Evolution 5730 Release Notes*	83-00005008
Installing and configuring hardware	*R/Evolution 5730 Getting Started Guide*	83-00005010
Configuring and managing storage	*R/Evolution 5000 Series Administrator's Guide*	83-00004298
Using the command-line interface (CLI)	*R/Evolution 5000 Series CLI Reference Manual*	83-00004297
Troubleshooting	*R/Evolution 5000 Series Troubleshooting Guide*	83-00004296

CHAPTER 1

R/Evolution 5000 Series Feature Overview

The R/Evolution 5000 Series of storage products comprises hardware and software architectures designed to rapidly evolve and to support your unique storage requirements. Modular and dynamic, the R/Evolution architecture delivers high levels of performance, manageability, and storage capacity density. Data integrity is achieved through a family of data protection and data management services. Highly configurable, the 5000 Series is flexible enough to meet your storage needs while built-in RAS features minimize equipment downtime.

Topics covered in this chapter include:

Storage Optimization

The R/Evolution architecture provides the tools to optimize for all the metrics of storage, including:

- **Performance – Inputs/Outputs per Second (IOPS), data rates (measured in Mbyte/sec, the amount of data processed), and response time.**

  The storage system can be configured to optimize the performance required for your specific application needs. For example, you can set the read-ahead cache size to a larger value for applications that read data in a primarily sequential pattern. There are a few types of write-intensive applications for which performance will improve on virtual disks set for write-through operation with cache mirroring disabled.

- **Manageability – The ability to control, configure, and allocate and deploy storage resources.**

  RAIDar, the web-browser interface, is the primary interface for configuring and managing the system. A web server resides in each controller module. RAIDar enables you to manage the system from a web browser that is properly configured and that can access a controller module through an Ethernet

connection. RAIDar provides the ability to create, configure, and control volumes dynamically without interrupting operations. It provides complete control and reporting tools over every aspect of the storage system.

The command-line interface (CLI) is a secondary interface that can be accessed through an Ethernet connection, or from a local host through a serial port connection, or through the serial port on either controller module.

- **Capacity density – Measurement of the number of disks that can fit into a standard unit of rack space and the number of disks you can attach to a storage system.**

  By maximizing density and capacity, you can optimize total IOPS and greater overall throughput rates per individual rack. The 5000 Series features a dense, space-saving enclosure in a 2U form factor that holds up to 12 drives. An internal SAS interface allows up to nine disk enclosures (including the controller enclosure), delivering a total of 108 drives.

- **Data Protection Services – The ability to protect and ensure the integrity of data.**

  The snapshot service (AssuredSnap™) enables you to create and save snapshots of a volume, where each snapshot preserves the volume's data state at the point in time when the snapshot was created. The volume copy service (AssuredCopy™) provides the ability to copy a point-in-time data capture from one volume to another, providing additional protection against virtual disk failure and eliminating I/O contention when accessing the same blocks.

Configuration Flexibility

Choosing the right storage solution for your organization begins with a thorough review of your storage needs and current infrastructure. The R/Evolution 5000 Series of storage products is flexible enough to meet a wide range of storage needs:

- 5730 FC Controller Enclosure – 2- or 4-Gbit/sec FC host ports. Provides four host connections per controller module.
- Can be configured with a mix of SAS and SATA drives within the same enclosure to deliver the densest multi-tiered storage solution in the market.
- When populated with SAS drives, delivers the most cost-effective, high-performance primary storage.
- When configured with SATA drives, provides a cost-effective, high-capacity solution for near-line, low-duty cycle storage. Ideal for business needs such as email, disk backup, streaming data, or helping to address critical Windows backup issues.

- The 5730 can be used as Direct Attach Storage (DAS) or integrated into a Storage Area Network (SAN) as either dedicated or shared storage to servers in your configuration.
- Depending on your specific application needs, the storage system can be configured for performance, capacity, fault tolerance, or connectivity requirements.

Reliability, Availability, and Serviceability

R/Evolution storage systems contain built-in reliability, availability, and serviceability (RAS) features, which are features and initiatives designed to maximize equipment uptime and mean time between failures (MTBF), minimize downtime and the length of time necessary to repair failures, and eliminate or decrease single points of failure (SPOF) through redundancy.

- **Reliability – Features that help avoid and detect faults**

  A reliable design increases the chance that a system can operate for a long period of time before suffering any failure. For example, the storage system features a dedicated channel between a pair of controllers enables each RAID controller to be aware of its partner controller. Through this dedicated communications channel (heartbeat), the controller can share its status with its partner and force a failover if necessary. The channel is also used to broadcast write to both controllers' cache, providing additional data redundancy.

- **Availability – A measure of the degree to which a system is capable of performing its intended function**

  Mathematically, availability is the percentage of time a system is in a functioning state. For a data storage system, availability reflects to what extent the data hosts attached to the storage system can read and write data that is located on the storage system.

  Availability is achieved through two approaches: improving reliability and adding redundancy. If a failure occurs, adequate redundancy ensures that the system can continue to operate. Faced with a component failure, a data storage system must not only protect against data loss or corruption, but must also ensure that data can be read and written by the data hosts using the storage system.

  A system that employs RAS features to ensure that the storage system will not have any downtime is defined as *high availability.*

- **Serviceability – Various methods of diagnosing and notifying the system when problems arise and being able to service a component in the system without shutting down the entire operation**

  Options for diagnosing your system include recovery and debug utilities and diagnostic tools. If a problem should occur with a component, such as a controller, power supply, or drive, the 5000 Series of storage products enables you to remove it without interrupting the data flow.

The R/Evolution 5000 Series of storage products contains the following built-in RAS features:

- **Dual-controller configuration**

  Dual-controller configurations avoid SPOF. In a dual-controller configuration, both controllers actively and independently process I/O in the data path (known as *active-active* operation), and provide failover capability for the data path. Both controllers are connected to the same set of disk drives in the controller enclosure and to any attached expansion enclosures, providing better I/O performance and failure tolerance than a single controller.

  Ownership of virtual disks, and the volumes they contain, is automatically assigned and balanced between dual controllers. Dual controllers also mirror each others' cache and controller software configuration, and provide multipath support. System resources can be accessed through a connection to either controller module; if a controller module fails, its partner controller takes ownership of the resources of the failed controller. In an enclosure with redundant controller modules, the failure of a single- controller module does not disable access to the storage.

- **Power redundancy**

  Power redundancy is achieved through two independent load-sharing power supplies. In the event of a power supply failure, or the failure of the power source, to maintain optimal cooling, the storage system can operate continuously on a single power supply. Greater redundancy can be achieved by connecting the power and cooling modules to separate circuits.

- **Superior cache backup design**

  Because batteries are unpredictable and fault-sensitive technology, as well as harmful to the environment, they have been replaced with far superior, super-capacitor EcoStor™ technology to protect RAID controller write cache in case of power failure. With up to a ten-year life span, EcoStor eliminates service calls for battery replacement, new battery inventory management, and the issues of battery disposal, as well as the periodic replacement downtime associated with batteries. Additionally, it offers improved customer experience upon installation or power

restore as it does not require lengthy battery charging time. The system operates in high-performance write-back cache mode within minutes of installation versus hours with batteries.

The EcoStor super-capacitor pack and compact flash memory in each controller module provide unlimited cache memory backup time. The super-capacitor pack provides energy for backing up unwritten data in the write cache to the compact flash in the event of a power failure. After the pending writes are flushed to the compact flash, the super-capacitor pack continues to provide a trickle charge to the cache in the event that the outage was just brown-out. Unwritten data in compact flash memory is automatically committed to disk media when power is restored.

- **SimulCache™**

  SimulCache leverages redundant RAID controllers, thus eliminating the performance degradation associated with conventional cache mirroring. By automatically broadcasting write data to the other controller's cache, the primary latency overhead is eliminated and bandwidth requirements are reduced on the primary cache. This exceptionally high bandwidth, low latency, active-active write-back cache capability approaches the performance of a dual- independent cache mode operation, providing the advantage of improved data protection options without sacrificing application performance or end-user responsiveness.

- **Interconnected host ports**

  The host ports on both controllers are used when creating active-active configurations. Internal connections of these ports increase data availability by facilitating controller failover and multipath connection to hosts.

Even with all of the built-in RAS features of the storage system and when following the best practices described in this guide, rare failures that disrupt storage access can still occur. It is important to implement best practices of system design in all areas, not just disk storage. Designing a highly available system requires careful planning, rigorous testing, and a continual effort to ensure network reliability.

CHAPTER 2

Planning Your Storage Architecture

Selecting the best storage architecture for a particular environment can be challenging. An effective approach for designing a R/Evolution storage solution for your environment is to answer the questions presented in this chapter.

Does your storage configuration require:

- Dual controllers or single controller?
- Direct attach or switch attach connections to the data host?
- Fault-tolerance or high performance?
- Connection to a single data host or multiple data hosts?
- SAS or SATA disk drives?
- Loop or point-to-point topology?

This chapter provides information to help you select the best storage architecture for your environment. Best practices for installing the storage system, along with representative sample cabling diagrams, are presented in “Hardware Configuration” on page 25. For detailed steps on how to configure your system based on the information presented in this chapter, refer to the *Administrator’s Guide* and the *CLI Reference Guide*.

Dual Controllers or Single Controller?

Although you can purchase a single-controller configuration, it is best practice to use the dual-controller configuration to ensure data availability. However, under certain circumstances, a single-controller configuration can be used as an overall redundant solution.

Dual Controller

A dual-controller configuration improves application availability because, in the unlikely event of a controller failure, the affected controller fails over to the surviving controller with no interruption to the flow of data. The failed controller can be replaced without shutting down the storage system, thereby providing further

increased data availability. An additional benefit of dual controllers is increased performance as storage resources can be divided between the two controllers, enabling them to share the task of processing I/O operations.

Controller failure results in the surviving controller:

- Taking ownership of all RAID sets
- Managing the failed controller's cache data
- Restarting data protection services
- Assuming the host port characteristics of both controllers

The dual-controller configuration takes advantage of patented SimulCache technology. By automatically "broadcasting" one controller's write data to the other controller's cache, the primary latency overhead is eliminated and bandwidth requirements are reduced on the primary cache. Any power loss situation will result in the immediate writing of cache data into both controllers' compact flash devices, eliminating any data loss concerns. The broadcast write implementation provides the advantage of enhanced data protection options without sacrificing application performance or end-user responsiveness.

Single Controller

A single-controller configuration provides no redundancy in the event that the controller fails; therefore, the single controller is a potential SPOF. Multiple hosts can be supported in this configuration (up to four for direct attach) and more if a Fibre Channel switch is used. Because there is no second controller, the host port interconnect is disabled. In this configuration, each host can have 2- or 4-Gbit/sec access to the storage resources. If the controller fails, the host loses access to the storage.

The single-controller configuration is less expensive than the dual-controller configuration. It is a suitable solution in cases where high availability is not required and loss of access to the data can be tolerated for short periods of time. A single-controller configuration is also an appropriate choice in storage systems where redundancy is achieved at a higher level, such as a two-node cluster. For example, a two-node cluster where each node is attached to a controller enclosure with a single controller and the nodes do not depend upon shared storage. In this case, the failure of a controller is equivalent to the failure of the node to which it is attached.

Another suitable example of a high-availability storage system using a single-controller configuration is where a host uses a volume manager to mirror the data on two independent single-controller storage systems. If one storage system fails, the other storage system can continue to serve the I/O operations. Once the failed controller is replaced, the data from the survivor can be used to rebuild the failed system.

Direct Attach or Switch Attach Connections?

There are two basic methods for connecting storage to data hosts: Direct Attach Storage (DAS) and Storage Area Networks (SANs). The option you select depends on the number of hosts you plan to connect and how rapidly you need your storage solution to expand.

Direct Attach

Direct Attach Storage (DAS) uses a direct connection between a data host and its storage system. The DAS solution of connecting each data host to a dedicated storage system is straightforward, and the absence of storage switches can minimize cost. Like a SAN, a DAS solution can also share a storage system but it's limited by the number of ports on the system. Host port interconnects are always enabled, except in configurations where fault-tolerance is less important than higher performance.

A powerful feature of the storage system is its ability to support eight direct attach single-port data hosts, or four direct attach dual-port data hosts without requiring storage switches.

If the number of connected hosts is not going to change or increase beyond four, then the DAS solution is appropriate. However, if the number of connected hosts is going to expand beyond the limit imposed by the use of DAS, it is best to implement a SAN.

Tip – It is a best practice to use a dual-port connection to data hosts when implementing a DAS solution.

Switch Attach

A switch attach solution, or Storage Area Network (SAN), places a storage switch between network servers and storage systems. This storage strategy tends to use storage resources more effectively and is commonly referred to as *storage consolidation.* A SAN solution shares a storage system among multiple servers using switches, and reduces the total number of storage systems required for a particular environment, at the cost of additional element management (switches) and path complexity. Host port interconnects are always disabled. There is an exception to this rule: host port interconnects are enabled for applications where fault-tolerance is required and highest performance is not required, and when switch ports are at a premium.

Using switches increases the number of servers that can be connected. Essentially, the maximum number of data hosts that can be connected to the SAN becomes equal to the number of available storage switch ports. Storage switches generally include the ability to manage and monitor the FC networks they create, which can reduce storage management workloads in multiple server environments. A SAN is suitable for environments where:

- There is going to be 20% growth each year
- There need to be hosts in different locations
- There is an additional need for security and performance that a switch can provide

---

Tip – It is a best practice to use a switched SAN environment anytime more than four hosts will be used or when growth in required storage or number of hosts is expected.

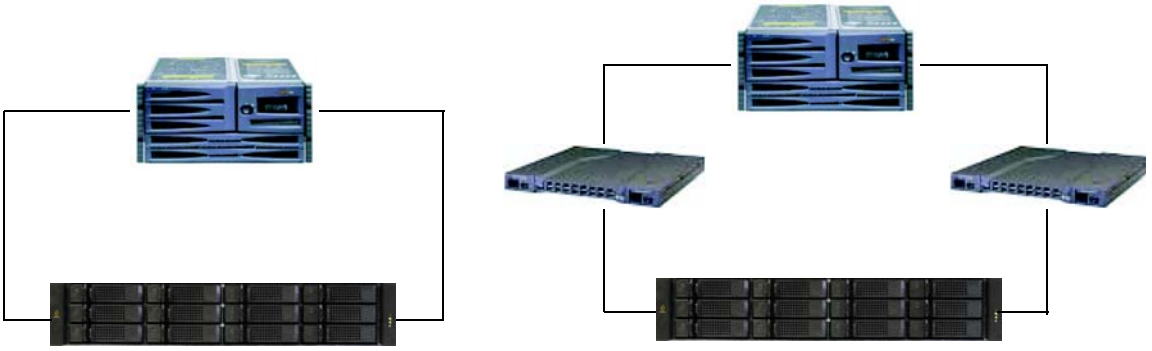


Figure 2-1 DAS and SAN Storage Configurations

Fault Tolerance or Performance?

Depending on whether fault tolerance (where redundant components are designed for continuous processing) or performance is more important to your storage system, the host port interconnects need to be enabled or disabled, which is done through RAIDar. In an FC storage system, the host port interconnects act as an internal switch to provide data-path redundancy.

When availability is more important than performance, the host port interconnects should be enabled to connect the host ports in controller A to those in controller B. When the interconnects are enabled, the host has access to both controllers' mapped volumes. This dual access makes it possible to create a redundant configuration without using an external switch.

If one controller fails in this configuration, the interconnects remain active so hosts can continue to access all mapped volumes without the intervention of host-based failover software. The controllers accomplish this by means of FC target multi-ID: while a controller is failed over, each surviving controller host port presents its own port WWN and the port WWN of the interconnected, failed controller host port that was originally connected to the loop. The mapped volumes owned by the failed controller remain accessible until it is removed from the enclosure.

When the host port interconnects are disabled, volumes owned by a controller are accessible from its host ports only. This is the default.

When controller enclosures are attached directly to hosts and high availability is required, host port interconnects should be enabled. Host port interconnects are also enabled for applications where fault tolerance is required and highest performance is not, and when switch ports are at a premium.

When controller enclosures are attached through one or more switches, or when they are attached directly but performance is more important than fault tolerance, host port interconnects should be disabled.

Tip – It is a best practice to enable host port interconnects when controller enclosures are attached directly to hosts and high availability is required, or when switch ports are at a premium and fault tolerance is required.

Note – Fault tolerance and performance are affected by cache settings as well. See "Cache Configuration" on page 48 for more information.

One or Multiple Data Hosts?

The number of data hosts that can effectively share a storage system depends on several factors, such as the type of host application, bandwidth requirements, and the need for concurrent IOPS. Because most applications have moderate performance needs, it is feasible to have several hosts sharing the same controller. The storage system supports eight direct attach single-port data hosts, or four direct-attach dual-port data hosts without requiring storage switches. Determine how much storage is currently accessible to these data hosts and then plan for that total capacity as the minimum amount of disk capacity needed.

Combining storage switches with a controller enclosure creates a SAN, increasing the number of data hosts that can be connected. Essentially, the maximum number of hosts that can be connected to the SAN becomes equal to the number of available storage switch ports. The storage system architecture can support a maximum of 64 data hosts, although connecting a large number of hosts can affect performance. Increasing the number of controller enclosures makes more performance and capacity available within a storage network for sharing among the servers connected to the SAN. A SAN also provides great flexibility in how storage capacity can be allocated among data hosts and eliminates cabling changes when reallocation of storage becomes necessary. Storage switches generally include the ability to manage and monitor the FC networks they create, which can reduce storage management workloads in multiple data host environments.

SAS or SATA Disk Drives?

The type of disk drive you choose is dependent on your specific application. For example, if you use the Information Lifecycle Management (ILM) strategy, you need to decide what data should reside where in the physical infrastructure, and how to optimize it for use by the appropriate application at the right time. Data is typically stored in tiers according to specific policies and migrated from one tier to another based on those criteria.

Tiered storage assigns data to progressively less expensive devices based on different categories, including levels of protection needed, performance requirements, frequency of use, and capacity. As a rule, tier 1, also called online data, which is data that must be accessed more frequently (such as mission-critical or top secret files) is stored on faster, but more expensive storage media. Tier 2, also called near-line data, (such as financial, seldom-used, or classified files) is stored on less expensive, but slower media.

Depending on your ILM strategy, for example, which includes performance, reliability, and capacity requirements, the storage system can be configured using Serial Attached SCSI (SAS) or Serial Advanced Technology Attached (SATA) disk drives.

SAS disk drives provide high performance, reliability, and availability and are best suited for random-access data and performance-intensive applications. Examples of these applications include eCommerce, digital cinema, life sciences and research; business management applications, such as SAP and Oracle; web services, email and messaging; and telecommunications.

SATA disk drives provide a lower-cost and higher-capacity alternative that is best suited to sequential-access data and secondary storage applications. These drives easily meet secondary (near-line or tier 2) storage requirements. Example applications include disk-to-disk, continuous data protection, virtual tape libraries, video security, and streaming data.

Tip – A best practice for determining your disk drive needs is to know your data and to fully understand what you need to do with it.

Mixing SAS and SATA Drives Within the Same Enclosure

The storage system allows for the mixing of SAS and SATA drives *within the same enclosure*. However, while a virtual disk can contain different *models* of disk drives, it cannot contain a mix of SAS and SATA drives. The controller safeguards against improperly combining SAS and SATA disk drives in a virtual disk. If your system contains multiple types of disk drives, the interface makes available different options. If you mix disk drives with different capacities, the smallest disk drive determines the logical capacity of all other disk drives in the virtual disk, regardless of RAID level. For example, if a RAID-5 virtual disk contains one 250-Gbyte disk drive and four 500-Gbyte disk drives, the capacity of the virtual disk is equivalent to approximately five 250-Gbyte disk drives.

Tip – To maximize capacity and reliability, it is best practice to use disk drives of the same size and rotational speed.

Note – Because of an advanced enclosure design, there are no restrictions on how the drives can be ordered in the chassis. That is, for example, there is no negative affect on performance if you install one column of SATA drives between two columns of SAS drives.

Loop or Point to Point Topology?

For R/Evolution FC storage systems, *topology* is defined as the path that data travels between devices: either through a series of connected devices (loop) or directly from one device to another (point-to-point). In a switch-attach configuration, either topology is supported but loop is preferred because the storage system uses multi-ID to impersonate a failed controller (that is, while a controller is failed over, each surviving controller host port presents its own port WWN and the port WWN of the interconnected, failed controller host port that was originally connected to the loop). In a direct-attach configuration, only loop is supported.

In a dual-controller FC storage system, the topology setting available to a host port depends on the system's host port interconnect setting (see "Fault Tolerance or Performance?" on page 19), and affects host access to volumes during failover, when their owning controller's host ports are inaccessible. This relationship is described in the following paragraphs.

Volumes can be mapped with access privileges through specific host ports to data hosts, with a LUN that identifies each mapping. Host access to volumes during a controller failover is determined by the storage system's host port interconnect and topology settings. For example, assume volumes are mapped through controller A's host ports and controller A fails over to controller B. The host can access controller A's volumes through controller B's host ports as follows:

- If all host ports are set to loop topology, both controllers' volumes are presented on controller B's host ports.
- If one controller fails in a switch attach configuration using loop topology, the host ports on the surviving controller present the port WWNs for both controllers. Each controller's mapped volumes remain accessible. For example, if controller B fails, data access is essentially unchanged because controller A already had access to all mapped volumes before the failure.
- If one or more host ports are set to point-to-point topology, controller B presents its volumes on half of its host ports and presents controller A's volumes on the remaining host ports.

- If one controller fails in a switch attach configuration using point-to-point topology, the surviving controller presents its mapped volumes on its primary host port (FC0) and the mapped volumes owned by the failed controller on the secondary port (FC1).

  Unlike loop topology where both controllers' mapped volumes are presented on all host ports of the surviving controller, point-to-point topology requires host-based failover software to redirect access to the failed controller's mapped volumes through the surviving controller's secondary port. Further redundancy can be added by linking the fabric switches together, however this approach requires detailed path management configuration through the host adapter software.

If host port interconnects are enabled, the paired ports are connected in a loop and must be set to use loop topology; you cannot set any ports to use point-to-point. Changing the topology setting for one host port automatically changes the setting for the paired port on the partner controller.

If host port interconnects are disabled, you can change the topology setting for each host port individually.

Note – In a switch attach connection, if you change from loop to point-to-point after already establishing a public loop connection, the switch might ignore subsequent attempts to perform point-to-point initialization.

CHAPTER 3

Hardware Configuration

This chapter presents best practice information for installing and connecting the controller and expansion controllers. It provides recommendations to supplement the installation procedures in the *Getting Started Guide* and any late-breaking information related to installation in the *Release Notes*. Sample cabling diagrams are presented for a number of configurations.

Topics covered in this chapter include:

General Installation Best Practices

This section lists general best practices to follow when installing your 5000 Series storage system.

- Ensure that your installation site meets all of the requirements in the *Site Planning Guide*. Plan carefully to ensure that you have the proper network, host, and rackmounting equipment for the system. Also gather information such as IP addresses and World Wide Names (WWNs) needed for system configuration.
- Use separate computers for management host and data host functions, so the data host's performance is not adversely affected by management operations. A management host can be directly or remotely attached to a controller enclosure to configure, monitor, and manage the system. A data host can be directly attached or attached through a switch, to a host port on the enclosure to read and write data.

Note – For a small storage configuration, the same computer can be used as a system's management host and data host.

- Use of a local management host with a serial connection guarantees communication with the controller module even if its IP address changes or is unknown, or if the Ethernet network suffers a temporary outage.

- To ensure power redundancy, connect the power and cooling modules in each enclosure to two separate circuits, such as one commercial circuit and one UPS. The fans in the power and cooling modules are driven from a common source on the midplane. If one power supply fails all fans continue to operate.
- To maintain proper airflow in the chassis, ensure that all slots are populated by either active components or air-management system I/O blanks or drive blanks.

Caution – If a drive module is removed and not replaced, the airflow is altered within the enclosure, which could cause overheating. Always have a replacement drive module or air management module (drive blank) on hand to immediately replace the one that was removed.

Network Connectivity

It is extremely important for the proper operation and reliability of the equipment that all network connectivity adhere to Ethernet and facility wiring standards IEEE 802.3 and EIA/TIA 568B. Make sure that the cabling and patch cords for your facilities are set up to these specifications, and protect the cables from excessive stress and damage.

Tip – The best practice, and the one that is recommended by all facilities wiring standards, is to test your structured cable system end-to-end with a quality cable test set. Adherence to these practices will help eliminate almost all connectivity issues.

Cabling

After you have determined your storage architecture needs according to the information described in “Planning Your Storage Architecture” on page 15, determine the best cabling configuration for your environment.

General cabling best practices include the following:

- Cleanly route FC, Ethernet, serial, and SAS cables and clearly label them at each end.
- Data host connections should ensure that redundant paths are split between host bus adapters (HBAs), system buses, and I/O modules to maximize performance and availability.

- If using switches, ensure they are installed and configured as described in the vendor's documentation. Use two switches instead of one zoned switch to avoid the switch being a SPOF.
- To improve serviceability for direct attach connections, connect a data host to one channel on the upper controller module (controller A) and to the other channel on the lower controller module (controller B). This makes the system accessible from either controller if one fails over, and enables the failed controller module to be hot-swapped.
- To improve accessibility, the Ethernet port on each controller module should be connected through different equipment to a common network for system management.

Example Business Configurations

The following configuration examples demonstrate cabling best practices for the specified connection. The standard default configuration for high-availability is shown in "Connection: High-Availability, Dual-Controller Through Two Switches to Two or More Dual-Port Data Hosts – Standard Default Configuration for High-Availability" on page 34.

For information about how failover works, depending on whether the topology is set to loop or point-to-point, see "Loop or Point to Point Topology?" on page 22.

Note – For clarity, the schematic illustrations of the controllers shown in this section show only relevant details such as host ports.

Connection: Single-Controller, Direct Attach, One Single-Port Data Host

- **Usage**: Benefits of 4-Gbit/sec speed and robust storage for video streaming/editing; individual server systems
- **Advantage**: Easy setup, low-cost
- **Example Business**: Small business with a single server system, video production

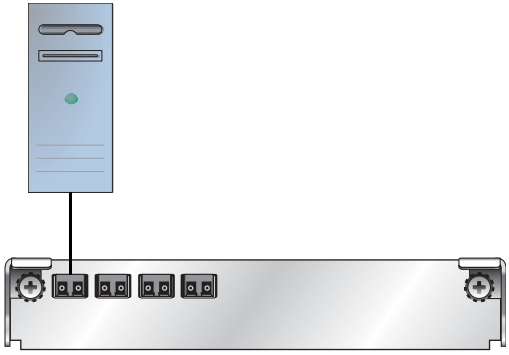


Figure 3-1 Single-Controller, Direct Attach Connection to One Single-Port Data Host

Connection: Single-Controller, Direct Attach, Two Single-Port Data Hosts

- **Usage**: Individual users have direct access to a file server; users can store all CAD work on robust storage
- **Advantage**: Easy setup, low-cost
- **Example Business**: Small business, architectural firm

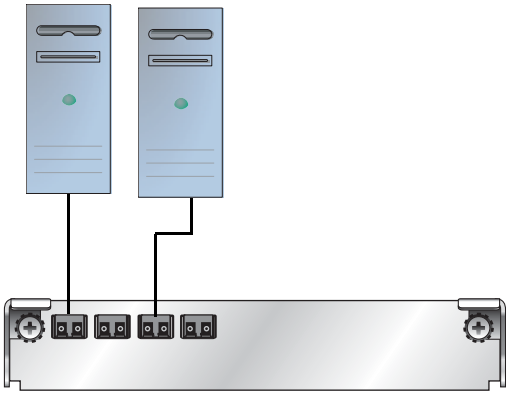


Figure 3-2 Single-Controller, Direct Attach Connection to Two Single-Port Data Hosts

Connection: High Availability, Dual-Controller, Direct Attach to One Dual-Port Data Host

In this configuration, the host has redundant connections to the volumes assigned to each controller. If a controller were to fail, the host maintains access to all the volumes through the host port on the surviving controller. This configuration requires that host port interconnects are enabled.

- **Usage**: File server with fault tolerance
- **Advantage**: Easy setup, low-cost, fault tolerance
- **Example Business**: Small to medium business, legal firm

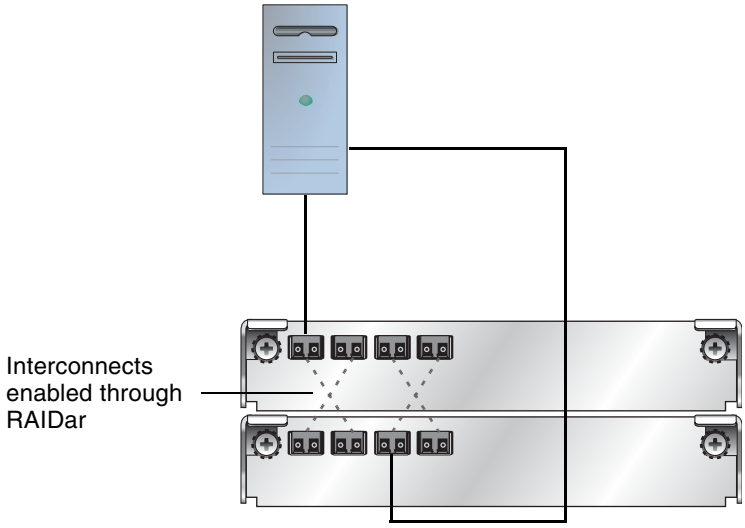


Figure 3-3 High-Availability, Dual-Controller, Direct Attach Connection to One Dual-Port Data Host

Connection: High-Availability, Dual-Controller, Direct Attach to Two Dual-Port Data Hosts

In this configuration, both hosts have redundant connections to volumes that are associated with each of the controllers. If a controller were to fail, the hosts maintain access to all of the volumes through the hosts ports on the surviving controller. This configuration requires that host port interconnects are enabled

- **Usage**: Clustered servers, direct connect to storage, classroom applications, classroom file server
- **Advantage**: Benefits of fault tolerance of dual-controller and clustered servers
- **Example Business**: Educational facility; small tech business

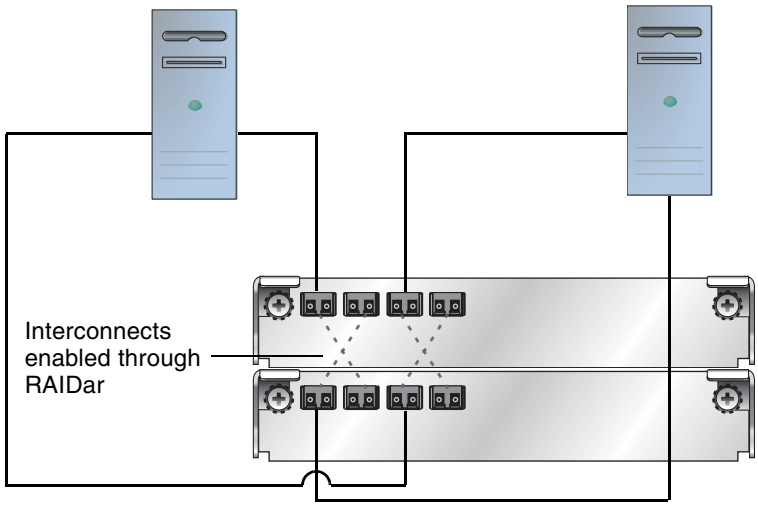


Figure 3-4 High-Availability, Dual-Controller, Direct Attach Connection to Two Dual-Port Data Hosts

Connection: High-Performance, Dual-Controller, Direct Attach to Two Dual-Port Data Hosts – Non-Fault Tolerant

This configuration shows a high-performance connection appropriate for high-bandwidth applications that do not require failover or path redundancy. Each host has two single-port FC HBAs or one dual-port FC HBA installed. For each host, one HBA port is connected to each controller. The host port interconnect is disabled. In this configuration, each host only has access only to the volumes assigned to the controller to which the HBA is connected; however, each host can use both HBA ports to transfer data to both controllers simultaneously at 2 or 4 Gbit/sec. If a controller fails, the host loses access to volumes owned by that controller.

- **Usage**: Benefits of 4-Gbit/sec speed and robust storage for video streaming/editing; individual user systems
- **Advantage**: Easy setup, low-cost, high performance
- **Example Business**: Video production

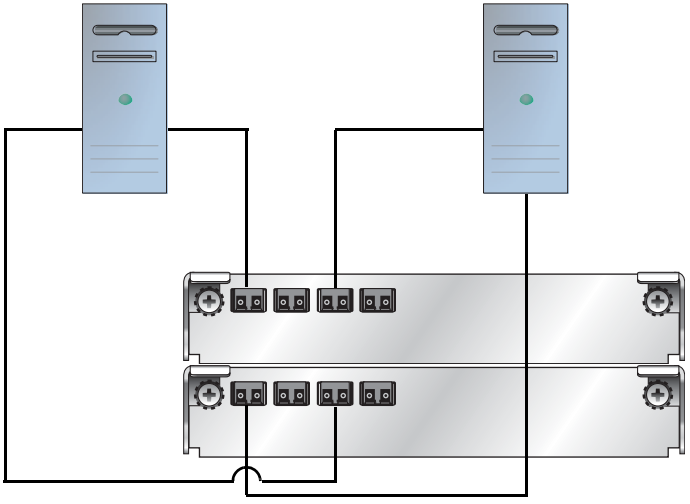


Figure 3-5 High-Performance, Dual-Controller, Direct Attach Connection to Two Dual-Port Data Hosts – Non-Fault Tolerant

Connection: High Availability, Dual-Controller, Through One Switch to One Dual-Port Data Host

This connection is for configurations in which there is a single switch. It requires that host port interconnects are disabled.

- **Usage**: Single switch
- **Advantage**: Low-cost; enables growth of more servers, which provides more applications; enables expansion of user account storage space for files and video
- **Example Business**: Internet Service Provider; Web commerce hosting company

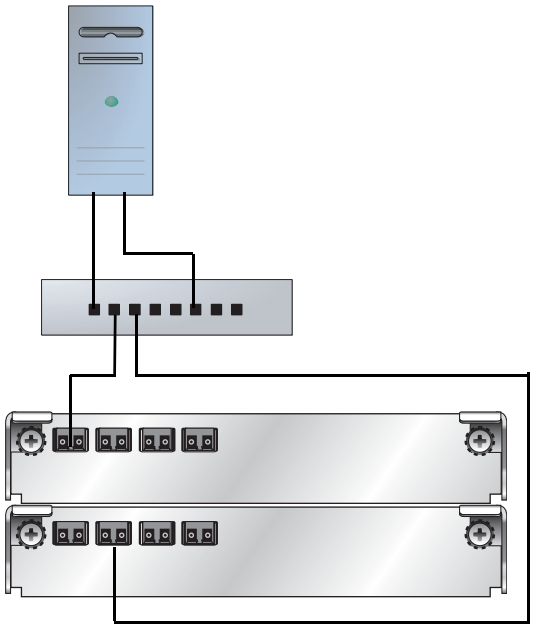


Figure 3-6 High-Availability, Dual-Controller Connection Through One Switch to One Dual-Port Data Host

Connection: High-Availability, Dual-Controller Through Two Switches to Two or More Dual-Port Data Hosts – Standard Default Configuration for High-Availability

This connection represents the standard default configuration for high-availability. The host port interconnects must be disabled.

- **Usage**: Cluster servers for fault tolerance of OS and server; dual switches for fault tolerance of switches; dual controllers for fault tolerance of storage (very high fault tolerance and also high speed)
- **Advantage**: Achieves higher uptime rate (99.999%) uptime) by eliminating many points of failure; enables more growth of additional cluster servers; enables more growth of data storage; offers additional security through switches
- **Example Business**: eCommerce company

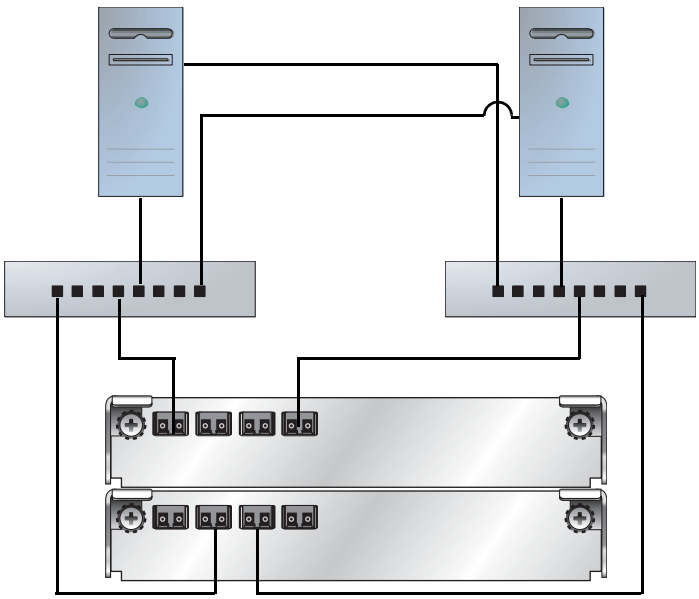


Figure 3-7 High-Availability, Dual-Controller Connection Through Two Switches to Two or More Dual-Port Data Hosts

Connection: High-Availability, Dual-Controller Through a Two-Zone Switch to Two or More Dual-Port Data Hosts

Each zone can be an independent FC switch. Note that the switch is a potential SPOF.

- **Usage**: Web server, multiple hosts handling web queues
- **Advantage**: Low cost; multiple web servers; switch access to allow multiple hosts
- **Example Business**: Mid-size Internet company (Web farm)

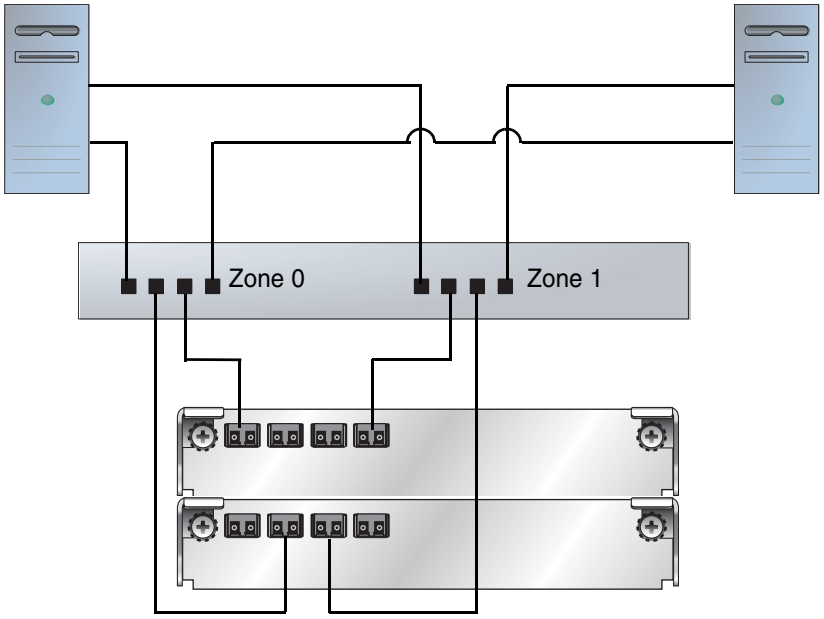


Figure 3-8 High-Availability, Dual-Controller Connection Through a Two-Zone Switch to Two or More Dual-Port Data Hosts

Connection: High-Availability, Dual-Controller Through a Four-Zone Switch to Two or More Dual-Port Data Hosts

Each zone can be an independent FC switch or two dual zoned FC switches. Multiple hosts can be attached to the switch zones. This configuration enables optimal connectivity for each dual port host because no FC single path to each controller is shared between the two hosts.

- **Usage**: High level of security; file server; web server; email server; SQL server
- **Advantage**: Switch access allow multiple hosts; switch adds extra security with zoning on storage ports, which allows volumes to be mapped to specific zones
- **Business**: Government (City Chamber of Commerce)

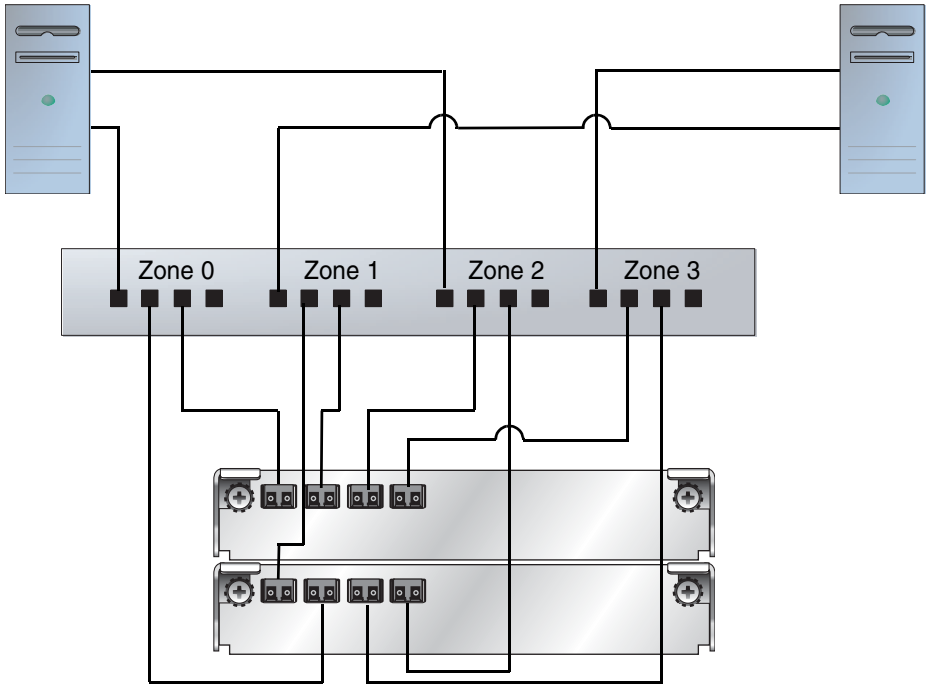


Figure 3-9 High-Availability, Dual-Controller Connection Through a Four-Zone Switch to Two or More Dual-Port Data Hosts

Configurations With Multiple Expansion Enclosures

You can connect a controller enclosure to up to two chains of up to four expansion enclosures each. Expansion enclosure balancing is recommended for maximizing performance. For every two expansion enclosures added, the connectivity should be split between each branch. The SAS expansion channel 0 and 1 are each referenced as an expansion chain.

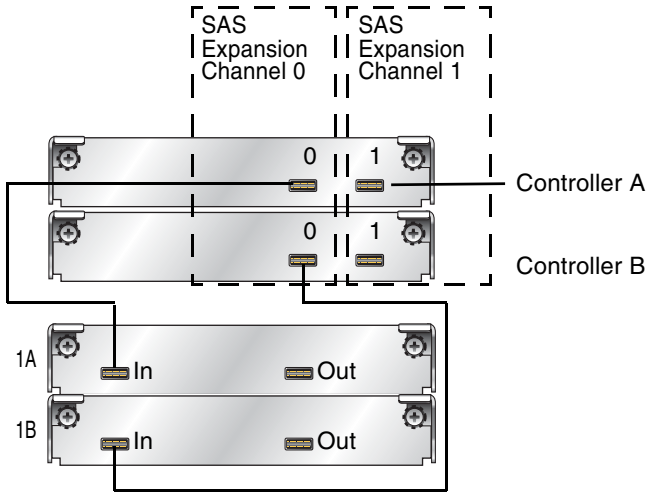


Figure 3-10 Cabling Connections Between Controller Enclosure and One Expansion Enclosure

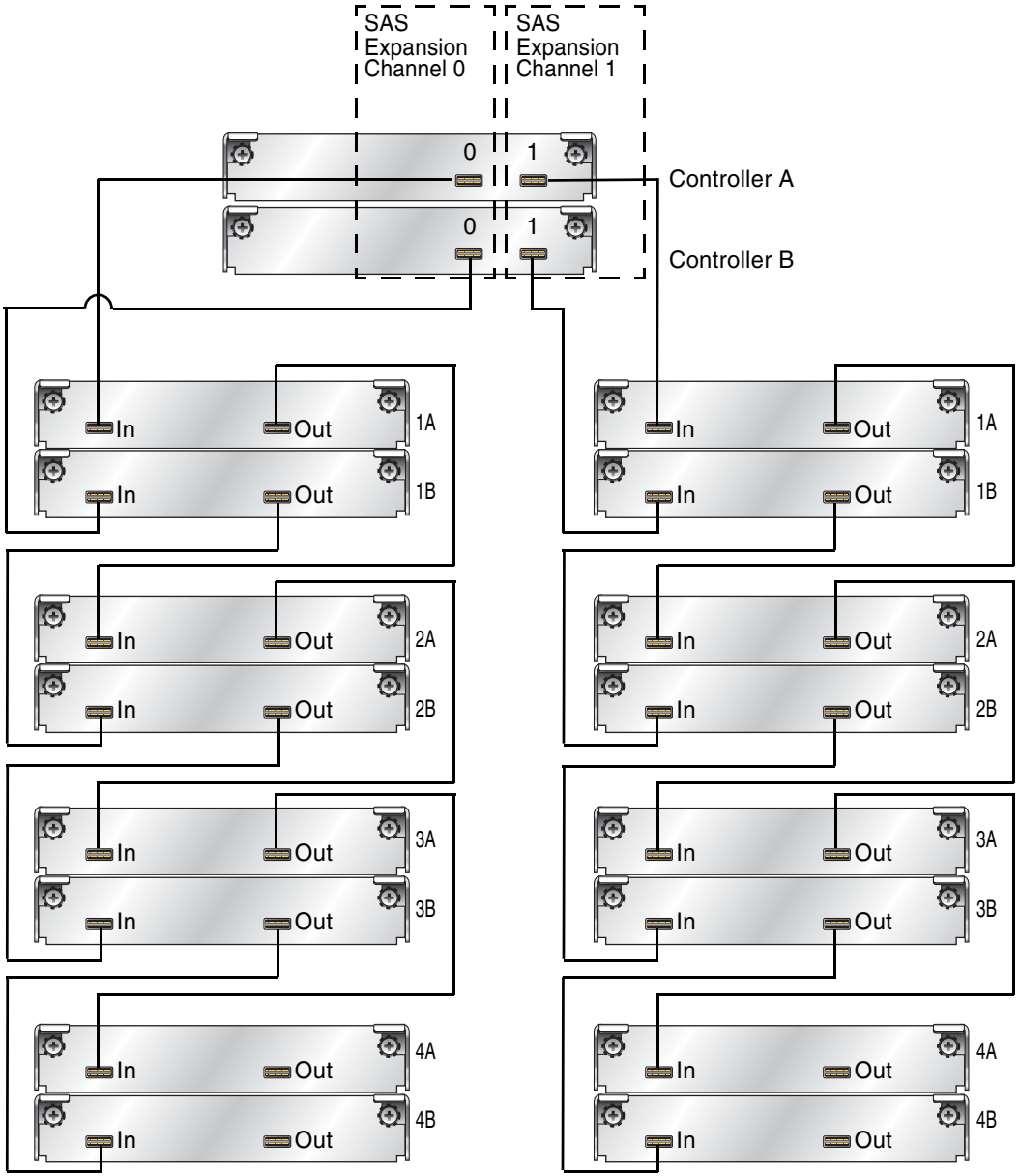


Figure 3-11 Cabling Connections Between Controller Enclosure and Eight Expansion Enclosures

CHAPTER 4

Storage Configuration

This chapter provides guidelines for using management software to administer and configure your storage system, including information on how to fine-tune and diagnose your system. Configuration recommendations are based on your application type. While RAIDar is the primary interface for storage management, you can also use a secondary interface, the command-line interface (CLI), to perform storage management.

For more detail, including the step-by-step procedures related to the topics contained in this chapter, refer to the *Administrator's Guide* and the *CLI Reference Manual*.

Topics covered in this chapter include:

General System Administration

This section introduces the management software and describes general storage configuration information that you should be aware of.

Management Software

RAIDar, the web-browser-interface, is the primary interface for configuring and managing the system. A web server resides in each controller module. RAIDar enables you to manage the system from a web browser that is properly configured and that can access a controller module through an Ethernet connection. RAIDar provides the ability to create, configure, and control volumes dynamically without interrupting operations. It provides users with complete control and reporting tools over every aspect of their storage system.

RAIDar supports the following browsers:

- Microsoft Internet Explorer 5.5 or later
- Mozilla Firefox 1.0.7 or later

The following guidelines ensure optimal performance:

- RAIDar uses pop-up windows to display various statistics and progress messages; pop-up windows must be enabled on your browser for proper operation.
- (Internet Explorer only) When web page caching is enabled in RAIDar (default), your browser must be set to *never* check for newer versions of stored pages.
- To optimize the display, a color monitor should be used and its color quality set to the highest setting.
- To ensure that animated status icons can be viewed, your browser should be set to play animations.

The CLI is a secondary interface that can be accessed through an Ethernet connection, or from a local host through a serial port connection, or through the serial port on either controller module. CLI scripts can be created using Perl, Expect, or similar tools.

Security

When you assign an IP addresses to each controller module to manage a system out-of-band, consider these security best practices:

- Keep the IP address on a private network rather than a publicly routable network.
- Use RAIDar to change service security settings to disable the ability to remotely connect to the system using network management services such as HTTP, Telnet, SMI-S, FTP, and SNMP, and in-band management services such as SES.

  The system's use of SSL ensures that communication between the system and hosts is secure. To secure user communication with RAIDar, enable the HTTPS service. To secure user communication with the CLI, enable the SSH service.
- Use RAIDar to configure user profiles to limit unauthorized access to the system, to system interfaces and services, and to system management features. Be sure to change the default passwords for the default users.

User Roles and Access Privileges

By default, the system provides three users that can access the system. In addition to these users, which you can modify, you can add 10 other users (13 maximum). The user configuration function enables you to define user roles by granting specific access privileges. For each user you can set a password and enable or disable access to the following system interfaces: WBI (RAIDar), CLI (command-line interface), and FTP.

Each user role is defined by an access level of either Monitor or Manage:

- Monitor – Enables access to functions on the Monitor menu
- Manage – Enables access to functions on the Monitor and Manage menus

Up to five Monitor users and only one Manage user can be logged in to each controller. RAIDar distinguishes users by their IP addresses. If you log in to RAIDar using multiple browser instances on the same management host, RAIDar considers all instances as a single user; actions you take in one instance are reflected in the other instances on the same host.

Each user is granted access privileges based on the following user types:

- Standard – Enables access to most functions.
- Advanced – In addition to enabling Standard functions, enables access to infrequently used administrative functions.
- Diagnostic – In addition to enabling Standard and Advanced functions, enables access to troubleshooting functions.

User configuration enables you to control which functions a user can access based on the user's role (assigned user type and access level).

Configuring the Management Interfaces

In addition to the first-time configuration tasks described in the *Getting Started Guide*, you should perform the following tasks to configure the management interfaces and your system:

- Set CLI and RAIDar preferences
- Configure RAIDar user access
- Set system information
- Set enclosure information
- Configure host port link speed, FC loop ID, and interconnect settings
- Configure LAN-related settings
- Configure the accessibility of management services
- Configure event notification
- Save the configuration

Event Notification

You can control how the system notifies you when specific events occur. You can enable or disable the following notification methods for selected event categories or individual events:

- Visual alerts, which can show a visual alert indicator or a popup window on the management host
- Email alerts, which sends an email message to designated users
- SNMP traps, which sends an SNMP trap to the designated trap host

The event categories are:

- Critical events, which might indicate system failure and require intervention. For example, a virtual disk is down.
- Warning events, which might require intervention although the system is still operating. For example, a virtual disk is critical.
- Informational events, which are expected to occur. For example, a virtual disk verification has completed.

Critical and warning events typically require some form of action whereas informational events are used to track specific behaviors when troubleshooting. Typically an administrator will want to be notified of all critical and warning events, but not informational events.

A Diagnostic manage user can select individual events to track or watch for a specific event, warning, or error; or to receive notification when a specific operation has started or completed, such as reconstruction or completion of initialization.

Statistics

Volume rate statistics, cumulative statistics, and individual volume statistics can help you interpret performance based on configuration of an individual element of your storage system, such as HBA, driver, SAN, or host operating system. The statistical information is useful to profile applications and their usage of a virtual disk, which can determine if additional virtual disks would increase performance and what RAID level fits your needs. You can analyze the performance of the same application using different RAID levels to determine which level gives you the best performance.

You can also view real-time volume statistics, disk error statistics, and disk usage statistics.

Real-time statistics pages show the performance of the system at 2-second intervals, as opposed to the rate and cumulative statistics pages which average performance numbers over a longer period.

Virtual Disks

A virtual disk (vdisk) is a group of disk drives configured with a RAID level. Each virtual disk can be configured with a different RAID level. A virtual disk can contain SATA drives or SAS drives, but not both (see “SAS or SATA Disk Drives?” on page 20 for more information). The controller safeguards against improperly combining SAS and SATA drives in a virtual disk. The system displays an error message if you choose drives that are not of the same type.

A system can have a maximum of 16 virtual disks per controller. For storage configurations with many drives it is recommended to create a few virtual disks each containing many drives, as opposed to many virtual disks each containing a few drives. Having many virtual disks is not very efficient in terms of drive usage when using RAID 3. For example, one 12-drive RAID-5 virtual disk has 1 parity drive and 11 data drives, whereas four 3-drive RAID-5 virtual disks each have 1 parity drive (4 total) and 2 data drives (only 8 total).

A virtual disk can be larger than 2 Tbyte. This can increase the usable storage capacity of configurations by reducing the total number of parity disks required when using parity-protected RAID levels. However, this differs from using volumes larger than 2 Tbyte, which requires specific operating system, HBA driver, and application program support.

Supporting large storage capacities requires advanced planning because it requires using large virtual disks with several volumes each or many virtual disks. To maximize capacity and drive usage (but not performance), you can create virtual disks larger than 2 Tbyte and divide them into multiple volumes with a capacity of 2 Tbyte or less.

The largest supported virtual disk configuration depends largely upon the cache optimization mode. The maximum capacity per virtual disk supported by the controller software is:

- 16 Tbyte with standard (random) optimization
- 64 Tbyte with super-sequential optimization

Tip – The best practice for creating virtual disks is to add them evenly across the both controllers. With at least one virtual disk assigned to each controller, both controllers are active. This active-active controller configuration allows maximum use of a dual-controller configuration's resources.

If you run out of free space in a virtual disk you can expand it by assigning additional disk drives. The virtual disk can continue to be used during the expansion; however, it can take days hours, perhaps days to complete, depending on the RAID level and size, drive speed, utility priority, and other processes running on the storage system. You can stop an expansion only by deleting a virtual disk.

Tip – Before starting an expansion, it is best practice to back up data so that if you need to delete a virtual disk, you can move the data into a new, larger virtual disk.

Chunk Size

When you create a virtual disk, you can use the default chunk size or one that better suits your application. The chunk (also referred to as stripe unit) size is the amount of contiguous data that is written to a virtual disk member before moving to the next member of the virtual disk. This size is fixed throughout the life of the virtual disk and cannot be changed. A stripe is a set of stripe units that are written to the same logical locations on each drive in the virtual disk. The size of the stripe is determined by the number of drives in the virtual disk. The stripe size can be increased by adding one or more drives to the virtual disk.

Available chunk sizes include:

- 16 Kbyte
- 32 Kbyte
- 64 Kbyte (default)

If the host is writing data in 16 Kbyte transfers, for example, then that size would be a good choice for random transfers because one host read would generate the read of exactly one drive in the volume. That means if the requests are random-like, then the requests would be spread evenly over all of the drives, which is good for performance.

If you have 16-Kbyte accesses from the host and a 64 Kbyte block size, then some of the hosts accesses would hit the same drive; each stripe unit contains four possible 16-Kbyte groups of data that the host might want to read, which is not an optimal solution.

Alternatively, if the host accesses were 128 Kbyte in size, then each host read would have to access two drives in the virtual disk. For random patterns, that ties up twice as many drives.

Tip – The best practice for setting the chuck size is to match the transfer block size of the application.

RAID Levels

An overview of supported RAID implementations is provided in the following table.

RAID Level	Cost	Performance	Protection Level
RAID 0 striping	N/A	Highest	No data protection
RAID 1 mirroring	High cost - 2X drives	High	Protects against individual drive failure
RAID 3 Block striping with dedicated parity drive	1 drive	Good	Protects against individual drive failure
RAID 5 Block striping with striped parity drive	1 drive	Good	Protects against any individual drive failure; medium level of fault tolerance
RAID 6 Block striping with multiple striped parity	2 drives	Good	Protects against multiple (2) drive failures; high level of fault tolerance
RAID 10 Mirrored striped array	High cost	High	Protects against certain multiple drive failures; high level of fault tolerance
RAID 50 Data striped across RAID 5	At least 2 drives	Good	Protects against certain multiple drive failures; high level of fault tolerance

---

Note – Non-RAID is supported for use when the data redundancy or performance benefits of RAID are not needed; no fault tolerance.

In general, the following RAID levels are recommended for the specific application (also see “Optimizing Performance for Your Application” on page 53):

- Databases – RAID 1
- Large file applications with few data requests – RAID 3
- Imaging applications with large file transfers with many requests – RAID 5

RAID 6

RAID technology has long been a standard in any storage environment with RAID 5 being the standard bearer. RAID 6 provides the required additional protection from physical drive failures at a very economical price point while maintaining performance expectations. However, the additional protection offered by RAID 6 does not mean much if the performance impact cripples the applications that are running on it. The following table shows the impact of RAID 6 versus RAID 5. When using this information, keep in mind that most applications will have an access pattern profile, for example, 80% reads, 20% writes. If those writes are evenly split between sequential and random, that means that 10% of the writes will suffer up to a 25% performance loss, but 90% of the application processing will only suffer up to a 5% performance impact.

	Normal Operation	**Single Disk Rebuild**	**Dual Disk Rebuild**
Sequential reads	Within 3-5% of RAID 5	Within 3-5% of RAID 5	Up to 25% drop compared to RAID 5 normal operations
Sequential writes	Within 3-5% of RAID 5	Within 3-5% of RAID 5	Up to 25% drop compared to RAID 5 normal operations
Random reads	Within 3-5% of RAID 5	Within 3-5% of RAID 5	Up to 25% drop compared to RAID 5 normal operations
Random writes	Up to 25% drop of RAID 5	Up to 25% drop of RAID 5	Up to 25% drop compared to RAID 5 normal operations

Note – RAID 5 cannot sustain a dual disk failure, so the performance impact of a dual disk rebuild under RAID 6 is minimal compared to the data loss that would occur under RAID 5.

Spares

When configuring virtual disks, you can add a maximum of four available drives to a redundant virtual disk (RAID 1, 3, 5, 6, and 50) for use as spares. If a drive in the virtual disk fails, the controller automatically uses the vdisk spare for reconstruction of the critical virtual disk to which it belongs. A spare drive must be the same type (SAS or SATA) as other drives in the virtual disk. You cannot add a spare that has insufficient capacity to replace the smallest drive in the virtual disk. If two drives fail in a RAID 6 virtual disk, two properly sized spare drives must be available before reconstruction can begin. For RAID 50 virtual disks, if more than one sub-disk becomes critical, reconstruction and use of vdisk spares occur in the order sub-vdisks are numbered.

You can designate a global spare to replace a failed drive in any virtual disk, or a vdisk spare to replace a failed drive in only a specific virtual disk. Alternatively, you can enable dynamic spares in RAIDar. Dynamic sparing enables the system to use any drive that is not part of a virtual disk to replace a failed drive in any virtual disk.

Tip – A best practice is to designate spare disk drive for use if a drive fails. Although using a vdisk spare is the most secure way to provide spares for your virtual disks, it is also expensive to keep a spare assigned to each virtual disk. An alternative method is to enable dynamic spares or to assign one or more unused drives as global spares.

Saving and Restoring Configuration Information

You can save your system configuration settings to a text file on the host system or anywhere accessible on your network. The configuration file contains all controller-dependent configuration information, including the following settings:

- FC host port
- Enclosure management
- Preferences and options
- Disk drive
- LAN
- Passwords
- Service security
- Event notification

The file does not include virtual disk, volume, or host-to-volume mapping information.

You can restore the configuration file to the same system to change its current configuration to the saved configuration, or to a different system to “clone” the first system's configuration. When you restore the file, you can specify using either:

- The system’s current IP and network settings
- New IP and network settings you enter
- The IP and network settings in the file

Tip – A best practice is to always save your configuration settings after you have made a change to the system and have confirmed that the change has the affect that you want. Always saving to a new file name, such as `saved_config`*`mmddyy`*`.config` enables you to easily identify and go back to a previous configuration.

Cache Configuration

Controller cache options can be set for individual volumes to improve a volume’s fault tolerance and I/O performance. This section describes the cache settings that can be set using RAIDar.

Read-Ahead Cache Settings

The read-ahead cache settings enable you to change the amount of data read in advance after two back-to-back reads are made. Read ahead is triggered by two back-to-back accesses to consecutive logical block address (LBA) ranges. Read ahead can be forward (that is, increasing LBAs) or reverse (that is, decreasing LBAs). Increasing the read-ahead cache size can greatly improve performance for multiple sequential read streams. However, increasing read-ahead size will likely decrease random read performance.

The default read-ahead size, which sets one chunk for the first access in a sequential read and one stripe for all subsequent accesses, works well for most users in most applications. The controllers treat volumes and mirrored virtual disks (RAID 1) internally as if they have a stripe size of 64 Kbyte, even though they are not striped.

Caution – The read-ahead cache settings should only be changed if you fully understand how your operating system, application, and HBA (FC) move data so that you can adjust the settings accordingly. You should be prepared to monitor system performance using the virtual disk statistics and adjust read-ahead size until you find the optimal size for your application.

The Read Ahead Size can be set to one of the following options: (See "Parameter Settings for Performance Optimization" on page 53 for application-specific recommendations for read ahead cache size.)

- **Default** – Sets one chunk for the first access in a sequential read and one stripe for all subsequent accesses. The size of the chunk is based on the block size used when you created the virtual disk (the default is 64 KB). Non-RAID and RAID 1 virtual disks are considered to have a stripe size of 64 KB.
- **Disabled** – Turns off read-ahead cache. This is useful if the host is triggering read ahead for what are random accesses. This can happen if the host breaks up the random I/O into two smaller reads, triggering read ahead. You can use the volume statistics read histogram to determine what size accesses the host is doing.
- **64, 128, 256, or 512 KB; 1, 2, 4, 8, 16, or 32 MB** – Sets the amount of data to read first, and the same amount is read for all read-ahead accesses.
- **Maximum** – Lets the controller dynamically calculate the maximum read-ahead cache size for the volume. For example, if a single volume exists, this setting enables the controller to use nearly half the memory for read-ahead cache.

Note – Only use Maximum when host-side performance is critical and disk drive latencies must be absorbed by cache. For example, for read-intensive applications, you will want data that is most often read in cache so that the response to the read request is very fast; otherwise, the controller has to locate which disks the data is on, move it up to cache, and then send it to the host.

Note – If there are more than two volumes, there is contention on the cache as to which volume's read data should be held and which has the priority; the volumes begin to constantly overwrite the other volume's data, which could result in taking a lot of the controller's processing power. Avoid using this setting if more than two volumes exist.

Cache Optimization can be set to one of the following options:

- **Standard** – Works well for typical applications where accesses are a combination of sequential and random. This method is the default.
- **Super-Sequential** – Slightly modifies the controller's standard read-ahead caching algorithm by enabling the controller to discard cache contents that have been accessed by the host, making more room for read-ahead data. This setting is not optimal if random accesses occur; use it only if your application is strictly sequential and requires extremely low latency.

Write-Back Cache Settings

Write back is a cache-writing strategy in which the controller receives the data to be written to disk, stores it in the memory buffer, and immediately sends the host operating system a signal that the write operation is complete, without waiting until the data is actually written to the disk drive. Write-back cache mirrors all of the data from one controller module cache to the other. Write-back cache improves the performance of write operations and the throughput of the controller.

When write-back cache is disabled, write through becomes the cache-writing strategy. Using write-through cache, the controller writes the data to the disk before signaling the host operating system that the process is complete. Write-through cache has lower throughput and write operation performance than write back, but it is the safer strategy, with minimum risk of data loss on power failure. However, write-through cache does not mirror the write data because the data is written to the disk before posting command completion and mirroring is not required. You can set conditions that cause the controller to switch from write-back caching to write-through caching as described in "Auto-Write-Through Trigger and Behavior Settings" on page 51.

In both caching strategies, active-active failover of the controllers is enabled.

You can enable and disable the write-back cache for each volume. By default, volume write-back cache is enabled. Because controller cache is backed by super-capacitor technology, if the system loses power, data is not lost. For most applications, this is the correct setting. But because back-end bandwidth is used to mirror cache and because this mirroring uses back-end bandwidth, if you are writing large chunks of sequential data (as would be done in video editing, telemetry acquisition, or data logging), write-through cache has much better performance. Therefore, you might want to experiment with disabling the write-back cache. You might see large performance gains (as much as 70 percent) if you are writing data under the following circumstances:

- Sequential writes

- Large I/Os in relation to the chunk size
- Deep queue depth

If you are doing any type of random access to this volume, leave the write-back cache enabled.

Caution – Write-back cache should only be disabled if you fully understand how your operating system, application, and HBA (FC) move data. You might hinder your storage system's performance if used incorrectly.

Auto-Write-Through Trigger and Behavior Settings

You can set the trigger conditions that cause the controller to change the cache policy from write-back to write-through. While in write-through mode, system performance might be decreased.

A default setting makes the system revert to write-back mode when the trigger condition clears. To ensure that this occurs and that the system doesn't operate in write-through mode longer than necessary, make sure you check the setting in RAIDar or the CLI.

You can specify actions for the system to take when write-through caching is triggered:

- Revert when Trigger Condition Clears – Switches back to write-back caching after the trigger condition is cleared. The default and best practice is Enabled.
- Notify Other Controller – In a dual-controller configuration, the partner controller is notified that the trigger condition is met. The default is Disabled.

Cache Mirroring Mode

In the default active-active mode, data for volumes configured to use write-back cache is automatically mirrored between the two controllers. Cache mirroring has a slight impact on performance but provides fault tolerance. You can disable cache mirroring, which permits independent cache operation for each controller; this is called *independent cache performance mode (ICPM)*.

The advantage of ICPM is that the two controllers can achieve very high write bandwidth and still use write-back caching. User data is still safely stored in nonvolatile RAM, with backup power provided by super-capacitors should a power failure occur. This feature is useful for high-performance applications that do not

require a fault-tolerant environment for operation; that is, where speed is more important than the possibility of data loss due to a drive fault prior to a write completion.

The disadvantage of ICPM is that if a controller fails, the other controller will not be able to fail over (that is, take over I/O processing for the failed controller). If a controller experiences a complete hardware failure, and needed to be replaced, then user data in its write-back cache is lost.

Data loss does not automatically occur if a controller experiences a software exception, or if a controller module is removed from the enclosure. If a controller should experience a software exception, the controller module goes offline; no data is lost, and it is written to disks when you restart the controller. However, if a controller is damaged in a nonrecoverable way then you might lose data in ICPM.

Caution – Data might be compromised if a RAID controller failure occurs after it has accepted write data, but before that data has reached the disk drives. ICPM should not be used in an environment that requires fault tolerance.

Cache Configuration Summary

The following guidelines list the general best practices to follow when configuring cache:

- For a fault-tolerant configuration, use the write-back cache policy, instead of the write-through cache policy.
- For applications that access both sequential and random data, use the standard optimization mode, which sets the cache block size to 32 Kbtye.

  For example, use this mode for transaction-based and database update applications that write small files in random order.
- For applications that access sequential data only and that require extremely low latency, use the super-sequential optimization mode, which sets the cache block size to 128 Kbtye.

  For example, use this mode for video playback and multimedia post-production video- and audio-editing applications that read and write large files in sequential order.

Parameter Settings for Performance Optimization

You can configure your storage system to optimize performance for your specific application by setting the parameters as shown in the following table. This section provides a basic starting point for fine-tuning your system, which should be done during performance baseline modeling.

Table 4-1 Optimizing Performance for Your Application

Application	RAID Level	Read Ahead Cache Size	Cache Optimization
Default	5 or 6	Default	Standard
HPC ((High-Performance Computing)	5 or 6	Maximum	Standard
MailSpooling	1	Default	Standard
NFS_Mirror	1	Default	Standard
Oracle_DSS	5 or 6	Maximum	Standard
Oracle_OLTP	5 or 6	Maximum	Standard
Oracle_OLTP_HA	1	Maximum	Standard
Random1	1	Default	Standard
Random5	5 or 6	Default	Standard
Sequential	5 or 6	Maximum	Super-Sequential
Sybase_DSS	5 or 6	Maximum	Standard
Sybase_OLTP	5 or 6	Maximum	Standard
Sybase_OLTP_HA	1	Maximum	Standard
Video Streaming	1 or 5 or 6	Maximum	Super-Sequential

Fastest Throughput Optimization

The following guidelines list the general best practices to follow when configuring your storage system for fastest throughput:

- Host interconnects should be disabled.
- Switches should be used if fault tolerance is also required.
- Host ports should be configured for 4 Gbit/sec.
- Virtual disks should be balanced between the two controllers.
- Disk drives should be balanced between the two controllers.
- Use SATA drives for email and streaming applications; otherwise use SAS.
- When creating virtual disks, assign them so that processing load is spread between the two controllers. Within each virtual disk, divide the drives between the two expansion branches to distribute the traffic.
- Host port activity should be balanced across ports. Ports 0 and 1 are in one group, and ports 2 and 3 are in another, and access should be balanced across the two groups.

Highest Fault Tolerance Optimization

The following guidelines list the general best practices to follow when configuring your storage system, for highest fault tolerance:

- If using a direct attach connection, host port interconnects must be enabled.
- If using a switch attach connection, host port interconnects are disabled and controllers are cross-connected to two physical switches.
- Multipath Input/Output (MPIO) should be used.

System Diagnostics

The serviceability features described in this section enable you to diagnose and troubleshoot your storage system. The diagnostic capabilities of the storage system enable you to troubleshoot down to and below the field replaceable unit (FRU) level.

- Core dumps (up to four) can be stored within the controller enclosure and accessed by knowledgeable personnel. These persist over power cycles and other outage events.
- Stack traces can be printed out or stored, showing from what subroutine a fault was generated. These stack traces also persist over power cycles and other outage events.

Different levels of diagnostics can be accessed depending on access privileges:

- **Dequarantine a virtual disk**

  The quarantining process prevents the controller enclosure from making a virtual disk critical and starting reconstruction when the missing drive is just slow to spin up, not properly seated in its slot, in an enclosure that is not powered up, or a member of an unknown virtual disk.

  The controller enclosure quarantines a virtual disk if it does not see all of the virtual disk's drives in these cases:

  - After restarting one or both controllers, typically after powering up the storage system
  - After inserting a disk drive that is part of a virtual disk from another controller/disk enclosure combination

  The virtual disk can be fully recovered if the missing disk drives can be restored.

- **Check PHY status**

  A Physical Layer Interface (PHY) is an interface in a device used to connect to other devices. Problem PHYs can cause a host or controller to continually rescan drives, which disrupts I/O or causes I/O errors. I/O errors can result in a failed drive, causing a virtual disk to become critical or causing complete loss of a virtual disk if more than one fails. You can troubleshoot PHY problems by isolating data path faults. When working with intermittent errors, you can reset PHY status so that you can observe error trend information.

- **Trust virtual disk** (disaster recovery only)

  The Trust Virtual Disk function attempts to bring a virtual disk back online after some disaster, such as a power outage or other problem that prevents enough disks in a virtual disk from being present when the controller module powers up or is rebooted. The virtual disk appears to be down or offline and the disks that

weren't ready or present when the controller module powers up or is rebooted are labeled "Leftover." The Trust Virtual Disk function brings a virtual disk back online by ignoring metadata that indicates the drives may not form a coherent virtual disk.

This function can force an offline virtual disk to be critical or fault-tolerant, or a critical virtual disk to be fault-tolerant. You might need to do this when:

- A drive was removed or was marked as failed in a virtual disk due to circumstances you have corrected (such as accidentally removing the wrong disk). In this case, one or more disks of a virtual disk can start up more slowly or might have been powered on after the rest of the disks in the virtual disk. This causes the date and time stamps to differ, which the storage system interprets as a problem.
- A virtual disk is offline because a drive is failing, you have no data backup, and you want to try to recover the data from the virtual disk. In this case, the Trust Virtual Disk function might work, but only as long as the failing drive continues to operate.

- **Save log information to a file**

  You can save the following types of log information to a file:

  - Device status summary, which includes basic status and configuration information.
  - Event logs from both controllers when in active-active mode.
  - Debug logs from both controllers when in active-active mode.
  - Boot logs, which show the startup sequence for each controller.
  - Up to four critical error dumps from each controller. These will exist only if critical errors have occurred.
  - Management controller traces, which trace interface activity between the controllers' internal processors and activity on the management processor.

Event and Debug Logs

Event logs capture reported events from components throughout the storage system. Each event consists of an event code, the date and time the event occurred, which controller reported the event, and a description of what occurred. You can use RAIDar or the CLI to view the event logs.

When instructed to do so by service personnel, Advanced users can set up and Diagnostic users can view the debug log that includes additional troubleshooting information. The debug log is used to capture information that will help Technical Support locate problems.

Additional Debug Utilities

In addition to the debug log, Diagnostic users have the following debug utilities available:

- **View error buffers** – Shows crash dump and boot information saved by the management controller. During normal operation, the management controller communicates with the storage controller. If there are problems with this communication, there is little information available to the LAN subsystem to show. In this case and under certain failure conditions, crash and boot buffer data can be examined. For normal operation, these buffers are empty.
- **View CAPI trace** – Shows the Configuration API (CAPI) commands sent and received by the management controller. For example, when creating a virtual disk, the request to create the virtual disk is shown and the reason why it failed. This panel provides detail for the underlying action that supports the failed function.
- **View management trace** – Shows a debug trace for the management controller. It traces interface activity between the controllers' internal processors and activity on the management processor.

CHAPTER 5

Data Protection Services

The data protection services offered by the R/Evolution architecture enable you to solve many data protection challenges by building complete solutions for business continuance, disaster recovery, and regulatory compliance.

When trying to determine the appropriate data protection service to deploy, consider the following key factors:

- **Recovery Time Objective (RTO)** – How long can you afford to be without the data; in what time frame must the data become available before the financial viability of your company is put to risk?
- **Recovery Point Objectives (RPO)** – How much data can you afford to lose before the financial viability of your company is put to risk? This will vary for the critical versus non-critical data.
- **Data Classification** – Some data is more critical to the operational and financial health of your company than other data, so no one service is the solution for all data types. Classifying data (for example, mission critical, critical, non-critical), and determining RTO and RPO requirements for each class of data will help in the selection of the appropriate data protection service. Multiple services are often selected and deployed across an enterprise.

Each of these factors has specific tradeoffs that must be resolved to implement the appropriate data management service. The following table summarizes the available technologies, related recovery objectives, and the type of protection each technology provides.

Table 5-1 Data Protection Service Summary

Feature	Function	Data Loss (RPO)	Outage (RTO)	Protection
RAID	Method for storing data across a number of disk drives to achieve data redundancy	None	None	Disk drive failure
Snapshots	Point-in-time logical image of a physical volume	Hours	Minutes	Logical data deletion or corruption up to the most recent snapshot
Volume Copy	Complete physical and independent copy of a volume	Hours to days	Minutes	Logical data loss or corruption
Backup	Traditional tape-based	Days-weeks	Hours-days	Local or wide area disaster, full system recovery
Archive	Traditional tape-based		Days-weeks	Long-term data retention and preservation

Tip – It's a best practice to give careful consideration before selecting a data protection service solution that fits your needs and environment.

It is also important to understand the implementation of the data protection technology to be able to build effective data protection solutions and identify which functions are most critical for your organization. The following table describes the implementation of the data protection service.

Table 5-2 Data Protection Services Implementation

	Snapshot (copy on write)	Volume Copy
Snapshot requires original copy of data	Yes; the unchanged data is accessed from the original copy	No; once the copy is complete, the image is a full allocation copy of the data. It is dependent on the source data only while the copy is in process.
Space-efficient	Yes; space is required only for changed data	No; the new copy is a full allocation copy of the data
I/O and CPU performance overhead on the system with original copy of the data	None; hardware-based so impact to host system is negligible to minimal	None; hardware-based so impact to host system is negligible to minimal
Write overhead on the original copy of the data	High; first write to data block results in additional write	No direct impact; depending on what is copied and how, might experience some of the same impact as snapshots
Protection against logical data errors	Yes; changes can be rolled back to the original copy	Not directly; copy is of a snapshot and so has whatever data existed in the snapshot.
Protection against physical media failures of the original data	None; original copy must exist	Yes; once the copy is complete, it is a fully allocated complete copy of the source data. There are no ties to the original data.

Snapshot Service

The system's licensed snapshot service enables you to keep near-line (tier 2) backups of data for data management and protection purposes. Each snapshot preserves the volume's data state at the point in time when the snapshot was created. Snapshot technology is fast, efficient, and provides the best method for protecting your vital business data in real time.

- Significantly reduces backup windows by creating fast, frequent point-in-time copies of data
- Provides instant access to backup data and recovery for files, folders, or volumes
- Provides support for regulatory requirements, business continuance, and rapid application development

A snapshot is a virtual volume. While really a set of pointers to data that is located on a different volume (a master volume and/or a snap pool), a snapshot behaves like a volume in that it can be mapped to data hosts and the mapping can be assigned a LUN and be made accessible as read-only or read/write, depending on the intended purpose of the snapshot.

Snapshots can be taken of master volumes only. A master volume is a standard volume that has been enabled for snapshots. You can either create a master volume directly or convert a standard volume to a master volume.

The snapshot service uses the single copy-on-write function to capture only data that has changed. That is, if a block is to be overwritten on the master volume, and a snapshot depends on the existing data in the block being overwritten, the data is copied from the master volume to the snap pool before the data is changed. All snapshots dependent on that older data are able to access it from the same location on the snap pool, which reduces the impact of snapshots on master volume writes. In addition, only a single copy-on-write operation is performed on the master volume.

Snap Pools

Master volumes are associated with a snap pool, which contains pre-allocated reserve space for the snapshot data. A snap pool is a hidden volume and is never exposed to data hosts. A snap pool and associated master volumes must be owned by the same controller. However, the snap pool and its master volumes do not need to be in the same virtual disk.

Because the snap pool and its master volumes do not need to be on the same virtual disk, depending on your application, the snap pool can use secondary storage, including drives of a different type, a different speed or configured in a different manner than the master volumes. For example, a master volume can be configured as a RAID-5 virtual disk of SAS disks, and the snap pool can be placed on a RAID-0 virtual disk of SATA disks. Alternatively, a master volume can be on newer, high-speed disks and the snap pool on older, slower disks.

In another example is if the snapshots are not considered critical to the overall business, then the snap pool can be put on secondary storage that is not fault-tolerant (for example, RAID 0). In this case, you are willing to risk losing the snapshot data in exchange for better space utilization on the disk.

However, if snapshots are considered highly critical, then the snap pool should be created using a fault tolerant configuration. Whether this is on the same virtual disk as the master volume or on a separate virtual disk is up to you. Having both on the same virtual disk might affect performance.

A case where you will likely want to create the snap pool on a different virtual disk is when you are converting an existing standard volume into a master volume and no space exists on the virtual disk to create the snap pool.

Snap Pool Size

The minimum size of a snap pool is 1 Gbyte. To help you accurately set a snap pool's size, consider the following:

- **Snap-pool reserve space.** A snap pool requires 1,500 Mbyte of reserve space for internal use.
- **What is the master volume size, and how much will master volume data change between snapshots?** The size of the master volume and the average amount of data change should be factored into snap pool sizing. Each snapshot requires space in the snap pool. The amount of space needed depends on the interval between snapshots and the number of updates to the master volume. If the interval between snapshots is long, the amount of data written to the snap pool will likely be greater. If the interval between snapshots is short, the likelihood of a large number of changes to the data is less.
- **How many snapshots is the system going to retain?** Determine the number of snapshots that will be retained for each master volume snapped; for example, one for each day of the week. Once the retention limit is reached, older snapshots will be deleted and replaced with the most current snapshot. This number is dependent upon the configuration limits of your system.

- **How many snapshots will be modified?** Snapshots can be mounted as read-only or read-write. Determine if the snapshots for a given volume will be mounted as. read-write and actually written to; also determine the number of retained snapshots that will be affected.
- **How much modified (write) data will the snapshots have?** Of the snapshots that will be mounted as read-write and actually written to, factor in the average amount of data that will be modified.
- **How much extra capacity should the snap pool have as a safety margin?** In case actual capacity use exceeds the estimate, add a 25% margin.

For example, assume the following values are estimated to calculate the snap pool size for a 10-Gbyte volume:

- Snap-pool reserve space = 1,500 Mbyte
- Volume size = 10,000 Mbyte
- Average percent of change = 0.05%
- Number of snapshots retained = 4
- Number of modified snapshots = 4
- Average write data = 1,000 Mbyte
- Snap pool size = 6,750 Mbyte
- Safety margin = 0.25%
- Snap pool size = 8,438 Mbyte

The sizing formula is:

```
(reserve-space + (volume-size x avg-change x snapshots-retained) +
(snapshots-modified x avg-write-data)) x (1 + margin) = snap-pool-size
```

If you substitute the values from the above example into the snap pool sizing formula, the snap pool size is as follows:

1500 + (10,000 x 0.05 x 4) + (4 x 1,000) = 6,750 x 1.25 = 9,375 Mbyte750 + (10,000 x 0.05 x 4) + (4 x 1,000) = 6,750 x 1.25 = 8,438 Mbyte

Thresholds and Policies

Each snap-pool has three threshold levels — warning, error, and critical — representing decreasing free space in the snap pool. Each threshold has an associated policy that specifies what the system will do when that threshold is reached. You can set the warning and error threshold values and the error and critical policies. The warning threshold must be less than the error threshold, which must be less than 99%. RAIDar help and the *Administrator's Guide* describe policies that can be enacted when a threshold is reached.)

Snapshots can be manually deleted when they are no longer needed or automatically deleted via a snap pool policy. When a snapshot is deleted all data uniquely associated with that snapshot is deleted and associated space in the snap pool is freed for use.

By default, when the error threshold is reached, the system automatically deletes the oldest snapshot to prevent the critical threshold from being reached.

Snapshot Rollback

Snapshot functionality supports a rollback feature. This feature rolls the master volume back to the state represented by a specified snapshot; that is, it replaces the master volume data with the snapshot data. Additionally, you can choose to have the rollback include data modified in the snapshot since it was created.

With most implementations of snapshot, the rollback must be completed before the data becomes available to the application. With the R/Evolution storage system's implementation of snapshot, the data becomes immediately available to the application while the actual rollback completes in the background. This unique capability further improves application data availability.

Before performing a rollback of the modified data, the master volume must be unmounted from the host operating system because the host maintains information about the volume and must flush its cache of all data pertaining to the master volume. After you initiate the rollback, the host file system can view and access the reverted master volume without any stale data causing inadvertent overwrites and data corruption. In addition, when rolling back to the modified data for a snapshot, the snapshot must also be unmounted from the host. In the case of the snapshot, the volume cannot be remounted or accessed until the rollback is 100% complete. Accessing a snapshot prior to rollback completion results in master volume data corruption.

A best practice is to take a snapshot of the master volume before initiating a rollback. This preserves the data as it existed on the master volume prior to the rollback. This snapshot provides the administrator a backup of the master volume and preserves data that would otherwise be lost. When the rollback is initiated, all snapshots (both older and newer) are kept.

The system's rollback functionality provides both a rollback and a roll-forward capability. That is, if a master volume is rolled back to an earlier point in time, since all snapshots (earlier and later) are still preserved, the master volume can be subsequently rolled back to an even earlier snapshot or rolled forward to a later snapshot.

The rollback operation is performed as a background operation. While the rollback is in process, the master volume can continue to process read and write operations, and additional snapshots can be taken of it. You can delete snapshots at any time, including when the associated snap pool is reaching capacity and you want to free some space, the maximum number of snapshots is reached and you want to delete older snapshots, or you no longer need the data associated with the snapshot. If another rollback operation is initiated, it is queued until the first rollback has completed.

The following illustration shows the difference between rolling back the master volume to the data that existed when a specified snapshot was created (preserved), and rolling back only the modified data.

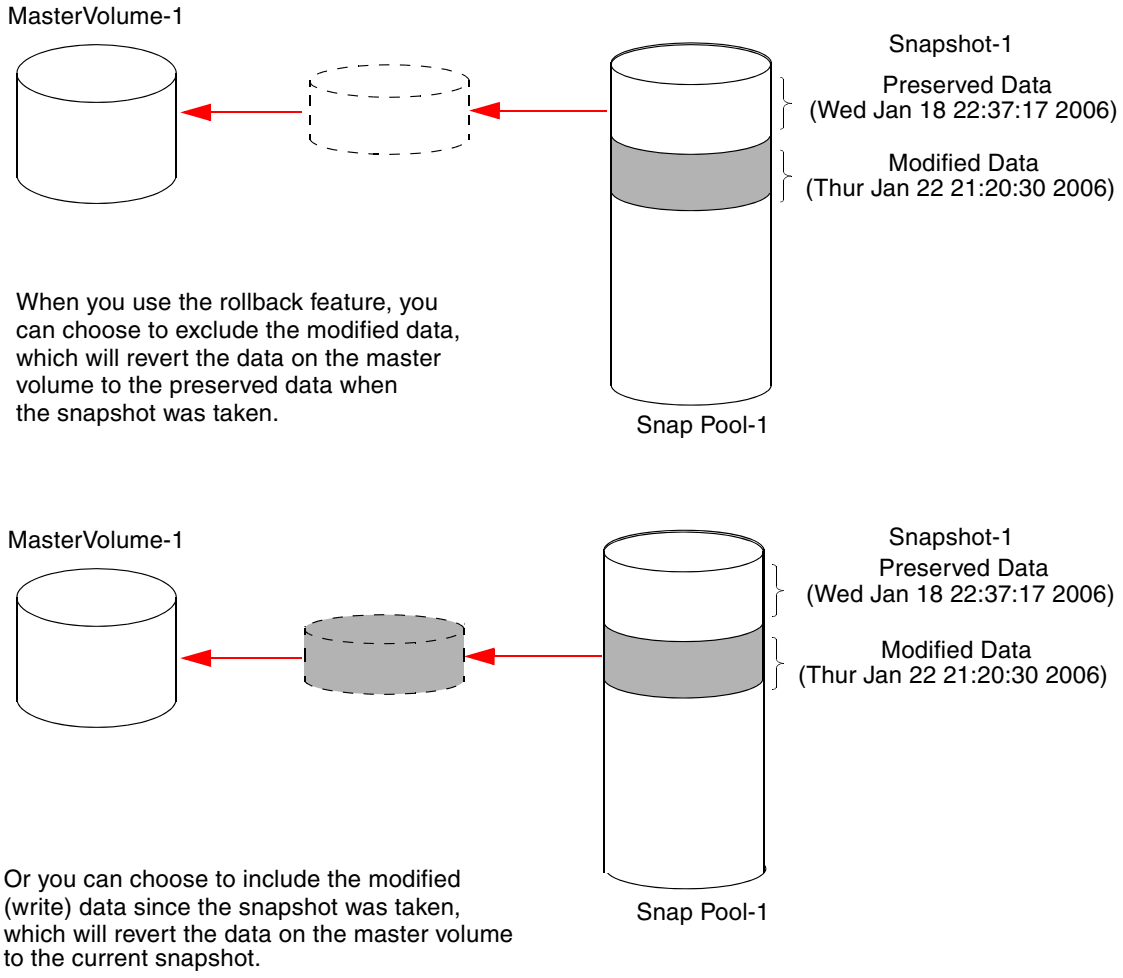


Figure 5-1 Rolling Back the Master Volume

Snapshot Reset

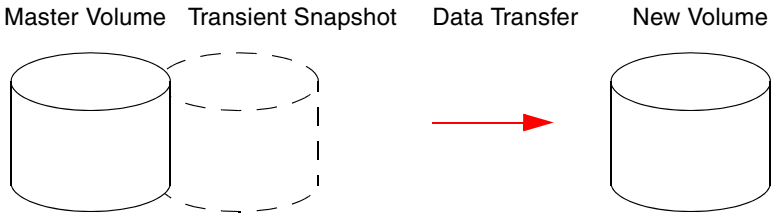
You can reset a snapshot to replace its content with the current data state of the associated master volume. The selected snapshot is replaced with a current snapshot having the same characteristics, such as name and LUN. The snapshot data is stored in the snap pool associated with the master volume. Before being reset, a snapshot must be unmounted from hosts.

Volume Copy Service

While a snapshot is a point-in-time logical copy of a volume, the licensed volume copy service creates a complete, physical and independent copy of a volume within a storage system. It is an exact copy of a master or a snapshot volume as it existed at the time the action was initiated, consumes the same amount of space as the source volume, and is independent from an I/O perspective. Volume independence is a key distinction of a volume copy (versus snapshot, which is a logical copy and dependent on the source volume). Benefits include:

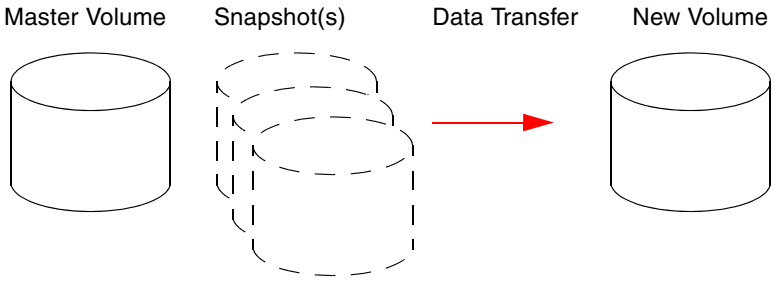
- **Additional Data Protection** – An independent copy of a volume (versus logical copy through snapshot) provides additional data protection against a complete master volume failure. If the source master volume fails, the volume copy can be used to restore the volume to the point in time the volume copy was taken.
- **Non-disruptive Use of Production Data** – With an independent copy of the volume, resource contention and the potential performance impact on production volumes is mitigated. Data blocks between the source and the copied volumes are independent (versus shared with snapshot) so that I/O is to each set of blocks respectively; application I/O transactions are not competing with each other when accessing the same data blocks.

Volume Copy from a Master Volume



1. Volume copy request is made with a master volume as the source.
2. A new volume is created for the volume copy, and a hidden, transient snapshot is created.
3. Data is transferred from the transient snapshot to the new volume.
4. On completion, the transient volume is deleted and the new volume is a completely independent copy of the master volume, representing the data that was present when the volume copy was started.

Volume Copy from a Snapshot



1. A master volume exists with one or more snapshots associated with it. Snapshots can be in their original state or they can be modified.
2. You can select any snapshot to copy, and you can specify that the modified or unmodified data be copied.
3. On completion, the new volume is a completely independent copy of the snapshot. The snapshot still remains, though you can choose to delete it.

Figure 5-2 Volume Copy From a Master Volume and a Snapshot

Some guidelines to keep in mind when performing a volume copy include:

- The virtual disk selected for the volume copy must be on the same controller.
- The virtual disk selected for the volume copy must have free space that is at least as large as the mount of space allocated to the original volume.

  A new volume will be created using this free space for the volume copy.
- The virtual disk for the volume copy does not need to have the same attributes (such as drive type, RAID level) as the volume being copied.
- Once the copy is complete, the new volume will no longer have any ties to the original.