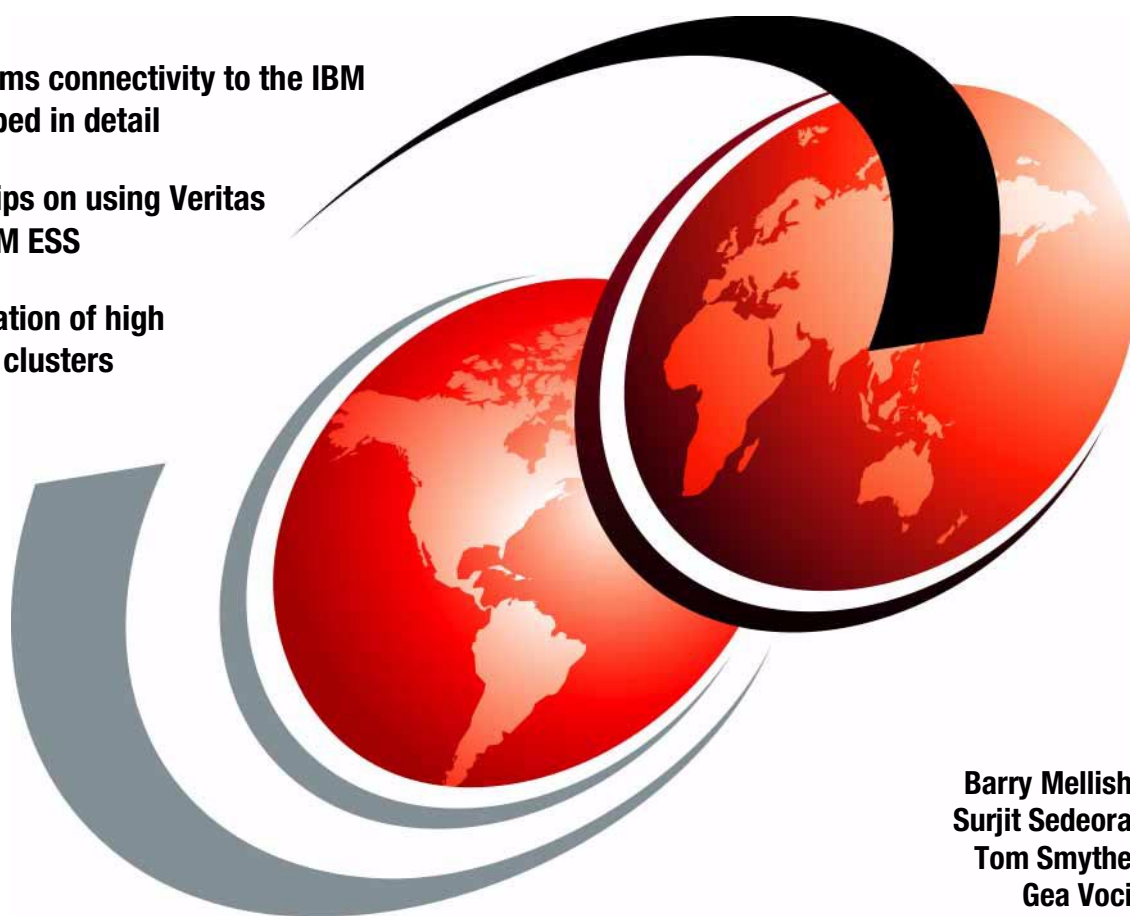


ESS Solutions for Open Systems Storage: Compaq AlphaServer, HP, and SUN

Open Systems connectivity to the IBM
ESS described in detail

Hints and tips on using Veritas
with the IBM ESS

Implementation of high
availability clusters



Barry Mellish
Surjit Sedeora
Tom Smythe
Gea Voci

ibm.com/redbooks

Redbooks



International Technical Support Organization

**ESS Solutions for Open Systems Storage:
Compaq AlphaServer, HP, and SUN**

March 2001

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix B, "Special notices" on page 119.

First Edition (March 2001)

This edition applies to the IBM Enterprise Storage Server E and F models and the microcode that was current at 12/15/00.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

© Copyright International Business Machines Corporation 2001. All rights reserved.

Note to U.S Government Users – Documentation related to restricted rights – Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	vii
Tablesix
Prefacexi
The team that wrote this redbookxi
Comments welcomexiii
Chapter 1. Enterprise Storage Server Overview	1
1.1 Configuring the ESS	3
1.1.1 ESS Specialist	4
1.1.2 ESS — Storage Allocation	5
1.1.3 ESS — Open System Storage	7
1.1.4 ESS — Modify Host Systems	9
1.1.5 ESS — Configure Host Adapter Ports	12
1.1.6 ESS — Add Volumes	17
1.1.7 ESS — Modify Volume Assignments	20
Chapter 2. Introduction to ESS connectivity	25
2.1 Fibre Channel connectivity to the ESS	25
2.2 FC-SW and the switch	26
2.2.1 Terms	26
2.2.2 Zoning benefits	26
2.2.3 Zoning example	27
2.2.4 Switch configuration	28
2.3 Host multi-pathing to the ESS	30
2.4 Volume management	33
2.4.1 Volume manager terms	33
2.4.2 Host volume layout and ESS pre-fetch buffers	38
Chapter 3. Compaq V4.0F	41
3.1 ESS configuration	41
3.2 Compaq configuration	42
3.2.1 Compaq systems tested	42
3.2.2 Compaq Host Optical Fibre Adapter	42
3.2.3 IBM SAN Fibre Channel Switch	42
3.2.4 Software versions	42
3.2.5 Additional software	42
3.3 How to check the Compaq configuration	43
3.3.1 System and cards firmware revision	43
3.3.2 Operating system version	43

3.3.3	Cluster software version	43
3.3.4	Patches installed on the system	44
3.3.5	Disk configuration	44
3.4	Tru64 cluster configuration	47
3.4.1	Example — disk service creation	47
3.4.2	Example — Apache disk service creation	51
3.5	LSM and ADVFS sample configuration	58
3.5.1	LSM disk initialization and disk group creation	58
3.5.2	LSM volume creation	60
3.5.3	LSM mirrored volume creation	61
3.5.4	Advanced File System creation and mounting	62
3.6	Compaq V4.0F — fibre connection to ESS configuration restrictions	62
3.6.1	System supports only eight LUNs	62
3.6.2	Disks are not seen at the console prompt	63
3.6.3	No boot or swap device supported	63
3.6.4	LSM and Advfs restrictions	63
3.7	Tru64 UNIX log files	63
3.7.1	References to Compaq documentation	63
 Chapter 4. Compaq V5.0A		65
4.1	ESS configuration	66
4.2	Compaq configuration	66
4.2.1	Compaq systems tested	66
4.2.2	Compaq Host Optical Fibre Adapter	66
4.2.3	IBM SAN Fibre Channel Switch	66
4.2.4	Compaq storage	66
4.2.5	Software versions	67
4.3	How to check the Compaq configuration	67
4.3.1	System and cards firmware revision	67
4.3.2	Operating system version	67
4.3.3	Cluster software version	67
4.3.4	Disk configuration	67
4.4	ADVFS sample configuration	69
4.5	Compaq V5.0A — fiber connection to ESS configuration restrictions	70
4.5.1	No boot from ESS volumes	70
4.5.2	On GS systems cannot find console commands	70
4.5.3	All ESS volumes seen with same ID/LUN from all cluster nodes	71
4.5.4	File command gives the LUN number in decimal	71
4.5.5	Termination of fibre cards	71
4.6	Tru64 UNIX log files	71
4.6.1	References to Compaq documentation	72

Chapter 5. IBM ESS and HP Servers	73
5.1 Pre-installation planning	73
5.2 Hardware connectivity: Fibre Channel - Arbitrated Loop (FC-AL)	74
5.2.1 Failover/failback test.	75
5.3 Hardware connectivity: simple switched fabric	76
5.4 High Availability tests	77
5.4.1 Test results.	79
5.4.2 Tuning recommendations	80
5.4.3 Supported servers and software	81
Chapter 6. IBM ESS and Sun Enterprise Servers	83
6.1 Topics covered in this chapter	84
6.2 Features supported in this release	84
6.3 Supported components	84
6.4 Caveats and limitations	85
6.5 Required modifications	86
6.5.1 All HBAs	86
6.5.2 Emulex	87
6.5.3 JNI	88
6.6 Boot messages	90
6.6.1 Emulex	90
6.6.2 JNI	92
6.7 Sun Veritas Volume Manager	95
6.7.1 Creating a filesystem under Veritas	95
6.7.2 Sun Veritas and ESS logical volumes	109
6.7.3 ESS identification under Veritas	110
6.8 Known Issues	111
Appendix A. Test suite details	113
A.1 I/O workload integrity	113
A.2 ESS exception tests	113
A.2.1 Test E.1: ESS warm start	113
A.2.2 Test E.2: ESS failover/failback	114
A.2.3 Test E.3: ESS cluster quiesce/resume	115
A.2.4 TEST E.4: Cable pulls	116
A.3 Host clustering function.	116
A.3.1 Test C1: Manual application package switch-over	116
A.3.2 Test C2: Application process failure test	116
A.3.3 Test C3: Test of each primary Heartbeat LAN	117
Appendix B. Special notices	119
Appendix C. Related publications	123
C.1 IBM Redbooks	123

C.2 IBM Redbooks collections	123
C.3 Other resources	123
C.4 Referenced Web sites	124
How to get IBM Redbooks	125
IBM Redbooks fax order form	126
Glossary	127
Index	139
IBM Redbooks review	141

Figures

1. Storage Area Network (SAN)	1
2. ESS basic overview	2
3. ESS logical configuration	3
4. ESS Specialist	5
5. Graphical Storage Allocation	6
6. Open System Storage	7
7. Open System Storage with assigned volumes	8
8. Modify Host Systems	9
9. Modify Host Systems — Host Type	10
10. Modify Host Systems — WWPN	11
11. Modify Host Systems — Add	12
12. Configure Host Adapter Ports	13
13. Configure Host Adapter Ports — Fiber Channel Access Mode	14
14. Configure Host Adapter Ports — SCSI	15
15. Configure Host Adapter Ports — SCSI Bus Configuration	16
16. Add Volumes (1 of 2) with host selected	17
17. Add Volumes (1 of 2) with host and DA selected.	18
18. Add Volumes (2 of 2)	19
19. Modify Volume Assignments	21
20. Modify Volume Assignments — Assign volume(s)	22
21. Modify Volume Assignments — duplicate target(s) and/or LUN(s)	23
22. Simple zone	27
23. Multiple zones	28
24. Switch configuration commands	29
25. Brocade/IBM switch GUI	30
26. Single path host	31
27. Multi-path host	32
28. Disk platter with partitions	33
29. ESS rank and disks	34
30. Logical volume example	35
31. Subdisks created from logical volume	36
32. Subdisks within plexes.	37
33. Plexes within volumes	38
34. ESS and Compaq AlphaServer Cluster	41
35. ESS and Compaq cluster.	65
36. Basic test system, ESS configured in FC-AL mode.	74
37. Simple switched configuration	76
38. High Availability configuration that was tested.	78
39. - Storage Area Network	83
40. Volume Manager Storage Administrator	96

41. VMSA controller view	97
42. VMSA disk properties	98
43. VMSA right-click on rootdg to get Disk Group menu	99
44. VMSA New Volume view	100
45. VMSA Assign Disks view	101
46. VMSA New Volume with disk information	102
47. VMSA Add File System view	104
48. VMSA Mount Details view	105
49. VMSA New Volume with disk and filesystem info	106
50. VMSA filesystem/volume creation in progress.	107
51. VMSA filesystem/volume creation complete	108
52. View of /etc/vfstab entries	109
53. VMSA Enclosure view	110

Tables

1. Host view of devices presented	24
2. Log files and the utilities that use them	63
3. Log files and the utilities that use them	71
4. JNI configuration file names.	88

x ESS Solutions for Open Systems Storage

Preface

This IBM Redbook is designed to help you install, tailor, and configure the IBM Enterprise Storage Server (ESS) when attaching Compaq AlphaServer running Tru64 UNIX, HP and Sun hosts. We describe the results of a 5-week project that took place in San Jose in October and November 2000. This book does not cover Compaq AlphaServer running Open VMS. Rather, the project was focused on settings required to give optimal performance, device driver levels. As such, the book is intended for the experienced UNIX professional who has a broad understanding of storage concepts, but is not necessarily an expert in each of the areas discussed.

First, Chapter 1 provides a broad overview of the ESS. Then, Chapter 2 explains general connectivity issues such as switch zoning, multi-pathing, and volume management. Chapters 3 and 4 cover the attachment of Compaq AlphaServers running Tru64 UNIX (V4 and V5 respectively), both for standalone and clustered configurations. Chapter 5 discusses HP servers. Chapter 6 discusses Sun systems, and also covers various aspects of using some of the Veritas suite of products.

The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization San Jose Center.

Barry Mellish is a Project Leader at the International Technical Support Organization, San Jose Center. He has coauthored four previous redbooks and has taught many classes on storage subsystems. He joined IBM UK sixteen years ago, and before recently joining the ITSO, he worked as a senior storage specialist in the Disk Expert team in EMEA.

Surjit Sedeora is a highly experienced UNIX professional. Arriving in the USA from India in 1975, he first developed Mathematical Geophysical Modeling techniques. In 1986 he was hired by NASA, where he worked extensively on the UNIX systems installed there. He then left NASA to work for Cray Research Inc. at the Lawrence Livermore Labs. in California. This was followed by a position with Information Technology (IT) of the California Department of Forestry and Fire (CDF) Protection, where he worked for over eight years. He then carried out consulting work for Hewlett Packard as their Response Center Engineer, where he gained considerable experience in Solaris 2.x and AIX 4.3.1 also. Finally, in August 1999, he began working for IBM Global Services, where he is now engaged in a wide variety of consulting assignments.

Tom Smythe is a team member with the New Computing Technologies group under the IBM Global Services umbrella located in Schaumburg, Illinois (SDC Midwest). He has over a decade of experience with UNIX systems, which includes several years building "home-grown" UNIX servers as well as providing local/remote support for those systems. He also spent a couple years supporting a mix of some 200+ Sun and HP servers and workstations in a subnetted and switched network architecture — an end-user environment consisting mostly of engineers. Today, his main role is in support of an IBM customer running batch and statistical analysis on a dozen NUMA-Q servers with anywhere from 4 to 48 CPUs each and a mix of both EMC and ESS storage totalling over 20TB's. His knowledge and experience were instrumental in the planning and implementation of the ESSs attached to the NUMA-Qs.

Gea voci is an ITS IT Specialist based in Milano (Italy). She worked for 10 years in Digital (then Compaq). Her main activities were supporting Tru64 UNIX and cluster solutions, teaching courses and new product introduction. She joined IBM in 1999 and now supports Open Systems UNIX platforms, mainly for the Italian Telecom Company. Her areas of interest have grown to include HP and Sun architectures.

Thanks to the following people for their invaluable contributions to this project:

Ken Fung, IBM SSG, San Jose

Chuck Grimm, IBM SSG, San Jose

Phil Michell, IBM SSG, San Jose

Mike Janini, IBM SSG, San Jose

Robert Moon, IBM SSG, San Jose

Dominick Nguyen, IBM SSG San Jose

Jack Flynn, IBM SSG San Jose

Edward Ng, IBM SSG San Jose

Sergio Y Okado, IBM Brazil.

Comments welcome

Your comments are important to us!

We want our Redbooks to be as helpful as possible. Please send us your comments about this or other Redbooks in one of the following ways:

- Fax the evaluation form found in “IBM Redbooks review” on page 141 to the fax number shown on the form.
- Use the online evaluation form found at ibm.com/redbooks
- Send your comments in an Internet note to redbook@us.ibm.com

Chapter 1. Enterprise Storage Server Overview

The IBM Enterprise Storage Server (ESS) provides high speed and high availability within a single storage system to computers around the world. This chapter gives the system administrator with a brief overview of the features and configuration of the ESS.

For Open Systems servers, the ESS is capable of utilizing Fibre Channel point-to-point and Fibre Channel switched (FC-SW), Fibre Channel arbitrated loop (FC-AL), SCSI, and combinations thereof. In fact, a single ESS can provide storage for multiple servers, as illustrated in Figure 1.

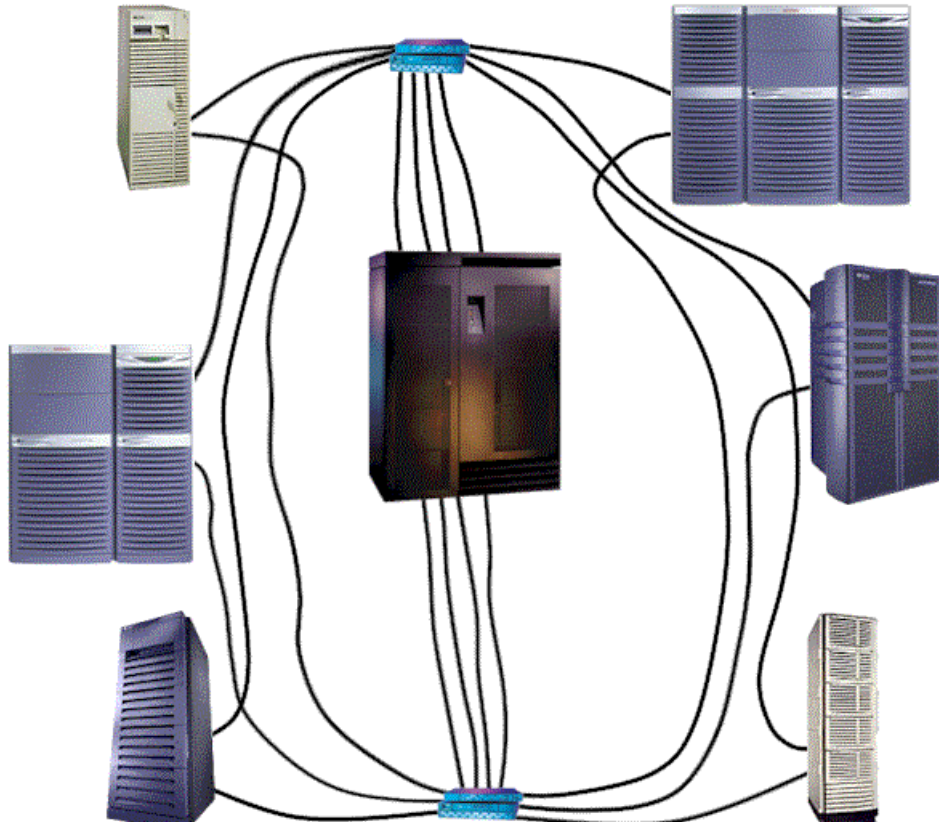


Figure 1. Storage Area Network (SAN)

Each ESS utilizes two RS6000s acting as Cluster 1 and Cluster 2. These two clusters are the central core to the success of the ESS. Following the data path will help explain why.

The data path begins at the RAID array or rank and moves across the SSA loop to the Device Adapters (DAs). Each SSA loop is attached to one DA in each of the clusters to provide redundancy in the event of a single component failure (see Figure 2).

From the DAs, the data moves into the Cache for retrieval if called on again. At the same time, it is passed through the Cluster Processor Complex to the Cluster Interconnect where the appropriate Host Adapter(s) (HAs) send the data on its way to the appropriate server.

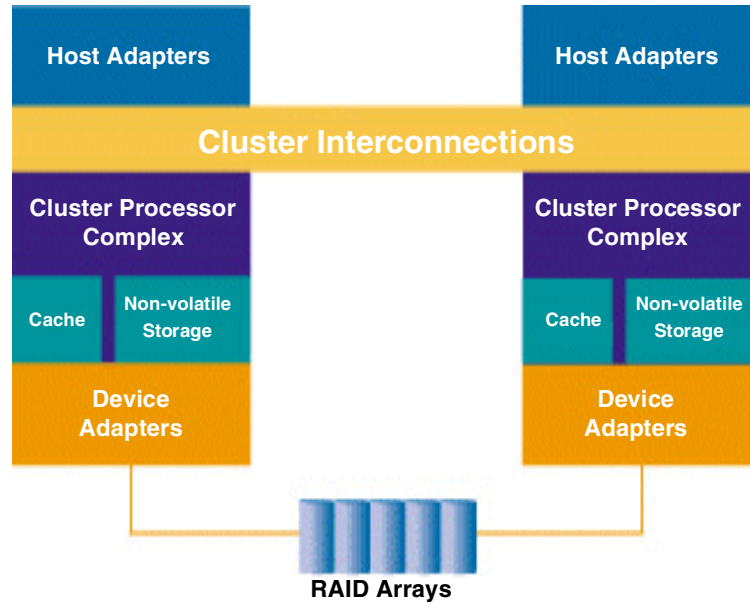


Figure 2. ESS basic overview

Writes from the attached host system enter both the cluster cache and the non-volatile store (NVS). Once the data is secured in two places, the write complete is issued back to the host and the data is destaged to disk.

Within the ESS, each rank is comprised of 8 hard drives of the same logical size. Also, ranks must be added in pairs to a loop. Up to 8 ranks may be on a loop, for a total of 32 ranks maximum per ESS with expansion cabinet.

The sizes available are 9.1 GB, 18.2 GB, and 36.4 GB. Utilizing the 36.4 GB hard drives, a single ESS, with expansion cabinet, can provide over 11 TB of RAID storage.

The ESS is capable of supporting both RAID-5 and Just a Bunch Of Disks (JBOD) simultaneously. However, only one can be used on any rank at a time. For example, if an ESS were configured with 4 ranks per loop, one of those ranks could be dedicated as JBOD storage, while the remaining three could be configured with RAID storage. Finally, the HAs can be either SCSI or Fibre Channel. Figure 3 provides a logical view of the internal layout of the ESS.

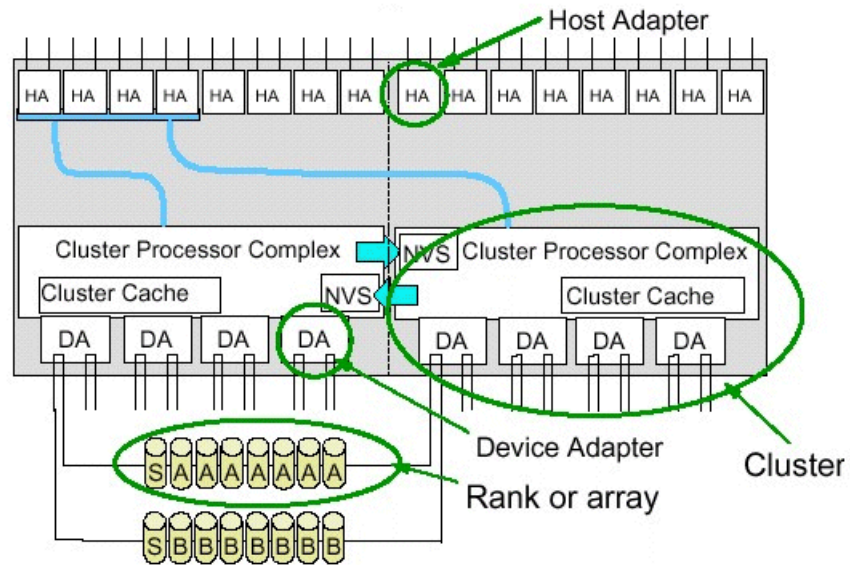


Figure 3. ESS logical configuration

1.1 Configuring the ESS

This section is not intended as a replacement for other books on configuring the ESS. Rather, it is meant as a basic guide or refresher for storage administrators with experience on the ESS.

1.1.1 ESS Specialist

The IBM ESS comes with a network based tool that allows managing and monitoring of the ESS. This product is called the ESS Specialist. It is a Web based tool the administrator uses to configure and administer the ESS. The Specialist uses a secure network connection and normally runs from a PC. A browser (such as Netscape Navigator, Microsoft Internet Explorer, or Sun Hot Java) that supports Java 1.1.6 or higher is required. The ESS Specialist provides the customer with the ability to:

- Monitor error logs
- View the status of the ESS
- View and update the configuration
- Add, delete, or modify host systems
- Configure host ports (both SCSI and fibre)
- Create RAID or JBOD logical volumes
- Add logical volumes and reassign logical volumes among the attached hosts
- View communication resource settings, such as TCP/IP configuration, pager lines, and users
- View cluster LIC levels
- Select an authorization level for each user

Isolating the ESS network connections using a firewall or stand-alone network is highly recommended.

If using an NT system to launch the ESS Specialist, verify that the *virtual memory total page file size* of the console PC is between 140 MB and 190 MB. Smaller configurations may cause the PC to run out of virtual memory.

You can launch the Specialist by entering the hostname or IP address for either of the ESS clusters in the Location or URL window of the browser (see Figure 4).

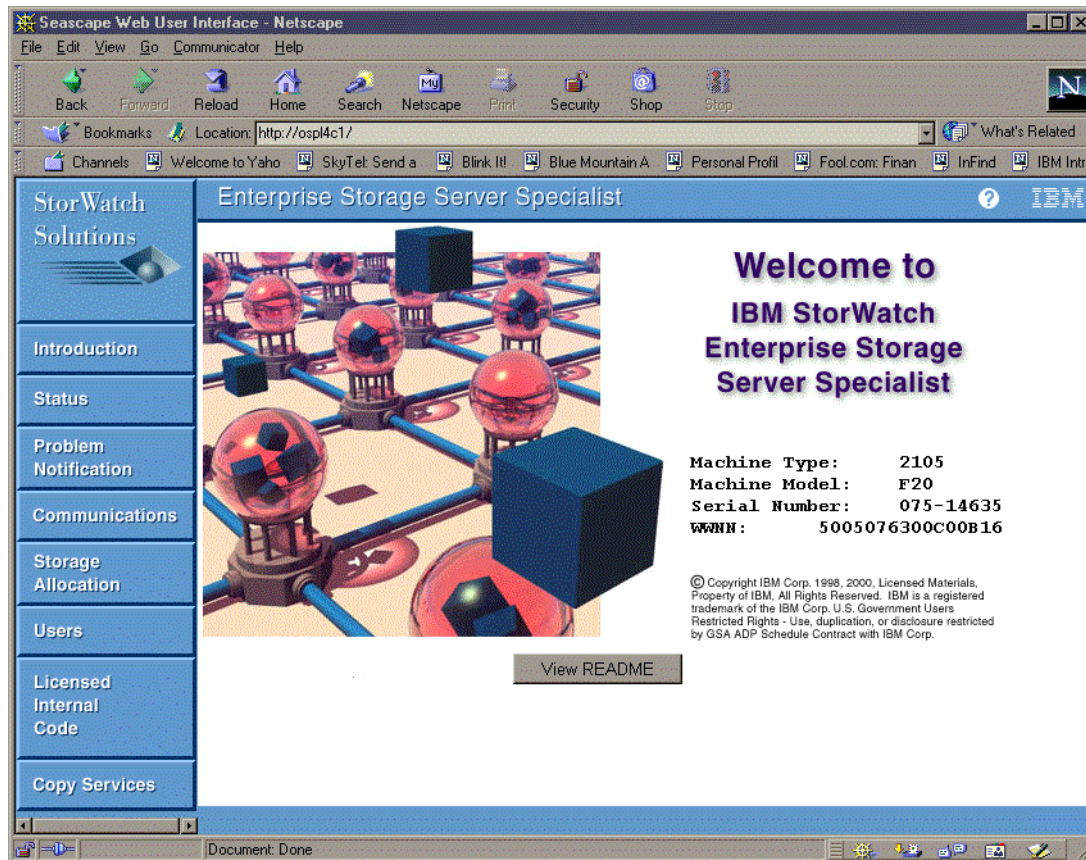


Figure 4. ESS Specialist

1.1.2 ESS — Storage Allocation

Selecting the Storage Allocation button on the left will present the administrator with a login pop-up window. After successfully entering a valid login and password, the administrator will be required to accept several security certificate questions. After the final certificate has been accepted, the Storage Allocation — Graphical View will be displayed (see Figure 5).

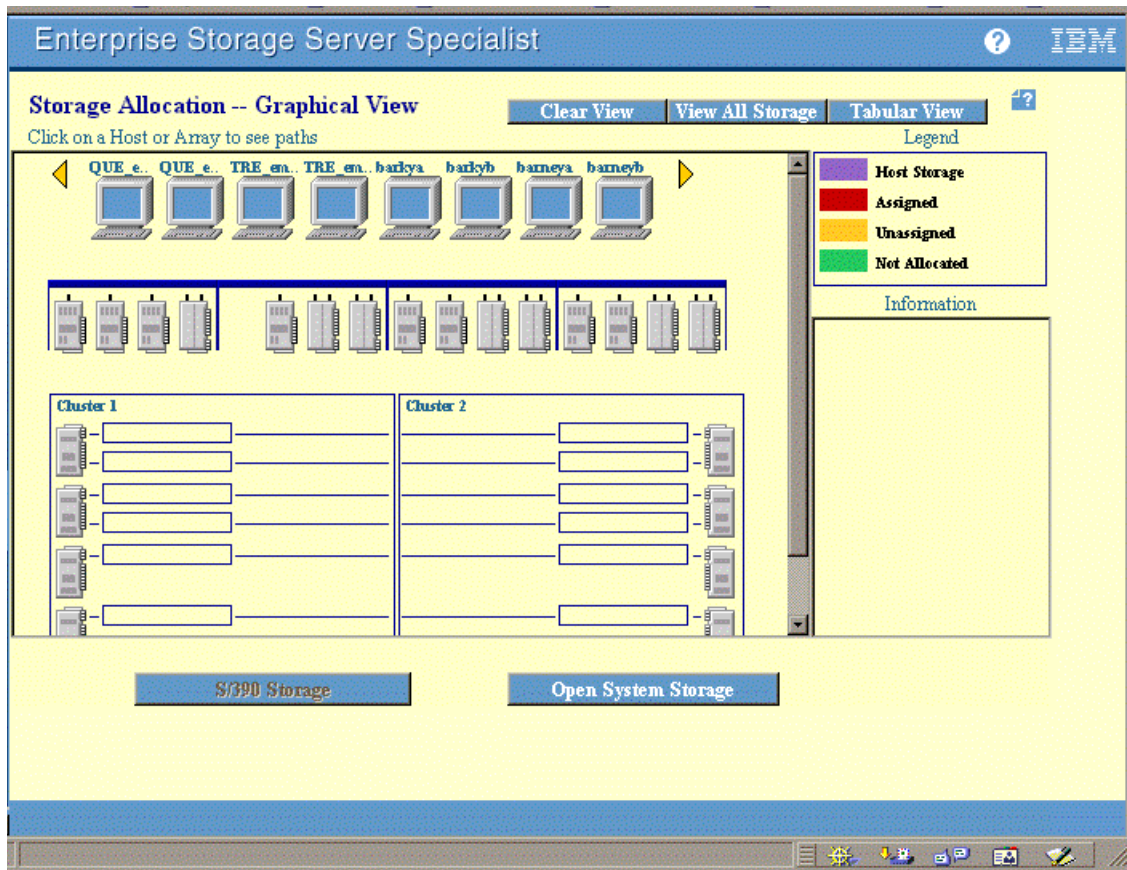


Figure 5. Graphical Storage Allocation

If any ESCON HAs are installed in the ESS bays, then the S/390 Storage button will be clickable. In this example, no ESCON HAs are installed. Rather, both Fibre Channel and SCSI HAs are installed in this ESS.

To see what hosts are attached to which HAs, clusters, and ranks, simply click on a host across the top row. This will display the pathway(s) assigned to the host selected starting with the HAs and moving down through the clusters, DAs and finally, the ranks. The color coded Legend on the top right will be useful in determining which ranks are candidates for establishing additional LUNs. To view another host, select the Clear View button before selecting the next host.

It is highly recommended a detailed log be maintained pertaining to the storage space within the ESS and how it is allocated. As systems are brought online and others removed, the layout of the arrays could easily become confusing. This would be especially true should a storage administrator move on to another position leaving the ESS behind for the new storage administrator to “figure out.”

1.1.3 ESS — Open System Storage

Selecting the Open System Storage button will produce a screen similar to the one depicted in Figure 6. From this view, it is possible to add and remove hosts, configure HAs, setup disk groups, add and remove volumes, and modify how existing volumes are assigned.

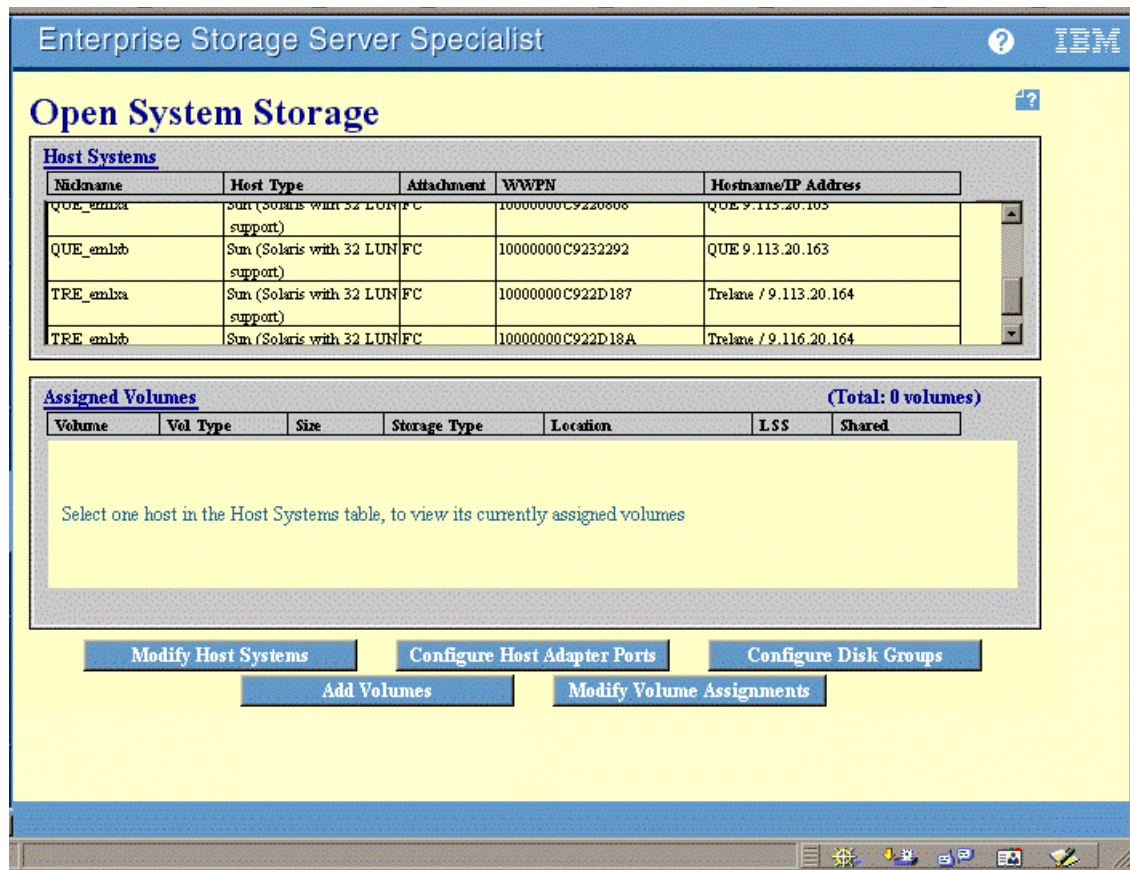


Figure 6. Open System Storage

Selecting a host in the Host Systems window displays the logical volumes assigned to that host in the Assigned Volumes window (see Figure 7).

The screenshot displays the 'Enterprise Storage Server Specialist' interface. The main window is titled 'Open System Storage'. It contains two primary tables:

Host Systems

Midname	Host Type	Attachment	WWPN	Hostname/IP Address
QUE_embxa	Sun (Solaris with 32 LUN support)	FC	10000000C9220806	QUE 9.113.20.105
QUE_embxb	Sun (Solaris with 32 LUN support)	FC	10000000C9232292	QUE 9.113.20.163
TRE_embxa	Sun (Solaris with 32 LUN support)	FC	10000000C922D187	Trelane / 9.113.20.164
TRE_embxb	Sun (Solaris with 32 LUN support)	FC	10000000C922D18A	Trelane / 9.116.20.164

Assigned Volumes (Total: 85 volumes)

Volume	Vol Type	Size	Storage Type	Location	LSS	Shared
13C-14635	Open System	01.0 GB	RAID Array	Device Adapter Pair 1 Cluster 2, Loop B Array 1, Vol 060	LSS: 011	Yes
13E-14635	Open System	01.0 GB	RAID Array	Device Adapter Pair 1 Cluster 2, Loop B Array 1, Vol 062	LSS: 011	Yes
1B0-14635	Open System	01.0 GB	RAID Array	Device Adapter Pair 1 Cluster 2, Loop B Array 1, Vol 062	LSS: 011	Yes

Below the tables are several control buttons: 'Modify Host Systems', 'Configure Host Adapter Ports', 'Configure Disk Groups', 'Add Volumes', and 'Modify Volume Assignments'.

Figure 7. Open System Storage with assigned volumes

1.1.4 ESS — Modify Host Systems

To add, remove, or modify an existing host, select the Modify Host Systems button. The display in Figure 8 will appear.



Figure 8. Modify Host Systems

To add a new host, fill in the fields within the Host Attributes window. Select an appropriate Host Type (see Figure 9). If an appropriate Host Type is not available, use the Host Type specified in the documentation provided with the ESS and/or the vendor's documentation. If none is available or specified, then the host being configured is **not** a supported host.

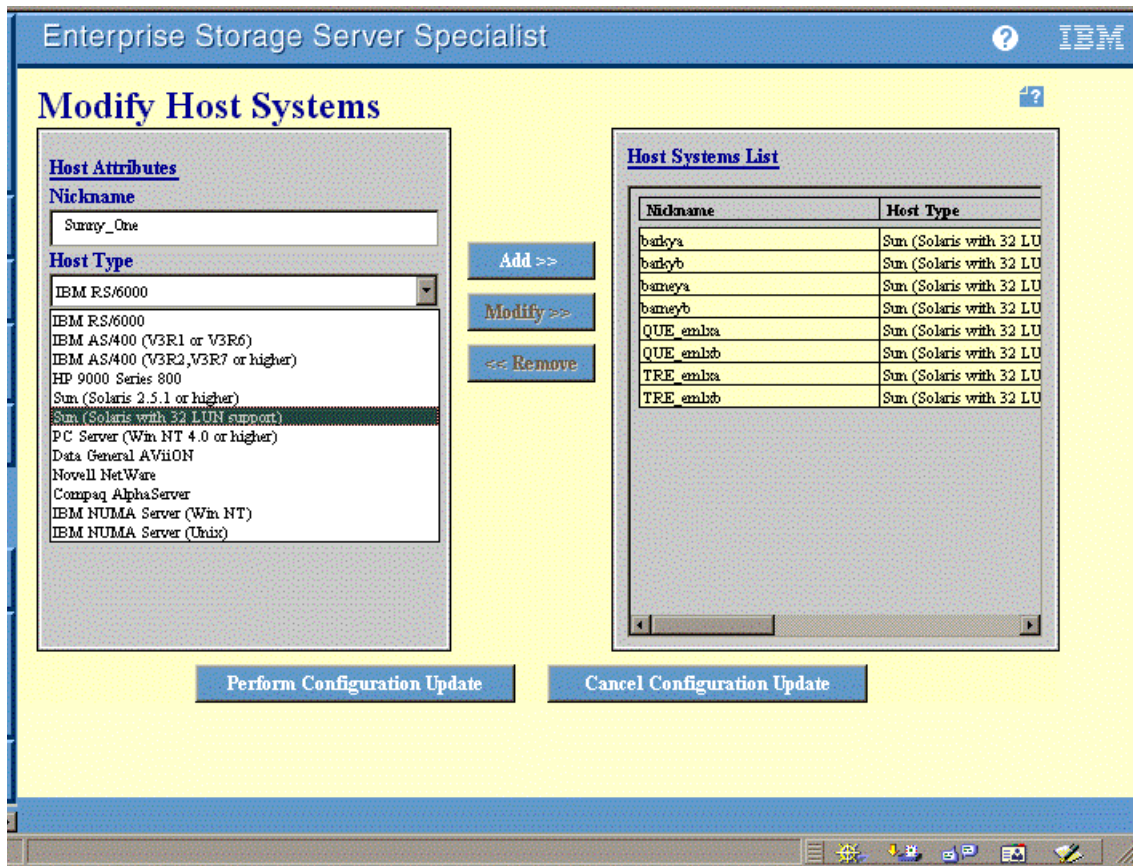


Figure 9. Modify Host Systems — Host Type

The Host Type configured is extremely important, as the ESS determines drive geometry, labels, targets, and LUNs available, and so on, based on the Host Type field.

If the host is a Fibre Channel attached host and DNS is implemented and available to the ESS (unlikely), then the hostname can be entered. Otherwise, you need to enter the IP address of the host in the Hostname/IP Address field (see Figure 10). Finally, enter the World-Wide Port-Name (WWPN) for the host fibre card. If the host has more than one fibre card, then create as many unique Host Systems as the host has fibre cards installed and attached to the ESS.

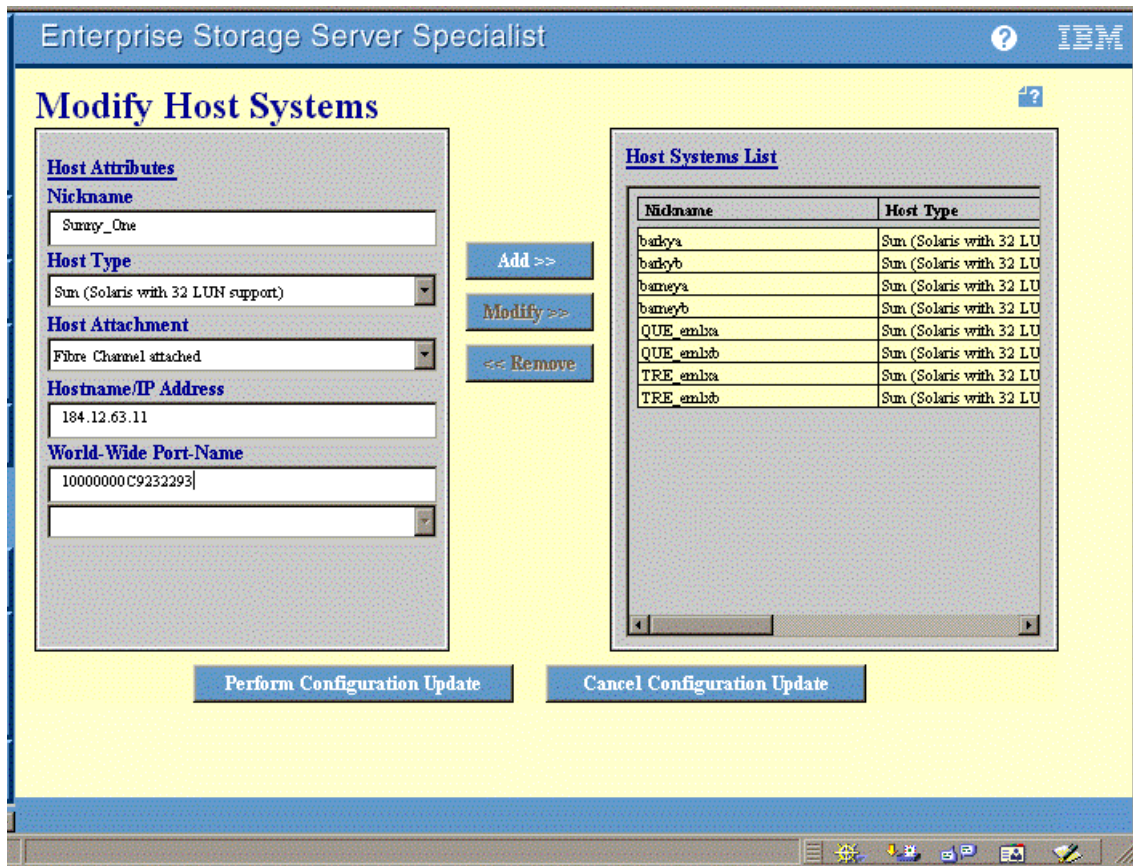


Figure 10. Modify Host Systems — WWPN

Create as many Host Systems as are necessary. After each one, simply press the Add button between the two windows. The newly created host will be displayed within the right window (see Figure 11).

Multiple additions, modifications, or removals can be performed within one configuration update. However, do not perform multiple operations such as Add, Modify, and Remove within a single configuration update. Rather, perform each type of operation independent of the others. None of the modifications, additions, or removals will take effect until the Perform Configuration Update button at the bottom of the screen is selected.

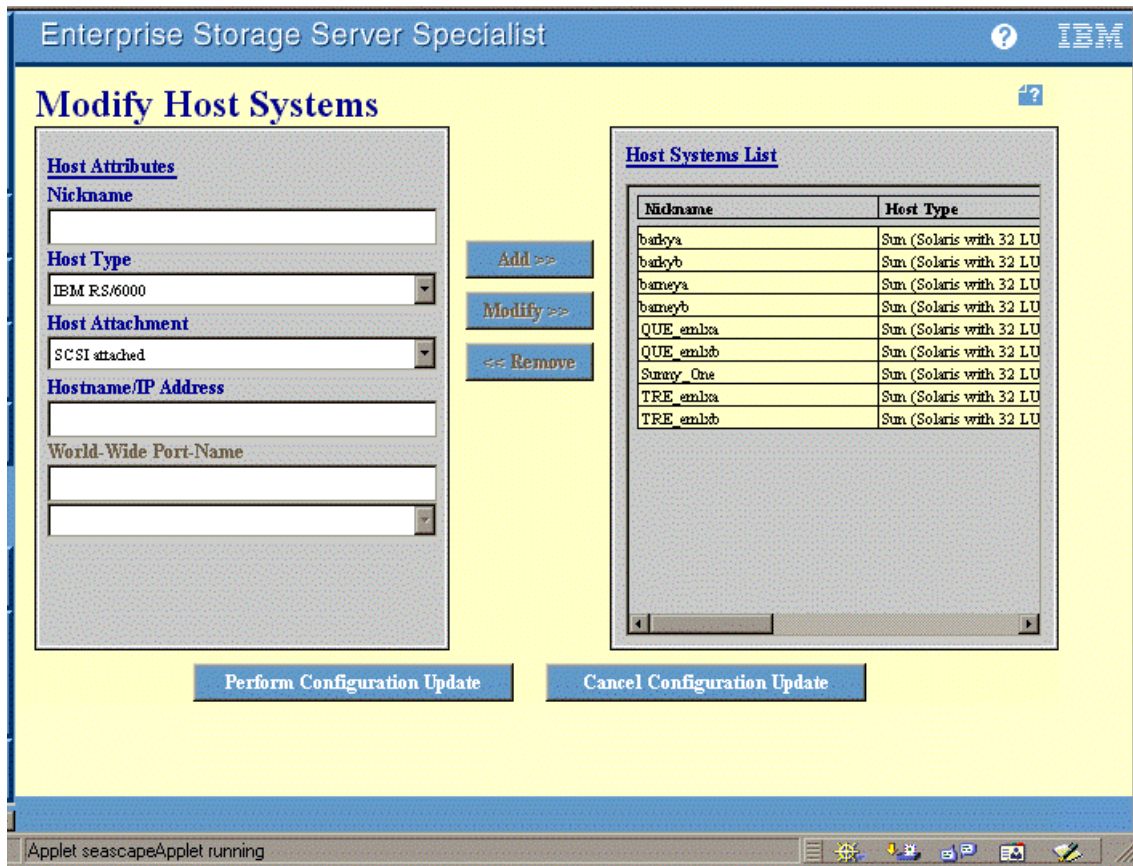


Figure 11. Modify Host Systems — Add

1.1.5 ESS — Configure Host Adapter Ports

Next, return to the Open System Storage Display (see Figure 6 on page 7) and select Configure Host Adapter Ports. This will bring up a display similar to the one in Figure 12.

The HAs are graphically represented beneath the Configure Host Adapter Ports title. SCSI HAs are identified by two ports located on the top of each HA in the view. ESCON and fibre HAs have a single port. ESCON and fibre can be differentiated by the detail on the HA representation. Also, by clicking on the HAs icon or by selecting the bay-adapter-port in the Host Adapter Port pull-down, different attributes will be displayed below the row of HAs.

Finally, only those adapter slots that are occupied will be visible. Empty adapter slots will not be visible.

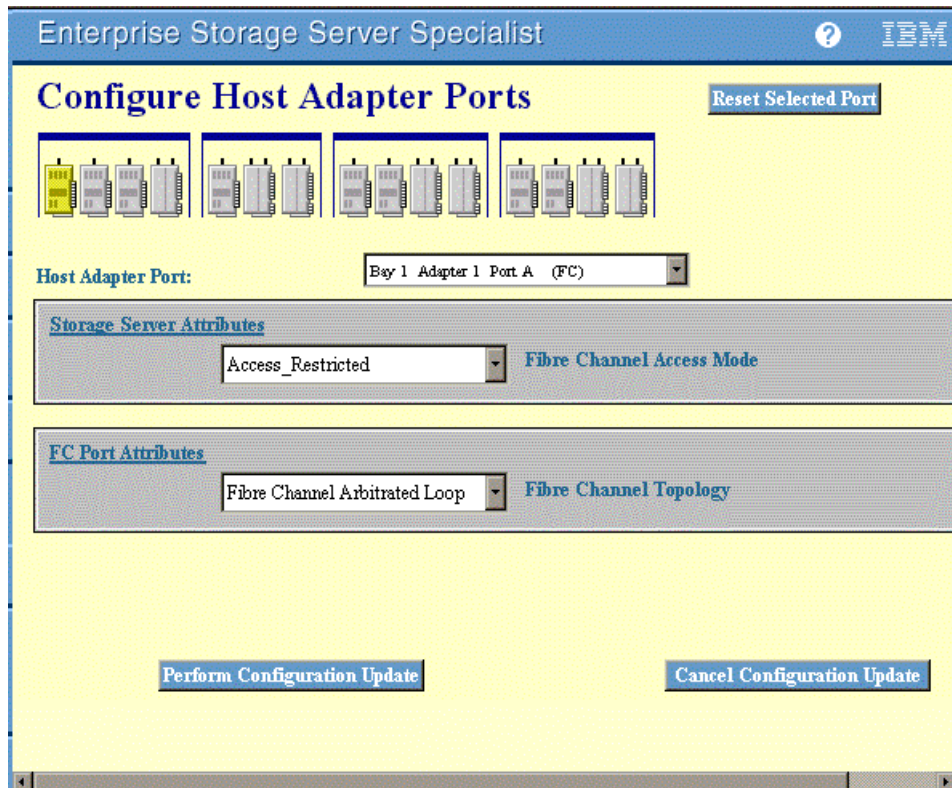


Figure 12. Configure Host Adapter Ports

Unlike SCSI, Fibre Channel allows all hosts in a FC-SW or FC-AL environment to view all storage available within the environment — assuming no zoning has been established within a switch or restrictions put in place elsewhere. In order to get around the “see everything” issue, the Fibre Channel HAs within the ESS can be configured with Access Restricted attributes. (See Figure 12)

Using Access Restricted mode, the ESS limits the visibility of the LUNs to only those WWPNs associated with each LUN. In effect, the ESS performs LUN masking to prevent other hosts from gaining access to LUNs that are not defined as available to those hosts.

For Fibre Channel HAs, it is not possible to change the Fibre Channel Access Mode from within the Specialist (see Figure 13) until the HA is set to service mode. This can be done by selecting Undefined for the Fibre Channel Topology pull-down and then selecting Perform Configuration Update.

Be aware that during the period of time the HA is undefined, all host access to LUNs across that HA will be suspended.

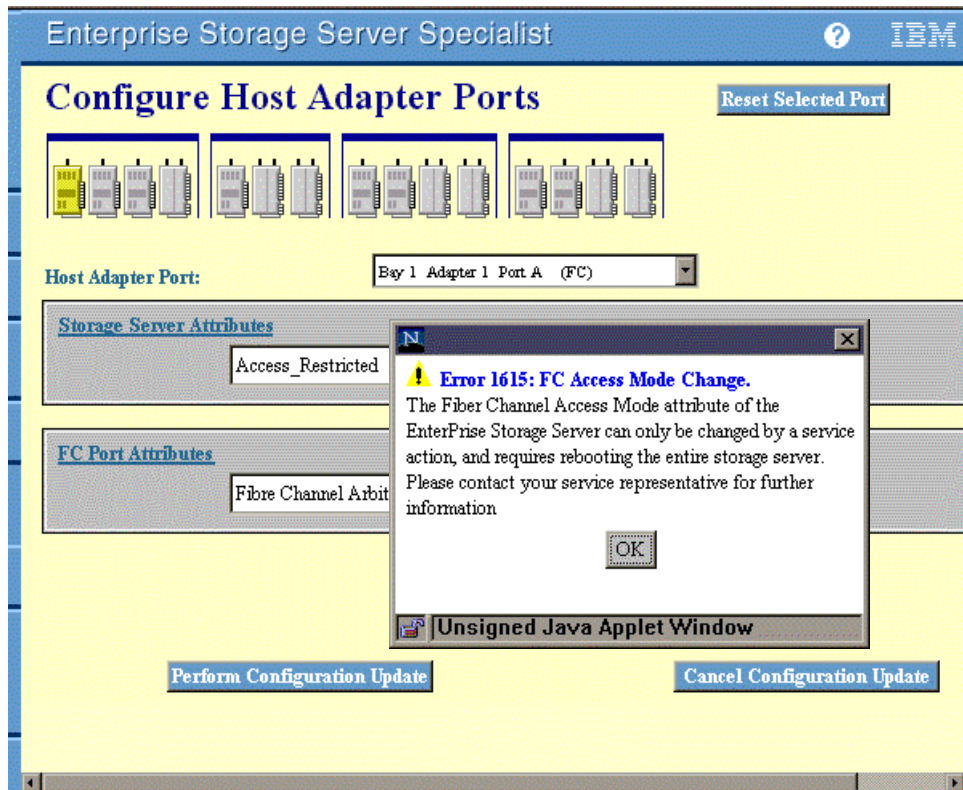


Figure 13. Configure Host Adapter Ports — Fiber Channel Access Mode

Also, if a Fibre Channel HA needs to be configured for FC-AL while currently configured as point-to-point, then the FC Port Attributes pull-down needs to be set to Undefined first, followed by selecting Perform Configuration Update. This will set the HA in service mode. After the update completes, it is then possible to change the pull-down to Fibre Channel Arbitrated Loop. Again, selecting Perform Configuration Update will enable the change.

SCSI HAs provide both A and B ports (see Figure 14). SCSI HAs provide the opportunity to *share* any of the logical volumes across two HA ports. This is done under the Second Bus Connection pull-down. Only SCSI HAs are available under this pull-down. If a second SCSI HA port is selected, it should not be located within the same bay as the first HA. Otherwise, if both HAs are located within the same bay, and that bay succumbs to a power supply failure, all pathways to the data will become unavailable.

When opting to access data from multiple paths, be sure the host system is capable of running IBM SDD, Veritas DMP, or some other vendor's version of multi-pathing software. Attempting to access the same logical volume without a multi-pathing solution installed **WILL** lead to disastrous results and may include the loss of all data on the logical volume.

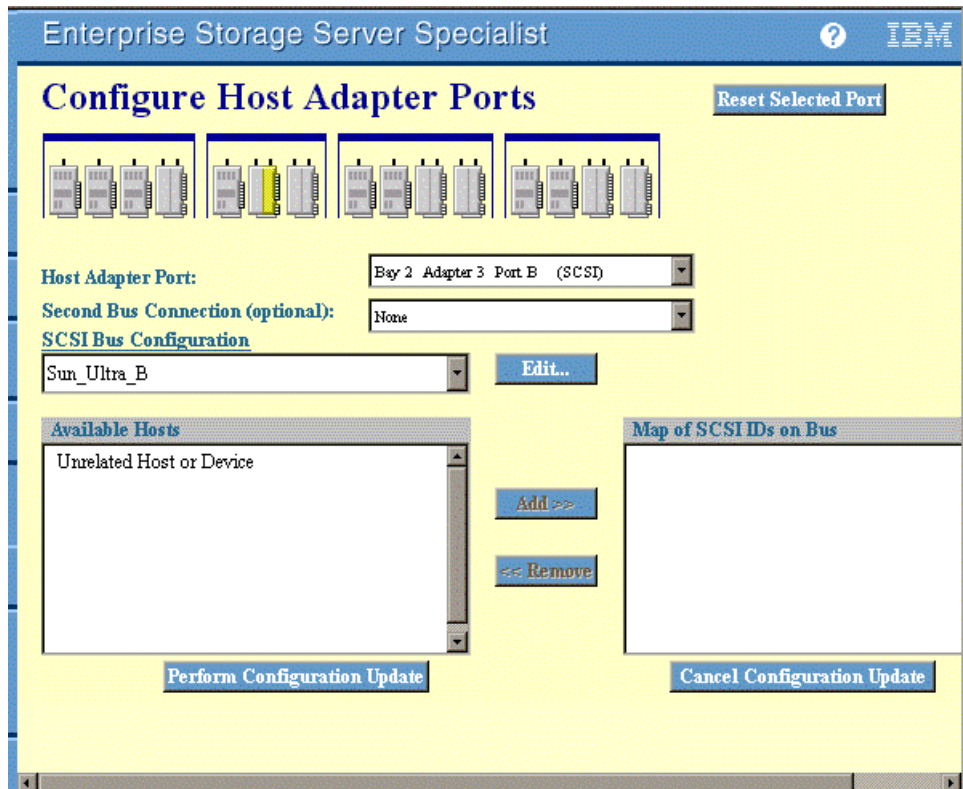


Figure 14. Configure Host Adapter Ports — SCSI

SCSI ports are always associated with a particular host as defined under the Modify Host Systems view. Under SCSI Bus Configuration, select the appropriate host type (see Figure 15). This should match the type of host being attached to the SCSI port.

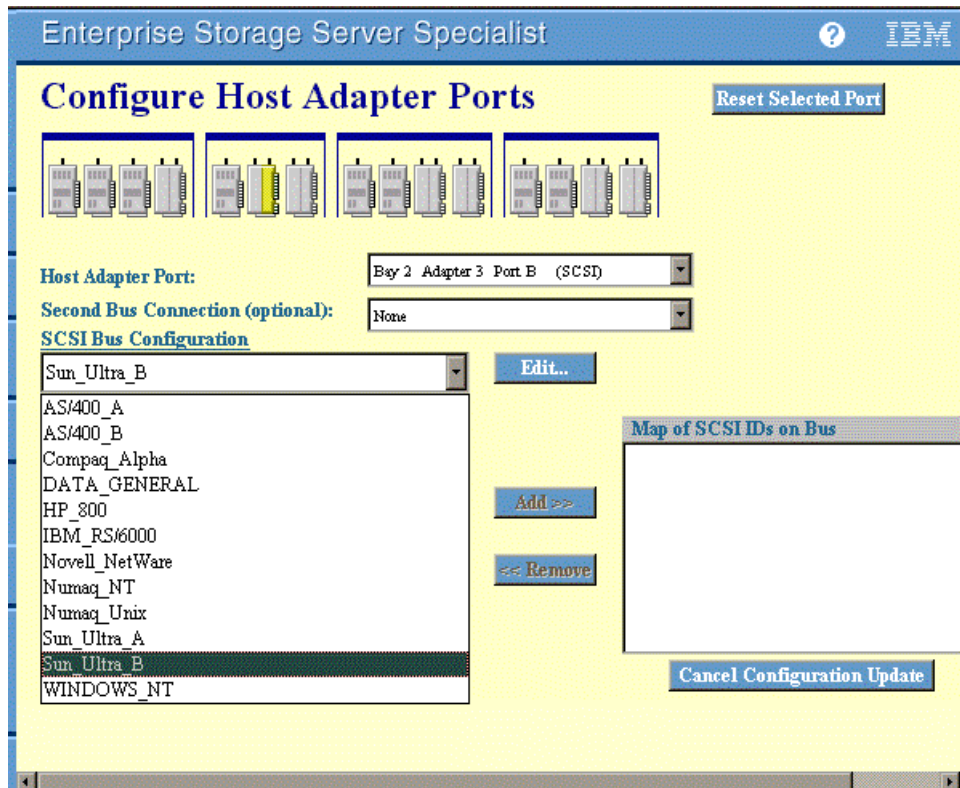


Figure 15. Configure Host Adapter Ports — SCSI Bus Configuration

Once the host type has been selected, a list of the default SCSI IDs not available on the bus will appear in the Map of SCSI IDs on Bus window. For each SCSI ID that should **not** be used for target and LUN assignment, change the values of the Unrelated Hosts or Devices pull-down within the Map of SCSI IDs on the Bus window.

If the number of unavailable SCSI IDs are greater than the number of available Unrelated Hosts or Devices in the right window, click on the same field in the left window and then the Add button. Continue this process until all the unavailable SCSI IDs have been added to the right window.

If a Second Bus Connection was selected above, the SCSI IDs blocked on the current SCSI port will also be blocked on the Second Bus Connection. It is possible to modify one of the currently available host types using the Edit button. However, if this is not recommended in any of the documentation provided by the host hardware vendor or IBM, these modifications will not be supported and may cause data loss.

1.1.6 ESS — Add Volumes

Return to the Open System Storage window (see Figure 6 on page 7) and select the Add Volumes button. This will bring up a display much like the Storage Allocation — Graphical View window in Figure 5 on page 6.

The Add Volumes view allows the creation of new volumes on the ranks within the ESS. Selecting one of the hosts across the top row displays a color coded representation of the HAs available to that host, and the disk space associated with that host (see Figure 16).

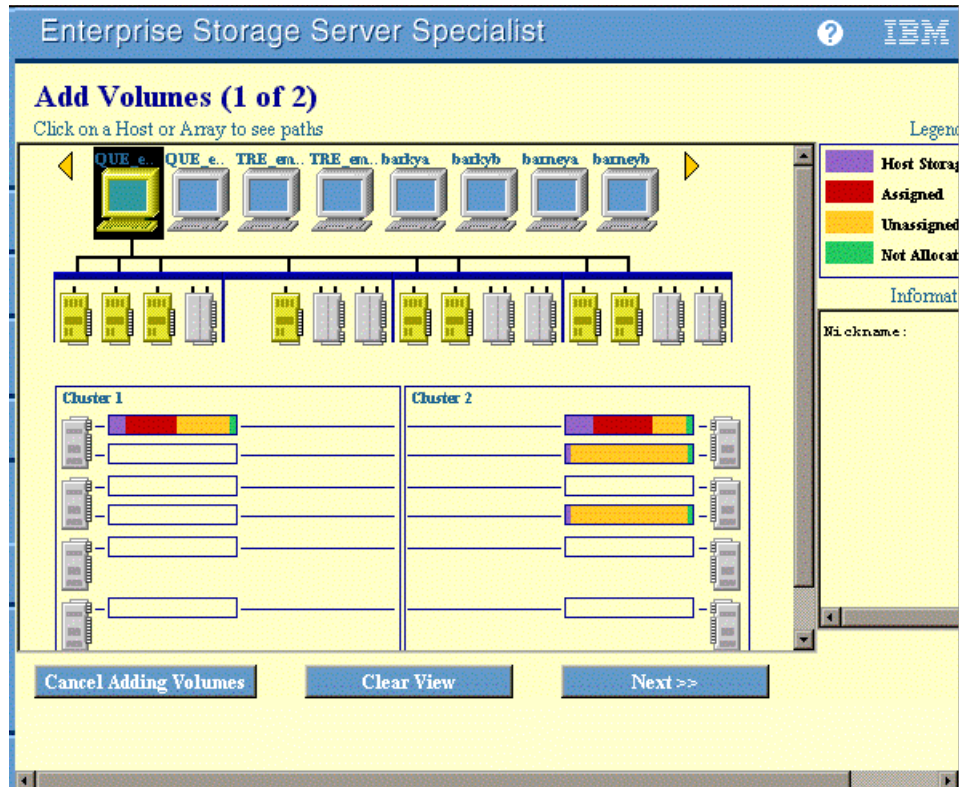


Figure 16. Add Volumes (1 of 2) with host selected

Selecting one of the HAs from those already highlighted displays a color-coded representation of all the storage within the ESS (see Figure 17).

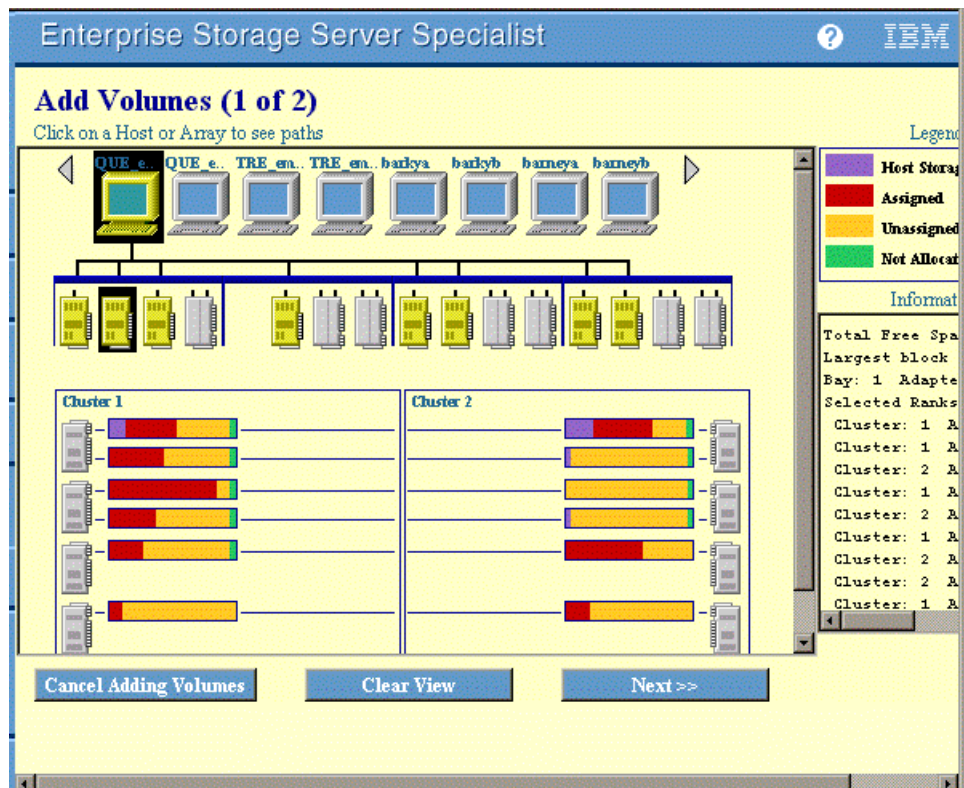


Figure 17. Add Volumes (1 of 2) with host and DA selected

Finally, pick one or more of the ranks with unallocated space before selecting the Next button at the bottom of the display.

- Please note that the location of the rank determines the cluster the rank is presented by. While the HAs can communicate with both clusters, only one cluster may communicate with any HA at a time.
- A form of switching takes place within the HA. When data requests are received that must be satisfied by both clusters, the HA stops communication with one cluster before opening communications with the other. For that reason, it is best to assign ranks to an HA from a single cluster.
- The ESS will function properly if an HA is assigned ranks from both clusters. However, performance will be degraded when compared to individual HAs attached to one cluster each.

The second display of the Add Volumes process will provide information on the total space available on the selected ranks as well as the largest possible volume size (see Figure 18).

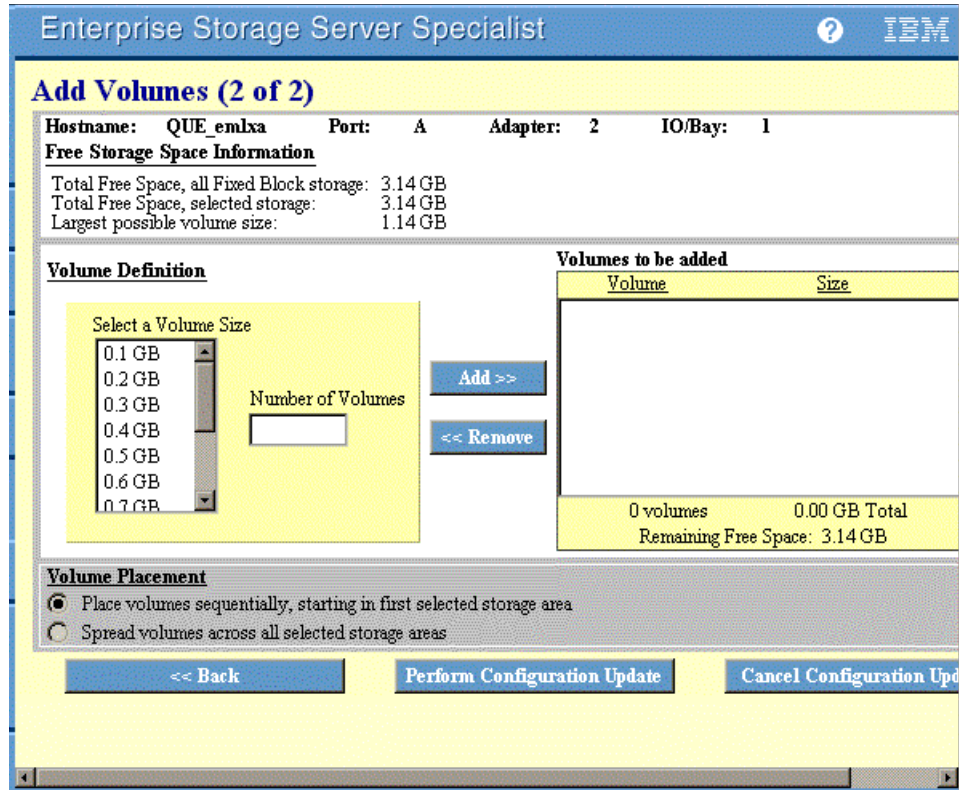


Figure 18. Add Volumes (2 of 2)

The Select Volume Size window provides the choices possible given the total largest possible volume size for the selected ranks. First, highlight one of the entries and enter the Number of Volumes just to the right. Then, select the Add button and the entry will be displayed in the Volumes to be added window. Continue this process until the storage desired has been reached or the available space on the selected ranks no longer provides the capacity to create volumes of the size desired.

The Volume Placement radio buttons allow volumes to be created along the selected ranks (top button) or across the selected ranks (bottom button). Spreading the I/O across the ranks provides greater bandwidth, as more of the DAs will be involved in data movement.

- *If* the application performs sequential I/O operations and *if* the application utilizes most, if not all, of the space available on the rank as a single large LUN, then spreading the I/O across the ranks *may* make sense. However, most storage layouts within the ESS will have several to dozens of smaller LUNs per rank with one or more hosts accessing each of the LUNs on the rank. This sets up a scenario whereby a single, heavy I/O LUN may and can cause a negative performance impact to the remaining LUNs on the associated ranks.
- Therefore, by reducing the number of ranks affected by any given LUN, it is possible to isolate the impact of any heavy I/O LUN to a single rank. The result of this configuration is to minimize the affected LUNs to those on the rank(s) occupied by the heavy I/O LUN.

At this point, selecting Perform Configuration Update will create the volumes as defined.

1.1.7 ESS — Modify Volume Assignments

From time-to-time, it may become necessary to modify the assignment of volumes. This is done through the Modify Volume Assignments display. From the Open System Storage window (see Figure 6 on page 7), select the Modify Volume Assignments button. A display similar to the one in Figure 19 should appear.

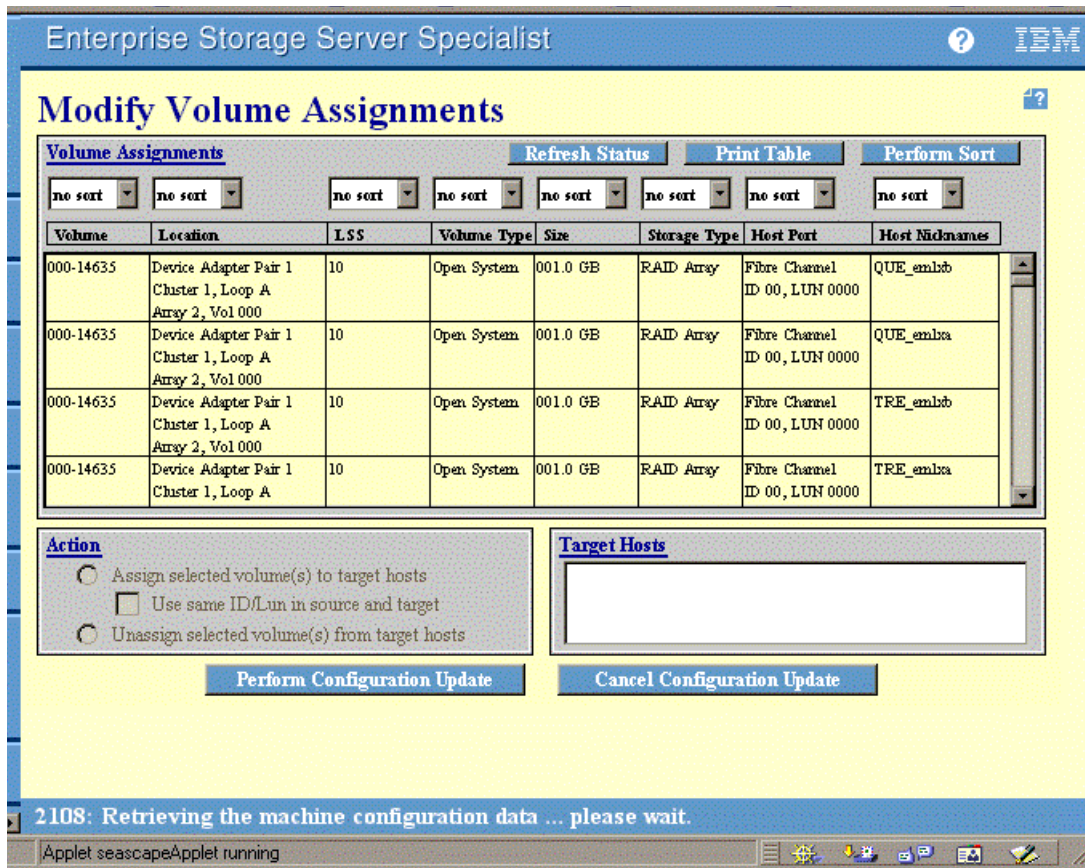


Figure 19. Modify Volume Assignments

Selecting one or more volumes provides the opportunity to associate those volumes with any host in the Target Hosts window. In fact, multiple hosts can be selected. This would benefit servers in a clustered environment. Also, hosts with more than two paths to the same data would be associated in this manner (see Figure 20).

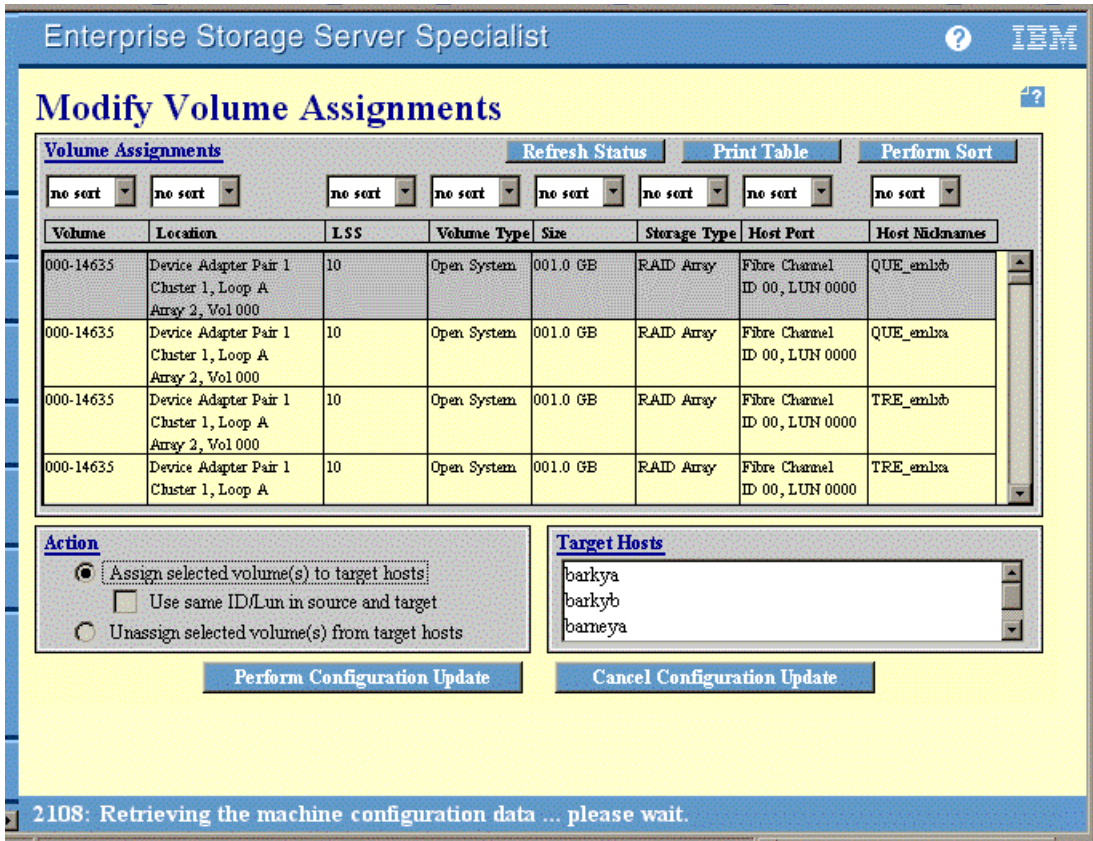


Figure 20. Modify Volume Assignments — Assign volume(s)

Within the Action window, it is possible to either assign the selected volumes to the selected hosts with or without matching the target/LUN assignment. It is also possible to remove the assignment of selected volumes from the selected hosts.

If an attempt is made to assign volumes using matching targets/LUNs that have already been assigned to one of the selected hosts, then the message in Figure 21 will be displayed.

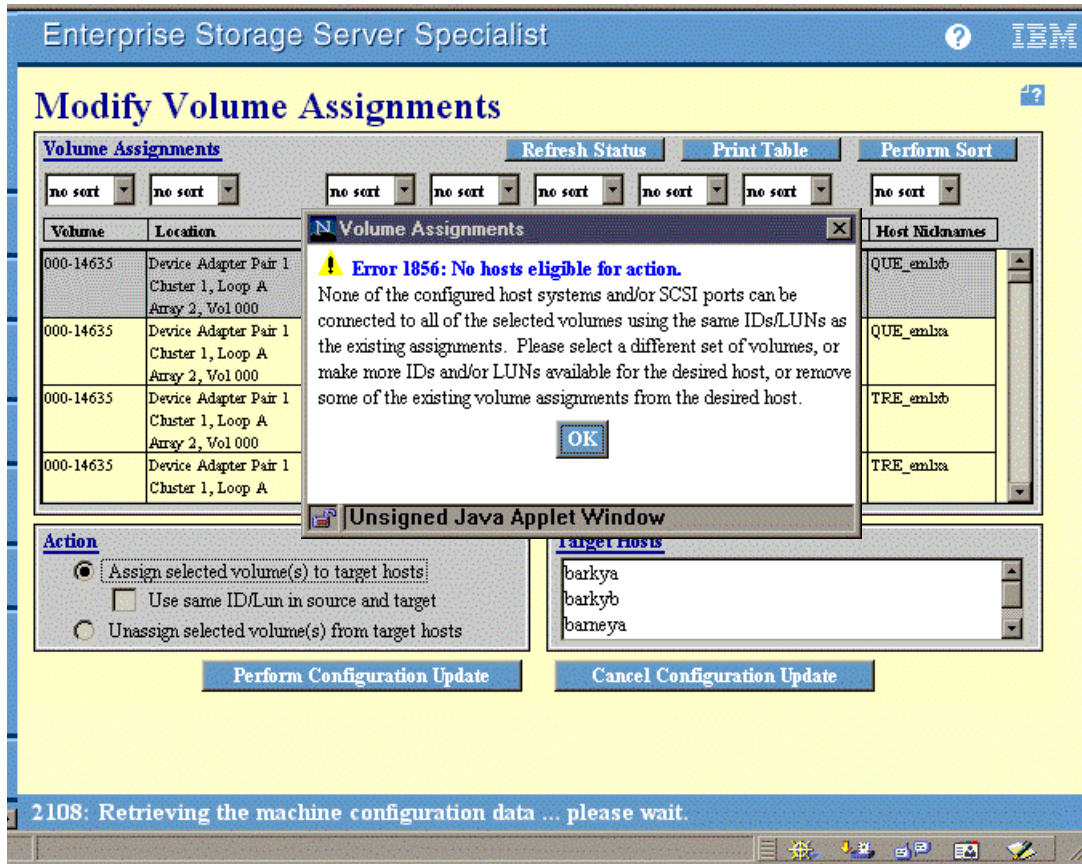


Figure 21. Modify Volume Assignments — duplicate target(s) and/or LUN(s)

The Modify Volume Assignments display allows the removal of assigned volumes, as well. Again, simply select the volumes to be removed, select the bottom radio button under Action, and select the appropriate host(s) under Target Hosts.

Multiple volumes may be selected using the Shift key while selecting the volumes at opposite ends of a range. Another option is to use the Ctrl key while selecting volumes that are interspersed among volumes that are not to be affected by the current modification.

On systems that will be sharing targets and LUNs, or systems that are built to be similar in appearance, it is often necessary or desirable to see the same targets/LUNs for the same volumes. Consider the following:

An ESS is configured with 3 volumes — the first with a size of 4 GB, the second with 8 GB, and the third with 16 GB. After the hosts and HAs are configured, the volumes are added or modified in differing sequences for each of five hosts. When each host views the size and label on the devices presented by the ESS, it is discovered that the hosts do not see the same device at the same target/LUN (see Table 1).

Table 1. Host view of devices presented

Server Name	Device Name	Target/LUN	Device Label	Device GB
Host A	sd10	0/0	Device 0	4
Host B	sd10	0/0	Device 2	16
Host C	sd10	0/0	Device 0	4
Host D	sd10	0/0	Device 1	8
Host E	sd10	0/0	Device 1	8

This situation could be avoided by configuring each of the volumes to each of the hosts using exactly the same sequence of steps during volume addition or modification on the ESS for each of the attached hosts.

Clustered hosts are usually unable to translate differing targets/LUNs for shared devices. Therefore, if the device on host A is not represented by the same device on host B, then it is highly likely the clustering software will be unable to bring the shared resource up.

For more information on configuring the ESS, please see the IBM Redbooks, *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420-00, and *IBM Enterprise Storage Server*, SG245465-00.

Chapter 2. Introduction to ESS connectivity

The ESS provides SCSI, FC-SW, and FC-AL connectivity to Open Systems platforms. This variety allows the administrator to allocate the type of connectivity necessary to meet a wide variety of needs — even multiple paths from the data within the ESS to the attached hosts.

2.1 Fibre Channel connectivity to the ESS

Traditionally, open system servers have used SCSI devices for disk access. Other devices have been tried, but none could provide the speed, versatility, and growth potential of SCSI — for example, multiple devices on a single connection, a wide bus for data transfer, and logical control with a mechanism to dictate controlling authority on the bus. All of these are features that allow for prioritization of the devices, high speed communications, and economical growth path.

While the ESS supports SCSI connectivity, the distance and device limitations are often sufficient reasons to examine other technologies. Also, with the advent of distributed storage, storage area networks (SANs), disaster recovery hot sites, single-image boot devices, and clustering, another medium for communications has risen to meet the growing demand — Fibre Channel.

Fibre Channel can be configured to use point-to-point connections such as fabric or switched fibre — FC-SW. Also, Fibre Channel can be configured to communicate via arbitrated loop — FC-AL. Both have advantages.

For environments that require a few TB of local storage using only a few physical connections, FC-AL provides a simple configuration. For example, NFS servers attached to a single ESS using FC-AL would be capable of exporting up to 11 TB of file system space to the workstation community. Also, it may turn out that a single host on a loop is able to transfer data faster across the fibre than the same host in a FC-SW environment.

In larger environments or environments that require features available under a switched configuration, implementing FC-SW becomes necessary. Multiple servers requiring access to the same devices, single point of failure resiliency, multiple pathways, and remote site support would best be served by attaching to an FC-SW environment.

As with any comparison of any product, it is possible for a few servers to take advantage of FC-AL and be configured to survive single point of failure

scenarios that do not include same-site disasters. However, even under these limited scenarios, FC-SW is ultimately the better choice. FC-AL quickly limits the growth of the environment, thus forcing the migration from FC-AL to FC-SW.

2.2 FC-SW and the switch

Switch configuration can be and is a book unto itself. However, what follows in this section will be a brief overview of what is possible using a switched or fabric to attach a host to the ESS.

2.2.1 Terms

The following are definitions of some relevant terms:

WWNNWorld Wide Node Name — the name given to a host or server.

WWPNWorld Wide Port Name — the name of an HBA or HA installed in a host or server.

Physical Port NumberThe number given to a physical port within a switch or on an HBA with more than one Fibre Channel interface.

PLOGIThe process by which an FC-AL port attempts to establish a connection via loop-login to another node on the loop.

FLOGIThe process by which an FC-SW port attempts to establish a connection via fabric-login to another node on the fabric.

Zone aliasA canonical name given to the eight hex number definition for a port. Can be used in lieu of the actual WWNN or WWPN within a switch that supports the C-name concept.

2.2.2 Zoning benefits

Zoning keeps devices that are not in the same zone from seeing each other. Whether a physical port number, an alias, or a WWNN/WWPN, zoning allows the resource to be grouped under a common definition that controls the flow of data and keeps it visible only to those members of that group or zone.

Zoning allows the administrator to build:

- Walls around systems with like operating environments or uses

- Isolated user groups within the fabric using a logical subset approach
- Unique areas separate from the rest of the fabric for testing, maintenance, data security, and so on
- Special purpose areas to be used for temporary jobs such as backups

Zoning also provides:

- The flexibility to manage a SAN and meet multiple operational objectives
- The ability to logically configure SAN resources for improved optimization
- Finer granularity to control environments and seclude those necessary for security
- Breadth of features to meet application demands

Zoning provides another level of segregation to the traffic that would otherwise be carried by an FC-AL environment. This additional segregation isolates data traffic only to the hosts within a zone. Used with LUN masking (within the ESS), the isolation of the logical volumes to the appropriate host(s) is ensured.

2.2.3 Zoning example

A simple zone would include one host and one ESS, as in Figure 22.

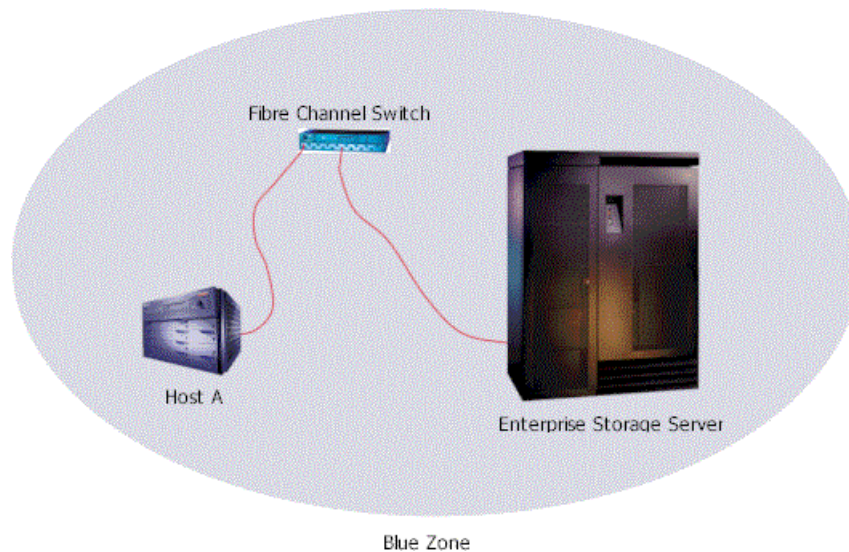


Figure 22. Simple zone

However, this type of configuration is neither very realistic nor very common in a switched environment. More frequently, there are several hosts and more than one storage server, as in Figure 23.

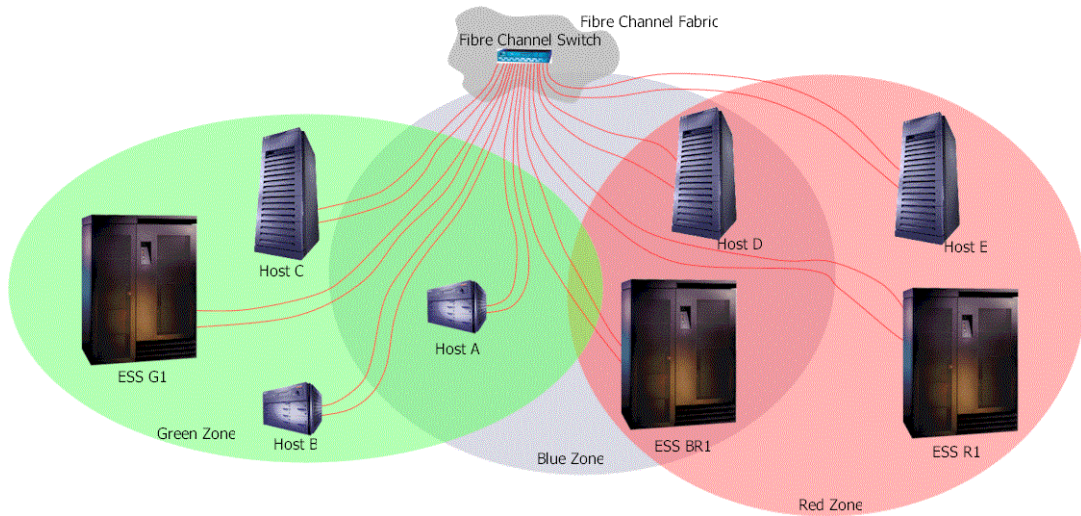


Figure 23. Multiple zones

In fact, there are probably multiple switches to attain redundancy at every level. Also, the number of pathways into the ESS in each of the zones would be multiplied to take advantage of the ESS's capabilities with regards to performance as well as availability.

2.2.4 Switch configuration

Each vendor provides multiple methods of switch configuration. The Brocade and IBM switches allow for front panel, network, and JAVA based configuration utilities.

An example of the commands necessary to create the zoning depicted in Figure 23 on page 28 within the Brocade 2400 and 2800 series switches or the IBM 2109 Series 8 and 16 port switches are available in Figure 24.

```
admin> cfgCreate "Main_cfg", "Green_zone; Blue_zone; Red_zone"
admin> zoneCreate "Green_zone", "ESS_G1; Host_A; Host_B; Host_C"
admin> zoneCreate "Blue_zone", "ESS_BR1; Host_A; Host_D"
admin> zoneCreate "Red_zone", "ESS_BR1; ESS_R1; Host_D; Host_E"
admin> aliCreate "ESS_G1", "10:00:00:00:C9:21:C0:CC;
10:00:00:00:C9:21:E2:8C"
admin> aliCreate "Host_A", "20:00:00:E0:69:40:61:B5;
20:00:00:E0:69:40:01:3B"
admin> aliCreate "Host_B", "20:00:00:E0:69:40:42:A8;
20:00:00:E0:69:40:A0:C7"
admin> aliCreate "Host_C", "20:00:00:E0:69:40:69:81;
20:00:00:E0:69:40:14:B2"
admin> aliCreate "ESS_BR1", "1,10; 1,11"
admin> aliCreate "Host_A", "20:00:00:E0:69:40:61:B5;
20:00:00:E0:69:40:01:3B"
admin> aliCreate "Host_D", "20:00:00:E0:69:40:58:01;
20:00:00:E0:69:40:29:E2"
admin> aliCreate "ESS_BR1", "1,10; 1,11"
admin> aliCreate "ESS_R1", "10:00:00:00:C9:21:B2:80;
10:00:00:00:C9:21:05:61"
admin> aliCreate "Host_D", "20:00:00:E0:69:40:58:01;
20:00:00:E0:69:40:29:E2"
admin> aliCreate "Host_E", "20:00:00:E0:69:40:A0:E9;
20:00:00:E0:69:40:38:30"
cfgEnable "Main_cfg"
zone config "Main_cfg" is in effect
admin> cfgSave
Updating flash ...
```

Figure 24. Switch configuration commands

Both switch ports and WWPNs can be used to establish members of a zone, alias, or configuration. For example, should the WWPN(s) be unknown for the ESS or any of the hosts, the switch domain — usually 1 for a single switch — and the switch port name can be used to identify the port to which the ESS or host is attached. However, while it is possible to make the configuration using only switch domain and switch port nomenclature, this is not the best way to make use of the features provided with the switch.

Using the WWPNs to configure the members of an alias allows the host HBA or ESS HA to be plugged into the switch in any location. The switch then takes care of the details outlined in the configuration to ensure the members only view the ports within their associated zones.

Aliases could be configured for each of the HBAs within a single host to allow only a part of the host capacity to exist within any zone instead of the entire host. This could be used as a sort of throttle to ensure that a particular host does not over-utilize the pathway(s) to any particular member in a zone.

It is also possible to configure a switch using built in GUIs (see Figure 25). See the vendor's documentation for more information on how to use the GUI and other switch-specific commands.



Figure 25. Brocade/IBM switch GUI

2.3 Host multi-pathing to the ESS

A simple connection to the ESS includes one host bus adapter (HBA), the cable connecting the HBA to an ESS HA (or to the Fibre Channel switch and then to the ESS). However, under this simple configuration, the host is exposed to several single points of failure (SPOF) — the ESS HA in one of the four bays, the cable (or cables and switch), and the HBA in the host (see Figure 26.).

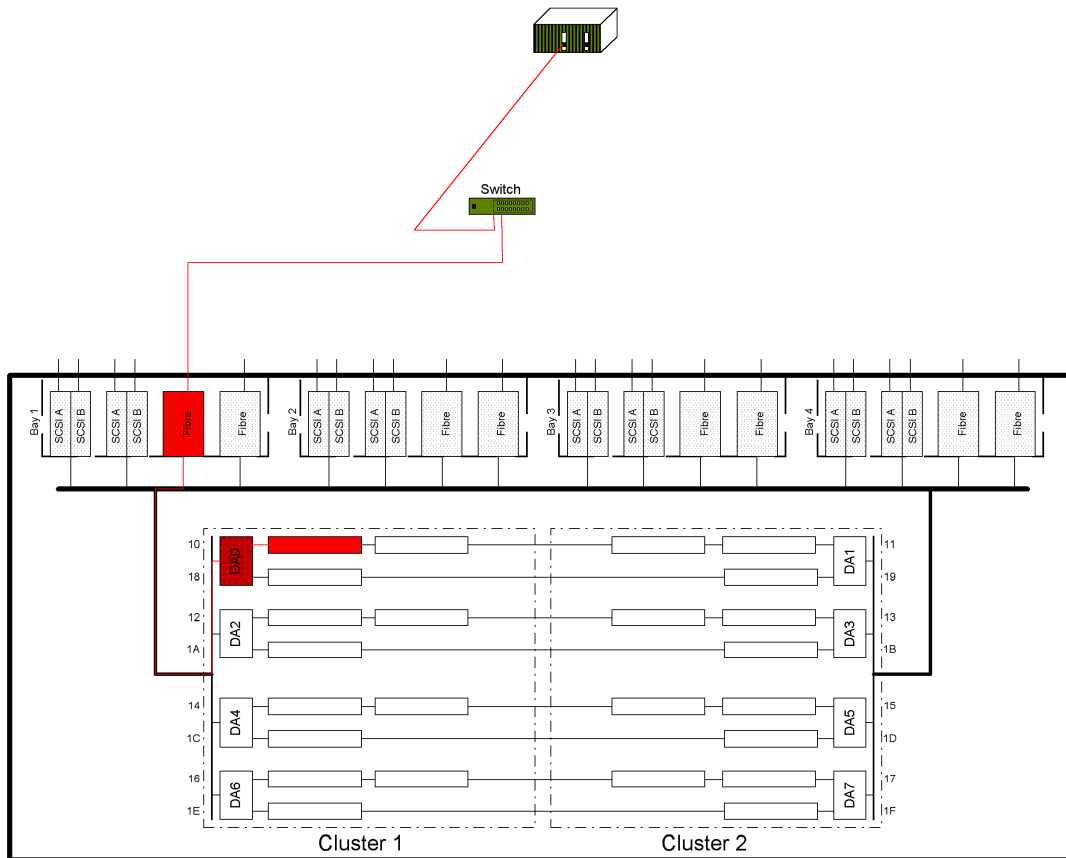


Figure 26. Single path host

Within the ESS, single points of failure have been eliminated, even down to the HA level. However, if a host is attached to a single HA and is not using some type of multi-pathing software to attach to a second HA, then failure of that single HA would cause a loss of communications between the ESS and the host.

To remedy this situation, it is recommended the host implement a second path to the data. Since the ESS can allow data to be shared across multiple HAs, the host simply duplicates HBAs, cables, and switches to access the same volumes from multiple paths (see Figure 27).



Figure 27. Multi-path host

It should be noted that a driver, software package, or other mechanism must be in place on the host end that is capable of managing multiple paths to the same data. Otherwise, data corruption *will* result in the event the host mounts the volumes on both paths simultaneously.

Typically, multi-pathing software or drivers allow a host to access the volumes from a single path. The alternate paths may be visible to the operator, but these alternate paths are almost never directly called by the operator. Instead, the software manages when the main or alternate path is utilized for data access.

As always, there are exceptions. But, for the most part, simply configuring the host to access the shared volumes via a single path is sufficient as the

multi-pathing software or drivers will take care of the rest. However, when configuring multi-pathing software or drivers, follow the vendor's recommendations.

With multi-pathing, the operating system may take advantage of the additional pathways to better distribute the I/O load. This *may* reduce the latency for the I/O requests. However, this is not guaranteed.

2.4 Volume management

A simple disk may contain up to 8 partitions. Unfortunately, with disk capacities increasing every few months, it has become increasingly difficult, if not impossible, to break up a large disk into manageable chunks using only the O/S based disk utilities (see Figure 28).

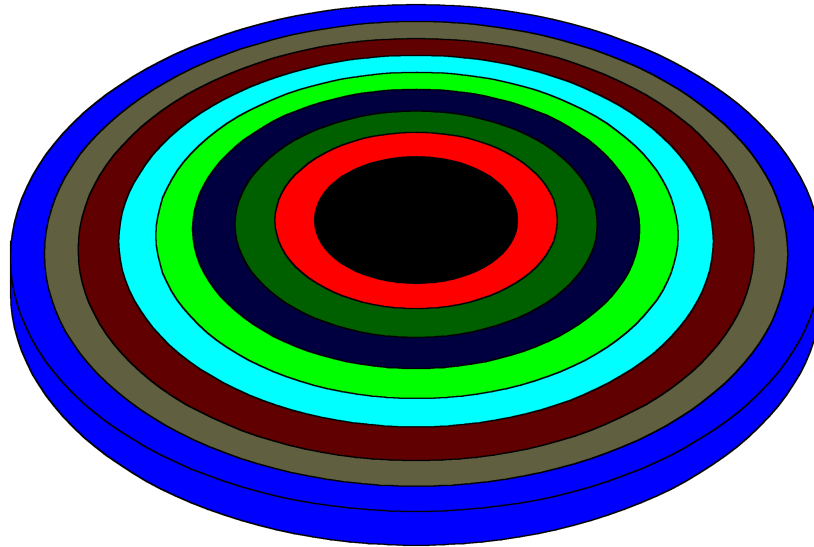


Figure 28. Disk platter with partitions

Directly opposite the requirement for small partitions is the need for extremely large storage spaces. Some applications require multi-gigabyte to terabyte storage areas that cannot be achieved with the largest disks available today. For both situations, volume management is the solution.

2.4.1 Volume manager terms

Before beginning an overview of volume management, it is necessary to understand a few terms:

DiskThe physical drive located within the ESS. These disks can be 9.1 GB, 18.2 GB, or 36.4 GB in size. In a RAID configuration, the individual disks are not visible to the host. Rather, the logical volume(s) configured on the ranks within the ESS are presented to the host. See Figure 29.

RankA group of 8 disks within the ESS. The basic building block of two ranks must be added to the ESS at a time — one rank to each cluster on the same loop. See Figure 29.



Figure 29. ESS rank and disks

Logical volumeThe virtual disk as presented by the ESS to the attached host. The size and type of a volume is limited to those definitions available within the ESS Specialist. See Figure 30.

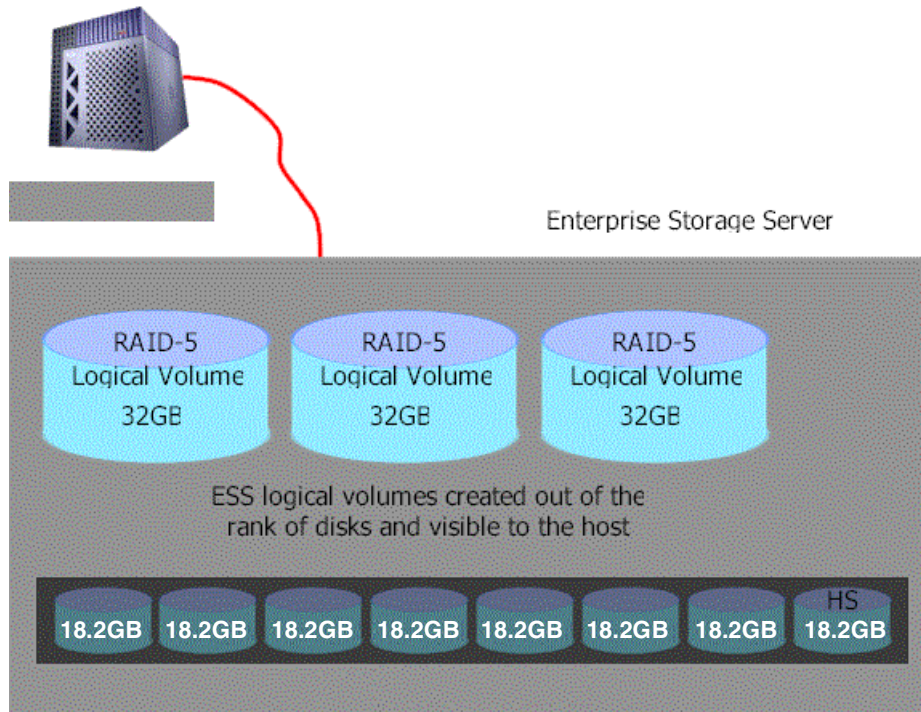


Figure 30. Logical volume example

Subdisk A portion of a managed logical volume that has been “sliced” from the managed logical volume and created as a separate entity within the volume manager application. A basic building block for the volume manager. See Figure 31.

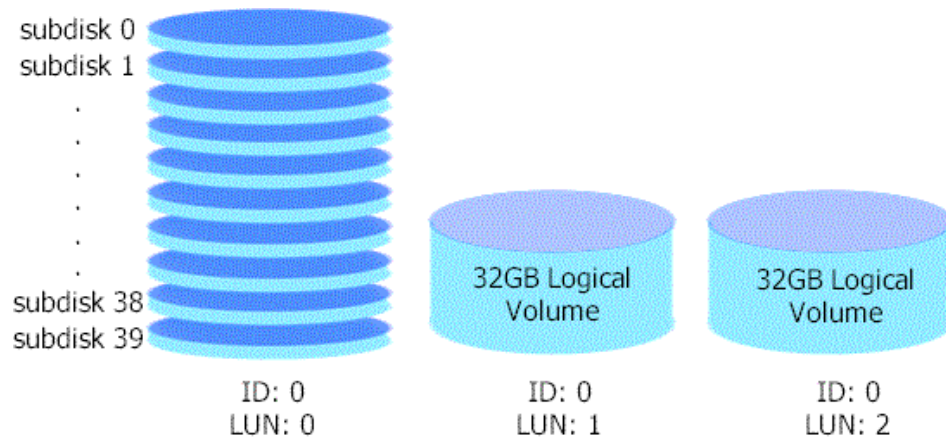


Figure 31. Subdisks created from logical volume

PlexSubdisks can be joined or combined to create larger objects called plexes. Plexes can be striped — data written in small chunks across all the sub-disks within the plex, or concatenated — data written sequentially on a sub-disk until full and then the next sub-disk until full and so on to the last sub-disk in the plex. See Figure 32.

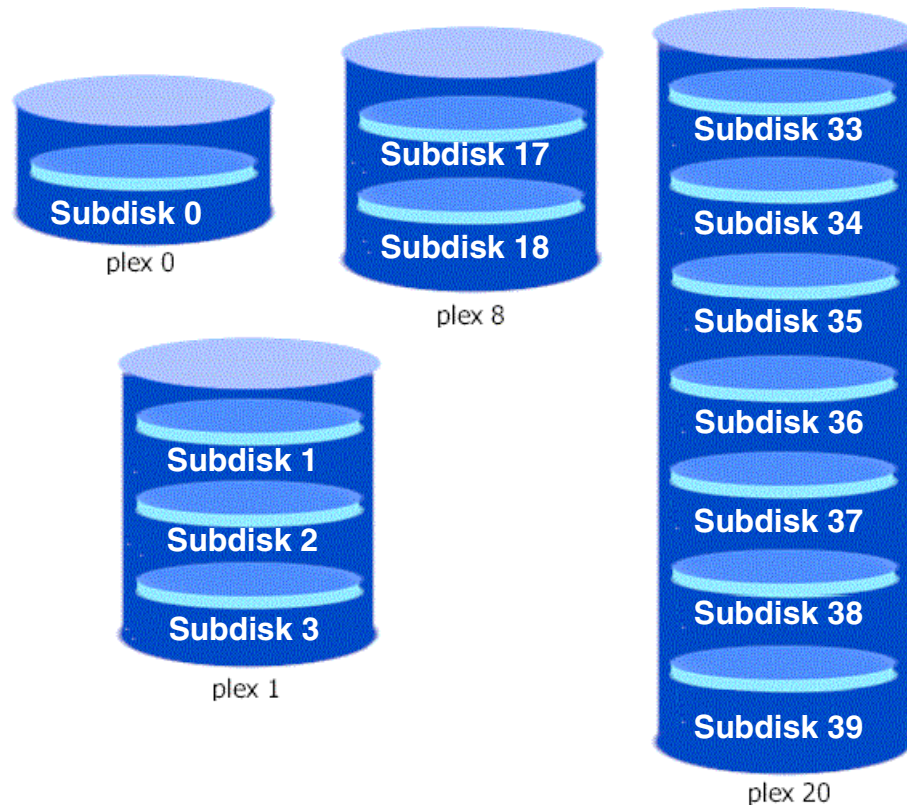


Figure 32. Subdisks within plexes

VolumeA plex cannot be used for file system creation or raw data storage. Rather, a “holder” called a volume is required to access the storage area from the O/S (See Figure 33). Once a plex is associated with a volume, it becomes possible to use the volume for either raw or file system storage. If more than one plex is associated with a volume, then a mirror of the plex(es) is created within the volume. In fact, some volume managers allow as many as 8 plexes within a single volume.

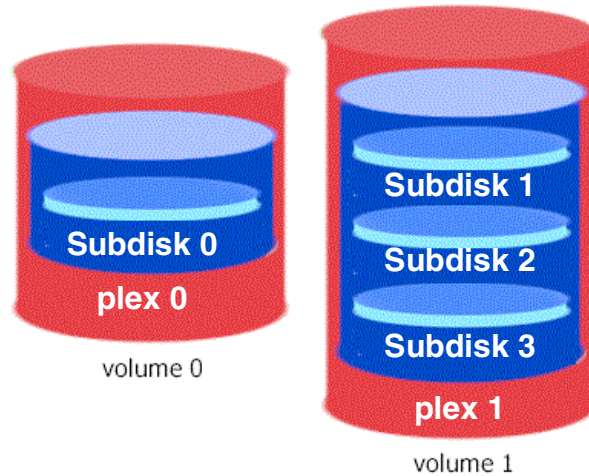


Figure 33. Plexes within volumes

2.4.2 Host volume layout and ESS pre-fetch buffers

The algorithms within the ESS anticipate when it would be most efficient to perform pre-fetching of additional sectors. When a series of consecutive reads has taken place, the ESS continues reading additional sectors to fill the pre-fetch buffers in anticipation of even more sequential reads. When the host requests data sequentially, the ESS is able to satisfy those requests from memory much faster — hundreds to thousands of times faster — than by retrieving the information from disk.

If an environment is 65/35 random/sequential I/O or greater random I/O, then striping the volumes across ranks *could* be beneficial. However, for most environments, the ESS will provide better performance to the attached hosts by simply using the largest possible slice or subdisk from a given logical volume to satisfy the total storage requirement.

In any case, a properly designed trial run using “real data” is the best method available to determine the appropriate volume configuration for each environment.

Different volume managers use different naming conventions for the components that make up a volume. However, the end result is always the same: the volume manager — a software application — manages logical volumes that have been brought under volume manager control. These logical volumes can be split into many smaller volumes or, combined together to create a few large volumes.

To learn more about volume management, see the vendor’s volume manager user guide.

Also, see *Implementing Fibre Channel Attachment on the ESS* SG246113-00, and *Implementing the Enterprise Storage Server in Your Environment*, SG245420-00.

Chapter 3. Compaq V4.0F

In this chapter we will discuss the configuration we tested with Compaq Alpha systems and Tru64 UNIX V4.0F and Cluster V1.6.

Figure 34 shows the configuration that we used for our testing.

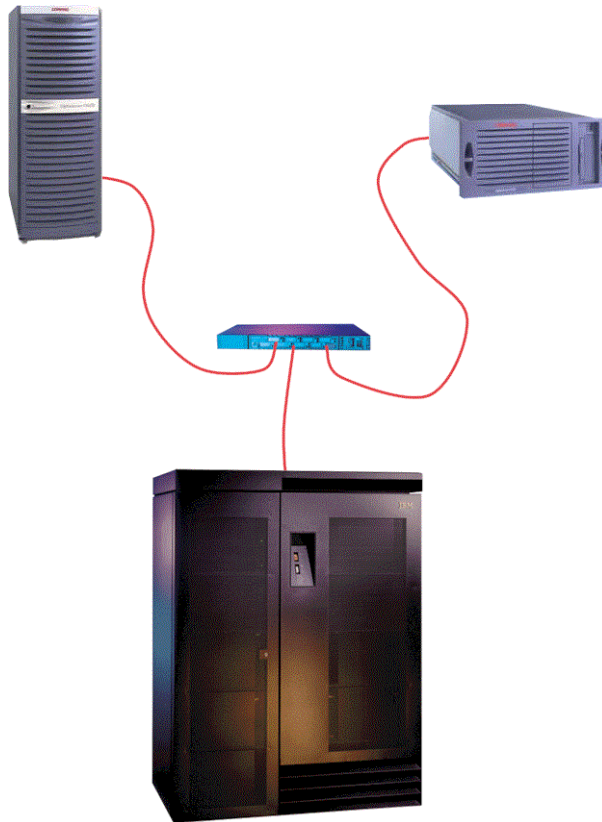


Figure 34. ESS and Compaq AlphaServer Cluster

3.1 ESS configuration

Here is the ESS configuration that was used:

- OS Level 4.3.2.15
- Code EC SC01013 kinit1013
- Emulex LP7000
- Latest firmware version for all cards

3.2 Compaq configuration

The following sections cover the Compaq configuration that was used.

3.2.1 Compaq systems tested

These are the systems we used for testing:

- Compaq AlphaServer DS20E 500 MHz SRM Console:V5.8-2
- Compaq AlphaServer GS60 6/700 MHz SRM Console:V5.8-2

3.2.2 Compaq Host Optical Fibre Adapter

The following list details the adapters we used for our testing of the systems. Any later revision of the cards should be satisfactory to use, but we have not tested them. As always with SANs, please check the manufacturer's Web site for the latest information regarding patches:

- KGPSA-BC Driver Rev 1.21 F/W Rev 2.22X1(1.13)
- KGPSA-CA Driver Rev 1.21 F/W Rev 3.02A1(1.11)
- KGPSA-CA Driver Rev 1.21 F/W Rev 3.01(1.31)

3.2.3 IBM SAN Fibre Channel Switch

We used an IBM Fibre Channel Switch:

- IBM 2109 Model S08 (PN 2109S08) F/W Rev 2.1.3
- Single Zone Configuration

3.2.4 Software versions

These are the software versions and the patch kit we used:

- Tru64 UNIX V4.0F patch kit DUV40FAS0004-20000613 OSF440
- Tru64 Cluster V1.6 patch kit DUV40FAS0004-20000613 TCR16

3.2.5 Additional software

We created some services that used the following additional software:

- Advanced File System Utilities r440
- Logical Storage Manager r444
- Apache HTTP server for Digital UNIX 1.3.9

3.3 How to check the Compaq configuration

In the following sections we explain how to find all the information you will need to check the state of the Compaq system configuration.

3.3.1 System and cards firmware revision

From the console prompt, use the command:

```
>>>show version
version V5.8-2, 21-Jul-2000 17:16:08
```

To check the system version and Fibre Channel adapter card version, you can simply read the boot messages (if boot messages scroll too rapidly, you will find them recorded in the `/var/adm/messages` file):

```
#view /var/adm/messages
```

Here is how the firmware version is referenced:

```
Sep 21 14:17:10 osplcpq-ds20 vmunix: Firmware revision: 5.8
```

Here is how the Fibre Channel host adapter is referenced:

```
Sep 21 14:17:10 osplcpq-ds20 vmunix: KGPSA adapter: Driver Rev 1.09 : F/W
Rev 2.22X1(1.13) : wwn 1000-0000-c922-d469
```

3.3.2 Operating system version

Use the command `uname`:

```
#uname -a
OSF1 osplcpq-ds20 V4.0 1229 alpha
```

Revision 1229 corresponds to V4.0F, or again in the `/var/adm/messages` file:

```
Sep 21 14:17:09 osplcpq-ds20 vmunix: Digital UNIX V4.0F (Rev. 1229);
```

3.3.3 Cluster software version

The output of the following command will show you if the TruCluster is installed, the central column will report “installed” if the product is installed, and a blank value if it is not; the output of the command varies if you have an Available Server or a Production Server environment. For both configurations, the subsets must end with 160.

```
#setld -i | grep TCR
TCRASE160    installed  TruCluster Available Server Software
TCRCMS160    installed  TruCluster Cluster Monitor
TCRCOMMON160 installed  TruCluster Common Components
```

```
TCRCNF160    installed  TruCluster Configuration Software
TCRMAN160    installed  TruCluster Reference Pages
```

3.3.4 Patches installed on the system

The utility dupatch is the one that manages the system and cluster patches:

```
# dupatch
    * Previous session logs saved in session.log.[1-25]
Tru64 UNIX Patch Utility (Rev. 27-04)
=====
    - This dupatch session is logged in /var/adm/patch/log/session.log
Main Menu:
1) Patch Installation
2) Patch Deletion
3) Patch Documentation
4) Patch Tracking
5) Patch Baseline Analysis/Adjustment
h) Help on Command Line Interface
q) Quit
Enter your choice: 4
Tru64 UNIX Patch Utility (Rev. 27-04)
=====
    - This dupatch session is logged in /var/adm/patch/log/session.log
Patch Tracking Menu:
-----
1) List installed patches
2) List installed patch files
3) List patch kit information on installed patches
b) Back to Main Menu
q) Quit
Enter your choice: 3
    Patches installed on the system came from following patch kits:
-----
    - DUV40FAS0002-19991116 OSF440
    - DUV40FAS0004-20000613 OSF440
    - DUV40FAS0004-20000613 TCR160
```

3.3.5 Disk configuration

There are no special operations that have to be done on the Compaq system to make it see the ESS volumes; just be sure that the host Fibre Channel is already configured (if it is not, you need to follow the instructions that come with the card and install the driver). Make all the hardware connections, configure the ESS, and then reboot the system. If the disks are seen at boot time, all the special files are created automatically. No kernel rebuild is necessary.

Once the ESS is configured, to check if the disks are seen correctly from the Compaq system, you must wait for the system to boot in order to have the driver of the HBA loaded. During boot time, you should see the following messages (if boot messages scroll too rapidly, you will find them recorded in the `/var/adm/messages` file):

Note: This is an extract of the complete boot message:

```
emx0 at pci1 slot 7 KGPSA-BC : Driver Rev 1.21 : F/W Rev 2.22X1(1.13) : wwn
1000-0000-c922-d469
emx0: emx_assign_fcp_id: nport at DID 0x21300 assigned tgt id 10 - out of
range for CAM
scsi16 at emx0 slot 0
rz128 at scsi16 target 0 lun 0 (LID=0) (IBM 2105F20 1013)
rzb128 at scsi16 target 0 lun 1 (LID=1) (IBM 2105F20 1013)
rzc128 at scsi16 target 0 lun 2 (LID=2) (IBM 2105F20 1013)
rzd128 at scsi16 target 0 lun 3 (LID=3) (IBM 2105F20 1013)
rze128 at scsi16 target 0 lun 4 (LID=4) (IBM 2105F20 1013)
rzf128 at scsi16 target 0 lun 5 (LID=5) (IBM 2105F20 1013)
rzg128 at scsi16 target 0 lun 6 (LID=6) (IBM 2105F20 1013)
rzh128 at scsi16 target 0 lun 7 (LID=7) (IBM 2105F20 1013)
```

After the boot completes, login as root and check if all the special files were created.

All special files are under the `/dev` directory; the disk block special file starts with `rz` and the character disk special file starts with `rrz`.

Compaq Tru64 UNIX uses the following syntax to identify the disk's special files:

`[r]rz[L][B][P]`

L = LUN letter; if the LUN is 0, no letter; from LUN 2 to 7, the letters b to h

B = bus number * 8 + target number

P = disk partition from a to h

Example: From the output of the boot log, we have a disk at LUN 0, bus SCSI 16, target 0, so the special files will be:

`[r]rz[LUN 0][16 * 8 + 0][a-h] = rz128a, rz128b, rz128c, rz128d, rz128e, rz128f, rz128g, rz128h, rrz128a, rrz128b, rrz128c, rrz128d, rrz128e, rrz128f, rrz128g, rrz128h`

We use the `file` command to check the major number. If there is a description of the disk, you will not find the word IBM, but only the model of the disk (use the information from the output of the previous command):

```
# file /dev/rrz*128c
/dev/rrz128c: character special (8/262146) SCSI #16 2105F20 disk #1024
(SCSI ID #0) (SCSI LUN #0)
/dev/rrzbl128c: character special (8/262210) SCSI #16 2105F20 disk #1025
(SCSI ID #0) (SCSI LUN #1)
/dev/rrzcl128c: character special (8/262274) SCSI #16 2105F20 disk #1026
(SCSI ID #0) (SCSI LUN #2)
/dev/rrzdl128c: character special (8/262338) SCSI #16 2105F20 disk #1027
(SCSI ID #0) (SCSI LUN #3)
/dev/rrzel128c: character special (8/262402) SCSI #16 2105F20 disk #1028
(SCSI ID #0) (SCSI LUN #4)
/dev/rrzfl128c: character special (8/262466) SCSI #16 2105F20 disk #1029
(SCSI ID #0) (SCSI LUN #5)
/dev/rrzgl128c: character special (8/262530) SCSI #16 2105F20 disk #1030
(SCSI ID #0) (SCSI LUN #6)
/dev/rrzhl128c: character special (8/262594) SCSI #16 2105F20 disk #1031
(SCSI ID #0) (SCSI LUN #7)
```

If a special file does not have a disk associated with it, the output of the `file` command will have only the part concerning the major number; no disk model will be displayed:

```
/dev/rrzh128c: character special (8/262594)
```

Finally, check if the disks have a valid disk label, we do this with the `disklabel` command:

```
# disklabel rz128
# /dev/rrz128a:
type: SCSI
disk: 2105F20
label:
flags: dynamic_geometry
bytes/sector: 512
sectors/track: 64
tracks/cylinder: 30
sectors/cylinder: 1920
cylinders: 1017
sectors/unit: 1953152
rpm: 7200
interleave: 1
trackskew: 0
cylinderskew: 0
headswitch: 0          # milliseconds
```

```

track-to-track seek: 0 # milliseconds
drivedata: 0
8 partitions:
#      size      offset  fstype  [fsize bsize  cpg] #
a:    131072      0    unused      0    0 # (Cyl.  0 - 68*)
b:    262144    131072  unused      0    0 # (Cyl. 68*- 204*)
c:    1953152      0    unused      0    0 # (Cyl.  0 - 1017*)
d:         0      0    unused      0    0 # (Cyl.  0 - -1)
e:         0      0    unused      0    0 # (Cyl.  0 - -1)
f:         0      0    unused      0    0 # (Cyl.  0 - -1)
g:    819200    393216  unused      0    0 # (Cyl. 204*- 631*)
h:    740736    1212416  unused      0    0 # (Cyl. 631*- 1017*)

```

If there is no disklabel on the disk you must use the following command to write it to the disk:

```
# disklabel -rw rzh128 shark
```

Substitute rzh128 with your specific disk, and always specify the rz file without the partition.

Now the disks can be used by Advanced File System (AdvFS), LSM, and Tru64 Cluster

3.4 Tru64 cluster configuration

To set up the cluster software and service, refer to the Compaq manuals:

- TruCluster Software Installation
- TruCluster Administration

3.4.1 Example — disk service creation

To set up a disk service, the shared storage must be visible to all the members of the cluster from the file, disklabel commands. If you are using the storage through a file system or LSM, you must configure these parts on only one system (the cluster will then share the definitions).

This service is called `gs60_mount` and is made of two Advfs file systems. One goes directly to the storage. The second one is made of an LSM mirrored volume. Respectively, the two Advfs file systems are called `shark4_dm#s4` mounted on `/s4` and `shark7_dm#s7` mounted on `/s7` (LSM volume name `/dev/vol/sharkdg/vol01`).

```

#asemgr
ASE Main Menu
  a) Managing the ASE      -->

```

- m) Managing ASE Services -->
- s) Obtaining ASE Status -->
- x) Exit ?) Help

Enter your choice: **m**

Managing ASE Services

- c) Service Configuration -->
- r) Relocate a service
- on) Set a service on line
- off) Set a service off line
- res) Restart a service
- s) Display the status of a service
- a) Advanced Utilities -->
- q) Quit (back to the Mai
- x) Exit ?) Help

Enter your choice [q]: **c**

Service Configuration

- a) Add a new service
- m) Modify a service
- o) Modify a service without interrupting its availability
- d) Delete a service
- s) Display the status of a service
- c) Display the configuration of a service
- q) Quit (back to Managing ASE Services)
- x) Exit ?) Help

Enter your choice [q]: **a**

Adding a service

Select the type of service:

- 1) NFS service
- 2) Disk service
- 3) User-defined service
- 4) Tape service
- q) Quit without adding a service
- x) Exit ?) Help

Enter your choice [1]: **2**

You are now adding a new disk service to ASE.

A disk service consists of a disk-based application and disk configuration that are failed over together. The disk configuration can include UFS filesystems, AdvFS filesets, LSM volumes, or raw disk information.

Disk Service Name

The name of a disk service must be a unique service name. Optionally, an IP address may be assigned to a disk service. In this case, the name must be a unique IP host name set up for this service and present in the local hosts database on all ASE members.

Enter the disk service name ('q' to quit): **gs60_mount**


```

Assign an IP address to this service? (y/n): n
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end): shark4_dm#s4
AdvFS domain `shark4_dm` has the following volume(s):
/dev/rzd128c
Is this correct (y/n) [y]: y
      Mount Point
The mount point is the directory on which to mount `shark4_dm#s4`
If you do not want it mounted, enter "NONE".
Enter the mount point or NONE: /s4
      AdvFS Fileset Read-Write Access and Quota Management
Mount `shark4_dm#s4` fileset with read-write or read-only access?
  1) Read-write
  2) Read-only
Enter your choice [1]: 1
You may enable user, and group and fileset quotas on this file system by
specifying the full pathnames for the quota files. Quota files must reside
within the fileset. Enter "none" to disable quotas.
User quota file path [/s4/quota.user]:
Group quota file path [/s4/quota.group]:
      AdvFS Mount Options Modification
Enter a comma-separated list of any mount options you want to use for
the `shark4_dm#s4` fileset (in addition to the defaults listed in the
mount.8 reference page). If none are specified, only the default mount
options are used.
Enter options (Return for none):
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end):shark7_dm#s7
AdvFS domain `shark7_dm` has the following volume(s):
/dev/vol/sharkdg/vol01
Is this correct (y/n) [y]: y

```

```

Following is a list of device(s) and pubpath(s) for disk group sharkdg:
DEVICE PUBPATH
rzg128 /dev/rzgL28g
rzh128 /dev/rzhL28g
Is this correct (y/n) [y]: y
      Mount Point
The mount point is the directory on which to mount `shark7_dm#s7`.
If you do not want it mounted, enter "NONE".
Enter the mount point or NONE: /s7
      AdvFS Fileset Read-Write Access and Quota Management
Mount `shark7_dm#s7` fileset with read-write or read-only access?
  1) Read-write
  2) Read-only
Enter your choice [1]: 1
You may enable user, and group and fileset quotas on this file system by
specifying the full pathnames for the quota files. Quota files must reside
within the fileset. Enter "none" to disable quotas.
User quota file path [/s7/quota.user]:
Group quota file path [/s7/quota.group]:
      AdvFS Mount Options Modification
Enter a comma-separated list of any mount options you want to use for
the `shark7_dm#s7` fileset (in addition to the defaults listed in the
mount.8 reference page). If none are specified, only the default mount
options are used.
Enter options (Return for none):
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
  For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end):
Modifying user-defined scripts for `gs60_mount`:
  1) Start action
  2) Stop action
  3) Add action
  4) Delete action
  x) Exit - done with changes
Enter your choice [x]: x
      Selecting an Automatic Service Placement (ASP) Policy
Select the policy you want ASE to use when choosing a member
to run this service:
  b) Balanced Service Distribution
  f) Favor Members

```

```

r) Restrict to Favored Members
? ) Help
Enter your choice [b]: f
Selecting an Automatic Service Placement (ASP) Policy
Select the favored member(s) IN ORDER for service 'gs60_mount':
1) ds20-ee1
2) gs60-ee1
q) No favored members
? ) Help
Enter a comma-separated list [q]: 2,1
Selecting an Automatic Service Placement (ASP) Policy
Do you want ASE to relocate this service to a more highly favored member
if one becomes available while this service is running (y/n/?): n

Enter 'y' to add Service 'gs60_mount' (y/n): y
Adding service...
Starting service...
Service gs60_mount successfully added...

```

Once the disk service is configured, you will see the storage mounted only on the cluster member that is managing the service.

3.4.2 Example — Apache disk service creation

Before creating the service, the shared disks must be visible to all the members of the cluster, and on one node you have to configure AdvFS (and eventually LSM) to have a file system used by the Apache service. Then run `asemgr` to define the disk service with an IP address (the IP address must be already configured in the `/etc/hosts` file on all the cluster members that will run the service). Apache must be installed on both systems; following this, you will find information on how we installed it.

The service start script and stop script must run the Apache stop and start script. In this case, the installed script is `/var/ase/startup/apache_svc`, and you pass to it the start option or stop option.

In the following example, the service is called `apachetst` and we included in the service two Advfs domains: one that holds the HTML pages (`shark2_dm#s2` mounted on `/s2`), and the other the temporary data (`shark3_dm#s3` mounted on `/s3`; this second domain is optional). We first added the disk service to the cluster, and relocated the service a couple of times to check that the disks were managed correctly by both members of the cluster. Then we added and edited the start script (the best way to do this it to add the default script and then edit it via `asemgr`), and checked that it was working fine. Then we edited the stop script.

Remember to manually stop the Apache service before editing the stop script; otherwise, the service will not stop correctly.

3.4.2.1 Apache setup

The Apache Web server can be downloaded from www.apache.org. It was found at <http://httpd.apache.org/dist/binaries/digitalunix>. The level installed on the test machine was 1.3.9. The gzipped file was 2.5MB.

Once you download it and run `gzip -d <filename>` to uncompress it, untar the resulting file. It creates its own directory relative to the directory from which you run `tar -xvf <filename>`.

By default, it installs in `/usr/local/apache`. You can choose to install it elsewhere, but first read the documentation for modifying the `httpd.conf` file.

For the test, the Apache server was installed on `/usr/local/apache` on both systems, and ESS storage was used for the document (HTML) directory.

An alias was set up in the Apache startup service. It was through this alias that clients accessed the Web pages from whichever server currently had it mounted.

3.4.2.2 Apache httpd.conf

The following edits need to be made to the `httpd.conf` file. Bold text is the option text, which should remain as shown. Italicized text represents descriptive text of the information to change/edit. This is because the `httpd` daemon is started external to the normal `inittab` or `inet` startup.

ServerType *standalone*

This points to the location of the HTML on the shared storage.

DocumentRoot */mountpoint/base_html_doc_directory*

This is the block that enables access to the directory structure below the `<html doc directory>`, and above the default line (exactly as it appears here, with brackets).

`<Directory "mountpoint/base_html_doc_directory">`

The block continues with access options.

`</Directory>`

This concludes the block.

The option `ServerRoot` should be left unchanged as long as the server is installed in the default location.

ServerRoot `"/usr/local/apache"`

Other options, such as the **Port** on which the server runs, any **Alias** or **ScriptAlias**, and security options, may be set up at the option of the installer.

3.4.2.3 Apache start/stop script

We used the file `/var/ase/examples_unsupported/apache_svc` that is installed by the cluster software, copied it under `/var/ase/startup`, and modified it. The modifications are in bold.

The lines that start with a '#' sign are comments and are therefore omitted.

```
# 8<-----8<----- Start Custom variables
svcName="apachetst"           # Servicename
TCR_ADMIN="root"              # Account to receive ASE mail
DIR="/usr/local/apache/logs"  # Directory for logfiles
ACTION=$1                     # Action (either start or stop)
LOG="$DIR}/${ACTION}_${svcName}.$$" # Destination for script output
#LOG="/dev/console"
FUSER="/usr/sbin/fuser"      # Command to use for closing open
files
START_APPCMD="/usr/local/apache/bin/httpd" #-d /cludemo/apache -f
conf/httpd.conf"
# Application startup cmd
STOP_APPCMD="zapdaemon httpd" # Application stop cmd
APPDIR="/usr/local/apache"    # Application home directory
ADVFS_DIRS="/usr/local/apache" # Application directories to
close open files on
export START_APPCMD START_APPCMD2 STOP_APPCMD STOP_APPCMD2 APPDIR
export ADVFS_DIRS
## 8<-----8<----- End Custom variables
```

3.4.2.4 The asemgr session

Here is a simple login of the `asemgr` session:

```
# asemgr
TruCluster Available Server (ASE)
ASE Main Menu
  a) Managing the ASE          -->
  m) Managing ASE Services    -->
  s) Obtaining ASE Status     -->
  x) Exit                      ?) Help
```

Enter your choice: **m**

```
Managing ASE Services
  c) Service Configuration  -->
  r) Relocate a service
  on) Set a service on line
  off) Set a service off line
  res) Restart a service
  s) Display the status of a service
  a) Advanced Utilities  -->
  q) Quit (back to the Main Menu)
  x) Exit                ?) Help
Enter your choice [q]: c
```

```
Service Configuration
  a) Add a new service
  m) Modify a service
  o) Modify a service without interrupting its availability
  d) Delete a service
  s) Display the status of a service
  c) Display the configuration of a service
  q) Quit (back to Managing ASE Services)
  x) Exit                ?) Help
Enter your choice [q]: a
```

```
Adding a service
Select the type of service:
  1) NFS service
  2) Disk service
  3) User-defined service
  4) Tape service
  q) Quit without adding a service
  x) Exit                ?) Help
Enter your choice [1]: 2
```

You are now adding a new disk service to ASE.
A disk service consists of a disk-based application and disk configuration that are failed over together. The disk configuration can include UFS filesystems, AdvFS filesets, LSM volumes, or raw disk information.

Disk Service Name

The name of a disk service must be a unique service name. Optionally, an IP address may be assigned to a disk service. In this case, the name must be a unique IP host name set up for this service and present in the local hosts database on all ASE members.

```
Enter the disk service name ('q' to quit): apachetst
Assign an IP address to this service? (y/n): y
```

```

Checking to see if apachetst is a valid host...
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
      For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end): shark2_dm#s2

AdvFS domain 'shark2_dm' has the following volume(s):
/dev/rzb128c
Is this correct (y/n) [y]: y
      Mount Point
The mount point is the directory on which to mount 'shark2_dm#s2'.
If you do not want it mounted, enter "NONE".
Enter the mount point or NONE: /s2
      AdvFS Fileset Read-Write Access and Quota Management
Mount 'shark2_dm#s2' fileset with read-write or read-only access?
1) Read-write
   2) Read-only
Enter your choice [1]: 1
You may enable user, and group and fileset quotas on this file system by
specifying the full pathnames for the quota files. Quota files must reside
within the fileset. Enter "none" to disable quotas.
User quota file path [/s2/quota.user]:
Group quota file path [/s2/quota.group]:
      AdvFS Mount Options Modification
Enter a comma-separated list of any mount options you want to use for
the 'shark2_dm#s2' fileset (in addition to the defaults listed in the
mount.8 reference page). If none are specified, only the default mount
options are used.
Enter options (Return for none):
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
      For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end): shark3_dm#s3
AdvFS domain 'shark3_dm' has the following volume(s):
/dev/rzc128c

```

```

Is this correct (y/n) [y]: y
      Mount Point
The mount point is the directory on which to mount 'shark3_dm#s3'.
If you do not want it mounted, enter "NONE".
Enter the mount point or NONE: /s3
      AdvFS Fileset Read-Write Access and Quota Management
Mount 'shark3_dm#s3' fileset with read-write or read-only access?
  1) Read-write
  2) Read-only
Enter your choice [1]: 1
You may enable user, and group and fileset quotas on this file system by
specifying the full pathnames for the quota files. Quota files must reside
within the fileset. Enter "none" to disable quotas.
User quota file path [/s3/quota.user]:
Group quota file path [/s3/quota.group]:
      AdvFS Mount Options Modification
Enter a comma-separated list of any mount options you want to use for
the 'shark3_dm#s3' fileset (in addition to the defaults listed in the
mount.8 reference page). If none are specified, only the default mount
options are used.
Enter options (Return for none):
      Specifying Disk Information
Enter one or more device special files, AdvFS filesets, or LSM volumes
to define the disk storage for this service.
  For example:
Device special file:      /dev/rz3c
AdvFS fileset:           domain1#set1
LSM volume:              /dev/vol/dg1/vol01
To end the list, press the Return key at the prompt.
Enter a device special file, an AdvFS fileset, or an LSM volume as storage
for this service (press 'Return' to end):
Modifying user-defined scripts for 'apachetst':
  1) Start action
  2) Stop action
  3) Add action
  4) Delete action
  x) Exit - done with changes

Enter your choice [x]: x
Selecting an Automatic Service Placement (ASP) Policy
Select the policy you want ASE to use when choosing a member
to run this service:
  b) Balanced Service Distribution
  f) Favor Members
  r) Restrict to Favored Members
  ?) Help

Enter your choice [b]: f

```


Selecting an Automatic Service Placement (ASP) Policy
 Select the favored member(s) IN ORDER for service 'apachetst':

- 1) ds20-ee1
- 2) gs60-ee1
- q) No favored members

?) Help

Enter a comma-separated list [q]: 1,2

Selecting an Automatic Service Placement (ASP) Policy

Note: the editing session of both start/stop scripts are omitted for editing issues, in the following paragraphs you'll find the final version of the scripts.

3.4.2.5 Service start script

The lines that start with a '#' sign are comments and are therefore omitted.

```
PATH=/sbin:/usr/sbin:/usr/bin
export PATH
ASETMPDIR=/var/ase/tmp

if [ $# -gt 0 ]; then
    /var/ase/startup/apache_svc start
#     svcName=$1           # Service name to start
else
    svcName=
fi
#
# Any non zero exit will be considered a failure.
#
exit 0
```

3.4.2.6 Service stop script

The lines that start with a '#' sign are comments and are therefore omitted.

```
PATH=/sbin:/usr/sbin:/usr/bin
export PATH
ASETMPDIR=/var/ase/tmp
if [ $# -gt 0 ]; then
    /var/ase/startup/apache_svc stop
#     svcName=$1           # Service name to stop
else
    svcName=
fi
case "${MEMBER_STATE}" in
BOOTING)      # Stopping ${svcName} as ASE member boots.
    ;;
RUNNING)     # This is a true stop of ${svcName}.
    ;;

```

```
esac
exit 0
```

3.5 LSM and ADVFS sample configuration

LSM must be installed on all the members of the cluster and must be configured (`rootdg` must be present on all nodes of the cluster).

For detailed information refer to the system administrator manuals:

- *Using Logical Storage Manager in a Cluster*, Order Number: AA-RHGYB-TE this Compaq book can be obtained from the Web site:

http://www.tru64unix.compaq.com/faqs/publications/cluster_doc/cluster_51/HTML/ARHGYC

- *Advanced File System and Utilities for Digital UNIX*, Order Number: AA-QTPZA-TE. This Compaq book can be obtained from the Web site:

http://tru64unix.compaq.com/faqs/publications/base_doc/DOCUMENTATION/V40G_SUP_PDF/AD

Following you will find an example showing how to create a separate disk group for the ESS volumes and how to build small LSM volumes. We also added an example on how to create mirrored volumes. Then we are going to build an Advanced File System (Advfs) on top of the LSM volumes.

Once Advfs and LSM are set up, you can pass the name on the file system to the `asemgr` utility during a service definition.

3.5.1 LSM disk initialization and disk group creation

A utility called `voldiskadd` automatically adds a disk to a disk group, and if the disk group is not present, it creates one.

In the following example, we add the disk `rz128` to the disk group `sharkdg`. Since it is not already present, the utility asks us if we want to create it. We say *yes* and define the internal LSM disk name equal to the physical name; this avoids confusion when we need to track down the disk's configuration.

Then we are going to add a second disk (`rzd128`) to the same disk group.

We also added a third disk (`rze128`), but for editing reasons, the output of that operation is not included; there are no differences with the `rzd128` output.

```
# voldiskadd
Add or initialize a disk
Menu: LogicalStorageManager/Disk/AddDisk
```

Use this operation to add a disk to a disk group. You can select an existing disk group or create a new disk group. You can also initialize a disk without adding it to a disk group, which leaves the disk available for use as a replacement disk. This operation takes, as input, a disk device or partition and a disk group (or none to leave the disk available for use as a replacement disk). If you are adding the disk to a disk group, you will be asked to give a name to the disk.

Select disk device to add [<disk/partition name>,list,q,?] **rz128**
 You can choose to add this disk to an existing disk group, to create a new disk group, or you can choose to leave the disk available for use by future add or replacement operations. To create a new disk group, select a disk group name that does not yet exist. To leave the disk available for future use, specify a disk group name of "none".

Which disk group [<group>,none,list,q,?] (default: rootdg) **sharkdg**
 There is no active disk group named sharkdg.

Create a new group named sharkdg? [y,n,q,?] (default: y) **y**
 You must now select a disk name for the disk. This disk name can be specified to disk removal, move, or replacement operations. If you move the disk, such as between host bus adapters, the disk will retain the same disk name, even though it will be accessed using a different disk device name.

Enter disk name [<name>,q,?] (default: sharkdg01) **rz128**
 The requested operation is to initialize disk device rz128 and to create a new disk group named sharkdg containing this disk. The disk will be named rz128 within the disk group.
 Approximate maximum number of physical disks that will be added to the sharkdg diskgroup.

Number of disks [e.g. 5,10,30,60,q,?] (default: 10) **10**
 Continue with operation? [y,n,q,?] (default: y) **y**
 Disk initialization for rz128 completed successfully.
 Add or initialize another disk? [y,n,q,?] (default: n) **y**
 Add or initialize a disk

Menu: LogicalStorageManager/Disk/AddDisk
 Use this operation to add a disk to a disk group. You can select an existing disk group or create a new disk group. You can also initialize a disk without adding it to a disk group, which leaves the disk available for use as a replacement disk. This operation takes, as input, a disk device or partition and a disk group (or none to leave the disk available for use as a replacement disk). If you are adding the disk to a disk group, you will be asked to give a name to the disk.

Select disk device to add [<disk/partition name>,list,q,?] **rz128**
 You can choose to add this disk to an existing disk group, to create a new disk group, or you can choose to leave the disk available for use by future add or replacement operations. To

create a new disk group, select a disk group name that does not yet exist. To leave the disk available for future use, specify a disk group name of "none".

Which disk group [<group>,none,list,q,?] (default: rootdg) **sharkdg**

You must now select a disk name for the disk. This disk name can be specified to disk removal, move, or replacement operations. If you move the disk, such as between host bus adapters, the disk will retain the same disk name, even though it will be accessed using a different disk device name.

Enter disk name [<name>,q,?] (default: sharkdg01) **rzbl28**

The requested operation is to initialize disk device rzbl28 and to add this device to disk group sharkdg as disk rzbl28.

Continue with operation? [y,n,q,?] (default: y) **y**

Once the disks are defined in LSM, we can see them using the commands *voldisk* and *volprint*.

```
# voldisk list
DEVICE      TYPE      DISK      GROUP      STATUS
rz128       sliced   rz128     sharkdg    online
rz33        sliced   rz33      rootdg     online
rzd128      sliced   rzd128    sharkdg    online
rze128      sliced   rze128    sharkdg    online
rzg128      sliced   -         -          online
rzh128      sliced   -         -          online

# volprint -g sharkdg -ht
DG NAME      GROUP-ID
DM NAME      DEVICE      TYPE      PRIVLEN   PUBLEN   PUBPATH
V NAME      USETYPE     KSTATE    STATE     LENGTH   READPOL   PREFPLEX
PL NAME      VOLUME      KSTATE    STATE     LENGTH   LAYOUT    ST-WIDTH
MODE
SD NAME      PLEX        PLOFFS    DISKOFFS  LENGTH   DISK-NAME  DEVICE
dg sharkdg   973124814.1279.osplcpq-ds20
dm rz128     rz128       sliced    1024     1952112  /dev/rrz128g
dm rzd128    rzd128      sliced    1024     1952112  /dev/rrzd128g
dm rze128    rze128      sliced    1024     1952112  /dev/rrze128g
```

3.5.2 LSM volume creation

To build the volumes, we are going to use the *volassist* utility. Remember that in LSM, there are always several ways to obtain the same thing. In this case, we are going to use a top-down utility that will automatically create for us several LSM objects.

We are going to pass to the utility the name and size of the disk on which we want the volume to be created. The volume name is *vol101*, the size is 50 MB, and we are going to put it on disk *rz128*.

```

# volassist -g sharkdg make vol01 50m rz128
# volprint -g sharkdg -ht
DG NAME          GROUP-ID
DM NAME          DEVICE      TYPE      PRIVLEN  PUBLEN  PUBPATH
V NAME          USETYPE     KSTATE    STATE    LENGTH  READPOL  PREFPLEX
PL NAME          VOLUME      KSTATE    STATE    LENGTH  LAYOUT   ST-WIDTH
MODE
SD NAME          PLEX        PLOFFS    DISKOFFS LENGTH  DISK-NAME  DEVICE
dg sharkdg      973124814.1279.osplcpq-ds20
dm rz128        rz128       sliced    1024     1952112 /dev/rrz128g
dm rzd128       rzd128      sliced    1024     1952112 /dev/rrzd128g
dm rze128       rze128      sliced    1024     1952112 /dev/rrze128g
v vol01         fsgen       ENABLED   ACTIVE   102400   SELECT   -
pl vol01-01     vol01       ENABLED   ACTIVE   102400   CONCAT   -          RW
sd rz128-01     vol01-01    0         0        102400   rz128    rz128

```

3.5.3 LSM mirrored volume creation

Following are two ways of mirroring a volume:

- Mirror an existing LSM volume.
- Define at the volume creation time that it is mirrored.

3.5.3.1 Mirror an existing LSM volume

Following you will find the commands needed to create LSM mirrored volumes.

```

# volassist -g sharkdg mirror vol01 rzd128
# volprint -g sharkdg -ht
DG NAME          GROUP-ID
DM NAME          DEVICE      TYPE      PRIVLEN  PUBLEN  PUBPATH
V NAME          USETYPE     KSTATE    STATE    LENGTH  READPOL  PREFPLEX
PL NAME          VOLUME      KSTATE    STATE    LENGTH  LAYOUT   ST-WIDTH
MODE
SD NAME          PLEX        PLOFFS    DISKOFFS LENGTH  DISK-NAME  DEVICE
dg sharkdg      973124814.1279.osplcpq-ds20
dm rz128        rz128       sliced    1024     1952112 /dev/rrz128g
dm rzd128       rzd128      sliced    1024     1952112 /dev/rrzd128g
dm rze128       rze128      sliced    1024     1952112 /dev/rrze128g
v vol01         fsgen       ENABLED   ACTIVE   102400   SELECT   -
pl vol01-01     vol01       ENABLED   ACTIVE   102400   CONCAT   -          RW
sd rz128-01     vol01-01    0         0        102400   rz128    rz128
pl vol01-02     vol01       ENABLED   ACTIVE   102400   CONCAT   -          RW
sd rzd128-02    vol01-02    0         102400   102400   rzd128    rzd128
LSM volume mirroring at creation time
# volassist -g sharkdg make vol02 50m mirror=yes
# volprint -g sharkdg -ht

```

DG NAME	GROUP-ID						
DM NAME	DEVICE	TYPE	PRIVLEN	PUBLEN	PUBPATH		
V NAME	USETYPE	KSTATE	STATE	LENGTH	READPOL	PREFPLEX	
PL NAME	VOLUME	KSTATE	STATE	LENGTH	LAYOUT	ST-WIDTH	
MODE							
SD NAME	PLEX	PLOFFS	DISKOFFS	LENGTH	DISK-NAME	DEVICE	
dg sharkdg	973124814.1279.osplcpq-ds20						
dm rz128	rz128	sliced	1024	1952112	/dev/rrz128g		
dm rzd128	rzd128	sliced	1024	1952112	/dev/rrzd128g		
dm rze128	rze128	sliced	1024	1952112	/dev/rrze128g		
v vol01	fsgen	ENABLED	ACTIVE	102400	SELECT	-	
pl vol01-01	vol01	ENABLED	ACTIVE	102400	CONCAT	-	RW
sd rz128-01	vol01-01	0	0	102400	rz128		rz128
v vol02	fsgen	ENABLED	ACTIVE	102400	SELECT	-	
pl vol02-01	vol02	ENABLED	ACTIVE	102400	CONCAT	-	RW
sd rz128-02	vol02-01	0	102400	102400	rz128		rz128
pl vol02-02	vol02	ENABLED	ACTIVE	102400	CONCAT	-	RW
sd rzd128-01	vol02-02	0	0	102400	rzd128		rzd128

3.5.4 Advanced File System creation and mounting

Once the LSM volumes are created, they can be used by the Advfs.

To create an Advanced File System, you need two commands **mkfdmn** and **mkfset**.

```
# mkfdmn /dev/vol/sharkdg/vol01 test1_dom
# mkfset test1_dom fset
# mount test1_dom#fset /mnt
# showfdmn test1_dom
      Id                Date Created  LogPgs  Domain Name
3a00b99f.000ed3c9  Wed Nov 1 16:47:27 2000      512  test1_dom
  Vol  512-Blks      Free  % Used  Cmode  Rblks  Wblks  Vol Name
   1L   102400     68032   34%   on    128   128
/dev/vol/sharkdg/vol01
```

3.6 Compaq V4.0F — fibre connection to ESS configuration restrictions

This section covers restrictions for Compaq V4.0F fibre connection to ESS configurations.

3.6.1 System supports only eight LUNs

The maximum number of LUNs that can be configured on a V4.0F system (in Fibre Channel) is 8. This limitation can be easily worked around using LSM.

You make 8 ESS volumes of the largest size possible, and then slice them up, thus building the desired number of LSM volumes.

The LSM volumes can be used for a file system or in raw mode.

3.6.2 Disks are not seen at the console prompt

ESS volumes connected via Fibre Channel are not seen at the console level, the command `show device` does not list them.

3.6.3 No boot or swap device supported

There is no support for boot or swap on ESS volumes connected via Fibre Channel.

3.6.4 LSM and Advfs restrictions

We found no ESS restrictions related to Advfs or LSM during our testing. Follow the Compaq official manuals on how to set up Advfs or LSM.

3.7 Tru64 UNIX log files

Table 2 shows the various log files on the system.

Table 2. Log files and the utilities that use them

Log files	Utilities that use these log files
<code>/var/adm/messages</code>	All the boot logs, Advfs and LSM messages
<code>/var/adm/binary.errlog</code>	All HW messages, read via the <code>dia</code> utility
<code>/var/adm/syslog.dated/current/daemon.log</code>	All TruCluster and daemon messages
<code>/var/adm/syslog.dated/current/kern.log</code>	Kernel messages

It is very useful to keep a window open with the `tail -f` of the `/var/adm/syslog.dated/current/daemon.log` file while configuring, modifying, or testing the cluster and the services. If you are having problems, you can see what they are related to.

3.7.1 References to Compaq documentation

Here are some of the more useful Internet Web sites for the Compaq documentation:

- <http://www.service.digital.com/patches/index.html>

- <http://www.compaq.com/support/>
- <http://www.unix.digital.com/cluster>
- <http://www.compaq.com/tru64unix>

Chapter 4. Compaq V5.0A

In this chapter we will discuss the configuration we tested with Compaq Alpha systems and Tru64 UNIX V5.0A and Cluster V5.0A.

At the time of writing (November 2000), it was not possible to boot from an ESS volume. Since this is a requirement for Cluster V5 we managed to configure the Compaq HSG80 with different volumes used by the Cluster File Systems (CFS), and boot from these volumes (that are seen correctly at the console level). Once the cluster was booted, we used the ESS volumes with CFS for user data.

Figure 35 shows the configuration we used.

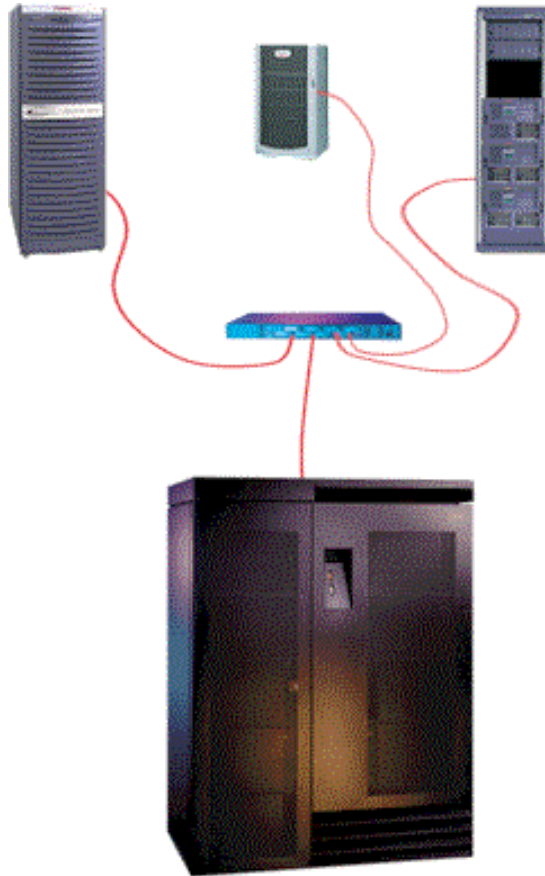


Figure 35. ESS and Compaq cluster

4.1 ESS configuration

Here is the ESS configuration that was used:

- OS Level 4.3.2.15
- Code EC SC01013 kinit1013
- Emulex LP7000
- latest firmware version for all cards

4.2 Compaq configuration

The following sections cover the Compaq configuration that was used.

4.2.1 Compaq systems tested

These are the systems we used for testing:

- Compaq AlphaServer ES20E 500 MHz SRM Console:V5.8
- Compaq AlphaServer GS60 6/700 MHz SRM Console:V5.8

4.2.2 Compaq Host Optical Fibre Adapter

The following list details the adapters we used for our testing of the systems. Any later revision of the cards should be satisfactory to use, but we have not tested them. As always with SANs, please check the manufacturers Web site for the latest information regarding patches.

- KGPSA-BC Driver Rev 1.21 F/W Rev 2.22X1(1.13)
- KGPSA-CA Driver Rev 1.21 F/W Rev 3.02A1(1.11)
- KGPSA-CA Driver Rev 1.21 F/W Rev 3.01(1.31)

4.2.3 IBM SAN Fibre Channel Switch

We used an IBM Fibre Channel Switch:

- IBM 2109 Model S08 (PN 2109S08) F/W Rev 2.1.3
- Single Zone Configuration

4.2.4 Compaq storage

We used the following Compaq storage:

- HSG80 Array Controller V85

4.2.5 Software versions

These are the software versions we used:

- Tru64 UNIX V5.0A
- Tru64 Cluster V5.0A

4.3 How to check the Compaq configuration

In the following sections we explain how to find all the information you will need to check the state of the Compaq system configuration.

4.3.1 System and cards firmware revision

The same commands used for V4 apply to V5.

4.3.2 Operating system version

The same commands used for V4 apply to V5.

4.3.3 Cluster software version

The output of the following command will show you if the TruCluster is installed.

```
# sysconfig -q clubase | grep version
cluster_version = TruCluster Server V5.0A (Rev. 354); 04/05/00 16:09
```

4.3.4 Disk configuration

There are no special operations that have to be done on the Compaq system to make it see the ESS volumes; just be sure that the host Fibre Channel is already configured (if it is not, you need to follow the instructions that come with the card and install the driver). Make all the hardware connections, configure the ESS, and then reboot the system. If the disks are seen at boot time, all the special files are created automatically. No kernel rebuild is necessary.

Once the ESS is configured to check if the disks are seen correctly from the Compaq system, you must wait for the system to boot in order to have the driver of the HBA loaded. After the boot completes, login as root and check if the ESS volumes are seen:

```
# hwmgr -view dev | more
HWID: Device Name          Mfg      Model      Location
-----
61: /dev/disk/dsk0c        DEC      HSG80      IDENTIFIER=1
62: /dev/disk/dsk1c        DEC      HSG80      IDENTIFIER=2
63: /dev/disk/dsk2c        DEC      HSG80      IDENTIFIER=3
64: /dev/disk/dsk3c        DEC      HSG80      IDENTIFIER=4
65: /dev/disk/dsk4c        DEC      HSG80      IDENTIFIER=5
66: /dev/disk/dsk5c        DEC      HSG80      IDENTIFIER=6
77: /dev/kevm
109: /dev/disk/dsk9c        COMPAQ   BB00911CA0 bus-0-targ-1-lun-0
110: /dev/disk/dsk10c       COMPAQ   BB00912301 bus-0-targ-2-lun-0
111: /dev/disk/dsk11c       COMPAQ   BB00911CA0 bus-0-targ-3-lun-0
112: /dev/disk/cdrom1c     DEC      RRD47      (C) DEC    bus-1-targ-4-lun-0
113: /dev/disk/dsk12c       IBM      2105F20    bus-2-targ-126-lun-0
114: /dev/disk/dsk13c       IBM      2105F20    bus-2-targ-126-lun-1
115: /dev/disk/dsk14c       IBM      2105F20    bus-2-targ-126-lun-2
116: /dev/disk/dsk15c       IBM      2105F20    bus-2-targ-126-lun-3
117: /dev/disk/dsk16c       IBM      2105F20    bus-2-targ-126-lun-4
118: /dev/disk/dsk17c       IBM      2105F20    bus-2-targ-126-lun-5
119: /dev/disk/dsk18c       IBM      2105F20    bus-2-targ-126-lun-6
120: /dev/disk/dsk19c       IBM      2105F20    bus-2-targ-126-lun-7
121: /dev/disk/dsk20c       IBM      2105F20    bus-2-targ-126-lun-8
```

With V5, the disk special file naming convention has changed.

Now all the disk special files are under 2 directories:

- /dev/disk
- /dev/rdisk

The first directory contains all block device special files, while the second contains all the character device special files.

Also, the name on the files has changed. Now all disks are simply called `dsk` followed by a number; you have to use the `hwmgr` command shown before to track the association between the file name and the device bus/target/LUN.

Finally, check if the disks have a valid disk label, we do this with the `disklabel` command:

```
# disklabel dsk20
# /dev/rrz128a:
type: SCSI
disk: 2105F20
label:
flags: dynamic_geometry
```

```

bytes/sector: 512
sectors/track: 64
tracks/cylinder: 30
sectors/cylinder: 1920
cylinders: 1017
sectors/unit: 1953152
rpm: 7200
interleave: 1
trackskew: 0
cylinderskew: 0
headswitch: 0 # milliseconds
track-to-track seek: 0 # milliseconds
drivedata: 0

```

```

8 partitions:
#      size      offset  fstype  [fsize bsize  cpg] #
a:    131072      0      unused      0    0 # (Cyl.  0 - 68*)
b:    262144    131072  unused      0    0 # (Cyl. 68*- 204*)
c:    1953152      0      unused      0    0 # (Cyl.  0 - 1017*)
d:         0      0      unused      0    0 # (Cyl.  0 - -1)
e:         0      0      unused      0    0 # (Cyl.  0 - -1)
f:         0      0      unused      0    0 # (Cyl.  0 - -1)
g:    819200    393216  unused      0    0 # (Cyl. 204*- 631*)
h:    740736    1212416  unused      0    0 # (Cyl. 631*- 1017*)

```

If there is no disk label on the disk, you must use the following command to write it to disk:

```
# disklabel -rw dsk30 ess
```

Now the disks can be used by AdvFS, LSM, and Tru64 Cluster.

4.4 ADVFS sample configuration

With Tru64 Cluster V5.0A, each domain that is configured is automatically seen as a cluster file system.

The commands to build an AdvFS domain are the same as for version 4.

Here is an example:

```

#mkfdmn /dev/disk/dsk30c dom_30
#mkfset dom_30 fset
#mount dom_30#fset /t30
#df
Filesystem          512-blocks      Used  Available Capacity  Mounted on
cluster_root#root   2309984         160978  2137168      8%  /
root2_domain#root   262144          86084   164080      35% /cluster/members
/member2/boot_partition
cluster_usr#usr      5330736        1176768  4122704      23% /usr
cluster_var#var      5508432         77548   5418288      2%  /var
/proc                0              0         0      100% /proc
dom_30#fset          1953152        1594282  343584       83% /t30

```

If you want the domain to be automatically mounted at boot time, add the following line to the file `/etc/fstab`:

```
dom_30#fset    /t30          advfs rw 0 2
```

To check which cluster node manages the domain you just created, issue the command `cfsmgr`. The output of the command will show you all the CFSs that are configured for the cluster, and you should find the following lines related to your domain:

```

Domain or filesystem name = dom_30#fset
Mounted On = /t30
Server Name = gs60
Server Statuapply OK

```

4.5 Compaq V5.0A — fiber connection to ESS configuration restrictions

This section covers Compaq V5.0A — fiber connection to ESS configuration restrictions.

4.5.1 No boot from ESS volumes

At the time of writing (November 2000), it was not possible to boot from an ESS volume. Since this is a requirement for Cluster V5, we managed to configure the Compaq HSG80 with different volumes used by the Cluster File Systems, and boot from these volumes (seen correctly at the console level). Once the cluster was booted, we used ESS volumes with CFS for user data.

4.5.2 On GS systems cannot find console commands

In order to be able to run all diagnostic commands, `wwidmgr`, `mc_cable`, `mc_diag`, and other diagnostic tools, you must first issue the command:

```
P0>>> set mode diag
```

4.5.3 All ESS volumes seen with same ID/LUN from all cluster nodes

It is necessary for all ESS volumes to be seen with the same ID/LUN from all the cluster nodes.

Be sure that when you create the ESS volumes for a cluster environment all nodes see the volumes with the same ID and LUN. The best thing is to define all the volumes for one node and then assign them to the other nodes. Do this by going to the StorWatch specialist and selecting **Storage**.

Select **Allocation -> Open Systems -> Modify Volume Assignment**. Select all the defined volumes and select both **Assign selected volumes to target host** and **Use same ID/LUN in source and target**. Finally, select **Perform Configuration Update**.

4.5.4 File command gives the LUN number in decimal

On V5 you can have up to 256 LUNs, be aware that the UNIX `file` command shows the LUN number in decimal format, while the ESS Specialist shows it in hexadecimal format.

4.5.5 Termination of fibre cards

All host fibre adapters that are plugged into the Compaq systems, but are not connected anywhere, must be terminated.

4.6 Tru64 UNIX log files

Table 3 shows the various log files on the system.

Table 3. Log files and the utilities that use them

Log files	Utilities that use these log files
<code>/var/adm/messages</code>	All the boot logs, Advfs and LSM messages
<code>/var/adm/binary.errlog</code>	All HW messages, read via the <code>dia</code> utility
<code>/var/adm/syslog.dated/current/daemon.log</code>	All TruCluster and daemon messages
<code>/var/adm/syslog.dated/current/kern.log</code>	Kernel messages

It is very useful to keep a window open with the `tail -f` of the `/var/adm/syslog.dated/current/daemon.log` file while configuring, modifying, or testing the cluster and the services, if you are having problems, you can see what they are related to.

4.6.1 References to Compaq documentation

Here are some of the more useful Internet Web sites for the Compaq documentation:

- <http://www.service.digital.com/patches/index.html>
- <http://www.compaq.com/support/>
- <http://www.unix.digital.com/cluster>
- <http://www.compaq.com/tru64unix>

Chapter 5. IBM ESS and HP Servers

Our prime objective was to demonstrate the capability and effectiveness of the ESS in an HP-UX environment and to provide the basic information to allow you to do this in your environment. We explored this by designing, implementing, and testing a High Availability (HA) system using the disk storage capacity and speed of the ESS for storing and restoring the information. The ESS uses state of the art technology for data transmission. For implementation of this, we used two HP servers for clustering and the ESS to test its compatibility and reliability with HP hardware and software.

We configured two servers in an HA configuration using HP-UX-11.00 and MC/ServiceGuard for clustering. The servers and software were installed and configured according to HP recommended practices. In addition to base HP-UX-11.00 we used the following additional software modules:

- Logical Volume Manager (LVM)
- MC/ServiceGuard for Clustering

5.1 Pre-installation planning

Using SWAP space for HP-UX on the ESS is not supported.

Installation suggestions include Journalled File System (JFS), also called Veritas File System (VXFS). HP-UX supports only HFS for `/stand`. The OnLine JFS software will not be able to perform on-line task on the `/stand`. We followed HP's instructions for the installation of Logical Volume Manager (LVM), OnLine JFS for on-line disk maintenance, and MC/ServiceGuard for clustering.

Please see 5.4.3, "Supported servers and software" on page 81 for the recommended hardware/software. The listed HW/SW has been thoroughly tested and certified for use. If you use unsupported hardware or software you may see unpredictable results. For the latest IBM supported hardware and software, please see the supported server Web site:

<http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm>

5.2 Hardware connectivity: Fibre Channel - Arbitrated Loop (FC-AL)

Figure 36 shows the configuration used for our first series of tests.

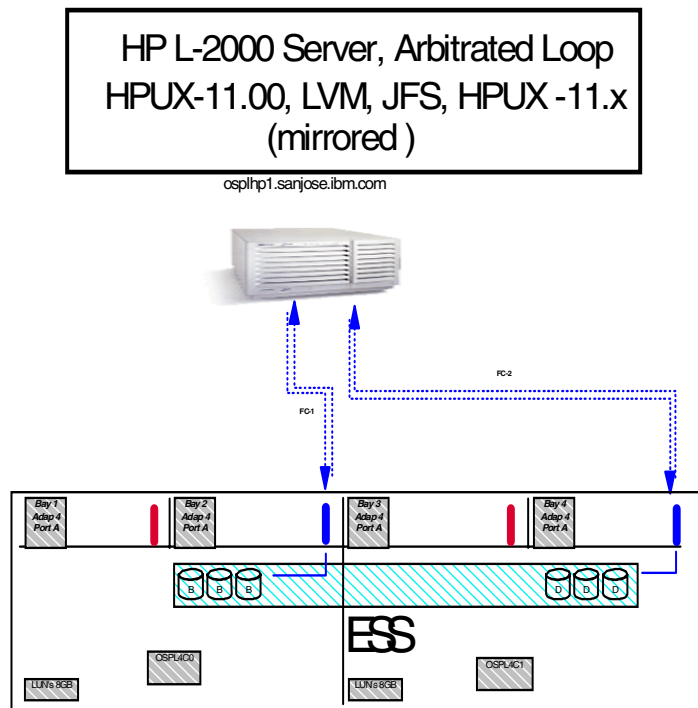


Figure 36. Basic test system, ESS configured in FC-AL mode

We used an HP L - 2000 server with HPUX-11.00. Two Fibre Channel connections from the HP server were made to the ESS. We used LUN masking within the ESS to enable both HBAs to see the same group of LUNs. This presents two images of the LUNs to the server operating system. HP-UX has a built in feature called pv-links that handles this, and only a single image of each LUN is seen by an application. This dual pathing arrangement is called alternate pathing, and in the event of a path failure, all I/O traffic will be routed along the remaining path. Although it was a point-to-point connection from the host to the ESS, the Fibre Channel adapters within the ESS were configured in FC-AL mode.

For detailed information on how to configure this setup, as well as LUN masking, please read *Fibre Channel attachment to the ESS*, SG24-6113. By running the `ioscan` command, we were able to examine how the LUNs were seen by the operating system.

Following are the results of the `ioscan` command:

```
Result of ioscan command.
disk 794 0/3/0/0.8.0.1.1.0.4 sdisk CLAIMED DEVICE IBM 2105F20
/dev/dsk/c20t0d4 /dev/rdisk/c20t0d4
disk 795 0/3/0/0.8.0.1.1.0.5 sdisk CLAIMED DEVICE IBM 2105F20
/dev/dsk/c20t0d5 /dev/rdisk/c20t0d5
disk 796 0/3/0/0.8.0.1.1.0.6 sdisk CLAIMED DEVICE IBM 2105F20
/dev/dsk/c20t0d6 /dev/rdisk/c20t0d6
disk 797 0/3/0/0.8.0.1.1.0.7 sdisk CLAIMED DEVICE IBM 2105F20
/dev/dsk/c20t0d7 /dev/rdisk/c20t0d7
disk 798 0/3/0/0.8.0.1.1.1.0 sdisk CLAIMED DEVICE IBM 2105F20
/dev/dsk/c20t1d0 /dev/rdisk/c20t1d0
```

5.2.1 Failover/failback test

The first test that we ran was a failover/failback test, using the setup shown in Figure 36. FC-1 and FC-2 are two Fibre Channel connections which are directly connected to the ESS. These Fibre Channel connections were made such that both FC-1 and FC-2 are logically pointing to the same disk storage.

First we physically disconnected the cable FC-1 from the server, and the operating system detected a timeout. All I/Os to the disks were routed through FC-2. We then reconnected FC-1 and then disconnected FC-2. We were able to observe, by running `iostat`, that I/O to the LUNs continued, this time using FC-1. Finally, we reconnected FC-2 and observed that both paths were used.

5.3 Hardware connectivity: simple switched fabric

Figure 37 shows the setup that we used for the next set of tests. This was a single HP host running HP-UX 11 with two HBAs connected to a single IBM 2109 switch. From the switch there are two paths to the ESS.

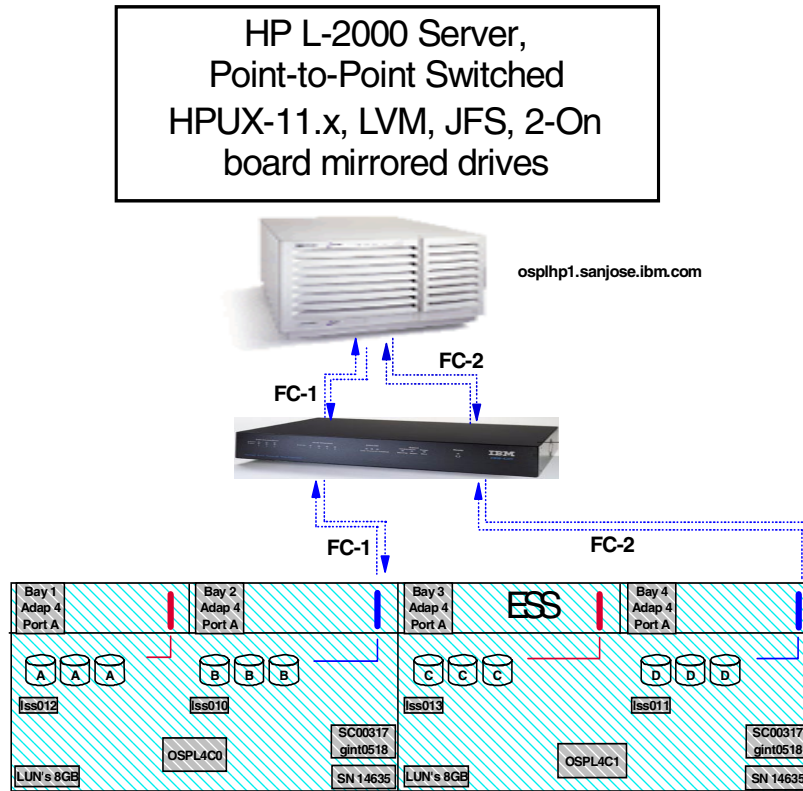


Figure 37. Simple switched configuration

For the initial tests, the switch was set up without zoning; one input port was linked to one output port. The ESS was configured so that both ports on the server were connected to the same set of LUNs. This means that we were not relying on the switch to handle any fibre link failure. As we were using a switch, the ESS ports were set up in point-to-point mode.

The tests that we ran using this configuration were similar to those used when we were setup in FC-AL mode. That is, they were a series of cable pull tests, checking to see that the host operating system was able to reroute the I/O along the alternate path.

First we physically disconnected the cable FC-1 from the server, and the operating system detected a timeout. All I/Os to the disks were routed through FC-2. We then reconnected FC-1 and then disconnected FC-2. We were able to observe, by running `iostat`, that I/O to the LUNs continued, this time using FC-1. Finally we reconnected FC-2 and observed that both paths were used.

We carried out this series of tests twice. First we did the tests removing the cables for the server to the switch. The second time we removed the cables from the switch to the ESS. In both cases the results were the same; I/Os to the ESS continued the failure, and alternate pathing was handled by pv-links within the operating system.

5.4 High Availability tests

Having established that basic connectivity and failover/ failback of fibre path was handled by the operating system for a single host, we embarked on testing a High Availability system using two hosts and clustering software. The configuration that we used can be seen in Figure 38.

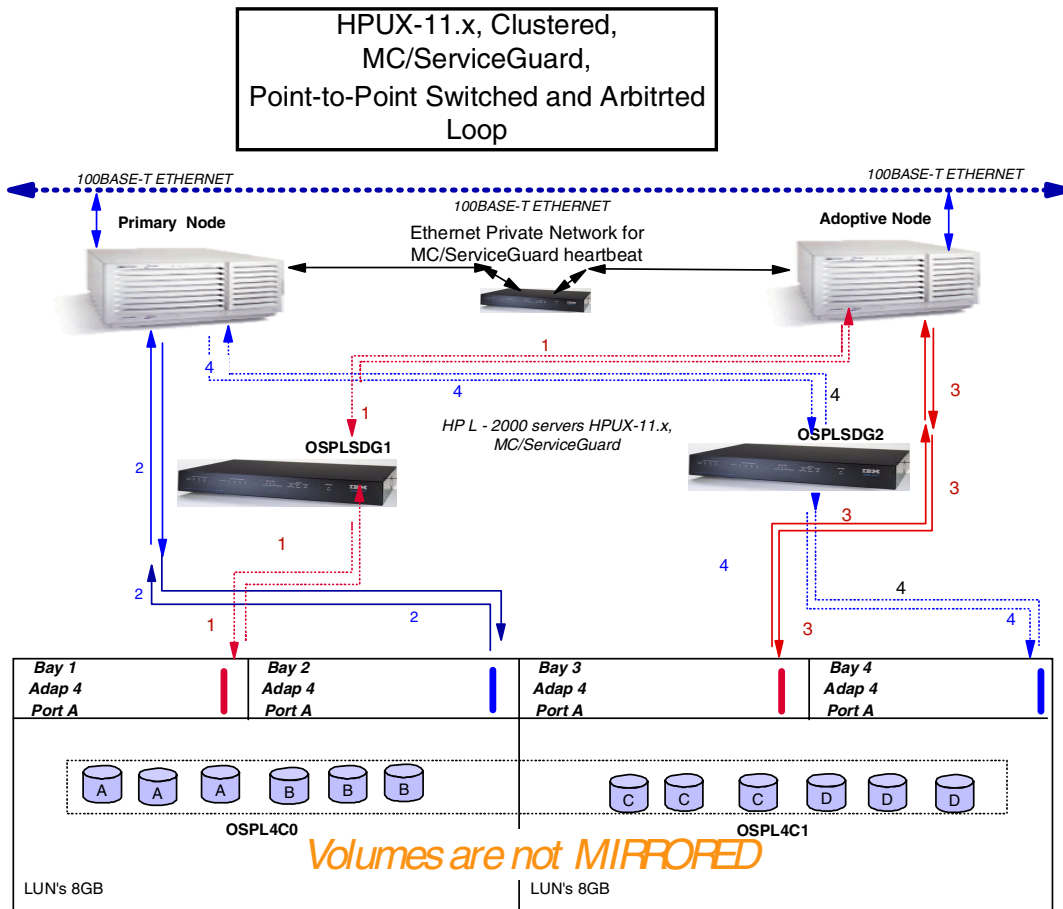


Figure 38. High Availability configuration that was tested

Topology and Host Configuration Details:

Throughout the testing, each host had one HBA attached to the ESS directly (utilizing the FCAL protocol), and one through the IBM switch (utilizing the PTP protocol). The ESS was configured with 256 volumes for MC/SG testing, and 450 volumes for stress testing. The switch was zoned such that each HBA was logically connected to one path into the ESS. This allowed each host to have two individual paths into the storage. The beta version of the A5158A driver was used.

Software:

- OS — HPUX-11.00, patch Bundle 0600 Support CDROM patch

- Bundles — superseded and/or downloaded for HW:
 - PHKL_21989
 - PHKL 18543
 - PHCO_20882
 - PHCO_21187
 - PHKL_22440
 - PHKL_22432
 - PHKL_20016
 - PHKL_22432
 - PHKL_21381

Hardware:

- Fibre Channel Switch IBM 2109 S16, Firmware 2.1.7
- Servers HP's L-2000 2
 - CPUs 2 — 400MHz
 - Physical memory — 3GB
 - Fibre Channel Host Adapter — HP A5158A with Beta driver
Rel.11:00:05

5.4.1 Test results

We ran a series of stress tests and other tests on the storage system. For full details of the tests, please see Appendix A, “Test suite details” on page 113. The following tests all passed:

- Single server disk stress, 96-hour test run
- Recovery time, 24-hour test run
- Disaster, 24-hour test run
- PV-links failover/failback — 20 cable pulls of various cables (failback occurred after approximately 15 to 30 seconds)
- ESS warm starts every 15 minutes for 24 hours
- ESS failover/failback
- ESS cluster quiesce/resume

Cluster tests

The following tests all passed:

- Manual package failover (MC/SG)
- Application process failover
- Test of primary LAN heartbeat
- Test of primary data LAN
- Full data network test
- Split-brain test
- TOC test
- Power outage test

- ESS microcode update, both concurrent and non-concurrent

5.4.2 Tuning recommendations

Based on the work that we did in order to satisfy the test requirements we have come to the following conclusions on the best tunable parameters in the HP-UX software

Note: You need to be an experienced HP-UX administrator before attempting to change kernel parameters

nbuf and bufpages: If you have static Input Output (I/O) buffer defined, a static buffer in HP-UX is usually defined by setting these two parameters to some appropriate number. These two parameters are defined as the size of an I/O buffer. But we feel that a dynamic I/O buffer will serve your needs better without occupying a large chunk (size of $nbuf * bufpages$) of your server's physical memory. On the flip side, a dynamic buffer will fluctuate between the value of `dbc_min_pct` and `dbc_max_pct`. A dynamic I/O buffer can be defined by setting both `nbuf` and `bufpages` to zero (0) and giving appropriate values to `dbc_min_pct` (lower limit) and `dbc_max_pct` (upper limit). These two parameters are measured in terms of percentage of the total physical memory of your server. We suggest that you should start with low numbers — for example, `dbc_min_pct = 5` and `dbc_max_pct = 15` or `20`. Then monitor your system's resources usage, and adjust according to your needs.

maxfiles: With the storage capacity of the ESS at your disposal, you will have the capability to run large applications, which may cause more than 60 files to be open concurrently. By default, this parameter is set to 60. However, care needs to be exercised with older K class machines, as they will not have the processing power to handle this many files.

maxvgs: By default this parameter is set to nine(9). Which will limit you to ten (0-9) Volume Group (VG)s. In order to keep storage organized you may need to create more than ten VGs. Your requirements will suggest the value of this parameter.

memswap_on: If your system is not used for real time computing. We suggest that you should turn off the memory swap, and create an interleaved device swap space of the size of twice the size of server's physical memory. If the server is running a database application then follow the database vendor's recommendations. Otherwise create device swap space approximately twice the size of server's physical memory. Followed by a kernel rebuild and reboot of the server.

5.4.3 Supported servers and software

The following are supported servers and software:

- HPUX 10.20 is supported on K class servers.
- HPUX 11.00 is supported for L/N class servers (64 bit only).
- HPUX 10.20 and 11.0 for D (requires additional beta test in customer site).
- HPUX 11.00 for K class servers (requires additional beta test in customer site).
- HPUX 11.00 for V class servers (requires additional beta test in customer site).
- HBA A3404 is supported with firmware version 38.22 in K class servers.
- HBA A5158 is supported with Fibre Channel Tachyon Driver B.11.00.05 (pre-release) in K/L/N class servers.
- SAN Fibre Channel Switch:
 - IBM 2109 Model S08 (PN 2109S08) F/W Rev 2.1.3, Single Zone Configuration.
 - McData Fibre Channel Switch: ED-5000 (PN 002-002120-200) Rev E, Single Zone Configuration.
- ESS Fibre Configuration: OS Level 4.3.2.15, Code EC SC01027 KINT1027 or G3.
- Milieux LP7000 Fibre Channel target adapter cards: (Feature Code 3022) latest firmware version for all cards.
- MC/Service Guard — Oracle Parallel Server is not supported.

Chapter 6. IBM ESS and Sun Enterprise Servers

The Enterprise Storage Server has several features that are desirable for Sun's Enterprise Servers — both within High Availability and distributed environments (see Figure 39).

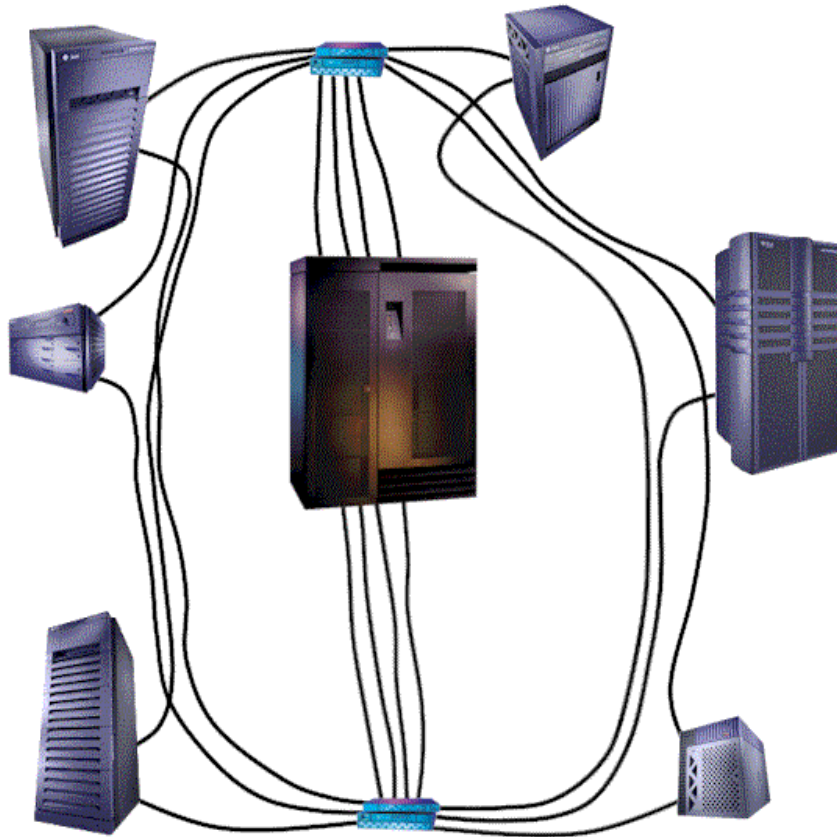


Figure 39. - Storage Area Network

However, several system variables need to be configured properly to successfully employ the ESS and the Sun Enterprise Server. While it is possible to determine these values through the process of “trial and error”, such a process is, to say the least, costly and time consuming.

The following sections will provide the integrator or system administrator with the changes necessary to bring the IBM ESS and Sun Enterprise Server(s) together and allow them to operate successfully.

6.1 Topics covered in this chapter

In the remainder of this chapter we cover the following topics:

- Section 6.2, “Features supported in this release” on page 84
- Section 6.3, “Supported components” on page 84
- Section 6.4, “Caveats and limitations” on page 85
- Section 6.5, “Required modifications” on page 86
- Section 6.6, “Boot messages” on page 90
- Section 6.7, “Sun Veritas Volume Manager” on page 95
- Section 6.8, “Known Issues” on page 111

6.2 Features supported in this release

The following functions are supported in this release:

- Direct attach via SCSI
- Direct attach via FC-AL
- Fibre Channel attach via IBM SAN Data Gateway (2108) to SCSI
- Fabric attach point-to-point via IBM SAN Fibre Channel Switch (2109)
- Sun Veritas Dynamic Multi Pathing
- Sun Veritas Clustering Software
- Sun Veritas Fast Filesystem
- Sun Veritas Volume Manager
- Sun Solaris 2.6
- Sun Solaris 7

6.3 Supported components

Supported components include the following:

- IBM ESS subsystem requirements:
 - ESS 2105 running O/S 4.3.2.15, code EC SC01027, KINT1027 or G3
 - ESS SCSI HAs (feature code 3002) with the latest firmware release, if configuring for SCSI
 - ESS Fibre Channel HAs (feature code 3022) with the latest firmware release, if configuring for Fibre Channel
- IBM 2109 SAN Fibre Channel Switches — Model S08 and Model S16, (PN 2109S08 and 2109S16 respectively) — running firmware revision 2.1.7, single zone configuration in FC-SW environments
- IBM 2108 SAN Data Gateway, in Fibre Channel to SCSI environments

- Brocade Silkworm Switches — Model 2400 and Model 2800 — running firmware 2.1.7, single zone configurations in FC-SW environments
- Any Sun Enterprise/Ultra Enterprise series servers except the Sun E/UE10000
- Operating systems supported:
 - Solaris 2.6 with Jumbo Patch 23 and all Sun recommended and security patches
 - Solaris 7 with Jumbo Patch 12 and all Sun recommended and security patches
- HBA cards supported:
 - Emulex
 - Emulex LP7000, driver rev 4.02d-COMBO, firmware revision 3.02A1 or later
 - Emulex LP8000, driver rev 4.02d-COMBO, firmware revision 3.02 or later
 - JNI
 - JNI FC64-1063 SBus 64 Bit, driver rev 2.5.9, firmware revision 13.3.7 or later
 - JNI FCI-1063 PCI 32 Bit,

6.4 Caveats and limitations

The following caveats apply to the implementation of ESS on Sun:

- Only the ESS and Sun hardware and software configurations listed under Section 6.3, “Supported components” on page 84 are supported.
- No data throughput speeds are expressed or implied. Overall performance, including throughput, is dependent on data/system load and other conditions that IBM does not control.
- The IBM subsystem device driver (SDD) is not available for the Solaris operating system.
- ESS JBOD disks or logical volumes are not supported as paging or swap devices.
- ESS JBOD disks or logical volumes are not supported as boot devices.
- ESS concurrent code load with ESS HA firmware upgrade is possible for systems running Sun Veritas DMP as long as the primary and alternate

path HAs are not in the same ESS bay and are not upgraded simultaneously.

- ESS concurrent code load without ESS HA firmware upgrade is possible for all Sun servers.
- Non-disruptive code loads may fail with large LUN configurations (more than 3,000) running heavy I/O.
- ESS to Sun connectivity via SCSI only supports differential SCSI.
- Point-in-time copies via FlashCopy are supported on the ESS but are not discussed in this book.
- Peer-to-peer Remote Copy (PPRC) is supported on the ESS but is not discussed in this book.
- Current Fibre Channel Host Type or LUNs created prior to ESS Code level G3 or KINT1013 must be deleted and re-configured using G3 or KINT1013 or later.
- For all HBAs, the maximum recommended number of logical volumes per adapter under FC-AL is 50.
- Use of any software and/or hardware not specifically stated as supported on supported platforms or any hardware and/or software on any non-supported platform is not supported.

6.5 Required modifications

For all O/Ss, install the appropriate driver for the HBA following the manufacturers instructions. When the driver install is complete, perform the modifications outlined below for the appropriate HBA(s).

6.5.1 All HBAs

Three variables are required in the `/etc/system` file. These modifications should be inserted above any `forceload` statements.

6.5.1.1 `sd_max_throttle`

```
set sd:sd_max_throttle = "calculated"
```

The `sd_max_throttle` variable assigns the default value `lpfc` will use to limit the number of outstanding commands per `sd` device. This value is global, affecting each `sd` device recognized by the driver. The maximum '`sd_max_throttle`' setting supported is 256. To determine the correct setting, perform the following calculation for each HBA:

$$256 / (\text{number of LUN's per adapter}) = \text{sd_max_throttle value}$$

For example, a server with two HBA's installed, 20 LUN's defined to HBA1, and 26 LUN's defined to HBA2.

$$\text{HBA1} = 256 / 20 = 12.8 \text{ and } \text{HBA2} = 256 / 26 = 9.8$$

Rounding down yields 12 for HBA1 and 9 for HBA2. In this example, the correct 'sd_max_throttle' setting would be the lowest value obtained or 9.

6.5.1.2 sd_io_time

```
set sd:sd_io_time = 120
```

The sd_io_time variable determines how long a queued job will wait for any sd device I/O to fail. Originally, sd_io_time is set to 60. This is too low for most configurations. Setting it to 120 provides the host more time to complete I/O operations.

```
set maxphys = 8388608
```

The maxphys value determines the maximum number of bytes that can be transferred with a SCSI transaction. The original value is too small to allow the Fibre Channel HBA(s) to run efficiently. Set this to 8388608.

6.5.2 Emulex

The Emulex HBAs require modifications in the /kernel/drv/lpfc.conf file. The following variables must be modified as specified below for all supported Emulex HBAs:

6.5.2.1 automap=1;

The automap variable is used to turn on or off the retention of SCSI IDs on the fibre. If automap is set, SCSI IDs for all FCP nodes without persistent bindings will be automatically generated. If new FCP devices are added to the network when the system is down, there is no guarantee that these SCSI IDs will remain the same when the system is booted again. If one of the above FCP binding methods is specified, then automap devices will use the same mapping method to preserve SCSI IDs between link down and link up. If automap is 0, only devices with persistent bindings will be recognized by the system. Set this to 1.

6.5.2.2 fcp-on=1;

The fcp-on variable controls whether or not Fibre Channel port access is enabled or not. Set this to 1 to enable FCP access.

```
lun-queue-depth="sd_max_throttle from /etc/system";
```

The lun-queue-depth variable determines how many requests can be accepted for each of the LUNs the host has access to. This value is global in nature as it affects all LUNs on the host. Set this to be equal to the sd_max_throttle value obtained from the /etc/system file. See Section , “set sd:sd_max_throttle =“calculated”” on page 86

6.5.2.3 network-on=0;

The network-on variable determines whether networking is enabled for the HBA. Networking will be enabled if set to 1, disabled if set to 0. This variable will be set during the installation of the driver via pkgadd. Verify it is set to 0.

6.5.2.4 topology=2;

The topology variable is used to let the lpfc driver know how to attempt to start the HBA. It can be set to start only one mode or to attempt one mode and then fail over to the other mode should the first mode fail to connect.

- 0x00 = attempt loop mode, if it fails attempt point-to-point mode
- 0x02 = attempt point-to-point mode only
- 0x04 = attempt loop mode only
- 0x06 = attempt point-to-point mode, if it fails attempt loop mode

Set the variable to point-to-point mode to run as an N_Port or FC-SW. Set the variable to loop mode to run as an NL_Port or FC-AL. The above setting reflects FC-SW only.

6.5.2.5 zone-rscn=1;

The zone-rscn variable allows the driver to check with the NameServer to see if an N_Port ID received from an RSCN applies. Setting zone-rscn to 1 causes the driver to check with the NameServer. If Soft Zoning is used, with Brocade Fabrics, this should be set to 1. Set this to 1.

6.5.3 JNI

Under the /kernel/drv directory, modifications will be necessary in the appropriate configuration file. Table 4 references the JNI HBA with its corresponding configuration file:

Table 4. JNI configuration file names

JNI HBA model	/kernel/drv filename
FC64-1063	fcaw.conf
FCI-1063	fca-pci.conf

The modifications necessary include these:

6.5.3.1 fca_nport

```
fca_nport = 0;
```

- or -

```
fca_nport = 1;
```

The `fca_nport` variable is used to setup either FC-AL or FC-SW. If false (0), then `fca` initializes on a loop. If true (1), then `fca` initializes as an N_Port and fabric operation is enabled. This variable can be overridden by `public_loop` (see below). For fabric, set this to 1.

6.5.3.2 public_loop = 0;

The `public_loop` variable can override the `fca_nport` variable. If `public_loop` is false (0), then `fca` initializes according to what `fca_nport` is set to. If true (1), then `fca` initializes as an NL_Port on a public loop and fabric operation is enabled via the FLPort of the switch. Also, if `public_loop = 1`, then `fca_nport` is overridden to be 0. Set this to 0.

6.5.3.3 ip_disable = 1;

The `ip_disable` variable allows the IP side of the driver to be enabled or disabled. If false (0), then the IP side of the driver is enabled. If true (1), then the IP side of the driver is completely disabled. Set this to 1.

6.5.3.4 scsi_probe_delay = 5000;

The `scsi_probe_delay` variable uses a 10 millisecond resolution to set the delay before SCSI probes are allowed to occur during boot. This allows time for the driver to build a network port list for target binding. Set this to 5000.

6.5.3.5 failover = 60;

The `failover` variable represents the number of seconds after a target is declared offline before the target is declared as failed and all pending commands are flushed back to the application. Using the IBM 2109 or the Brocade switch, set this to 60. If using a McData switch, set to 300.

6.6 Boot messages

Boot the system in reconfigure mode and bring it up in single-user (`boot -rs` or `reboot -- -rs`) to allow the host to discover the new HBA(s) installed. During the boot process, the WWPN(s) for the HBA(s) will be displayed and written to the messages file. This information will be necessary to configure the ESS to perform LUN masking of the logical volumes.

Armed with the WWPN(s) for the HBA(s), add/modify the logical volumes to the appropriate host type(s) within the ESS. Remember that each HBA within a host requires its own host type defined within the ESS to properly associate/restrict access to the logical volumes. If the host will be fabric attached versus loop attached, then make the appropriate entries within the switch. However, if the switch was configured using the switch domain and switch port number method, then only the changes on the ESS will be required. See Chapter 2, "Introduction to ESS connectivity" on page 25 for more information on configuring the switch.

Once the ESS (and switch, if required) have been configured, perform a second reconfigure boot. The second reconfigure boot will display the HBA information as well as the target/LUN information for the attached logical volumes (with verbosity or the appropriate variable in the driver configuration file turned on).

6.6.1 Emulex

During boot, the Emulex HBAs will load with information similar to the following:

```
Nov 16 14:09:58 que Emulex LightPulse FC SCSI/IP 4.02d-COMBO
Nov 16 14:09:58 que unix: NOTICE: lpfc0:031:Link Up Event received Data: 1
1 00
Nov 16 14:10:01 que unix: NOTICE: lpfc0: Firmware Rev 3.02A1 (D2D3.02A1)
Nov 16 14:10:01 que unix: NOTICE: lpfc0: WWPN:10:00:00:00:c9:22:68:68
WWNN:20:00:00:00:c9:22:68:68 DID 0x11600
Nov 16 14:10:01 que unix: NOTICE: Device Path for interface lpfc0:
Nov 16 14:10:01 que unix: PCI-device: fibre-channel@3, lpfc0
Nov 16 14:10:01 que unix: lpfc0 is /pci@6,4000/fibre-channel@3
Nov 16 14:10:01 que unix: NOTICE: lpfc1:031:Link Up Event received Data: 1
1 00
Nov 16 14:10:04 que unix: NOTICE: lpfc1: Firmware Rev 3.02 (D2D3.02)
Nov 16 14:10:04 que unix: NOTICE: lpfc1: WWPN:10:00:00:00:c9:23:22:92
WWNN:20:00:00:00:c9:23:22:92 DID 0x11400
Nov 16 14:10:04 que unix: NOTICE: Device Path for interface lpfc1:
Nov 16 14:10:04 que unix: PCI-device: fibre-channel@4, lpfc1
Nov 16 14:10:04 que unix: lpfc1 is /pci@6,4000/fibre-channel@4
```

This host has two Emulex HBAs installed that load at lpfc0 and lpfc1. The HBAs are running with firmware 3.02A1 and 3.02 respectively on driver 4.02d-COMBO.

The WWPNs are visible as 10:00:00:00:c9:22:68:68 for lpfc0 and 10:00:00:00:c9:23:22:92 for lpfc1. This information is required during configuration of the switch in FC-SW environments and for configuration of the ESS within the ESS Specialist for both FC-SW and loop environments.

During the second reconfigure boot, as the host discovers the logical volumes, information similar to the following will scroll by:

```
Nov 16 14:10:04 que unix: NOTICE: lpfc0: Acquired FCP/SCSI Target 0 LUN 0
Nov 16 14:10:04 que          D_ID 0x11800 WWPN:10:00:00:00:c9:21:c2:c3
WWNN:50:05:07:63:00:c0:0b:16
Nov 16 14:10:04 que
Nov 16 14:10:04 que unix: sd9040 at lpfc0:
Nov 16 14:10:04 que unix: target 0 lun 0
Nov 16 14:10:04 que unix: sd9040 is /pci@6,4000/fibre-channel@3/sd@0,0
Nov 16 14:10:04 que unix: <IBM-2105F20-1013 cyl 1015 alt 2 hd 30 sec
64>
```

The logical volume is discovered by the driver and the information pertaining to the WWNN and WWPN for the ESS and the associated HA, respectively, are displayed. The WWNN and WWPN are the names provided by the ESS to the host during the discovery and login process. If the ESS does not recognize the hosts WWPN, then these messages will be absent as well as any logical volumes that would have been presented by the ESS to this host.

The logical volume is assigned an sd device name that will be retained within the host. Also, the full pathname to the device is displayed along with information pertaining to the device type. IBM-2105F20 device type reflects that an ESS logical volume has been presented to the host. The actual number of cylinders, heads, and sectors presented will vary with the size of the logical volume.

- Please note that while it is possible to configure the ESS with a wide variety of logical volume sizes, the Solaris operating system may not know how to deal with some of the unusual disk types this will create. For example, the 14 GB disk type presented by the ESS requires a new disk type definition within the format utility. Otherwise, the utility will fail when attempting to format the disk with a bad magic message.
- Using logical volumes that are powers of 2 for all volumes greater than 10 GB seems to solve this problem on the Solaris operating system. However, only logical volumes of 16 GB and 32 GB were formatted

successfully. Larger logical volumes have not been tested and may fall outside the range of known disk types for the Solaris operating system. The larger logical volumes should be tested for compatibility with the format utility before using them in production environments.

If the host is running Sun Veritas Volume Manager and has licensed DMP, then a message similar to the following will appear:

```
NOTICE: vxvm:vxndmp: added disk array 14635
```

The number 14635 is the serial number for the ESS. Multiple numbers will be displayed if multiple ESSs are configured and attached to a host running DMP.

6.6.2 JNI

During boot, the JNI HBAs will load with information similar to the following:

```
Nov 18 19:11:29 barney fcaw: [ID 451854 kern.notice] fcaw0: JNI Fibre
Channel Adapter model FCW
Nov 18 19:11:29 barney fcaw: [ID 451854 kern.notice] fcaw0: 64-bit SBus 2:
IRQ 3: FCODE Version 13.3.7 [18c932]: SCSI ID 125: AL_PA 01
Nov 18 19:11:29 barney fcaw: [ID 451854 kern.notice] fcaw0: Fibre Channel
WWNN:100000E06940013B WWPN: 200000E06940013B
Nov 18 19:11:29 barney fcaw: [ID 451854 kern.notice] fcaw0: FCA SCSI/IP
Driver Version 2.5.9, August 22, 2000 for Solaris 7
Nov 18 19:11:29 barney fcaw: [ID 451854 kern.notice] fcaw0: All Rights
Reserved.
Nov 18 19:11:29 barney fcaw: [ID 580805 kern.info] fcaw0: < Total IOPB
space used: 1145024 bytes >
Nov 18 19:11:29 barney fcaw: [ID 580805 kern.info] fcaw0: < Total DMA space
used: 8458269 bytes >
Nov 18 19:11:29 barney fcaw: [ID 580805 kern.info] fcaw0: Resetting GLM...
Nov 18 19:11:36 barney fcaw: [ID 585631 kern.notice] NOTICE: fcaw0 NPORT
Initialization Complete, SID=40013B
Nov 18 19:11:37 barney fcaw: [ID 580805 kern.info] fcaw0: New Fabric
ParametersvReceived. Resetting...
Nov 18 19:11:37 barney fcaw: [ID 580805 kern.info] fcaw0: LINK DOWN
Nov 18 19:11:37 barney fcaw: [ID 585631 kern.notice] NOTICE: fcaw0 NPORT
Initialization Complete, SID=40013B
Nov 18 19:11:37 barney fcaw: [ID 580805 kern.info] fcaw0: LINK UP
(180002FF)
Nov 18 19:11:38 barney fcaw: [ID 580805 kern.info] fcaw0: Host: Port 011300
(100000E06940013B:200000E06940013B)
Nov 18 19:11:38 barney fcaw: [ID 580805 kern.info] fcaw0: Port 011900
(5005076300C00B16:10000000C921E28C) available.
```

```

Nov 18 19:11:41 barney fcaw: [ID 451854 kern.notice] fcaw1: JNI Fibre
Channel Adapter model FCW
Nov 18 19:11:41 barney fcaw: [ID 451854 kern.notice] fcaw1: 64-bit SBus 2:
IRQ 3: FCODE Version 13.3.7 [18c932]: SCSI ID 125: AL_PA 01
Nov 18 19:11:41 barney fcaw: [ID 451854 kern.notice] fcaw1: Fibre Channel
WWNN: 100000E06940013B WWPN: 200000E0694061B5
Nov 18 19:11:41 barney fcaw: [ID 451854 kern.notice] fcaw1: FCA SCSI/IP
Driver Version 2.5.9, August 22, 2000 for Solaris 7
Nov 18 19:11:41 barney fcaw: [ID 451854 kern.notice] fcaw1: All Rights
Reserved.
Nov 18 19:11:41 barney fcaw: [ID 580805 kern.info] fcaw1: < Total IOPB
space used: 1145024 bytes >
Nov 18 19:11:41 barney fcaw: [ID 580805 kern.info] fcaw1: < Total DMA space
used: 8458269 bytes >
Nov 18 19:11:41 barney fcaw: [ID 580805 kern.info] fcaw1: Resetting GLM...
Nov 18 19:11:48 barney fcaw: [ID 585631 kern.notice] NOTICE: fcaw1 NPORT
Initialization Complete, SID=4061B5
Nov 18 19:11:49 barney fcaw: [ID 580805 kern.info] fcaw1: New Fabric
Parameters Received. Resetting...
Nov 18 19:11:49 barney fcaw: [ID 580805 kern.info] fcaw1: LINK DOWN
Nov 18 19:11:49 barney fcaw: [ID 585631 kern.notice] NOTICE: fcaw1 NPORT
Initialization Complete, SID=4061B5
Nov 18 19:11:49 barney fcaw: [ID 580805 kern.info] fcaw1: LINK UP
(180002FF)
Nov 18 19:11:50 barney fcaw: [ID 580805 kern.info] fcaw1: Host: Port 011100
(100000E06940013B:200000E0694061B5)
Nov 18 19:11:50 barney fcaw: [ID 580805 kern.info] fcaw1: Port 011B00
(5005076300C00B16:10000000C921C0CC) available.

```

This host has two JNI HBAs installed that load at fcaw0 and fcaw1. The HBAs are running with firmware 13.3.7 and driver 2.5.9.

The WWPNs are visible as 200000E06940013B for fcaw0 and 200000E0694061B5 for fcaw1. This information is required during configuration of the switch in FC-SW environments and for configuration of the ESS within the ESS Specialist for both FC-SW and loop environments.

During the second reconfigure boot, as the host discovers the logical volumes, information similar to the following will scroll by:

```

Nov 18 19:11:56 barney fcaw: [ID 580805 kern.info] fcaw0: Target 0: Port
011900 (5005076300C00B16:10000000C921E28C) online.
Nov 18 19:11:56 barney fcaw: [ID 580805 kern.info] fcaw0: Target 0 Lun 0:
Port 011900 (5005076300C00B16:10000000C921E28C) present.
Nov 18 19:11:56 barney fcaw: [ID 580805 kern.info] fcaw1: Target 0: Port
011B00 (5005076300C00B16:10000000C921C0CC) online.

```

```

Nov 18 19:11:56 barney fcaw: [ID 580805 kern.info] fcaw1: Target 0 Lun 0:
Port 011B00 (5005076300C00B16:10000000C921C0CC) present.
Nov 18 19:11:56 barney scsi: [ID 193665 kern.info] sd60 at fcaw0: target 0
lun 0
Nov 18 19:11:56 barney genunix: [ID 936769 kern.info] sd60 is
/sbus@2,0/fcaw@2,0/sd@0,0
Nov 18 19:11:56 barney scsi: [ID 193665 kern.info] sd75 at fcaw1: target 0
lun 0
Nov 18 19:11:56 barney genunix: [ID 936769 kern.info] sd75 is
/sbus@6,0/fcaw@2,0/sd@0,0
Nov 18 19:11:56 barney scsi: [ID 365881 kern.info]          <IBM-2105F20-5766
cyl 12205 alt 2 hd 30 sec 64>
Nov 18 19:11:56 barney last message repeated 1 time

```

The logical volume is discovered by the driver and the information pertaining to the WWNN and WWPN for the ESS and the associated HA, respectively, are displayed. The WWNN and WWPN are the names provided by the ESS to the host during the discovery and login process. If the ESS does not recognize the hosts WWPN, then these messages will be absent as well as any logical volumes that would have been presented by the ESS to this host.

The logical volume is assigned an sd device name that will be retained within the host. Also, the full pathname to the device is displayed along with information pertaining to the device type. IBM-2105F20 device type reflects that an ESS logical volume has been presented to the host. The actual number of cylinders, heads, and sectors presented will vary with the size of the logical volume.

- Please note that while it is possible to configure the ESS with a wide variety of logical volume sizes, the Solaris operating system may not know how to deal with some of the unusual disk types this will create. For example, the 14 GB disk type presented by the ESS requires a new disk type definition within the format utility. Otherwise, the utility will fail when attempting to format the disk with a bad magic message.
- Using logical volumes that are powers of 2 for all volumes greater than 10 GB seems to solve this problem on the Solaris operating system. However, only logical volumes of 16 GB and 32 GB were formatted successfully. Larger logical volumes have not been tested and may fall outside the range of known disk types for the Solaris operating system. The larger logical volumes should be tested for compatibility with the format utility before using them in production environments.

If the host is running Sun Veritas Volume Manager and has licensed DMP, then a message similar to the following will appear:

NOTICE: vxvm:vxdump: added disk array 14635

The number 14635 is the serial number for the ESS. Multiple numbers will be displayed if multiple ESSs are configured and attached to a host running DMP.

6.7 Sun Veritas Volume Manager

ESS logical volumes appear and can be used just like any fibre-attached disk drive. The logical volumes can be formatted, partitioned, encapsulated under Sun Veritas Volume Manager as simple or sliced disks, or used as raw disks.

When using a logical volume as a file system — whether under Veritas control or just a disk partition — be sure to keep in mind the following:

- Large files — Solaris has the ability to “help” other functions and applications deal more intelligently with files over 2 GB in size. This is done during the mount phase with the large files option. Otherwise, some applications may not behave nicely when files grow larger than 2 GB in size.
- Logging — Sun’s Veritas Fast Filesystem allows for faster disk access, faster crash recovery, almost instantaneous filesystem creation, and a better method of handling large numbers of files on a filesystem. However, if the VxFS product is not installed, it is still possible to get the faster recovery and pseudo-journaling for the filesystem using the logging option during the mount phase of a UFS filesystem. Logging takes approximately 1 MB of disk space for each 1 GB capacity up to a maximum of 64 MB. This is definitely an option to turn on to reduce the `fsck` phase after a crash — especially with large filesystems.

6.7.1 Creating a filesystem under Veritas

The Sun Veritas Volume Manager and the Sun Veritas Volume Manager Storage Administrator provide the building blocks to administer any amount of disk space from any type of disk array attached to a Sun server — including the ESS.

The Storage Administrator (VMSA) should be run as the root account. Otherwise, changes to the disks, subdisks, and volumes will not be possible. Launching VMSA is done by entering `/opt/VRTSvmsa/bin/vmsa` assuming the application was installed in the default directory and the application bin directory is not in the search path. Next, enter the host, account name, and password for this session. This will display the standard VMSA view as seen in Figure 40.

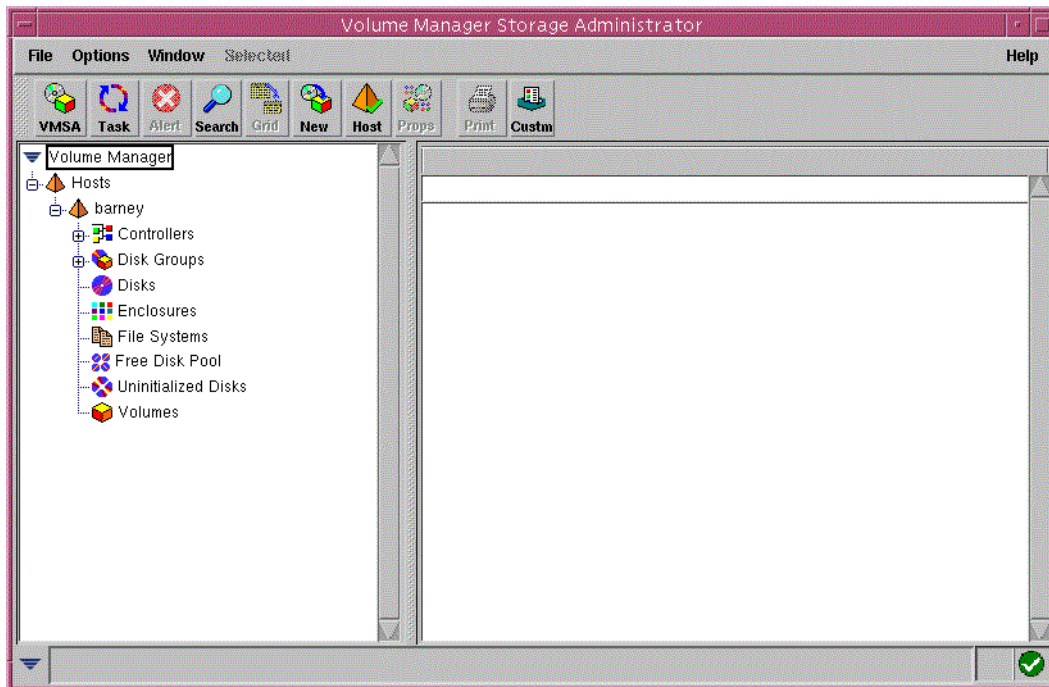


Figure 40. Volume Manager Storage Administrator

To open the Controllers icon in the left pane, simply double-click the line with the **Controllers** icon, or click the **plus sign** in the box to the left of the **Controllers** icon. On Sun Solaris systems, controller c0 is always the root or boot controller.

Clicking one of the other controllers under this view — in this case, either **c4** or **c5** — will produce a list of devices attached to that controller in the right pane. If any of these devices are under Veritas control, then the Disk Name and Disk Group columns will have that information. This will produce a view similar to the one in Figure 41.

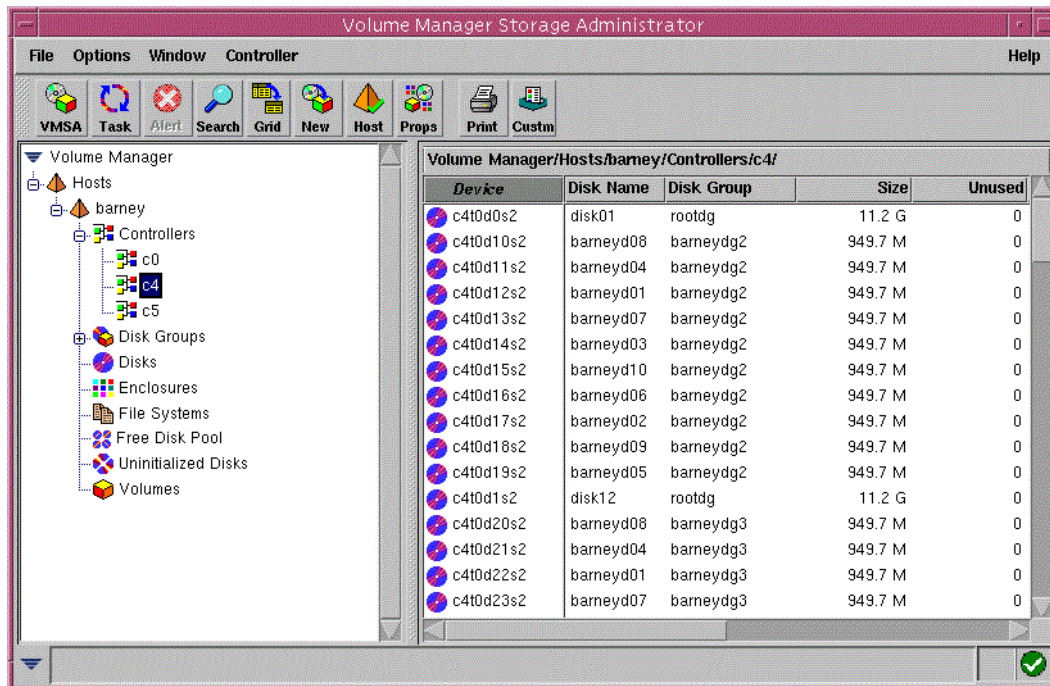


Figure 41. VMSA controller view

Right-clicking one of the devices in the right pane will produce a pop-up menu with several options including **Properties**. Select **Properties** to view a window similar to the one in Figure 42.

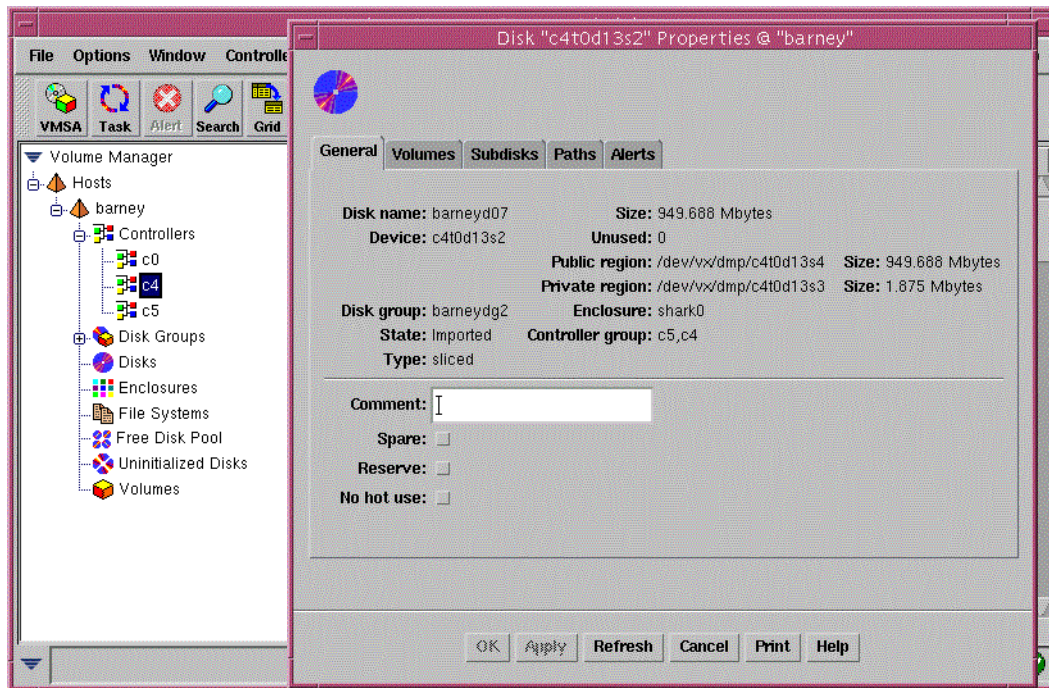


Figure 42. VMSA disk properties

Before creating any volume, it is important to know if and/or how the device(s) being contemplated for use are currently being used. The Disk "XXXXX" Properties view provides a great deal of information pertaining to the selected device.

Selecting any of the tabs across the top row also provides information pertaining to the associated device, such as Volumes, Subdisks, Paths, and Alerts. For example, within the General view above, the Controller group field near the center of the display shows both c5, c4. This is indicative of Sun Veritas DMP on the host, as both controllers point to the same set of logical volumes on the ESS.

All Sun Veritas Volume Manager (VxVM) installations come with a rootdg disk group by default. This disk group is so important that the rootdg disk group cannot be removed. The VxVM application will not work properly should the rootdg become corrupt or become missing for any reason. For that reason alone, only the boot disk, the chosen mirror for the boot disk (on an internal SCSI adapter), and any on-board drives (again, on an internal SCSI adapter)

that will be used for swap/paging space should be included in the rootdg disk group.

While it is possible to include ESS logical volumes in the rootdg disk group, it is highly recommended that all logical volumes and all fibre-attached storage of any type be associated with other disk groups. This will eliminate any possibility of damage or corruption to the Sun Veritas Volume Manager database(s) on the rootdg disk group should the fibre attached storage be unavailable, for any reason.

Right-click the **rootdg** icon or text to display the **Disk Group** pop-up menu similar to the one in Figure 43.

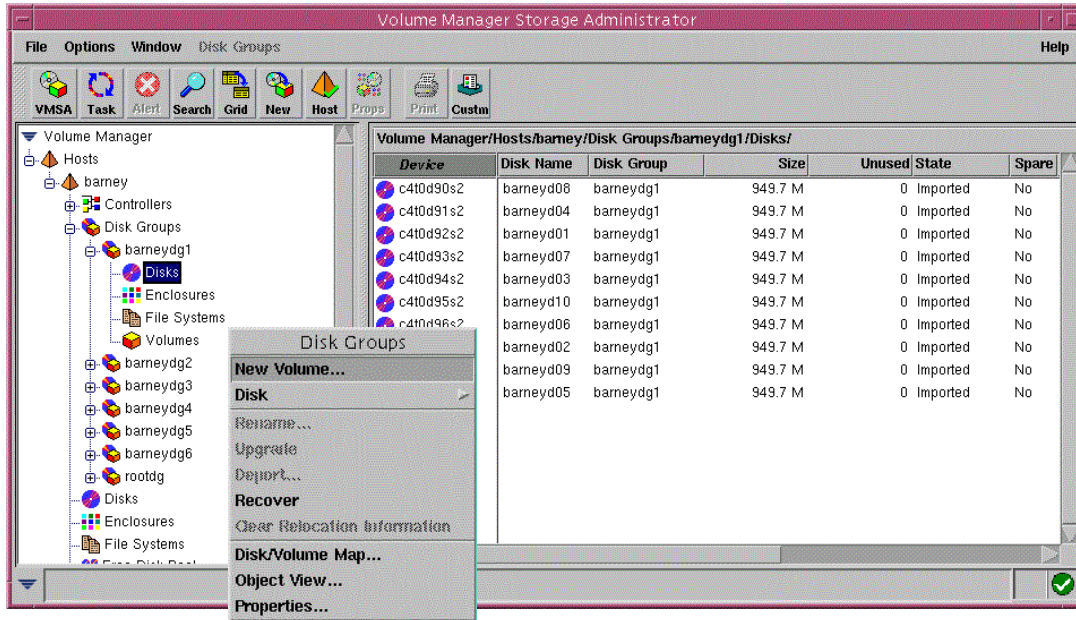


Figure 43. VMSA right-click on rootdg to get Disk Group menu

Selecting **New Volumes** from the pop-up menu will open a view similar to the one in Figure 44.

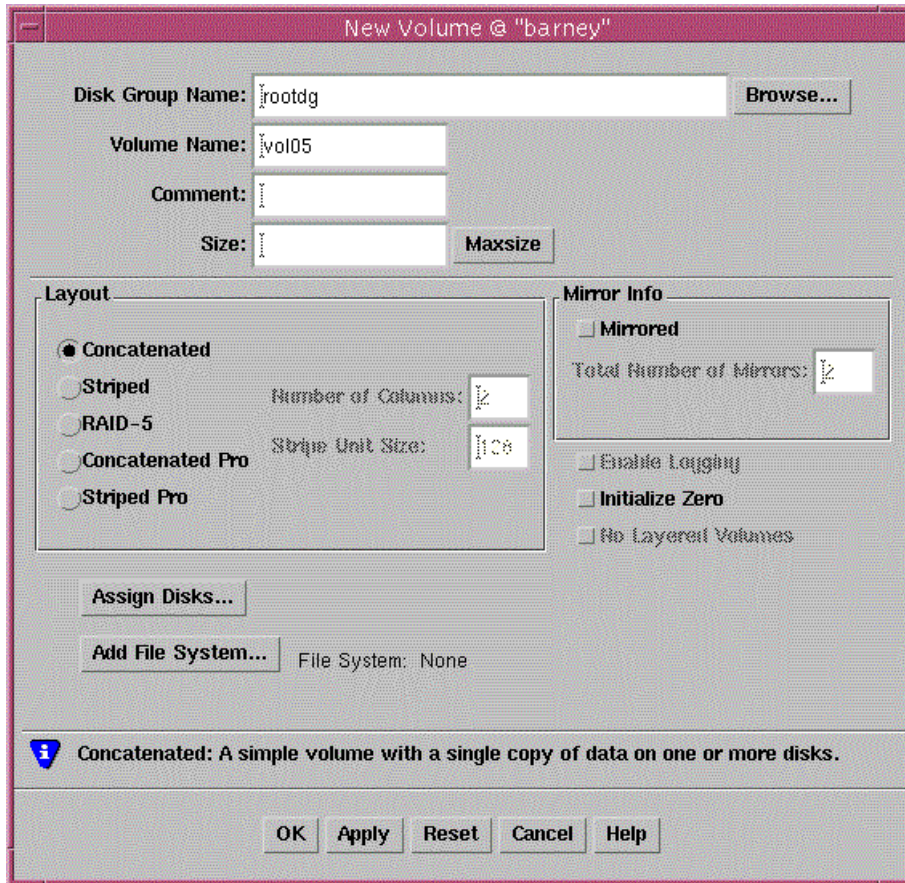


Figure 44. VMSA New Volume view

Even though the **rootdg** disk group was used to open this view, it is possible to change to another disk group by either typing in the new disk group name or using the **Browse** button to select from the disk groups configured on this host.

Also, the **Volume Name** can be renamed to anything the administrator desires. However, a naming convention that is illustrative of the use of the volume and/or the location of the devices that make up the volume is recommended. Using a naming convention that makes it visually obvious where the volume derives its storage (if possible) will only make the job as system administrator easier during changes, additions, or problem management in the future.

Select **Assign Disks** to get a listing of devices within the rootdg disk group. Expand the **Disk Group** icon in the left pane and then click the **rootdg** icon to display all the devices available in the rootdg disk group. What is visible should be a view similar to the one in Figure 45.

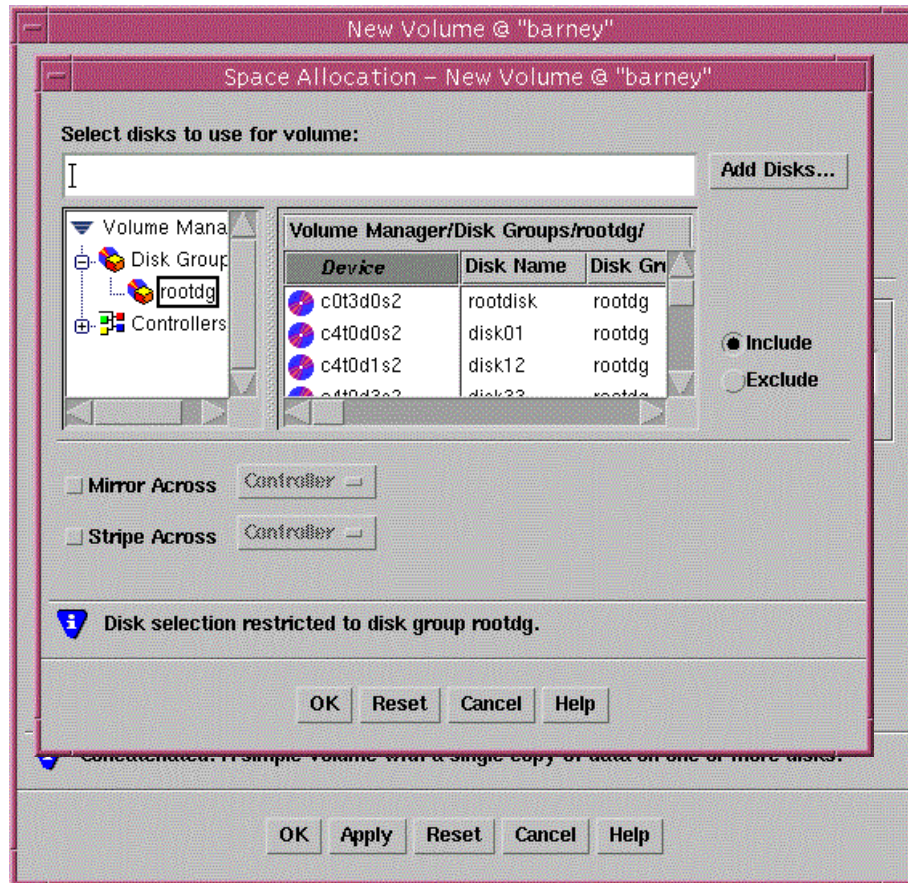


Figure 45. VMSA Assign Disks view

To find a device with available storage (assuming that the devices required for the new volume are not known yet), simply move the slider at the bottom of the right pane to the right. This will bring up the Available column, which will show how much space is available for use in the creation of a new volume on each of the devices.

In this example, a single disk is selected. Click the **OK** button to return to the New Volume view with the disk name displayed to the right of the **Assign Disks** button as in Figure 46.

Also in Figure 46, the **Maxsize** button can be clicked to utilize all the available free space on the disk for this new volume.

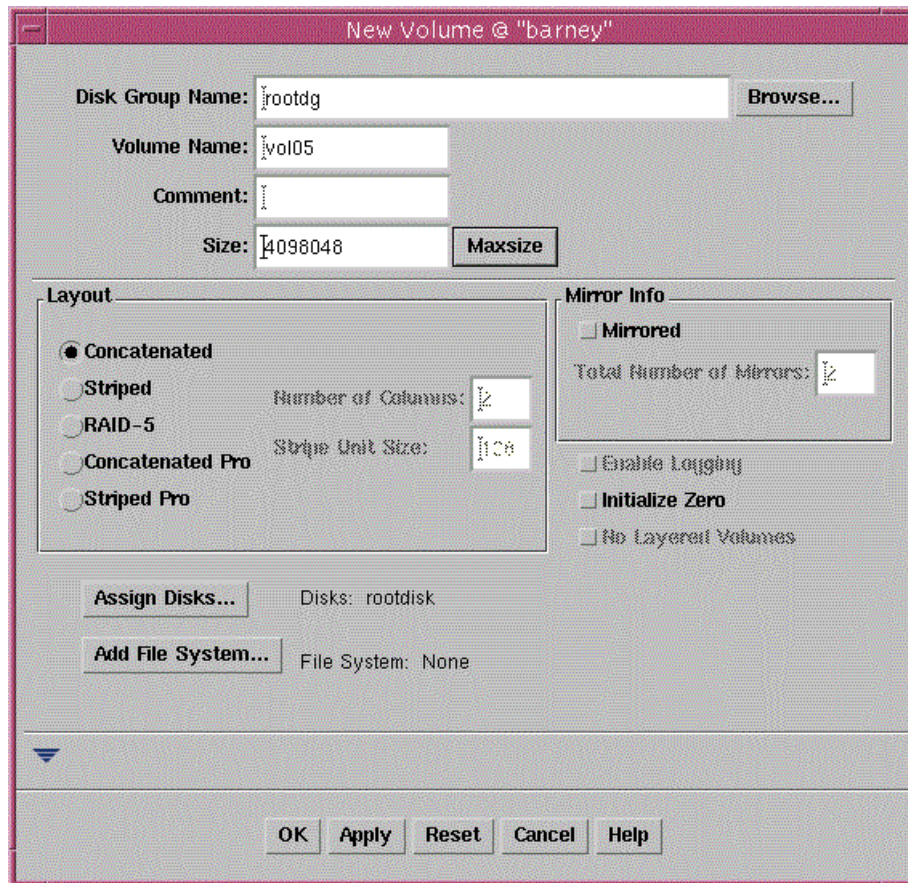


Figure 46. VMSA New Volume with disk information

Another option would be to enter some value, in 512-byte blocks, which is less than the space available on the selected disks.

One method of finding the total space available on a group of disks would be to select the disks and return to this view. Then, clicking **Maxsize** would display the maximum space available using a Concatenated layout as illustrated above. Use some value less than the system-obtained maximum size for the new volume.

Also, while it is possible to set up striped, mirrored, or both striped and mirrored volumes, great care must be taken by the administrator to ensure

that the layout of the volume will actually improve performance. It is entirely possible that a layout other than concatenated could be created that would actually degrade performance with regard to storage on the ESS.

The design of the disk read/write and cache algorithms within the ESS allows for faster data access, as long as most of those reads are sequential in nature. Even if the reads are non-sequential, the RAID-5 layout of the logical volumes on the ranks within the ESS allow more drives to be accessed than is possible using a JBOD configuration.

Suffice it to say that most volumes created using ESS logical volumes perform very well using the concatenated layout. Only during those special cases where a mirror is necessary for “snapshot” or off-line backups should anything other than simple concatenation be used.

If a filesystem is desired on the volume, it can be created and mounted within this session, as well. Clicking the **Add File System** button will produce a view similar to the one in Figure 47.

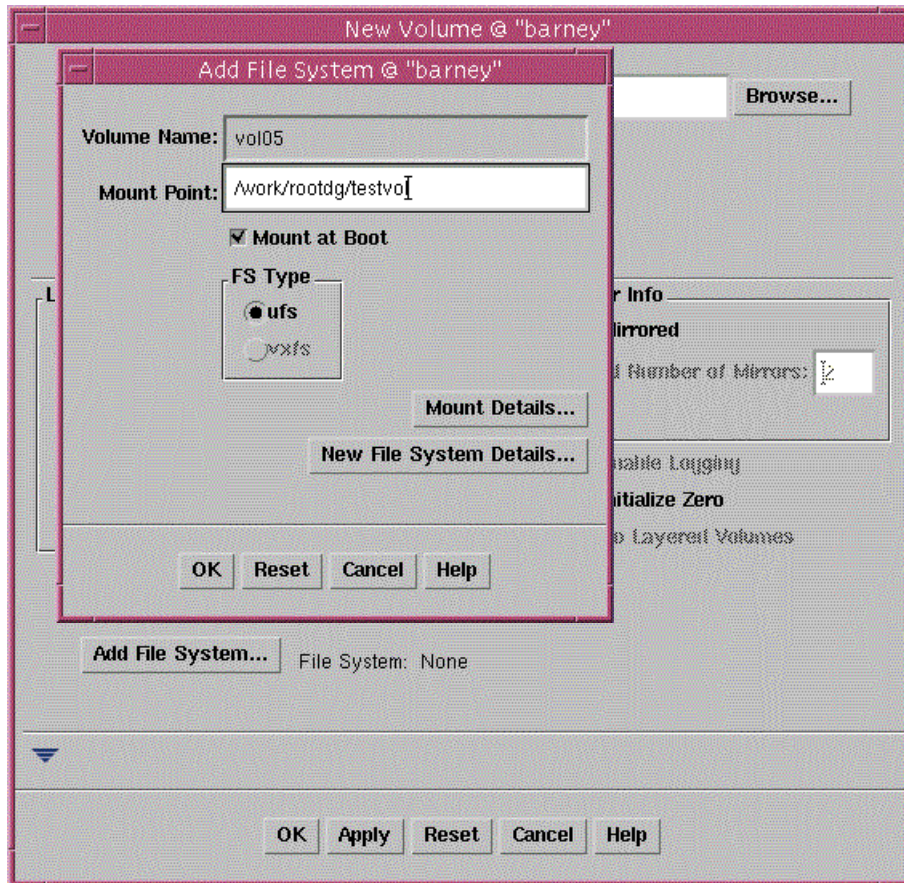


Figure 47. VMSA Add File System view

It is not possible to change the volume name within this view. However, the mount point can be entered, as well as whether or not an entry should go in the /etc/vfstab file for automatic mount during system boot.

Should any other filesystem type be available, such as VxFS, it would be possible to select it using one of the radio buttons. However, in this example, the only choice is the **UFS** filesystem type.

Mount arguments can be entered by clicking the **Mount Details** button. A view similar to the one in Figure 48 will be displayed.

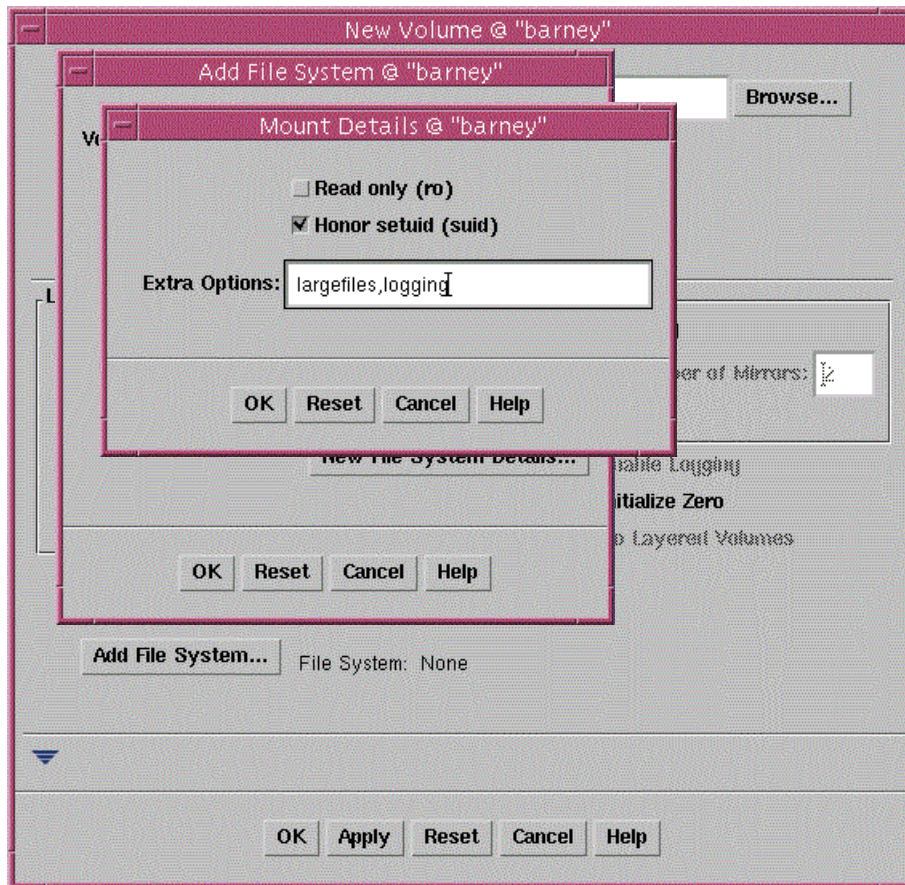


Figure 48. VMSA Mount Details view

For most applications, the defaults are fine. However, for large volumes, it is recommended that the large files option be used. Also, for all volumes, the logging option is *highly* recommended. See 6.7, “Sun Veritas Volume Manager” on page 95 and the operating system man pages for more information.

Click **OK** on each of the windows until the New Volume view is displayed with the filesystem information to the right of the **Add File System** button, as shown in Figure 49.

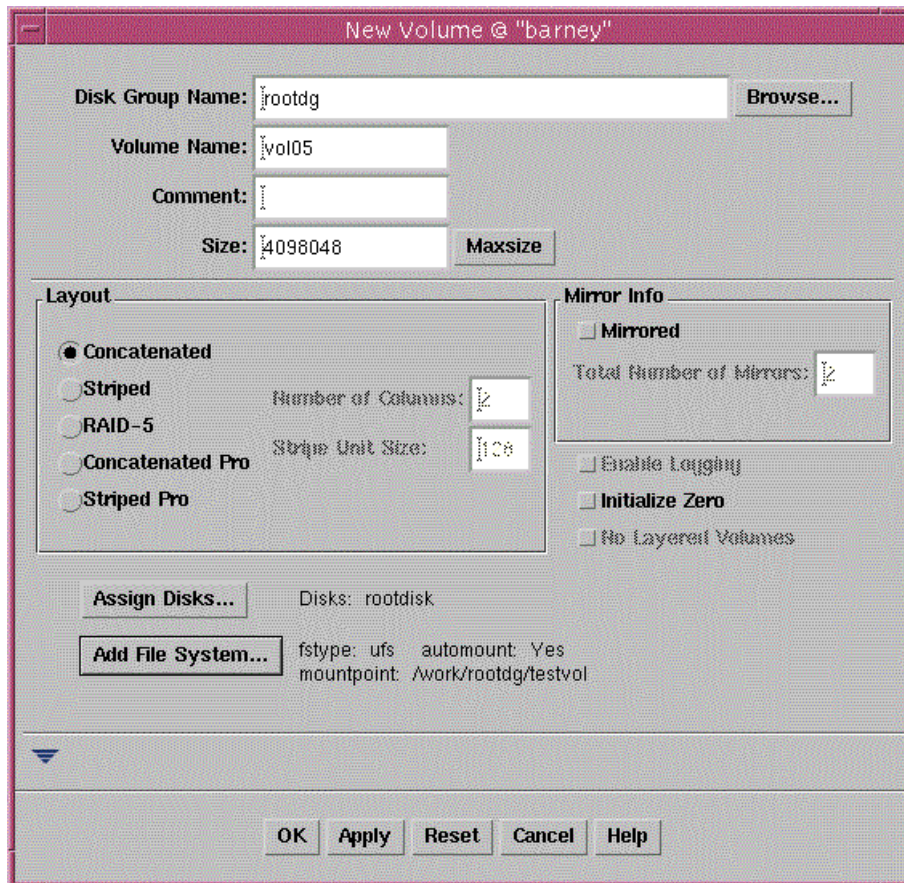


Figure 49. VMSA New Volume with disk and filesystem info

At this point, clicking the **OK** or **Apply** button will begin the volume and filesystem creation process. The **OK** button will cause the New Volume view to vanish, while the **Apply** button will retain the view and allow additional volumes to be created, as desired.

After clicking the **OK** button, the VMSA main view will be visible, as illustrated in Figure 50.

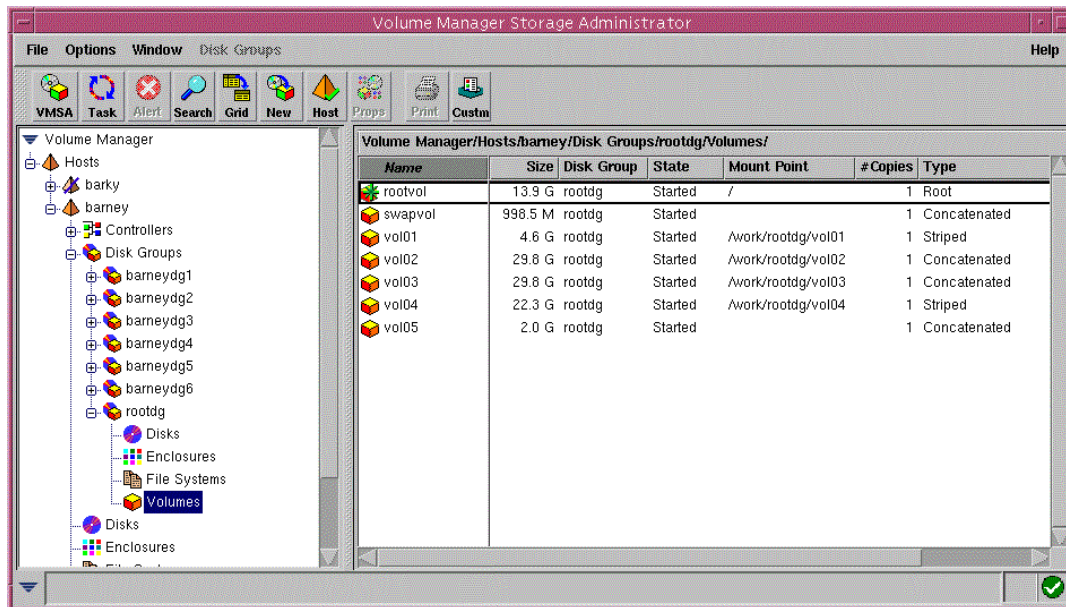


Figure 50. VMSA filesystem/volume creation in progress

Note that the new volume — vol05 — is still “under construction”, as the mount point is not yet visible. As soon as the volume has been created and mounted, the view will change to that in Figure 51.

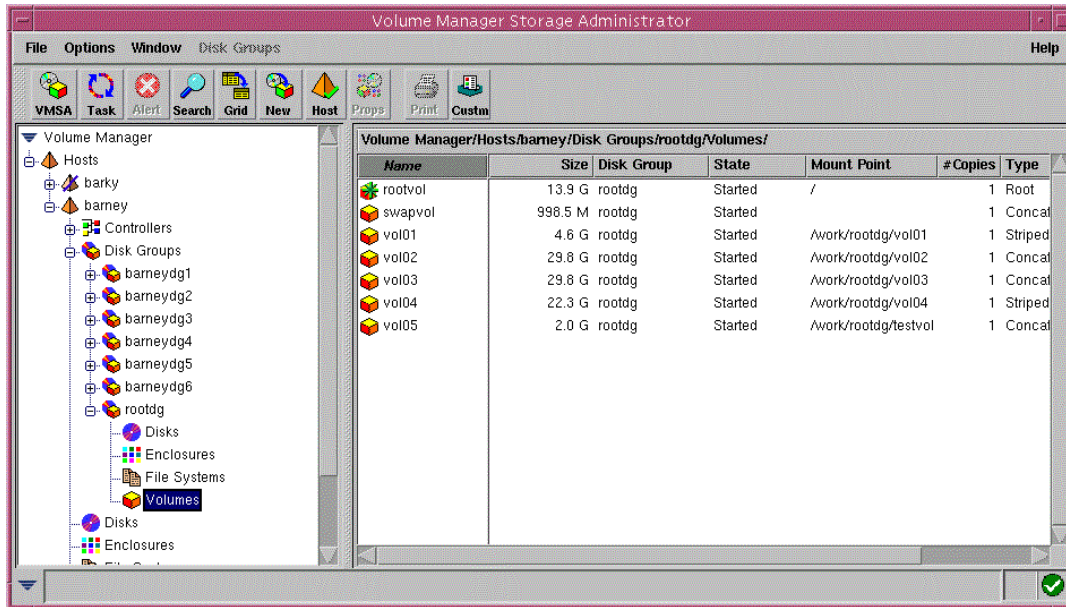


Figure 51. VMSA filesystem/volume creation complete

A quick look at the entry in the `/etc/vfstab` file verifies that the mount information is ready for the next system boot. The last line illustrates the entry, including the mount options for large files and logging (see Figure 52).

```

/dev/vx/dsk/barneydg6/vol06 /dev/vx/rdsk/barneydg6/vol06 /work/barneydg6/
vol06 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol07 /dev/vx/rdsk/barneydg6/vol07 /work/barneydg6/
vol07 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol08 /dev/vx/rdsk/barneydg6/vol08 /work/barneydg6/
vol08 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol09 /dev/vx/rdsk/barneydg6/vol09 /work/barneydg6/
vol09 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol10 /dev/vx/rdsk/barneydg6/vol10 /work/barneydg6/
vol10 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol11 /dev/vx/rdsk/barneydg6/vol11 /work/barneydg6/
vol11 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol12 /dev/vx/rdsk/barneydg6/vol12 /work/barneydg6/
vol12 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol13 /dev/vx/rdsk/barneydg6/vol13 /work/barneydg6/
vol13 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg4/vol01 /dev/vx/rdsk/barneydg4/vol01 /work/barneydg4/
vol01 ufs 3 yes largefiles,logging
/dev/vx/dsk/barneydg6/vol14 /dev/vx/rdsk/barneydg6/vol14 /work/barneydg6/
vol14 ufs 3 yes largefiles,logging
/dev/vx/dsk/rootdg/vol05 /dev/vx/rdsk/rootdg/vol05 /work/rootdg/tes
tvol ufs 3 yes largefiles,logging
~

```

Figure 52. View of /etc/vfstab entries

6.7.2 Sun Veritas and ESS logical volumes

Within the VxSA GUI, almost everything has options that become available when the device, volume, diskgroup, and so on, are right-clicked. The pop-up menus that appear provide the available commands for that member under the current conditions.

If additional disk groups are required, simply right-clicking the **Disk Groups** icon in the left pane will produce a pop-up menu that has, as one of its options, **New Disk Group**. Follow the prompts under this view to create the disk group desired. Repeat this process for as many disk groups as are required.

If the VxVM software was installed after the installation of the ESS and the logical volumes were brought under VxVM control, then it will not be necessary to initialize the ESS logical volumes. However, it may be necessary to remove the logical volumes from the rootdg. Under the **rootdg** icon, click the **Disks** icon to view the devices associate with the rootdg disk group.

Using the Shift-click and/or the Ctrl-click method, select the devices that will be moved to another disk group or to the Free Pool. Then, right-click one of

the selected devices and click **Move to Free Disk Pool**. The VxVM software will then move all the free devices within the rootdg to the Free Disk Pool area. This step is necessary, as VxVM will not allow a device within one disk group to be used in the creation of a volume in another disk group.

Once the devices are in the Free Disk Pool, they can be moved to any of the disk groups using the technique described above.

When the logical volumes are in the appropriate disk group, follow the steps outlined in 6.7.1, "Creating a filesystem under Veritas" on page 95 to create the volumes desired.

6.7.3 ESS identification under Veritas

When a system like the ESS is brought under Veritas control, it will show up under the **Enclosures** icon, as can be seen in Figure 53.

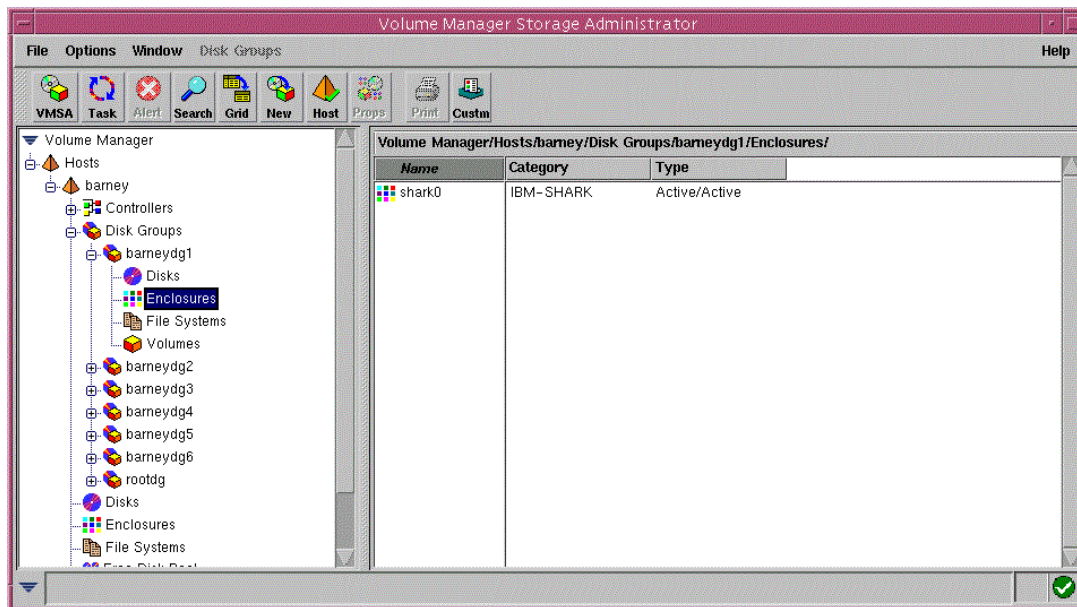


Figure 53. VMSA Enclosure view

The ESS will be given the name *shark0* to identify the array to the VxVM software. Multiple ESSs will be named consecutively.

6.8 Known Issues

The following are known problems or issues in this release of ESS code KINT1027.

QLogic HBAs cause switch resets when used in a zoned environment

Switch resets occur when a zoned host using QLogic HBAs is booted. Other hosts lose access to their respective ESS logical volumes during the switch reset. This can cause I/O time-outs, job failures, and system crashes, depending on system load. This is an industry-wide known problem with QLogic HBAs.

Workaround

Do not use QLogic HBAs in FC-SW environments. Only FC-AL connections between the ESS and the Sun server are recommended.

Running 'sys-unconfig' on a host causes problems with HBA drivers running on FC-SW protocol

During a 'sys-unconfig' of a host, the software attempts to configure the HBA driver as an ethernet port — for example, the Emulex is seen as `lpfcn0` — during the configuration phase. This causes the system to hang indefinitely, and the system becomes unbootable.

Workaround

Boot the host in single-user mode and remove the driver for the Fibre Channel HBA. Then reboot the host and continue with the system configuration process. When complete, reboot the host and install the driver for the Fibre Channel HBA. Verify the changes in the appropriate configuration and system files before rebooting.

Concurrent code load with high I/O and an ESS configured with 3,000 LUNs or more may fail

During concurrent code load of the ESS while the ESS is under heavy I/O and 3,000 or more LUNs are configured on the ESS — whether upgrading the HA firmware or not — the ESS may hang and require a reboot to recover.

Workaround

Reduce the I/O activity on the ESS during the period of the code load.

Appendix A. Test suite details

The testing was divided into three parts, which are described in this appendix:

- I/O workload integrity
- ESS exception tests
- Host clustering function test

A.1 I/O workload integrity

Configure the servers and ESS with operating system for representative storage volumes. Ensure that cabling and switches are properly installed and configured. A heavy I/O work load was generated in RAW, FORMATTED I/O, buffered and non-buffered I/O modes. Please see Chapter 5.4.1, “Test results” on page 79 for results and details.

A.2 ESS exception tests

Perform ESS exception tests as indicated below with Customer Simulated Operation (CSO) — an I/O intense application — running in the background.

A.2.1 Test E.1: ESS warm start

Performing a *warm start* will cause BOTH clusters to be unavailable for a brief period of time. This should be less than 30 seconds with the current ESS microcode. With host(s) connected and running an I/O exerciser such as IOGEN, IOMETER, or BLAST, perform the following at least 100 times for each operating system platform:

1. Run goodpath I/O for at least 1/2 hour.
2. Log into a cluster as service1.
3. Select **Utility Menu**.
4. Select **Trace/State Save**.
5. Select **Force a State Save**.

These steps will cause BOTH clusters to be warm started. The ESS will be unavailable for about 30 seconds, after which time the ESS should recover and I/O should resume.

Note: Wait at least 15 minutes before performing another warm start. A script may be used to automate this process.

A.2.2 Test E.2: ESS failover/failback

All I/O is handled by the remaining cluster when one of the clusters fails. This test simulates a cluster failure. With host(s) connected and running an I/O exerciser such as IOGEN, IOMETER, or BLAST, perform the following at least 5 times for each operating system platform:

1. Run goodpath I/O for at least 1/2 hour.
2. Find the power reset switches for cluster bay 1; open up the rear of the ESS. Look for the RPC cards (these cards are located at the lower middle of the ESS rack). The power reset switches for cluster bay 1 are located about halfway down in the center of the RPC card. There are two reset switches vertically located on each RPC card.
3. For cluster bay 1, depress the top switch simultaneously on each RPC card. This will cause a failover in cluster bay 1, and the resource will be fenced (It will not fail back automatically).
4. After about 5 minutes, log into service1 on cluster bay 2.
Note: This login process will take longer than it normally does when both clusters are active.
5. Select **Utility Menu**.
6. Select **Resource Management Menu**.
7. Select **Show Fenced Resources**. Confirm that a Fence and FailOver have occurred on Cluster 1.
8. Now use F3 to back up to the **Resource Management Menu**.
9. Select **Reset Fenced for a Resource**.
10. Select **Reset Fence** at the bottom of the screen and press Enter.
11. At the message screen, tab down to **cpcluster0 R1-T1 Cluster Bay 1** and use F7 to select **cpcluster0 R1-T1** (this will put a > by it). Press Enter.
12. After a confirmation message appears, press Enter again.
Note: This step can take up to 15 minutes to complete, and a **Reset Fence successful** message should be displayed. Verify that the message is for cluster 0, and press Enter.

Note: Wait at least one hour after one failover and failback before doing attempting this test again.

A.2.3 Test E.3: ESS cluster quiesce/resume

All I/O is handled by the remaining ESS cluster when one of the clusters is quiesced (generally to perform maintenance functions). This test includes performing a cluster quiesce and resume operation while continuing to run I/O. With host(s) connected and running an I/O exerciser such as IOGEN, IOMETER, or BLAST, perform the following to quiesce a cluster at least 5 times for each operating system platform:

1. Log into a cluster as service1.
2. Select **Utility Menu**.
3. Select **Resource Management Menu**.
4. Select **Quiesce a Resource**.
5. Move cursor down to select which cluster to quiesce. Select a cluster with F7 and then press Enter.
Note: Quiesce the cluster you are not logged into.
6. Move the cursor down to **Quiesce Selected Resource(s)** and press Enter.
7. Press Enter to clear the confirmation screen.
8. Continue to run I/O to the remaining cluster for at least one hour before performing the resume operation.
9. After running for at least one hour with a single active cluster, perform the following steps to resume the quiesced cluster:
10. Log into the cluster that is not quiesced
11. Select **Utility Menu**.
12. Select **Resource Management Menu**.
13. Select **Resume a Resource**.
14. Move cursor down to highlight cluster to be resumed. Use F7 to select, and then press Enter.
15. Move cursor down to highlight **Resume Selected Resource(s)**, then press Enter.
16. Press Enter to clear the confirmation screen.

A.2.4 TEST E.4: Cable pulls

Perform applicable cable pulls to simulate cable, HBA, GBIC, or switch failure. Repeat a minimum of 20 times for each operating system version.

1. Disconnect cabling between server and switch or ESS. Reconnect after 60 seconds.
2. Disconnect cabling between switch and ESS. Reconnect after 60 seconds.

A.3 Host clustering function

The tests in this next series are designed to test the host clustering function.

A.3.1 Test C1: Manual application package switch-over

This test has to be performed on each node in the host server cluster.

Test

Manually switch application package (CSO or I/O simulation) from primary node to backup node, and then back to primary node. Monitor package log and application log (if any). Logon and connect to application (if applicable). Include ESS exception tests while performing this test.

Results

Application starts up on alternate node with no observed problems and is available after switch-over.

A.3.2 Test C2: Application process failure test

This test has to be performed on each cluster package running on each cluster node.

Test

This test only applies if there is application service monitoring. Terminate the processes of monitored application service(s).

Results

Application starts up on alternate node with no observed problems, and is available after switch-over.

A.3.3 Test C3: Test of each primary Heartbeat LAN

This test has to be performed on each node separately.

Test

Check Heartbeat LAN against Local LAN Failover to Standby LAN card.
Monitor syslog for messages of primary HB LAN card failing, and standby LAN card picking up the active Ethernet Stack.

Results

Communication continues with no observed problems; netstat -rn shows previous IP on Failover Card.

Post action

Failback is done from standby card to primary card.

Appendix B. Special notices

This publication is intended to help experienced system administrators to install, tailor, and configure the IBM Enterprise Storage Server (ESS) when attaching Compaq AlphaServer running Tru64 UNIX, HP, and Sun hosts. The information in this publication is not intended as the specification of any programming interfaces that are provided by IBM. See the PUBLICATIONS section of the IBM Programming Announcement for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.


Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers

attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

IBM ®	Redbooks Logo 
Enterprise Storage Server	RS/6000
ESCON	S/390
IBM	SP
Manage. Anything. Anywhere.	StorWatch
Netfinity	System/390
Redbooks	

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere., The Power To Manage., Anything. Anywhere., TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others.

Appendix C. Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

C.1 IBM Redbooks

For information on ordering these publications see “How to get IBM Redbooks” on page 125.

- *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- *Implementing Fibre Channel Attachment on the ESS*, SG24-6113
- *Planning and Implementing an IBM SAN*, SG24-6116

C.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at ibm.com/redbooks for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
IBM System/390 Redbooks Collection	SK2T-2177
IBM Networking Redbooks Collection	SK2T-6022
IBM Transaction Processing and Data Management Redbooks Collection	SK2T-8038
IBM Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
IBM AS/400 Redbooks Collection	SK2T-2849
IBM Netfinity Hardware and Software Redbooks Collection	SK2T-8046
IBM RS/6000 Redbooks Collection	SK2T-8043
IBM Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

C.3 Other resources

These publications are also relevant as further information sources:

- *Enterprise Storage Server Configuration Planner*, SC26-7353
- *Introduction and Planning Guide 2105 Models E10, E20, F10, and F20*, GC26-7294

C.4 Referenced Web sites

These Web sites are also relevant as further information sources:

- <http://www.ibm.com/redbooks> IBM Redbooks home page
- <http://www.ibm.com/storage> IBM Storage home page
- <http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm>
ESS Supported servers page
- <http://httpd.apache.org/dist/binaries/digitalunix>
Index of Compaq binaries page
- <http://www.service.digital.com/patches/index.htm> Compaq Fix page
- <http://www.compaq.com/support> Compaq Support home page
- <http://www.unix.digital.com/cluster> Compaq Unix Cluster page
- <http://www.compaq.com/tru64unix> CompaqTru64Unix home page
- <http://www.elink.ibm.link.ibm.com/pbl/pbl> IBM Publications

How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** ibm.com/redbooks

Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

Send orders by e-mail including information from the IBM Redbooks fax order form to:

	e-mail address
In United States or Canada	pubscan@us.ibm.com
Outside North America	Contact information is in the "How to Order" section at this site: http://www.elink.ibm.com/pbl/pbl

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the "How to Order" section at this site: http://www.elink.ibm.com/pbl/pbl

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the "How to Order" section at this site: http://www.elink.ibm.com/pbl/pbl

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at <http://w3.itso.ibm.com/> and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at <http://w3.ibm.com/> for redbook, residency, and workshop announcements.

IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name _____ Last name _____

Company _____

Address _____

City _____ Postal code _____ Country _____

Telephone number _____ Telefax number _____ VAT number _____

Invoice to customer number _____

Credit card number _____

Credit card expiration date _____ Card issued to _____ Signature _____

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Glossary

Glossary

This glossary contains a list of terms used within this redbook.

A

allegiance. The ESA/390 term for a relationship that is created between a device and one or more channel paths during the processing of certain condition.

allocated storage. On the ESS, this is the space that you have allocated to volumes, but not yet assigned.

application system. A system made up of one or more host systems that perform the main set of functions for an establishment. This is the system that updates the primary DASD volumes that are being copied by a copy services function.

AOM. Asynchronous operations manager.

APAR. Authorized program analysis report.

array. An arrangement of related disk drive modules that you have assigned to a group.

assigned storage. On the ESS, this is the space that you have allocated to volumes, and assigned to a port.

asynchronous operation. A type of operation in which the remote copy XRC function copies updates to the secondary volume of an XRC pair at some time after the primary volume is updated. Contrast with synchronous operation.

ATTIME. A keyword for requesting deletion or suspension at a specific target time.

availability. The degree to which a system or resource is capable of performing its normal function.

B

bay. Physical space on an ESS rack. A bay contains SCSI, ESCON or Fibre Channel interface cards and SSA device interface cards.

backup. The process of creating a copy of data to ensure against accidental loss.

C

cache. A random access electronic storage in selected storage controls used to retain frequently used data for faster access by the channel.

cache fast write. A form of fast write where the subsystem writes the data directly to cache, where it is available for later destaging.

CCA. Channel connection address.

CCW. Channel command word.

CEC. Central electronics complex.

channel. (1) A path along which signals can be sent; for example, data channel and output channel. (2) A functional unit, controlled by the processor, that handles the transfer of data between processor storage and local peripheral equipment.

channel connection address (CCA). The input/output (I/O) address that uniquely identifies an I/O device to the channel during an I/O operation.

channel interface. The circuitry in a storage control that attaches storage paths to a host channel.

channel path. The ESA/390 term for the interconnection between a channel and its associated controllers.

channel subsystem. The ESA/390 term for the part of host computer that manages I/O communication between the program and any attached controllers.

CKD. Count key data. An ES/390 architecture term for a device that specifies the format of and access mechanism for the logical data units on the device. The logical data unit is a track that can contain one or more records, each consisting of a count field, a key field (optional), and a data field (optional).

CLIST. TSO command list.

cluster. See storage cluster.

cluster processor complex (CPC). The unit within a cluster that provides the management function for the storage server. It consists of cluster processors, cluster memory, and related logic.

concurrent copy. A copy services function that produces a backup copy and allows concurrent access to data during the copy.

concurrent maintenance. The ability to service a unit while it is operational.

consistency group time. The time, expressed as a primary application system time-of-day (TOD) value, to which XRC secondary volumes have been updated. This term was previously referred to as “consistency time”.

consistent copy. A copy of data entity (for example a logical volume) that contains the contents of the entire data entity from a single instant in time.

contingent allegiance. ESA/390 term for a relationship that is created in a controller between a device and a channel path when unit-check status is accepted by the channel. The allegiance causes the controller to guarantee access; the controller does not present the busy status to the device. This enables the controller to retrieve sense data that is associated with the unit-check status, on the channel path with which the allegiance is associated.

control unit address (CUA). The high order bits of the storage control address, used to identify the storage control to the host system.

Note: The control unit address bits are set to zeros for ESCON attachments.

CUA. Control unit address.

D

daisy chain. A method of device interconnection for determining interrupt priority by connecting the interrupt sources serially.

DA. Device adapter.

DASD. Direct access storage device. See disk drive module.

data availability. The degree to which data is available when needed. For better data availability when you attach multiple hosts that share the same data storage, configure the data paths so that data transfer rates are balanced among the hosts.

data sharing. The ability of homogenous or divergent host systems to concurrently utilize information that they store on one or more storage devices. The storage facility allows configured storage to be accessible to any attached host systems, or to all. To use this capability, you need to design the host program to support data that it is sharing.

DDM. Disk drive module

data compression. A technique or algorithm that you use to encode data such that you can store the encoded result in less space than the original data. This algorithm allows you to recover the original data from the encoded result through a reverse technique or reverse algorithm.

data field. The third (optional) field of a CKD record. You determine the field length by the data length that is specified in the count field. The data field contains data that the program writes.

data record. A subsystem stores data records on a track by following the track-descriptor record. The subsystem numbers the data records consecutively, starting with 1. A track can store a maximum of 255 data records. Each data record consists of a count field, a key field (optional), and a data field (optional).

DASD-Fast Write. A function of a storage controller that allows caching of active write data

without exposure of data loss by journaling of the active write data in NVS.

DASD subsystem. A DASD storage control and its attached direct access storage devices.

data in transit. The update data on application system DASD volumes that is being sent to the recovery system for writing to DASD volumes on the recovery system.

data mover. See system data mover.

dedicated storage. Storage within a storage facility that is configured such that a single host system has exclusive access to the storage.

demote. The action of removing a logical data unit from cache memory. A subsystem demotes a data unit in order to make room for other logical data units in the cache. It could also demote a data unit because the logical data unit is not valid. A subsystem must destage logical data units with active write units before they are demoted.

destage. (1) The process of reading data from cache. (2) The action of storing a logical data unit in cache memory with active write data to the storage device. As a result, the logical data unit changes from cached active write data to cached read data.

device. The ESA/390 term for a disk drive.

device address. The ESA/390 term for the field of an ESCON device-level frame that selects a specific device on a control-unit image. The one or two leftmost digits are the address of the channel to which the device is attached. The two rightmost digits represent the unit address.

device adapter. A physical sub unit of a storage controller that provides the ability to attach to one or more interfaces used to communicate with the associated storage devices.

device ID. An 8-bit identifier that uniquely identifies a physical I/O device.

device interface card. A physical sub unit of a storage cluster that provides the communication with the attached DDMs.

device number. ESA/390 term for a four-hexadecimal-character identifier, for example 13A0, that you associate with a device to facilitate communication between the program and the host operator. The device number that you associate with a subchannel.

device sparing. Refers to when a subsystem automatically copies data from a failing DDM to a spare DDM. The subsystem maintains data access during the process.

Device Support Facilities program (ICKDSF). A program used to initialize DASD at installation and perform media maintenance.

DFDSS. Data Facility Data Set Services.

DFSMSdss. A functional component of DFSMS/MVS used to copy, dump, move, and restore data sets and volumes.

director. See storage director and ESCON Director.

disaster recovery. Recovery after a disaster, such as a fire, that destroys or otherwise disables a system. Disaster recovery techniques typically involve restoring data to a second (recovery) system, then using the recovery system in place of the destroyed or disabled application system. See also recovery, backup, and recovery system.

disk drive module. The primary nonvolatile storage medium that you use for any host data that is stored within a subsystem. Number and type of storage devices within a storage facility may vary.

drawer. A unit that contains multiple DDMs, and provides power, cooling, and related interconnection logic to make the DDMs accessible to attached host systems.

DRAIN. A keyword for requesting deletion or suspension when all existing record updates from the storage control cache have been cleared.

drawer. A unit that contains multiple DDMs, and provides power, cooling, and related interconnection logic to make the DDMs accessible to attached host systems.

dump. A capture of valuable storage information at the time of an error.

dual copy. A high availability function made possible by the nonvolatile storage in cached IBM storage controls. Dual copy maintains two functionally identical copies of designated DASD volumes in the logical storage subsystem, and automatically updates both copies every time a write operation is issued to the dual copy logical volume.

duplex pair. A volume comprised of two physical devices within the same or different storage subsystems that are defined as a pair by a dual copy, PPRC, or XRC operation, and are in neither suspended nor pending state. The operation records the same data onto each volume.

E

ECSA. Extended common service area.

EMIF. ESCON Multiple Image Facility. An ESA/390 function that allows LPARs to share an ESCON channel path by providing each LPAR with its own channel-subsystem image.

environmental data. Data that the storage control must report to the host; the data can be service information message (SIM) sense data, logging mode sense data, an error condition that prevents completion of an asynchronous operation, or a statistical counter overflow. The storage control reports the appropriate condition as unit check status to the host during a channel initiated selection. Sense byte 2, bit 3 (environmental data present) is set to 1.

Environmental Record Editing and Printing (EREP) program. The program that formats and prepares reports from the data contained in the error recording data set (ERDS).

EREP. Environmental Record Editing and Printing Program.

ERP. Error recovery procedure.

ESCD. ESCON Director.

ESCM. ESCON Manager.

ESCON. Enterprise Systems Connection Architecture. An ESA/390 computer peripheral interface. The I/O interface utilizes ESA/390 logical protocols over a serial interface that configures attached units to a communication fabric.

ESCON Director (ESCD). A device that provides connectivity capability and control for attaching any two ESCON links to each other.

extended remote copy (XRC). A hardware- and software-based remote copy service option that provides an asynchronous volume copy across storage subsystems for disaster recovery, device migration, and workload migration.

ESCON Manager (ESCM). A licensed program that provides host control and intersystem communication capability for ESCON Director connectivity operations.

F

failover. The routing of all transactions to a second controller when the first controller fails. Also see cluster.

fast write. A write operation at cache speed that does not require immediate transfer of data to a DDM. The subsystem writes the data directly to cache, to nonvolatile storage, or to both. The data is then available for destaging. Fast write reduces the time an application must wait for the I/O operation to complete.

FBA. Fixed block address. An architecture for logical devices that specifies the format of and access mechanisms for the logical data units on the device. The logical data unit is a block. All blocks on the device are the same size (fixed size); the subsystem can access them independently.

FC-AL. Fibre Channel - Arbitrated Loop. An implementation of the fibre channel standard that uses a ring topology for the communication fabric.

FCS. See fibre channel standard.

fibre channel standard. An ANSI standard for a computer peripheral interface. The I/O interface defines a protocol for communication over a serial interface that configures attached units to a communication fabric. The protocol has two layers. The IP layer defines basic interconnection protocols. The upper layer supports one or more logical protocols (for example FCP for SCSI command protocols, SBCON for ESA/390 command protocols). **fiber optic cable.** A fiber, or bundle of fibers, in a structure built to meet optic, mechanical, and environmental specifications.

fixed utility volume. A simplex volume assigned by the storage administrator to a logical storage subsystem to serve as working storage for XRC functions on that storage subsystem.

FlashCopy. A point-in-time copy services function that can quickly copy data from a source location to a target location.

floating utility volume. Any volume of a pool of simplex volumes assigned by the storage administrator to a logical storage subsystem to serve as dynamic storage for XRC functions on that storage subsystem

G

GB. Gigabyte.

gigabyte. 1 073 741 824 bytes.

group. A group consist of eight DDMs. Each DDM group is a raid array.

GTF. Generalized trace facility.

H

HA. Home address, host adapter.

hard drive. A storage medium within a storage server used to maintain information that the storage server requires.

HDA. Head and disk assembly. The portion of an HDD associated with the medium and the read/write head.

HDD. Head and disk drive.

home address. A nine-byte field at the beginning of a track that contains information that identifies the physical track and its association with a cylinder.

host adapter. A physical sub unit of a storage controller that provides the ability to attach to one or more host I/O interfaces.

I

ICKDSF. See Device Support Facilities program.

identifier (ID). A sequence of bits or characters that identifies a program, device, storage control, or system.

IML. Initial microcode load.

initial microcode load (IML). The act of loading microcode.

I/O device. An addressable input/output unit, such as a direct access storage device, magnetic tape device, or printer.

I/O interface. An interface that you define in order to allow a host to perform read and write operations with its associated peripheral devices.

implicit allegiance. ESA/390 term for a relationship that a controller creates between a device and a channel path, when the device accepts a read or write operation. The controller guarantees access to the channel program over the set of channel paths that it associates with the allegiance.

Internet Protocol (IP). A protocol used to route data from its source to its destination in an Internet environment.

invalidate. The action of removing a logical data unit from cache memory because it cannot support continued access to the logical data unit on the device. This removal may be the result of a failure within the storage controller or a storage device that is associated with the device.

IPL. Initial program load.

ITSO. International Technical Support Organization.

J

JCL. Job control language.

Job control language (JCL). A problem-oriented language used to identify the job or describe its requirements to an operating system.

journal. A checkpoint data set that contains work to be done. For XRC, the work to be done consists of all changed records from the primary volumes. Changed records are collected and formed into a “consistency group”, and then the group of updates is applied to the secondary volumes.

K

KB. Kilobyte.

key field. The second (optional) field of a CKD record. The key length is specified in the count field. The key length determines the field length. The program writes the data in the key field. The subsystem uses this data to identify or locate a given record.

keyword. A symptom that describes one aspect of a program failure.

kilobyte (KB). 1 024 bytes.

km. Kilometer.

L

LAN. See local area network.

least recently used. The algorithm used to identify and make available the cache space that contains the least-recently used data.

licensed internal code (LIC).

(1) Microcode that IBM does not sell as part of a machine, but licenses to the customer. LIC is implemented in a part of storage that is not addressable by user programs. Some IBM products use it to implement functions as an alternative to hard-wired circuitry.

(2) LIC is implemented in a part of storage that is not addressable by user programs. Some IBM products use it to implement functions as an alternative to hard-wired circuitry.

link address. On an ESCON interface, the portion of a source, or destination address in a frame that ESCON uses to route a frame through an ESCON director. ESCON associates the link address with a specific switch port that is on the ESCON director. Equivalently, it associates the link address with the channel-subsystem, or controller-link-level functions that are attached to the switch port.

link-level facility. ESCON term for the hardware and logical functions of a controller or channel subsystem that allows communication over an ESCON write interface and an ESCON read interface.

local area network (LAN). A computer network located on a user's premises within a limited geographical area.

logical address. On an ESCON interface, the portion of a source or destination address in a frame used to select a specific channel-subsystem or control-unit image.

logical data unit. A unit of storage which is accessible on a given device.

logical device. The functions of a logical subsystem with which the host communicates when performing I/O operations to a single addressable-unit over an I/O interface. The same device may be accessible over more than one I/O interface.

logical disk drive. See logical volume.

logical subsystem. The logical functions of a storage controller that allow one or more host I/O interfaces to access a set of devices. The controller aggregates the devices according to the addressing mechanisms of the associated I/O interfaces. One or more logical subsystems exist on a storage controller. In general, the controller associates a given set of devices with only one logical subsystem.

logical unit. The SCSI term for a logical disk drive.

logical unit number. The SCSI term for the field in an identifying message that is used to select a logical unit on a given target.

logical partition (LPAR). The ESA/390 term for a set of functions that create the programming environment that is defined by the ESA/390 architecture. ESA/390 architecture uses this term when more than one LPAR is established on a processor. An LPAR is conceptually similar to a virtual machine environment except that the LPAR is a function of the processor. Also the LPAR does not depend on an operating system to create the virtual machine environment.

logical volume. The storage medium associated with a logical disk drive. A logical volume typically resides on one or more storage devices. A logical volume is referred to on an AIX platform as an hdisk, an AIX term for storage space. A host system sees a logical volume as a physical volume.

LSS. See logical subsystem.

LUN. See logical unit number.

least-recently used (LRU). A policy for a caching algorithm which chooses to remove the item from cache which has the longest elapsed time since its last access.

M

MB. Megabyte.

megabyte (MB). 1 048 576 bytes.

metadata. Internal control information used by microcode. It is stored in reserved area within disk array. The usable capacity of the array take care of the metadata.

million instructions per second (MIPS). A general measure of computing performance and, by implication, the amount of work a larger computer can do. The term is used by IBM and other computer manufacturers. For large servers or mainframes, it is also a way to measure the cost of computing: the more MIPS delivered for the money, the better the value.

MTBF. Mean time between failures. A projection of the time that an individual unit remains functional. The time is based on averaging the performance, or projected performance, of a population of statistically independent units. The units operate under a set of conditions or assumptions.

Multiple Virtual Storage (MVS). One of a family of IBM operating systems for the System/370 or System/390 processor, such as MVS/ESA.

MVS. Multiple Virtual Storage.

N

nondisruptive. The attribute of an action or activity that does not result in the loss of any existing capability or resource, from the customer's perspective.

nonvolatile storage (NVS). Random access electronic storage with a backup battery power source, used to retain data during a power failure. Nonvolatile storage, accessible from all cached IBM storage clusters, stores data during DASD fast write, dual copy, and remote copy operations.

NVS. Nonvolatile storage.

O

open system. A system whose characteristics comply with standards made available throughout the industry, and therefore can be connected to other systems that comply with the same standards.

operating system. Software that controls the execution of programs. An operating system may provide services such as resource allocation, scheduling, input/output control, and data management.

orphan data. Data that occurs between the last, safe backup for a recovery system and the time when the application system experiences a disaster. This data is lost when either the application system becomes available for use or when the recovery system is used in place of the application system.

P

path group. The ESA/390 term for a set of channel paths that are defined to a controller as being associated with a single LPAR. The channel paths are in a group state and are on-line to the host.

path-group identifier. The ESA/390 term for the identifier that uniquely identifies a given LPAR. The path-group identifier is used in communication between the LPAR program and a device to associate the path-group identifier with one or more channel paths. This identifier defines these paths to the control unit as being associated with the same LPAR.

partitioned data set extended (PDSE). A system-managed, page-formatted data set on direct access storage.

P/DAS. PPRC dynamic address switching.

PDSE. Partitioned data set extended.

peer-to-peer remote copy (PPRC). A hardware based remote copy option that provides a synchronous volume copy across storage subsystems for disaster recovery, device migration, and workload migration.

pending. The initial state of a defined volume pair, before it becomes a duplex pair. During this state, the contents of the primary volume are copied to the secondary volume.

pinned data. Data that is held in a cached storage control, because of a permanent error condition, until it can be destaged to DASD or until it is explicitly discarded by a host command. Pinned data exists only when using fast write, dual copy, or remote copy functions.

port. (1) An access point for data entry or exit. (2) A receptacle on a device to which a cable for another device is attached.

PPRC. Peer-to-peer remote copy.

PPRC dynamic address switching (P/DAS). A software function that provides the ability to dynamically redirect all application I/O from one PPRC volume to another PPRC volume.

predictable write. A write operation that can cache without knowledge of the existing formatting on the medium. All writes on FBA DASD devices are predictable. On CKD DASD devices, a write is predictable if it does a format write for the first record on the track.

primary device. One device of a dual copy or remote copy volume pair. All channel commands to the copy logical volume are directed to the primary device. The data on the primary device is duplicated on the secondary device. See also secondary device.

PTF. Program temporary fix.

R

RACF. Resource access control facility.

rack. A unit that houses the components of a storage subsystem, such as controllers, disk drives, and power.

random access. A mode of accessing data on a medium in a manner that requires the storage device to access nonconsecutive storage locations on the medium.

read hit. When data requested by the read operation is in the cache.

read miss. When data requested by the read operation is not in the cache.

recovery. The process of rebuilding data after it has been damaged or destroyed. In the case of remote copy, this involves applying data from secondary volume copies.

recovery system. A system that is used in place of a primary application system that is no longer available for use. Data from the application system must be available for use on the recovery system. This is usually accomplished through backup and recovery techniques, or through various DASD copying techniques, such as remote copy.

remote copy. A storage-based disaster recovery and workload migration function that can copy data in real time to a remote location. Two options of remote copy are available. See peer-to-peer remote copy and extended remote copy.

reserved allegiance. ESA/390 term for a relationship that is created in a controller between a device and a channel path, when a Sense Reserve command is completed by the device. The allegiance causes the control unit to guarantee access (busy status is not presented) to the device. Access is over the set of channel paths that are associated with the allegiance; access is for one or more channel programs, until the allegiance ends.

restore. Synonym for recover.

resynchronization. A track image copy from the primary volume to the secondary volume of only the tracks which have changed since the volume was last in duplex mode.

RVA. RAMAC Virtual Array Storage Subsystem.

S

SAID. System adapter identification.

SAM. Sequential access method.

SCSI. Small Computer System Interface. An ANSI standard for a logical interface to computer peripherals and for a computer peripheral interface. The interface utilizes a SCSI logical protocol over an I/O interface that configures attached targets and initiators in a multi-drop bus topology.

SCSI ID. A unique identifier assigned to a SCSI device that is used in protocols on the SCSI interface to identify or select the device. The number of data bits on the SCSI bus determines the number of available SCSI IDs. A wide interface has 16 bits, with 16 possible IDs. A SCSI device is either an initiator or a target.

Seascape architecture. A storage system architecture developed by IBM for open system servers and S/390 host systems. It provides storage solutions that integrate software, storage management, and technology for disk, tape, and optical storage.

secondary device. One of the devices in a dual copy or remote copy logical volume pair that contains a duplicate of the data on the primary device. Unlike the primary device, the secondary device may only accept a limited subset of channel commands.

sequential access. A mode of accessing data on a medium in a manner that requires the storage device to access consecutive storage locations on the medium.

server. A type of host that provides certain services to other hosts that are referred to as clients.

service information message (SIM). A message, generated by a storage subsystem, that is the result of error event collection and analysis. A SIM indicates that some service action is required.

sidefile. A storage area used to maintain copies of tracks within a concurrent copy domain. A concurrent copy operation maintains a sidefile in storage control cache and another in processor storage.

SIM. Service information message.

simplex state. A volume is in the simplex state if it is not part of a dual copy or a remote copy volume pair. Ending a volume pair returns the two devices to the simplex state. In this case, there is no longer any capability for either automatic updates of the secondary device or for logging changes, as would be the case in a suspended state.

SMF. System Management Facilities.

SMS. Storage Management Subsystem.

SRM. System resources manager.

SnapShot copy. A point-in-time copy services function that can quickly copy data from a source location to a target location.

spare. A disk drive that is used to receive data from a device that has experienced a failure that requires disruptive service. A spare can be pre-designated to allow automatic dynamic sparing. Any data on a disk drive that you use as a spare is destroyed by the dynamic sparing copy process.

SSA. Serial Storage Architecture. An IBM standard for a computer peripheral interface. The interface uses a SCSI logical protocol over a serial interface that configures attached targets and initiators in a ring topology.

SSID. Subsystem identifier.

stacked status. An ESA/390 term used when the control unit is holding for the channel; the channel responded with the stack-status control the last time the control unit attempted to present the status.

stage. The process of reading data into cache from a disk drive module.

storage cluster. A power and service region that runs channel commands and controls the storage devices. Each storage cluster contains both channel and device interfaces. Storage clusters also perform the DASD control functions.

storage control. The component in a storage subsystem that handles interaction between processor channel and storage devices, runs channel commands, and controls storage devices.

STORAGE_CONTROL_DEFAULT. A specification used by several XRC commands and messages to refer to the timeout value specified in the maintenance panel of the associated storage control.

storage device. A physical unit which provides a mechanism to store data on a given medium such that it can be subsequently retrieved. Also see disk drive module.

storage director. In an IBM storage control, a logical entity consisting of one or more physical storage paths in the same storage cluster. See also storage path.

storage facility. (1) A physical unit which consists of a storage controller integrated with one or more storage devices to provide storage capability to a host computer. (2) A storage server and its attached storage devices.

Storage Management Subsystem (SMS). A component of MVS/DFP that is used to automate and centralize the management of storage by providing the storage administrator with control over data class, storage class, management class, storage group, aggregate group and automatic class selection routine definitions.

storage server. A unit that manages attached storage devices and provides access to the storage or storage related functions for one or more attached hosts.

storage path. The hardware within the IBM storage control that transfers data between the DASD and a channel. See also storage director.

storage subsystem. A storage control and its attached storage devices.

string. A series of connected DASD units sharing the same A-unit (or head of string).

striping. A technique that distributes data in bit, byte, multibyte, record, or block increments across multiple disk drives.

subchannel. A logical function of a channel subsystem associated with the management of a single device.

subsystem. See DASD subsystem or storage subsystem.

subsystem identifier (SSID). A user-assigned number that identifies a DASD subsystem. This number is set by the service representative at the time of installation and is included in the vital product data.

suspended state. When only one of the devices in a dual copy or remote copy volume pair is being updated because of either a permanent error condition or an authorized user command. All writes to the remaining functional device are logged. This allows for automatic resynchronization of both volumes when the volume pair is reset to the active duplex state.

synchronization. An initial volume copy. This is a track image copy of each primary track on the volume to the secondary volume.

synchronous operation. A type of operation in which the remote copy PPRC function copies updates to the secondary volume of a PPRC pair at the same time that the primary volume is updated. Contrast with asynchronous operation.

system data mover. A system that interacts with storage controls that have attached XRC primary volumes. The system data mover copies updates made to the XRC primary volumes to a set of XRC-managed secondary volumes.

system-managed data set. A data set that has been assigned a storage class.

T

TCP/IP. Transmission Control Protocol/Internet Protocol.

TOD. Time of day.

Time Sharing Option (TSO). A System/370 operating system option that provides interactive time sharing from remote terminals.

timeout. The time in seconds that the storage control remains in a “long busy” condition before physical sessions are ended.

timestamp. The affixed value of the system time-of-day clock at a common point of reference for all write I/O operations directed to active XRC primary volumes. The UTC format is yyyy.ddd hh:mm:ss.thmiju.

track. A unit of storage on a CKD device that can be formatted to contain a number of data records. Also see home address, track-descriptor record, and data record.

track-descriptor record. A special record on a track that follows the home address. The control program uses it to maintain certain information about the track. The record has a count field with a key length of zero, a data length of 8, and a record number of 0. This record is sometimes referred to as R0.

TSO. Time Sharing Option.

U

Ultra-SCSI. An enhanced small computer system interface.

unit address. The ESA/390 term for the address associated with a device on a given controller. On ESCON interfaces, the unit address is the same as the device address. On OEMI interfaces, the unit address specifies a controller and device pair on the interface.

Universal Time, Coordinated. Replaces Greenwich Mean Time (GMT) as a global time reference. The format is yyyy.ddd hh:mm:ss.thmiju.

utility volume. A volume that is available to be used by the extended remote copy function to perform data mover I/O for a primary site storage control's XRC-related data.

UTC. Universal Time, Coordinated.

V

vital product data (VPD). Nonvolatile data that is stored in various locations in the DASD

subsystem. It includes configuration data, machine serial number, and machine features.

volume. An ESA/390 term for the information recorded on a single unit of recording medium. Indirectly, it can refer to the unit of recording medium itself. On a non-removable medium storage device, the terms may also refer, indirectly, to the storage device that you associate with the volume. When you store multiple volumes on a single storage medium transparently to the program, you may refer to the volumes as logical volumes.

vital product data (VPD). Information that uniquely defines the system, hardware, software, and microcode elements of a processing system.

VSAM. Virtual storage access method.

VTOC. Volume table of contents.

W

workload migration. The process of moving an application's data from one set of DASD to another for the purpose of balancing performance needs, moving to new hardware, or temporarily relocating data.

write hit. A write operation where the data requested is in the cache.

write miss. A write operation where the data requested is not in the cache.

write penalty. The term that describes the classical RAID write operation performance impact.

write update. A write operation that updates a direct access volume.

X

XDF. Extended distance feature (of ESCON).

XRC. Extended remote copy.

XRC planned-outage-capable. A storage subsystem with an LIC level that supports a software bitmap but not a hardware bitmap.

Index

A

ADVFS 69
algorithms 38
alias 29
AlphaServer 41, 42, 65
apache 42, 51, 52
array 7

B

Brocade 28
buffer 38

C

Cache 2
CFS 65
Cluster 43
 Interconnect 2
 Processor Complex 2
cluster 51
Cluster 1 1
Cluster 2 1
Compaq xi, 41, 42, 65, 119, 146
 configuration 42, 43, 66, 67
 Disk configuration 44
 host optical adapter 42
 Host optical fibre adapter 66
 systems tested 42
Configure 12
configuring the ESS 3
connectivity 25

D

DA 2
Device Adapter 2
directory 68
Disk 34
 configuration 67
disk
 group creation 58
 initialization 58
 label 46
 service creation 47
 shared 51
DNS 10

E

environment 25
error logs 4
ESCON 6, 12
ESS xi, 1, 2, 4, 10, 18, 20, 24, 119, 146
 configuration 66
expansion cabinet 2

F

FC port 14
FC-AL 1, 25
FC-F 1
FC-SW 25
Fibre Channel 1, 10, 13, 14, 44
firmware 43

G

GUI 30

H

HA 2, 3, 18, 30, 31
HBA 31, 45, 67
host
 cluster 24
 modify 9
 ports 4
Host Adapter 2
HP xi, 119, 146

I

I/O 20

J

Java 28
JBOD 3, 4
Just a Bunch Of Disk 3

L

latency 33
LIC 4
logical size 2
Logical volume 35
logical volumes 4
LSM 58, 60

LUN 13, 20, 23, 45, 71

M

modify assignment 20
multi-path 33
multi-pathing 30, 32
multiple path 31

N

NVS 2

O

Open Systems 7
Operating system
 version 43

P

patch 44
Perform Configuration Update 20

R

RAID 2, 3, 4
RAID level 5 3
rank 2, 34
restriction 62
RS6000 1

S

SAN 25
script 51, 53, 57
SCSI 1, 3, 4, 14, 16, 25
security 5
specialist 4
SPOF 30
SSA 2
storage allocation 5
SUN xi, 119, 146
Switch 66
switch 18, 28, 31, 42

T

TCP/IP 4
throttle 29
Tru64 xi, 42, 119, 146
Tru64 UNIX 65
TruCluster 43

U

utility 60

V

Veritas xi, 146
virtual 4
virtual disk 35
volassist 60
volume
 add 17
 layout 38
 management 33
 manager 39

W

Windows NT 4
WWNN 26
WWPN 10, 13, 26, 29

Z

Zone 66
zone 26
Zoning 26, 27

IBM Redbooks review

Your feedback is valued by the Redbook authors. In particular we are interested in situations where a Redbook "made the difference" in a task or problem you encountered. Using one of the following methods, **please review the Redbook, addressing value, subject matter, structure, depth and quality as appropriate.**

- Use the online **Contact us** review redbook form found at ibm.com/redbooks
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Document Number	SG24-6119-00
Redbook Title	ESS Solutions for Open Systems Storage: Compaq AlphaServer, HP, and SUN
Review	
What other subjects would you like to see IBM Redbooks address?	
Please rate your overall satisfaction:	<input type="radio"/> Very Good <input type="radio"/> Good <input type="radio"/> Average <input type="radio"/> Poor
Please identify yourself as belonging to one of the following groups:	<input type="radio"/> Customer <input type="radio"/> Business Partner <input type="radio"/> Solution Developer <input type="radio"/> IBM, Lotus or Tivoli Employee <input type="radio"/> None of the above
Your email address: The data you provide here may be used to provide you with information from IBM or our business partners about our products, services or activities.	<input type="checkbox"/> Please do not use the information collected here for future marketing or promotional contacts or other communications beyond the scope of this transaction.
Questions about IBM's privacy policy?	The following link explains how we protect your personal information. ibm.com/privacy/yourprivacy/



ESS Solutions for Open Systems Storage: Compaq AlphaServer, HP, and SUN

(0.2"spine)
0.17" <-> 0.473"
90 <-> 249 pages



ESS Solutions for Open Systems Storage: Compaq AlphaServer, HP, and SUN



**Open Systems
connectivity to the
IBM ESS described in
detail**

**Hints and tips on
using Veritas with
the IBM ESS**

**Implementation of
high availability
clusters**

This IBM Redbook is designed to help you install, tailor, and configure the IBM Enterprise Storage Server (ESS) when attaching Compaq AlphaServer running Tru64 UNIX, HP and Sun hosts. We describe the results of a 5-week project that took place in San Jose in October and November 2000. This book does not cover Compaq AlphaServer running Open VMS. Rather, the project was focused on settings required to give optimal performance, device driver levels. As such, the book is intended for the experienced UNIX professional who has a broad understanding of storage concepts, but is not necessarily an expert in each of the areas discussed.

First, Chapter 1 provides a broad overview of the ESS. Then, Chapter 2 explains general connectivity issues such as switch zoning, multi-pathing, and volume management. Chapters 3 and 4 cover the attachment of Compaq AlphaServers running Tru64 UNIX (V4 and V5 respectively), both for standalone and clustered configurations. Chapter 5 discusses HP servers. Chapter 6 discusses Sun systems, and also covers various aspects of using some of the Veritas suite of products.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:
ibm.com/redbooks**

SG24-6119-00

ISBN 0738418234