IBM Enterprise Storage Server

Understand the ESS characteristics and design

Discover its powerful performance and copy features

Learn the basics before planning and configuring

Gustavo A. Castets
Daniel Leplaideur
Jose Alcino Bras
Jason Galang

Redbooks

ibm.com/redbooks

International Technical Support Organization

SG24-5465-01

IBM Enterprise Storage Server

September 2001

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix G, “Special notices” on page 315.

Second Edition (September 2001)

This edition applies to the IBM 2105 Enterprise Storage Server model F. For other information you can see the Publications section of the IBM Announcement letter for the IBM Enterprise Storage Server and you can visit our Web site at www.storage.ibm.com.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. 471 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.

Contents

Figures

Tables

Preface

This IBM Redbook describes the Enterprise Storage Server (ESS), its design characteristics and advanced functions. This book will be useful for those who need to know the general characteristics of this powerful disk storage server, and also for those who need a more detailed understanding of how the ESS is designed and operates.

This redbook describes, in detail, the architecture, hardware, and advanced functions of the Enterprise Storage Server. It covers Fibre Channel and FICON attachment. The information will help the IT Specialist in the field to understand the ESS, its components, its overall design, and how all of this works together. This understanding will be useful for proper planning and the logical configuration at the time of installation. The IT Specialist will find useful information when availability and disaster recovery solutions are to be implemented, with the explanation of the advanced copy functions that ESS has available: FlashCopy, Peer-to-Peer Remote Copy, Extended Remote Copy, and Concurrent Copy. Also the performance features of the ESS are explained, making all of these topics very helpful for the IT Specialist for the optimization of resources in the computing center.

This book has been updated to support the announcement and general availability of the new Fibre Channel/FICON host adapters. Also, it has been updated to the ESS models F10 and F20, and to generally all that has been available since the initial announcement of Enterprise Storage Server.

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.

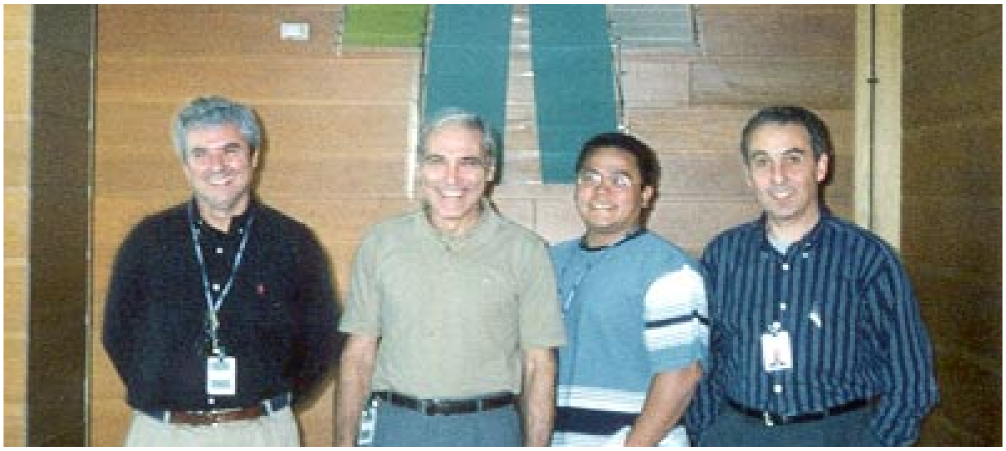


Figure 1. Authors: Daniel Leplaideur, Alcino Bras, Jason Galang, Gustavo Castets

Gustavo A. Castets is a Project Leader for Disk Storage Systems at the International Technical Support Organization, San Jose Center. Gustavo has worked for IBM for 22 years in many IT areas. Before joining the ITSO, Gustavo worked in Buenos Aires, Argentina, as a Storage Specialist.

Daniel Leplaideur joined IBM France in 1967 as Mathematician and Statistician to develop scientific subsystems and software packages. He has been a Systems Engineer in networking and large systems in the services, branch offices, and field support centers. He is now with the European Mainz Advanced Technical Support for ESS and Disaster Recovery. His job mainly deals with defining and implementing solutions for various complex environments.

Jose Alcino Bras is a Certified IT Specialist from IBM Brazil. He has 30 years of experience in IT technical business, of which the last six years were spent on storage solutions. He started as a Customer Engineer and then was a Systems Engineer working on large accounts. He has practical experience in all IBM disk and tape storage products. He is now working for the Advanced Technical Support (ATS) organization.

Jason Galang joined IBM Philippines in 1989 as an IT Specialist for both AS/400 and RS/6000. He has been doing country pre-sales technical support for AS/400 since 1995, and in 1999 he included the technical support for the RS/6000. Working in the RS/6000 environment he gained experience in storage solutions, mainly the ESS and its advanced functions.

The authors of previous editions of this book are:

Alison Pate, IBM UK
Cay-Uwe Kulzer, IBM Germany
Phil Norman, IBM UK
Roland Wolf, IBM Germany

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

Eneo Baborsky, IBM Italy
Marg Beal, IBM Poughkeepsie
Michael Benhase, IBM Tucson
Helen Burton, IBM Tucson
Henry Caudillo, IBM San Jose
Jerry Coale, IBM San Jose
Paul Coles, IBM UK
Arturo Dana Anello, IBM Argentina
Omar Escola, IBM Argentina
Amine Hajji, IBM San Jose
Nick Harris, ITSO Rochester
Rowell Hernandez, IBM ITSO Almaden
Lee La Frese, IBM Tucson
Phillip Mills, IBM San Jose
Phil Norman, IBM UK
Alison Pate, IBM UK
John Ponder, IBM Tucson
Dave Reeve, IBM UK
Bill Worthington, IBM San Jose

Comments welcome

Your comments are important to us!

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

- Use the online **Contact us** review redbook form found at:

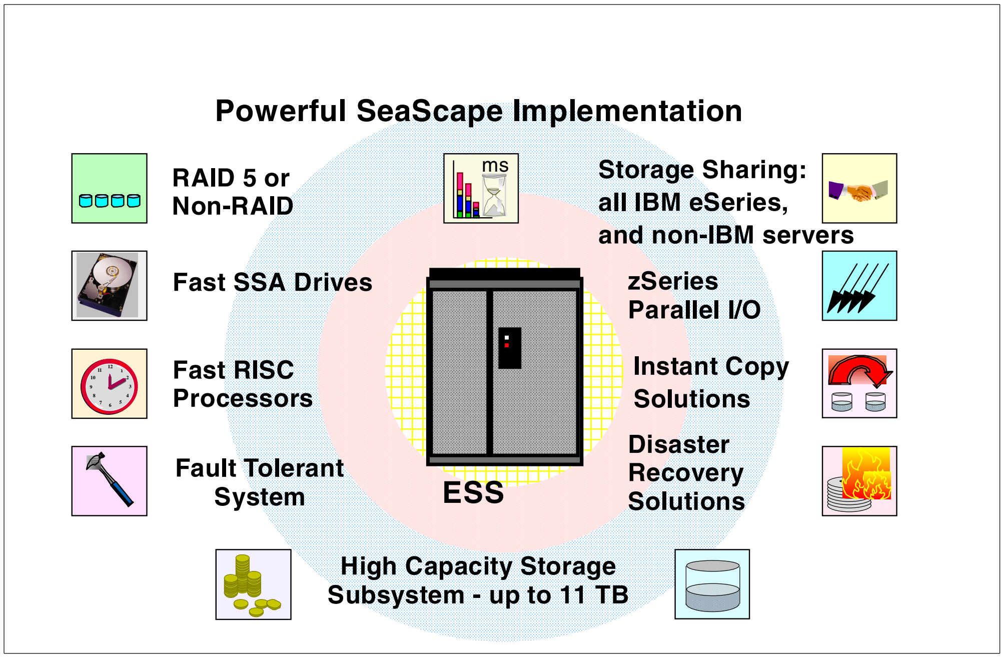
  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to the address on page ii.

Chapter 1. Introduction



Powerful SeaScape Implementation

RAID 5 or Non-RAID

Fast SSA Drives

Fast RISC Processors

Fault Tolerant System

Storage Sharing: all IBM eSeries, and non-IBM servers

zSeries Parallel I/O

Instant Copy Solutions

Disaster Recovery Solutions

ESS

High Capacity Storage Subsystem - up to 11 TB

Figure 2. IBM's Seascape Enterprise Storage Server

The Enterprise Storage Server (ESS) is IBM's most powerful disk storage server developed using IBM's Seascape architecture. The ESS provides un-matchable functions for all the e-business servers of the new IBM @server brand, and also for the non-IBM open servers. Across all of these environments the ESS features unique capabilities that allow it to meet the most demanding requirements of performance, capacity and availability that the computing business may require.

The Seascape architecture is the key to the development of IBM's storage products. Seascape allows IBM to take the best of the technologies developed by the many IBM laboratories and integrate them producing flexible and upgrade-able storage solutions. The Magstar Tape drives, the Magstar Virtual Tape Server, the award-winning Ultrastar disk drives, the Serial Interconnect technology, the 160 MB/sec. SSA adapters, Fibre Channel connectivity and the most recent FICON connectivity for the IBM @server zSeries servers, all of these are examples of the tremendous versatility of the Seascape architecture.

e-business continues to drive storage at an ever-increasing rate. e-commerce applications are deployed on heterogeneous servers and require high-function, high-performance and scalable storage servers to meet the demanding requirements of Enterprise Resource Planning and Business Intelligence applications. The IBM Enterprise Storage Server (ESS) has set new standards in function, performance, and scalability in these most challenging environments and is now enhanced to provide:

- Direct connection to Storage Area Networks (SANs)
- Advanced storage management functions which provide fast data duplication and high-performance backup and disaster recovery capability.

The Enterprise Storage Servers Models F10 and F20

Since its initial availability with the Models E10 and E20, the ESS is the storage server solution that offers exceptional performance, extraordinary capacity scalability, heterogeneous server connectivity, and an extensive suite of advanced functions to support customers mission-critical, high-availability, multi-platform environments. The ESS set a new standard for storage servers back in 1999 when it was first available, and since then it kept up with the pace of customers needs, adding more sophisticated functions to the initial set, increasing the connectivity options, and powering its performance features.

The Enterprise Storage Server Models F10 and F20 provide significantly improved levels of performance, throughput, and scalability while continuing to exploit performance innovations introduced with the Models E10 and E20, such as Parallel Access Volumes, Multiple Allegiance, and I/O Priority Queueing. These models are also supporting the introduction of many new enhancements that are dramatically improving the overall value of ESS in the marketplace and provide a strong base for strategic Storage Area Network initiatives. Among those enhancements are the new FICON and the new Fibre Channel long and short-wave attachment capabilities of the ESS. Additional enhancements include new levels of granularity for installed systems, remote copy services capabilities with synchronous and asynchronous options, and more cache options. Additional PCI buses, faster state-of-the-art IBM RISC microprocessors, plus up to 32 GB of cache memory all combine to provide unprecedented levels of performance and throughput for the F models. The increased bandwidth capabilities of the Models F10 and F20 allow for up to 100% greater throughput for sequential workloads, or 25% greater throughput for database workloads. This allows customers to install 25-100% larger configurations in the Model F20 than were previously possible with the Model E20, depending on access density and the nature of the workload, while maintaining industry leading performance. With these capabilities, the ESS is perfectly suited to enable customers to take advantage of the promise of SAN, while protecting the significant investments already made in storage and I/T infrastructures.

The Seascape design of the ESS has allowed customers to transition from the initial E Models to the more recent F Models and its associated features. Customers investment has been protected, and the upgrades to keep technologically current have been non-traumatic, because of the Seascape design of the ESS. And so it will go on, for future innovations.

1.1 Benefits

The ESS can help you achieve your business objectives in many areas; it provides a high-performance, high-availability subsystem with flexible characteristics that can be configured according to your requirements.

Storage consolidation

The ESS attachment versatility, and large capacity, enable you to consolidate your data from different platforms onto a single high performance, high availability box. Storage consolidation can be the first step towards server consolidation, reducing the number of boxes you have to manage and allowing you the flexibility to add or assign capacity when and where it is needed. ESS supports all the major server platforms, from the IBM @server series of e-business servers and IBM NUMA-Q, to the non-IBM Windows NT and different variations of UNIX servers, as shown in Figure 3. With a capacity of up to 11TB, and up to 32 host connections, an ESS can meet both your high capacity requirements and your performance expectations.

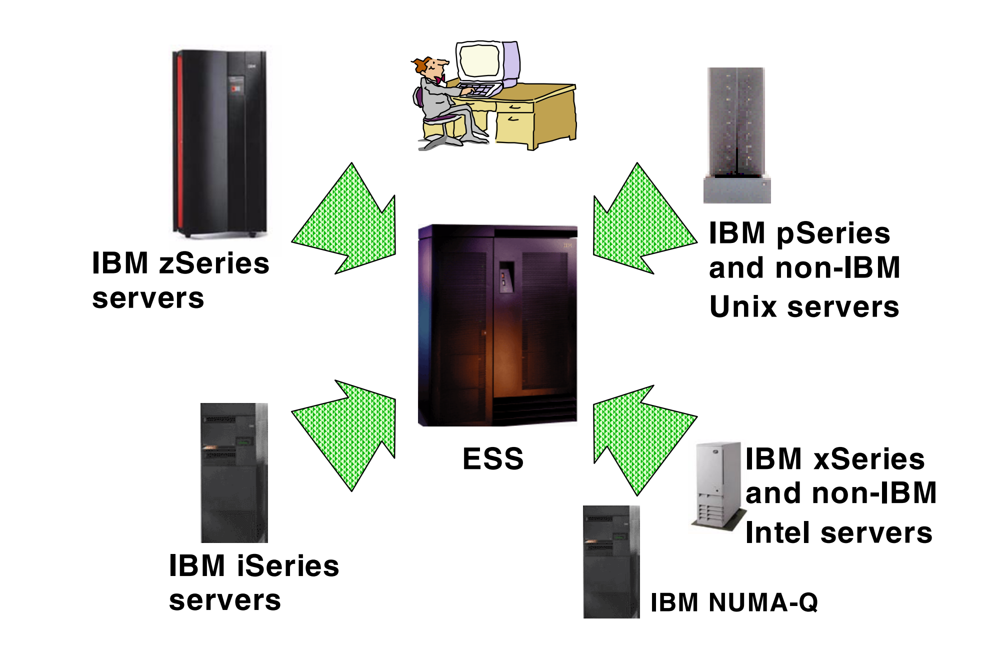

IBM zSeries servers
IBM pSeries and non-IBM Unix servers
ESS
IBM iSeries servers
IBM xSeries and non-IBM Intel servers
IBM NUMA-Q

Figure 3. ESS for storage consolidation

Performance

The ESS is a high performance designed storage solution and takes advantage of IBM's leading technologies. In today's world, you need business solutions that can deliver high levels of performance continuously every day, day after day. You also need a solution that can handle different workloads simultaneously, so you can run your business intelligence models, your large databases for enterprise resource planning (ERP), and your online and internet transactions alongside each other. Figure 4, on page 4 shows some of the performance enhancing capabilities of the ESS.

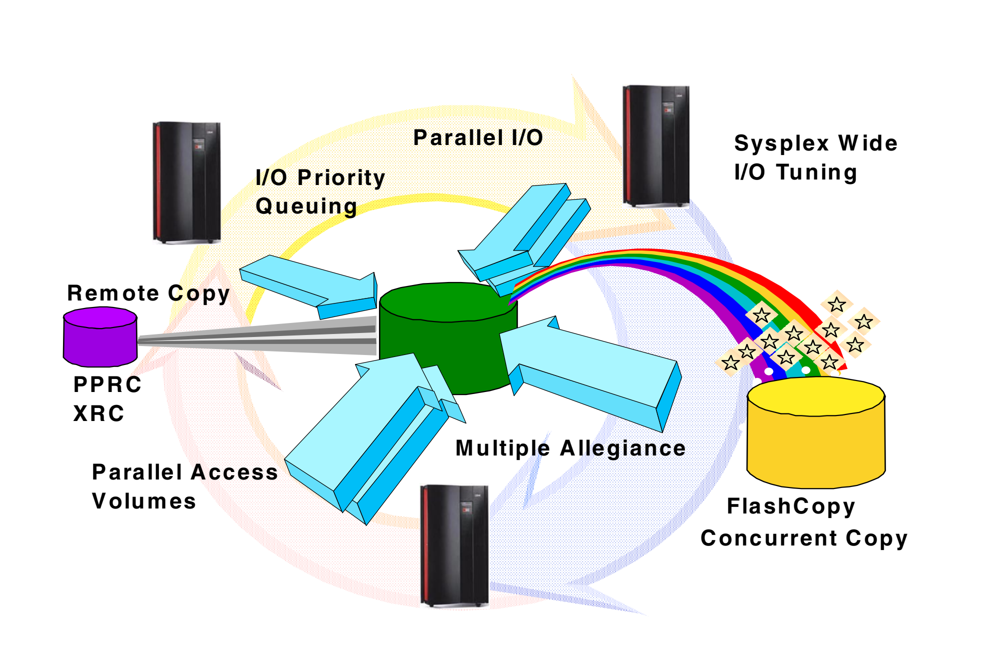


Figure 4. Enterprise Storage Server capabilities

The ESS is designed to provide the highest performance, for the different type of workloads, even when mixing those dissimilar workload demands. For example, zSeries 900 servers and open systems, put very different workload demands on the storage system. A server like the zSeries 900 typically has an I/O profile that is very cache-friendly, and takes advantage of the cache efficiency. On the other side, an open system server does an I/O that is often very cache-unfriendly, because most of the hits are already solved in the buffers defined in the host server. So when open system servers do I/Os, most of the times they will be solved in the hard disk drive (HDD) and not in the cache. For the zSeries 900 type of workload, the ESS has a large cache and, most important, it has very sophisticated cache algorithms. For the workloads that do not take so much benefit on the cache efficiency, the ESS has in its back-end the SSA high performance disk adapters.

Parallel Sysplex I/O management

In the zSeries 900 Parallel Sysplex environments, the z/OS Workload Manager (WLM) controls where work is run and optimizes the throughput and performance of the total system. The ESS provides the WLM with more sophisticated ways to control the I/O across the Sysplex. These functions, described in detail later in this book, include parallel access to both single system and shared volumes and the ability to prioritize the I/O based upon WLM goals. The combination of these features significantly improve performance in a wide variety of workload environments.

Disaster recovery and availability

The Enterprise Storage Server is designed with no single point of failure. It is a fault tolerant storage subsystem, which can be maintained and upgraded concurrently with customer operation. Some of the functions that contribute to these attributes of the ESS are shown in Figure 5.

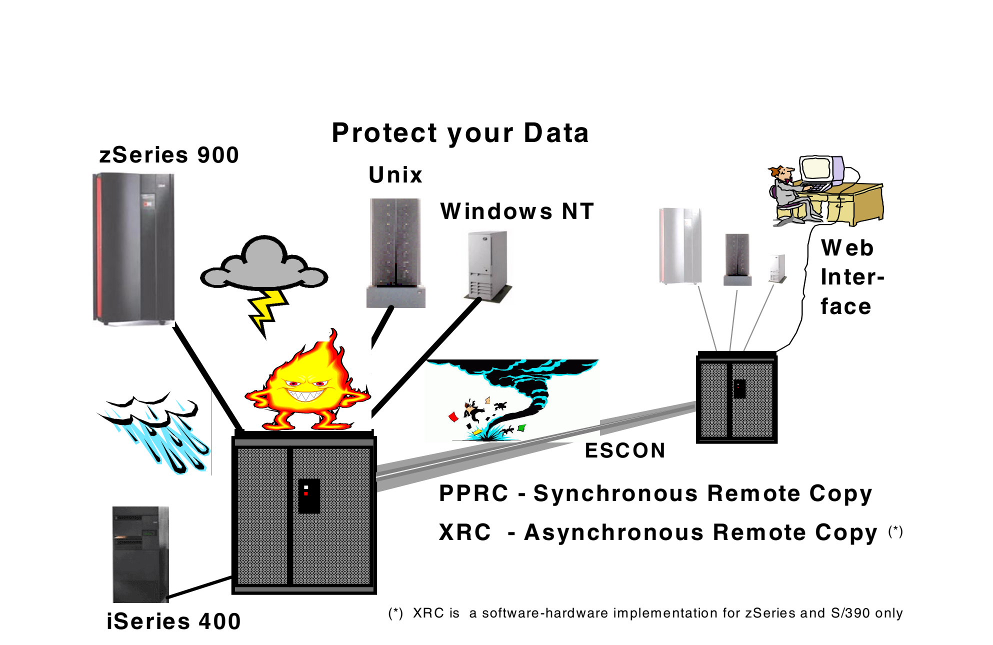

Protect your Data
zSeries 900
Unix
Windows NT
Web Inter-face
ESCON
PPRC - Synchronous Remote Copy
XRC - Asynchronous Remote Copy (*)
iSeries 400
(*) XRC is a software-hardware implementation for zSeries and S/390 only

Figure 5. Disaster recovery and availability

The Peer-to-Peer Remote Copy (PPRC) function is a hardware solution that enables the shadow-mirroring of application system data from one site, to a second site system. Updates made on the primary logical volumes are synchronously shadowed on the secondary site logical volumes. For the open system environments, management of the PPRC setup is done using the IBM TotalStorage ESS Specialist (ESS Specialist, for short) Web interface. The Enterprise Storage Server also provides a command line interface (CLI) for invocation and management of the PPRC functions through batch processes and scripts.The command line interface can be run from AIX, HP-UX, Solaris, and Windows NT servers. This way, solutions for disaster are available for open system platforms using a simple and easy-to-use interface. For the zSeries 900 environments, PPRC can also be managed using TSO commands (for z/OS) or the ICKDSF utility. PPRC together with Geographically Dispersed Parallel Sysplex (GDPS) for z/OS environments, lead the industry in high availability solutions.

The Extended Remote Copy (XRC) is a combined hardware and software disaster recovery solution for the z/OS environments. The asynchronous characteristics of XRC makes it suitable for the very long distance implementations. XRC offers the highest levels of data integrity and availability for disaster recovery, workload movement, and disk migration.

Instant Copy

Customers still need to take backups to protect data from logical errors and disasters. For all environments, taking backups of user data usually takes a long time. Backups are often taken outside prime shift because of the impact to normal operations. Databases must be closed to create consistency and data integrity, and online systems are normally shut down. To minimize the impact of these type of activities, the Enterprise Storage Server has FlashCopy. FlashCopy creates a physical point-in-time copy of data and makes it possible to access both the source and target copies immediately. By creating an “instant” copy, FlashCopy enables applications using either the source or the target to operate with only a minimal interruption to perform the FlashCopy.

On all server platforms, FlashCopy may be controlled using the copy services Web interface that the IBM TotalStorage ESS Specialist provides. Under z/OS, FlashCopy can also be invoked using DFSMSdss, and TSO commands. Under VSE/ESA the support is given by the SNAP command. And z/VM provides guest support for FlashCopy.

For the open systems the Enterprise Storage Server also provides in addition a command line interface (CLI) for invocation and management of FlashCopy functions through batch processes and scripts from AIX, Sun Solaris, HP-UX, and Windows NT and 2000 environments.

Storage Area Network

The IBM ESS is a comprehensive Storage Area Network (SAN) disk storage system that provides solutions to today’s most demanding business intelligence, e-business, server consolidation and ERP customer requirements.

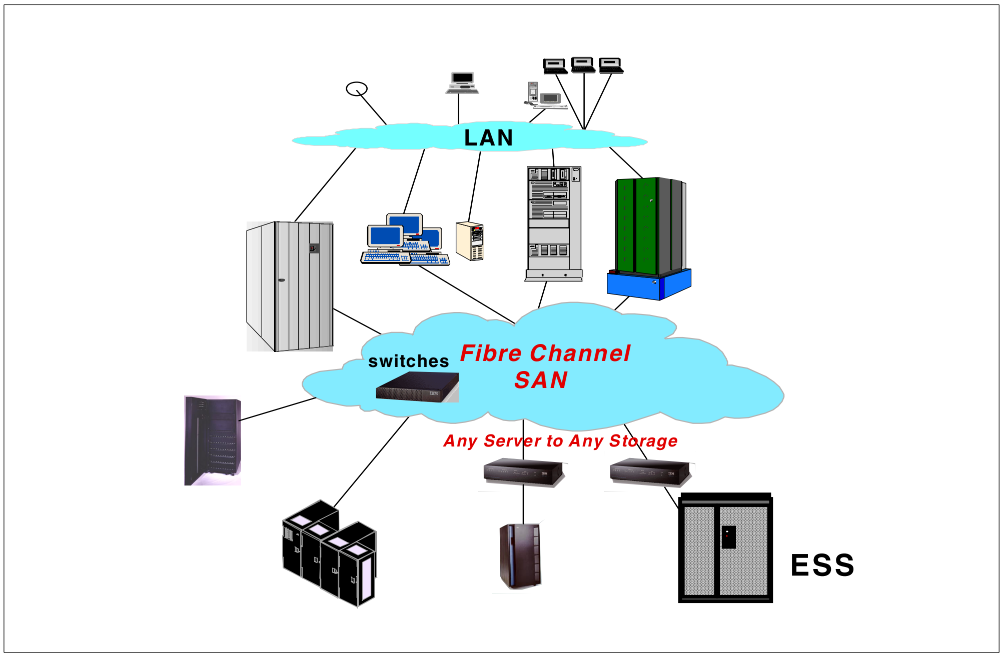


Figure 6. Storage Area Network (SAN)

The Enterprise Storage Server continues to deliver on its SAN strategy, initiated with the previous ESS Models E10 and E20. Now the ESS Models F10 and F20 provide up to sixteen 100 MB/sec. native Fibre Channel short-wave or long-wave host adapters. Each single port adapter supports Fibre Channel Protocol (FCP) in a direct point-to-point configuration, point-to-point to a switch (fabric) configuration, or Fibre Channel-Arbitrated Loop (FC-AL) in a private loop configuration. Fabric support includes the IBM SAN Fibre Channel Switches (2109 Model S08 and S16), McDATA Enterprise Fibre Channel Directors ED-5000 and ED-6064(IBM 2032-001 and 064), Inrange FC/9000 Fibre Channel Directors (IBM 2042-001 and 128), and IBM Fibre Channel Storage Hub (2103-H07). All these SAN connectivity options make the ESS the unquestionable choice when customers plan for their SANs.

Native FICON

FICON is a new interface available in the ESS. It is used for attaching natively to the zSeries 900 servers and the S/390 G5 and G6 servers, and is based on the industry standard Fibre Channel architecture. FICON has many improvements over ESCON and also relieves many ESCON limitations. With the native FICON interface, the ESS Model F becomes an invaluable disk storage solution for the zSeries 900 and S/390 customers.

1.2 Terminology

Before starting to look at the hardware, architecture, and configuration characteristics of the Enterprise Storage Server, we will see the terminology that we use in this book.

1.2.1 Host attachment

Figure 7, and Figure 8 show the elements involved in the attachment of the storage server to the host. This is what we generally refer as host attachment.

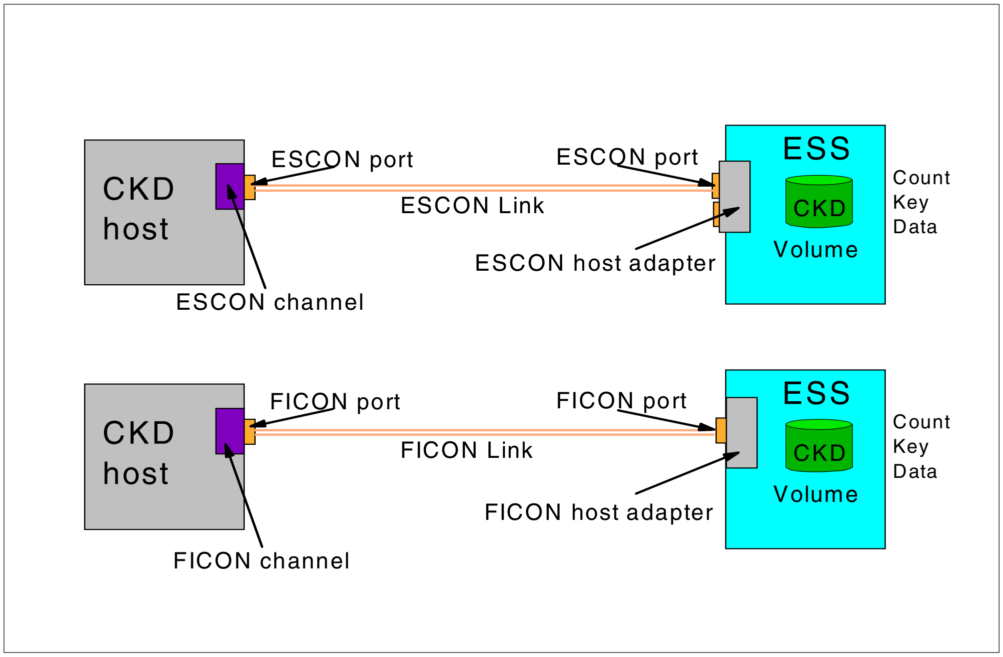


Figure 7. ESCON and FICON attachment terminology

ESCON channel
The ESCON channel is a hardware feature on the zSeries and S/390 servers that controls data flow over the ESCON link. An ESCON channel is usually installed on an ESCON channel card which may contain up to four ESCON channels

ESCON host adapter
The host adapter (HA) is the physical component of the storage server used to attach one or more host I/O interfaces. The ESS can be configured with ESCON host adapters. The ESCON host adapter is connected to the ESCON channel and accepts CCWs (channel command words) that are sent by the host system.The ESCON host adapter has two ports for ESCON link connection.

ESCON port
The ESCON port is the physical interface into the ESCON channel. An ESCON port has an ESCON connector interface. You have an ESCON port wherever you plug in an ESCON link.

ESCON link
An ESCON link is the fiber connection between the zSeries server and the storage server. An ESCON link can also exist between a zSeries processor and an ESCON Director (fiber switch), and between an ESCON Director and the ESS (or other ESCON capable devices).

FICON channel
The FICON channel is a hardware feature on the zSeries 900 and IBM 9672 G5 and G6 servers, that controls data flow over the FICON link. A FICON channel is usually installed on a FICON channel card which contains up to two FICON channels.

FICON host adapter
The host adapter (HA) is the physical component of the ESS used to attach one or more host I/O interfaces. The ESS can be configured with FICON host adapters. The FICON host adapter is connected to the FICON channel and accepts the host CCWs (channel command words) that are sent by the host system.The FICON host adapter has one port for FICON link connection.

FICON port
The FICON port is the physical interface into the FICON channel. A FICON port has a FICON connector interface. You have a FICON port wherever you plug in a FICON link.

FICON link
A FICON link is the fiber connection between the zSeries server and the storage server. A FICON link can also exist between a zSeries processor and a FICON Director (switch), and between a FICON switch and the ESS (or other FICON capable devices)

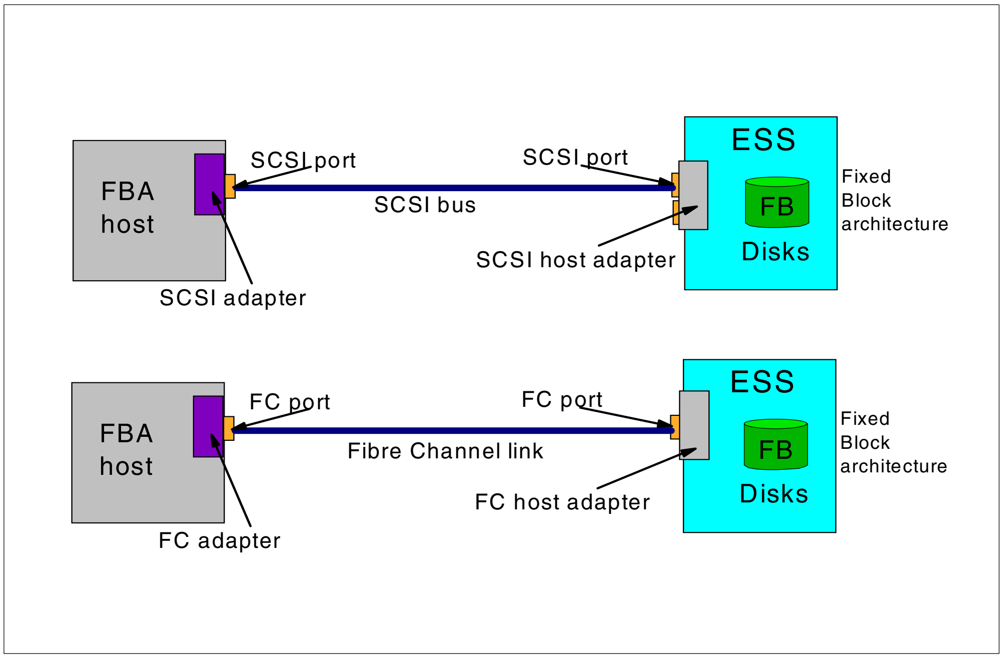


Figure 8. SCSI and Fibre Channel attachment terminology

SCSI adapter

A SCSI adapter is a card installed in a host system. It connects to the SCSI bus through a SCSI connector. There are different versions of SCSI, some of which can be supported by the same adapter. The protocols that are used on the SCSI adapter (the command set) can be either SCSI-2 or SCSI-3.

SCSI port

A SCSI port is the physical interface into which you connect a SCSI cable. The physical interface varies, depending on what level of SCSI is supported.

SCSI bus

The SCSI bus is the path linking all the devices that are chained together on the same SCSI adapter. Each device on the bus is connected to the next one by a SCSI cable, and at the last device on the bus, there is a terminator.

SCSI host adapter

The host adapter (HA) is the physical component of the storage server used to attach one or more host I/O interfaces.The ESS can be configured with SCSI host adapters. The SCSI host adapter is connected to the SCSI bus and accepts the SCSI commands that are sent by the host system.The SCSI host adapter has two ports for SCSI bus connection.

SCSI

SCSI (Small Computer Systems Interface) is the protocol that the SCSI adapter cards use. Although SCSI protocols can be used on Fibre Channel (then called FCP) most people mean the parallel interface when they say SCSI.

Fibre Channel adapter

A Fibre Channel adapter (FC adapter) is a card installed in a host system. It connects to the Fibre Channel (fibre) through a connector. The Fibre Channel adapter allows data to be transferred over fibre links at very high speeds (currently 100 MB/s) and over greater distances than SCSI. According to its characteristics and its configuration, they allow the server to participate in different connectivity topologies.

Fibre Channel

Some people refer to Fibre Channel as the Fibre version of SCSI. Fibre Channel (FC) is capable of carrying IPI traffic, IP traffic, FICON traffic, FCP (SCSI) traffic, and possibly traffic using other protocols, all at the same level in the standard FC transport. Two types of cable can be used: copper and fiber. Copper for shorter distances and fiber for the longer distances.

Fibre Channel host adapter

The host adapter (HA) is the physical component of the storage server used to attach one or more host I/O interfaces.The ESS can be configured with Fibre Channel host adapters. The Fibre Channel host adapter is connected to the fibre link and accepts the upper layer commands (more than one protocol is supported by the Fibre Channel standard) that are sent by the host system.The Fibre Channel host adapter has one port for fibre connection.

FCP

Fibre Channel Protocol. When mapping SCSI to the Fibre Channel transport (FC-4 Upper Layer) then we have FCP.

1.2.2 Data architecture

CKD

Count key data (CKD) is the disk data architecture used by zSeries and S/390 servers. Because data records can be variable length, they all have a count field that indicates the record size. The key field is used to enable a hardware search on a key, however, this is not generally used for most data anymore. ECKD is a more recent version of CKD that uses an enhanced S/390 channel command set.

The commands used by CKD are called Channel Command Words (CCWs); these are equivalent to the SCSI commands.

DASD

Acronym for *Direct Access Storage Device*. This term is common in the z/OS environments to designate a volume. This *volume* may be a physical disk or, more typically, a logical disk drive spread across multiple physical disks.

Fixed Block Architecture (FBA or FB)

Fibre Channel and SCSI use a fixed block architecture, that is, the logical disk drive is arranged in fixed size blocks or sectors. With an FB architecture the location of any block can be calculated to retrieve that block. The concept of tracks and cylinders also exists, because on a physical disk we have multiple blocks per track, and a cylinder is the group of tracks that exists under the disk heads at one point in time without doing a seek.

Hard disk drive (HDD) and disk drive module (DDM)

Also referred to as *disk drives*. The HDD is the primary nonvolatile storage medium that is used for any host data that is stored within a storage server. These are the round-flat rotating plates coated with a magnetic substance that store the data on their surface. The DDMs are the hardware replaceable units that pack the HDDs.

Logical disk drive

See *logical volume*.

Logical volume

The storage medium associated with a *logical disk drive*. A logical volume typically resides on one or more DDMs. For the ESS the logical volumes are defined at logical configuration time. For CKD the logical volume size is defined by the device emulation mode (3390 or 3380 track format). For FB hosts, the size is 100 MB to the maximum capacity of a rank. For theIBM @server iSeries 400 servers the size corresponds to the 9337 or 2105 emulated volume models.

Notes

The AIX operating system views a logical volume as a logical disk or a hard disk (hdisk), an AIX term for storage space.

For the FB hosts, the ESS logical volume sizes (also referred as LUN sizes) can be configured in increments of 100 MB.This increased granularity of LUN sizes enables improved storage management efficiencies, especially for Windows NT systems which have a limited number of LUNs and therefore, require full exploitation of the rank capacity.

Logical device

The logical disk drives in the ESS are seen by the host as its logical devices and will be pointed using an addressing scheme that depends on the attachment setup. For FICON and ESCON attachments it will be the ESCON or FICON unit address of a 3390 emulated device. For open systems with SCSI attachment it will be the SCSI target, LUN assigned to a logical volume. For open systems with Fibre Channel attachment it will be the Fibre Channel adapter, LUN assigned to a logical volume.

Logical unit

The Small Computer System Interface (SCSI) term for a *logical disk drive*.

Logical unit number (LUN)

The SCSI term for the field in an identifying message that is used to select a l*ogical unit* on a given target. The LUNs are the virtual representation of the logical disk drives as seen and mapped from the host system.

SCSI ID

An unique identifier (ID) 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.

1.2.3 Server platforms

The following are the different terms used to designate all that is related to the servers that process the applications and control the storage devices.

Note

Most of the terms described here, where applicable through out this book, are also used as *generic associations* of similar systems. These similar systems grouped under generic words, may be generically similar but not specifically identical because they have some variations. Support by the specific systems, of the described ESS features and functions, must be checked against the appropriate sources of information, that are also referred to in this book.

CKD server

The term *CKD server* is used to refer in a generic way to the zSeries 900 servers and the S/390 servers that are ESCON or FICON connected to the ESS. In these environments the data in the ESS for these servers is organized in CKD format.

These servers run the z/OS, OS/390, MVS/ESA, z/VM, VM/ESA, VSE/ESA, Linux and TPF family of operating systems.

ESCON host

A CKD server that uses ESCON channels to connect to the ESS

FB server

The term *FB server* is used to refer in a generic way to the hosts that attach SCSI or Fibre Channel to the storage server. For these servers, the data in the ESS is organized according to the Fixed Block Architecture characteristics. These servers run the Windows NT, Windows 2000, Novell NetWare, DYNIX/ptx, IBM AIX, the OS/400 operating systems, and the non-IBM variants of UNIX.

Fibre Channel host

An FB server that uses Fibre Channel adapter cards to connect to the ESS

FICON host

A CKD server that uses FICON channels to connect to the ESS

Intel-based servers

The term *Intel-based server* (or *Intel* servers) is used to refer to all the different server makes that run on Intel processors. This is a generic term that includes servers running Windows NT and Windows 2000, as well as Novell Netware, and DYNIX/ptx. These servers are the IBM Netfinity and IBM PC Server families of servers, the IBM NUMA-Q servers, the most recent e-business IBM @server xSeries family of servers, and the various non-IBM server makes available on the market that run using Intel processors.

iSeries

The term *iSeries* (or *iSeries 400*) is used to refer to the IBM @server iSeries 400 family of enterprise e-business servers. The iSeries 400 is the successor to the AS400 family of processors. This term is also used in a generic way to include the IBM AS/400 family of servers. These servers run the OS/400 operating system.

Mainframe

The term *mainframe* is used to refer in a generic way to the most recent enterprise e-business zSeries 900 family of processors of the IBM @server brand, and the previous IBM 9672 G5 and G6 (Generation 5 and 6, respectively) S/390 processors. The zSeries servers run under the z/Architecture, and the 9672 and previous servers run under the S/390 (System/390) architecture.

Open systems

The term *open systems* (or *open servers*) is used in a generic way to refer to the systems running Windows NT, Windows 2000, Novell NetWare, DYNIX/ptx, as well as the systems running IBM AIX, IBM OS/400, and the many variants of non-IBM UNIX operating systems.

pSeries

The term *pSeries* is used to refer to the IBM @server pSeries family of enterprise e-business servers. The pSeries are the successors to the RS/6000 and RS/6000 SP family of processors. This term is also used in a generic way to include the IBM RS/6000 and the IBM RS/6000 SP family of servers. These servers run the AIX operating system.

SCSI host

An FB server that uses SCSI adapters to connect to the ESS

UNIX servers

The term *UNIX servers* is used in a generic way to refer to the servers that run the different variations of the UNIX operating system. This includes the IBM eserver pSeries, and the RS/6000 and the RS/600o0 SP family of servers, running IBM AIX. It also includes the non-IBM servers running the various variants of UNIX, like the Sun servers running Solaris, the HP9000 Enterprise servers running HP-UX, the Data General AViiON servers running DG/UX, the Compaq Alpha servers running Open VMS or Tru64 UNIX.

xSeries

The term *xSeries* is used to refer to the IBM eserver xSeries family of enterprise e-business servers. The xSeries are the new servers in the IBM Netfinity and PC Server family of servers. This term is also used in a generic way to include the IBM Netfinity and IBM PC Server families of servers. These servers run operating systems on Intel processors.

z/Architecture

The IBM 64-bit real architecture implemented in the new IBM eserver zSeries 900 enterprise e-business family of servers. This term is also used in a generic way to include the predecessor IBM S/390 architecture.

z/OS

This IBM operating system is highly integrated with the z/Architecture microcode and hardware implementation of the zSeries family of processors. It is the evolution of the OS/390 operating system.

The term z/OS is also used in a generic way, to encompass the OS/390 and the MVS/ESA operating systems.

z/VM

This IBM operating system is the evolution of the VM/ESA operating system.

The term z/VM is also used in a generic way, encompassing the VM/ESA operating system.

zSeries

The term *zSeries* (or *zSeries 900*) is used to refer to the IBM eserver zSeries 900 family of enterprise e-business servers. The zSeries 900 servers, with their z/Architecture, are the architectural evolution of the S/390 servers. These new servers run the z/OS, z/VM, OS/390, VM/ESA, VSE/ESA and TPF operating systems.

The term *zSeries* is also used in a generic way to include the S/390 families of IBM processors. This includes the latest S/390 9672 G5 and G6 processors.

1.2.4 Other terms

Array
A group of disk drive modules (DDMs) that are arranged under a relationship. Also referred to as a *rank*. For the ESS, an array is a group of up to eight DDMs.

Backend
All the hardware components of the storage server that are connected and functionally relate, starting from the cluster processors and down to the DDMs. In the ESS, these refer basically the SSA device Adapters, the internal DDMs, and the buses that communicate them to the cluster processors.

Cluster
A partition of a storage server that is capable of performing all functions of a storage server. When a cluster fails in a multiple-cluster storage server, any remaining clusters in the configuration can take over the processes of the cluster that fails (fail-over).

Controller image
This is used in the zSeries 900 (and S/390) environments to designate a logical subsystem that is accessed with an ESCON or with a FICON I/O interface. Also referred as *logical control unit* (LCU). One or more controller images exist in a physical controller. Each image appears to be an independent controller, but all of them share the common set of hardware facilities of the physical controller. In the Enterprise Storage Server, the CKD LSSs are viewed as a logical control units by the operating system. The ESS can emulate 3990-3, 3990-6, or 3990-3 TPF controller images.

Destage
The process of writing modified user data from cache to the disk arrays.

Frontend
All the hardware components of the storage server that are connected and functionally relate, starting from the cluster processors and up to the host adapters.

Staging
The process of moving data from an offline or low-priority device back to an online or higher priority device. In the ESS this is the movement of data records from the disk arrays to the cache.

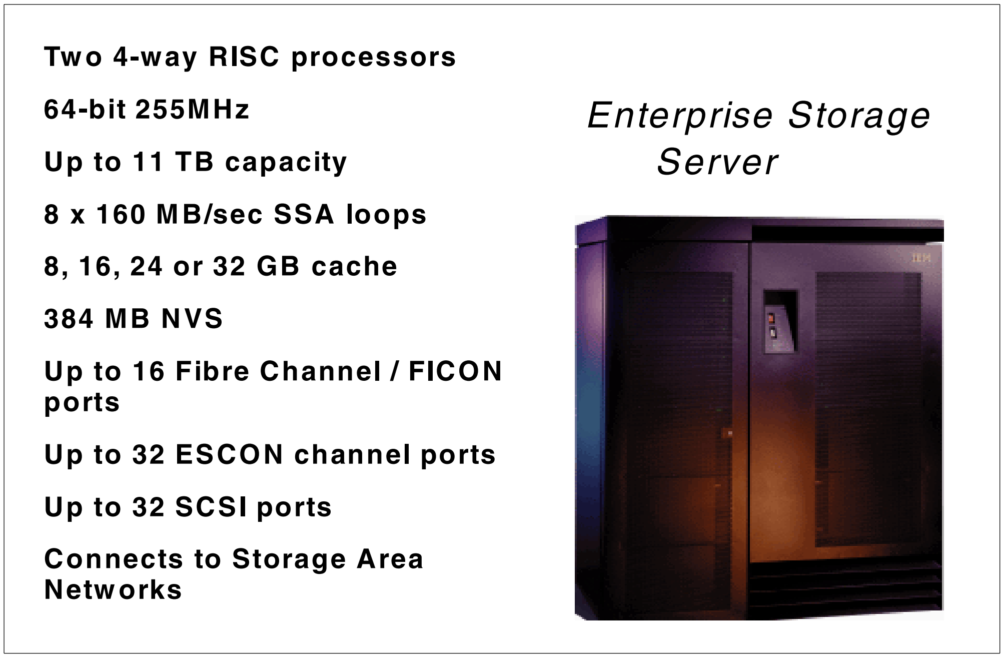
Storage server
A unit that manages attached storage devices and provides access to that storage and storage-related functions for one or more attached host systems.

Chapter 2. Hardware

The Enterprise Storage Server is an IBM Seascape Architecture disk system implementation for the storage and management of enterprise data. It is a solution that provides the outboard intelligence required by Storage Area Network (SAN) solutions, off-loading key functions from host servers and freeing up valuable processing power for applications. The ESS also provides Advanced Copy Services for backup-recovery and disaster-recovery situations. As a comprehensive SAN storage solution, the ESS provides the management flexibility to meet the fast paced changing requirement of today's business.

This chapter covers the physical hardware components of the Enterprise Storage Server Model F. This includes the models and the expansion rack. We also describe the internal device adapters, the different host connections and power characteristics.

2.1 Enterprise Storage Server overview



Two 4-way RISC processors

64-bit 255MHz

Up to 11 TB capacity

8 x 160 MB/sec SSA loops

8, 16, 24 or 32 GB cache

384 MB NVS

Up to 16 Fibre Channel / FICON ports

Up to 32 ESCON channel ports

Up to 32 SCSI ports

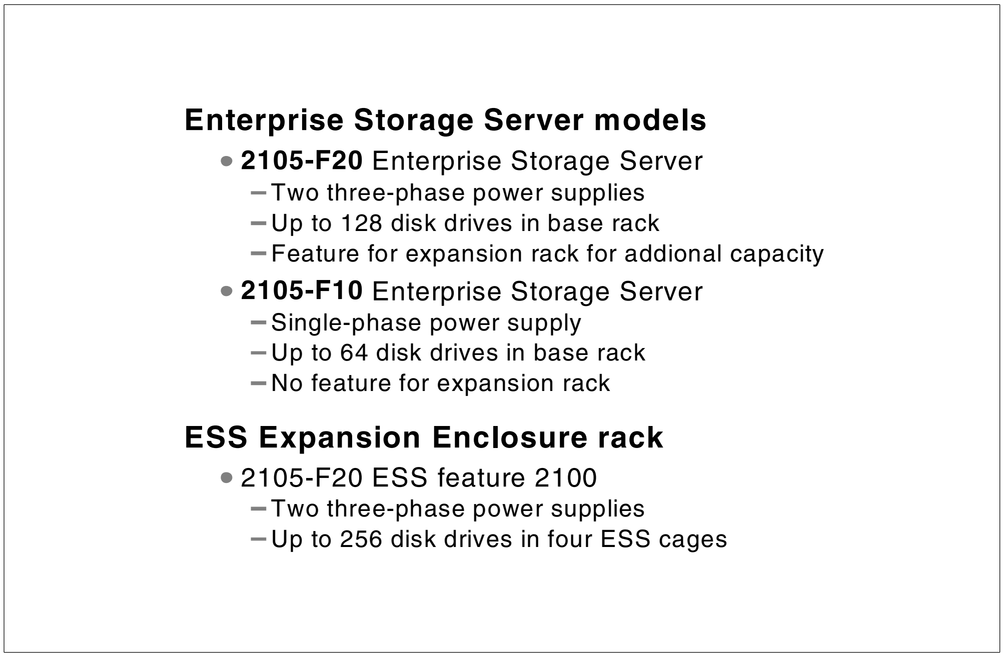
Connects to Storage Area Networks

Enterprise Storage Server

Figure 9. IBM Enterprise Storage Server overview

The IBM Enterprise Storage Server (ESS) is a high performance, high availability, and high capacity disk storage subsystem. It is a member of the Seascape family of storage servers. It contains two clusters each having a 4-way 64-bit RISC processor. It also has 384 MB of non-volatile storage (NVS), and options of either 8, 16, 24 or 32 GB of cache. The ESS has a capacity that scales up to 11.2 Terabytes. An expansion rack can be attached to the side of the ESS for the larger capacities. For servers with ESCON or FICON channels, like the IBM zSeries, the ESS can connect to them via 32 ESCON links or 16 FICON links. For servers with SCSI or Fibre Channel adapters, the ESS can connect to them either via 32 SCSI ports or 16 Fibre Channel ports. The ESS also connects to Storage Area Networks (SANs), and by means of a gateway it can also participate on a Network Attached Storage (NAS) environments.

2.2 ESS Models and Expansion Enclosure



Enterprise Storage Server models

- **2105-F20** Enterprise Storage Server
  - Two three-phase power supplies
  - Up to 128 disk drives in base rack
  - Feature for expansion rack for addional capacity
- **2105-F10** Enterprise Storage Server
  - Single-phase power supply
  - Up to 64 disk drives in base rack
  - No feature for expansion rack

ESS Expansion Enclosure rack

- 2105-F20 ESS feature 2100
  - Two three-phase power supplies
  - Up to 256 disk drives in four ESS cages

Figure 10. ESS models

2.2.1 Enterprise Storage Server models

The ESS has two models available, the model F10 and the model F20. Both models can be populated with the same internal disks. Both models also have the same clusters, RISC processors, host attachments and SSA adapters. The basic difference between the F10 and the F20 models is in the power supply, and the resulting scalability limit:

- The IBM 2105-F20 Enterprise Storage Server

  This model has two three-phase supplies and supports the full complement of 128 disk drives in two cages.

- The IBM 2105-F10 Enterprise Storage Server

  This model has two single-phase power supplies and, because of this, has limited capacity in terms of disk drive arrays. Only 64 disk drives can be installed in the base ESS rack. This uses a single cage.

  A 2105-F10 can be upgraded to 2105-F20.

The Enterprise Storage Server initially was introduced with two earlier models, the E10 and the E20. The E models and the new F models do not differ in external appearance but they do differ in the throughput and performance features as well as in the attachment capabilities. Because of its Seascape Architecture design, your can upgrade your currently installed model E to a model F.

2.2.2 ESS Expansion Enclosure (feature 2100)

The ESS Expansion Enclosure rack attaches to the model F20 only (feature 2100 of the 2105-F20) and uses two three-phase power supplies. Two power line cords are required for the ESS Expansion Enclosure, and they have the same requirements as the ESS base rack line cords. Up to four ESS cages can be installed in the Expansion Enclosure rack, and this gives a maximum of 256 disk drives for the expansion rack. This brings the total disk drive capacity of the model F20, when configured with Expansion Enclosure rack, to 384.

2.3 Photograph of the ESS

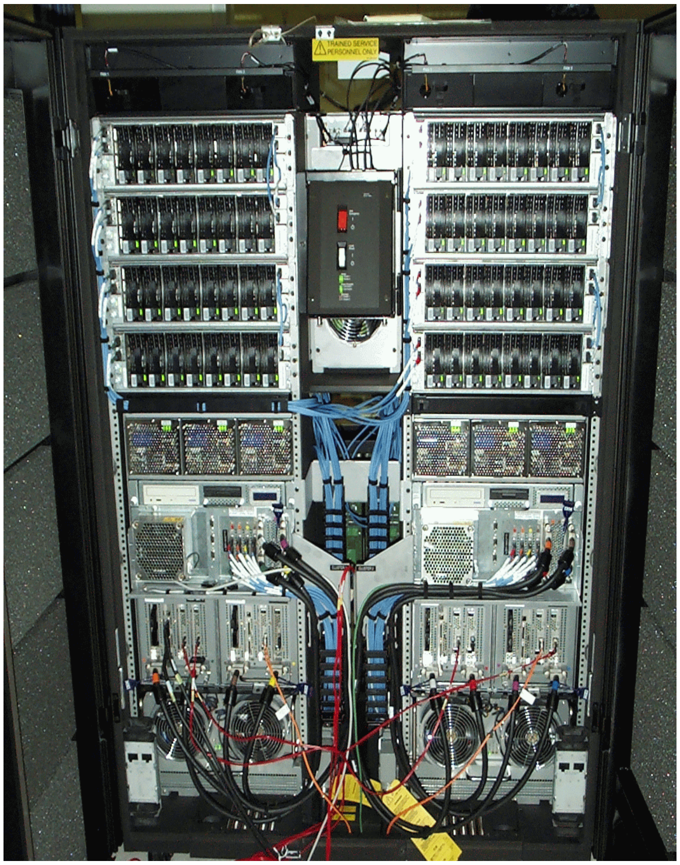


Figure 11. Photograph of the ESS (with front covers removed)

Figure 11 shows a photograph of a 2105-F20 Enterprise Storage Server with the front covers removed. At the top of the frame are the disk drives; beneath, the DC power supplies. And directly under these are the RISC processors. Just below the processors, and the Device Adapters, are the Host Adapters. And at the bottom of the frame are the AC power supplies and batteries.

The ESS in this photo has two cages holding the HDDs. If this would be a 64 or fewer disk drives configuration, then the up right portion of this ESS would have no cage in it. The photo clearly shows the two clusters, one on each side of the frame.

The height of the ESS is 70.7 inches (1.793 meters), width is 54.4 inches (1.383 meters), and depth is 35 inches (0.89 meters), without its top cover. The ESS requires a raised floor environment.

For larger configurations the ESS attaches an expansion rack that is the same size as the ESS, and stands next to the ESS base frame shown in Figure 11.

2.4 ESS major components

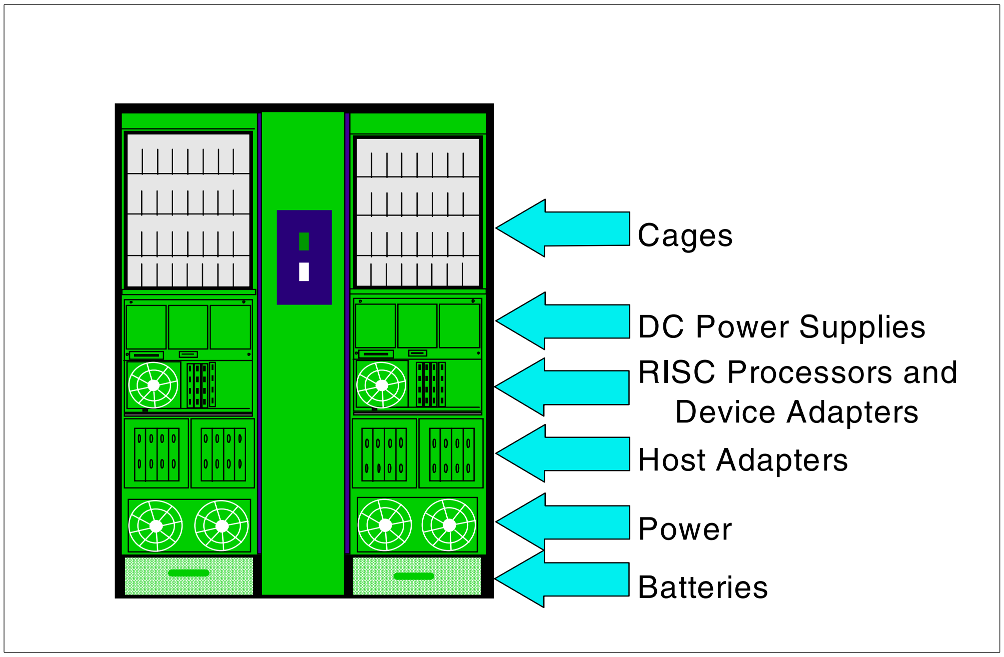


Figure 12. ESS major components

The diagram in Figure 12 shows a 2105 Enterprise Storage Server and its major components. As you can see, the ESS base rack consists of two clusters, each with their own power supplies, batteries, host adapters, device adapters, processors and hard disk drives.

At the top of each cluster is an ESS cage. Each cage provides slots for up to 64 disk drives, 32 in front and 32 at the back. If this were a 2105-F10 model, it would have only one cage located above on the left cluster.

In the following sections, we will look at the ESS major components in detail.

2.5 ESS cages

Cages

- F20, 1 or 2 cages
- F10, 1 cage
- Up to 64 disk drives per cage
- Disk drives installed in sets of 8 only (8-packs)
- 8 disk drives = 1 RAID rank
  - 6 + Parity + Spare or
  - 7 + Parity

Expansion Enclosure rack

- 1 to 4 ESS cages per expansion rack

Cage - up to 64 disk drives
(32 front/32 back)

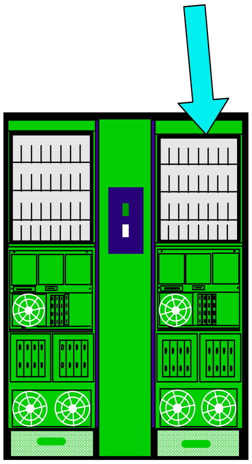


Figure 13. ESS cages

2.5.1 Cages

The high capacity ESS cages provide some common functions to the large set of disk they accommodate.

Disks are installed in the cages in groups of 8. These are called *disk 8-packs* (*8-pack* for short). The cage provides the power connections for each 8-pack which comes packaged into a case and slides into a slot in the cage.

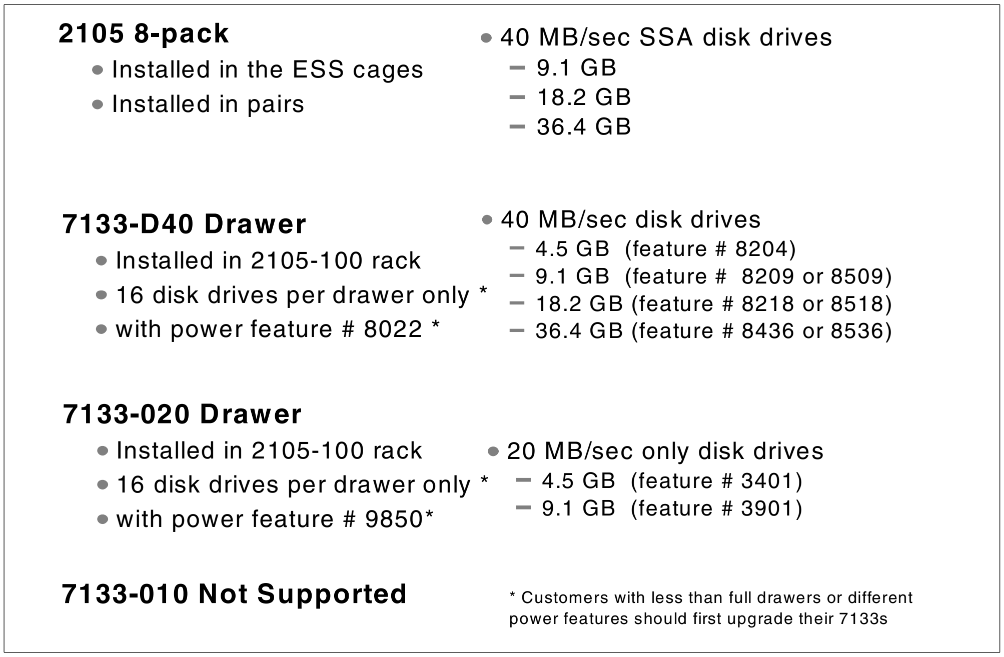
Each group of 8 disk drives is configured as a RAID *rank* of either 6 Data + Parity + Spare, or 7 Data + Parity. Or it can be configured as JBOD (Just a Bunch Of Disks)—No Parity. See Section 3.2, “Rank (array) types” on page 47.

The 8-packs are added in sets of two. The front of the left cage is filled-up first from bottom to top, then the back, again filled from bottom to top. Once the left cage is full, then the right cage is filled-up in the same order. Empty slots have a protective flap that controls the airflow over the disks.

2.5.2 Expansion Enclosure rack

The ESS Expansion Enclosure rack can hold from 1 to 4 cages. Each cage can contain up to 64 disk drives in sets of 8, giving to the configuration of the ESS an additional capacity of 256 disk drives in 4 cages.

2.6 ESS disks



2105 8-pack

- Installed in the ESS cages
- Installed in pairs

- 40 MB/sec SSA disk drives
  - 9.1 GB
  - 18.2 GB
  - 36.4 GB

7133-D40 Drawer

- Installed in 2105-100 rack
- 16 disk drives per drawer only *
- with power feature # 8022 *

- 40 MB/sec disk drives
  - 4.5 GB (feature # 8204)
  - 9.1 GB (feature # 8209 or 8509)
  - 18.2 GB (feature # 8218 or 8518)
  - 36.4 GB (feature # 8436 or 8536)

7133-020 Drawer

- Installed in 2105-100 rack
- 16 disk drives per drawer only *
- with power feature # 9850*

- 20 MB/sec only disk drives
  - 4.5 GB (feature # 3401)
  - 9.1 GB (feature # 3901)

7133-010 Not Supported

* Customers with less than full drawers or different power features should first upgrade their 7133s

Figure 14. ESS disks

The maximum disk drive capacity of the Enterprise Storage Server Model F20 is 384 — with 128 disk drives in the base unit and 256 disk drives in the expansion rack. Using 36GB disk drives and RAID-5, this gives a total usable capacity of approximately 11TB.

The minimum available configuration of the 2105-F10 or F20 is 420 GB. This capacity can be configured with 64 disk drives of 9.1 GB contained in eight 8-packs, or with 32 disk drives of 18.2 GB contained in four 8-packs, using one ESS cage. All 8-packs must be ordered and installed in pairs.

Also for protecting your investment in existing 7133 Serial Disks, the 7133 models D40 and 020 can be attached to the Enterprise Storage Server to exploit the ESS advanced functions.

The disk drives installed in the ESS are the latest state-of-the-art magneto resistive head technology disk drives. They support all the advanced disks functions, such as predictive failure analysis.

2.6.1 2105 8-pack

The ESS 8-pack is the basic unit of capacity within the ESS base and expansion racks. As mentioned before, these 8-packs are ordered and installed in pairs. Each 8-pack can be configured as a RAID-5 rank (6+P+S or 7+P) or as a JBOD (Just a Bunch Of Disks).

If the choice is RAID ranks then the IBM TotalStorage ESS Specialist will normally configure the 8-packs of the base rack as 6+P+S and the 8-packs of the expansion rack as 7+P (*Note*: when you specify reserved loops for 7133 disk attachment, you have the possibility of having 8-packs with a 7+P configuration in the base ESS frame).

Currently you have the choice of three different disk drive capacities for use within an 8-pack:

- 9.1 GB/ 10,000 RPM disks
- 18.2 GB/ 10,000 RPM disks
- 36.4 GB/ 10,000 RPM disk

The eight disk drives assembled in each 8-packs are all of the same capacity. Each disk drive uses the 40 MB/sec. SSA interface on each of the four connections to the loop.

It is required that all disk drives attached to the same SSA loop should be of the same capacity. The reason for this is that sparing operations take place within the loop not within RAID rank. See Section 3.23, “Sparing” on page 82.

You cannot mix 8-packs and 7133 drawers on the same loop. You must reserve specific loops (using the Reserve Loop feature 9904) for 7133 attachment exclusively.

2.6.2 Disk intermixing

Statement of Direction (SOD): IBM plans expanded flexibility with support for intermix of 8-pack features within an Enterprise Storage Server. See announcement letter 101-037 from February 2001.

This statement represents IBM's current intent and is subject to change or withdrawal.

2.6.3 7133-D40 drawers

The Enterprise Storage Server provides investment protection by supporting the attachment of your existing 7133 Serial DIsks. The 7133-D40 drawer can be attached to the ESS base unit using new 2105 model 100 racks. The 7133-D40 drawers in a 7015-R00 rack cannot be attached directly to an ESS. They must be removed and installed in the 2105-100 racks. This restriction is because the ESS cannot manage the power sequencing of the 7015-R00 rack.

Up to twelve 7133 Model D40 drawers installed in 2105-100 racks can be attached to the ESS. Mixing of 8-packs and 7133 drawers in the same loop is not allowed. SSA loops in the ESS must be reserved for 7133 attachment. ESS Reserve Loop feature #9904 reserves 2 loops for up to six 7133 drawers. Feature #9904 must be ordered twice to attach seven to twelve 7133 drawers. Reserve Loop feature #9904 is available only when you order the machine. An ESS Expansion Enclosure (feature 2100) cannot be part of the ESS when attaching 2105-100 racks.

All disk drives on an SSA loop must be of the same capacity. This is because the sparing operations take place within the loop not within the rank (See Section 3.23, “Sparing” on page 82). Note that the capacity of the 7133 disk drives does not have to match the capacity of the disk drives in the base ESS.

The 7133-D40 drawer must contain a full complement of 16 disks of the same capacity. The 7133-D40 drawers must also be equipped with power feature 8022.

Configurations with less than full drawers or different power features should first be upgraded.

The disk drives supported by ESS in the 7133-D40 drawer are:

- 4.5 GB - 7,200 RPM (feature 8204)
- 9.1 GB - 7,200 or 10,000 RPM (features 8209 or 8509 respectively)
- 18.2 GB - 7,200 or 10,000 RPM (features 8218 or 8518 respectively)
- 36.4 GB - 7,200 or 10,000 RPM (features 8436 or 8536 respectively)

See Section 4.5, “Mixing with 2105-100 racks” on page 96 for additional considerations when attaching 7133 disk to the ESS.

2.6.4 IBM 7133-020 drawer

The 7133-020 drawers are supported only when installed in the 2105-100 racks.

The 7133-020 operates only at 20MB/sec. on its SSA interface, and should not be mixed on the same loop as the 40MB/sec. disks. Mixing the disks will impact the performance of all the disks in the loop.

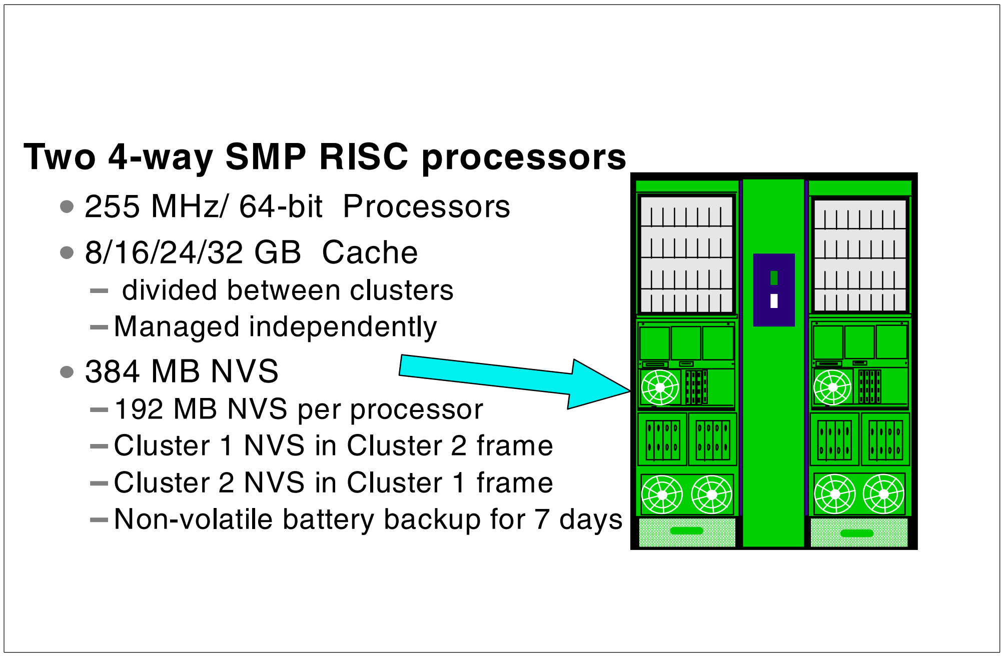
The disk drives supported by the ESS in the 7133-020 drawers are:

- 4.5 GB 7,200 RPM (feature 3401)
- 9.1 GB 7,200 RPM (feature 3901)

The same general considerations described for the 7133-D40 models also apply to the 7133-020 models. The 7133-020 drawers must also be equipped with power feature # 9850.

See Section 4.5, “Mixing with 2105-100 racks” on page 96 for additional considerations when attaching 7133 disk to the ESS.

2.7 ESS RISC processors



Two 4-way SMP RISC processors

- 255 MHz/ 64-bit Processors
- 8/16/24/32 GB Cache
  - divided between clusters
  - Managed independently
- 384 MB NVS
  - 192 MB NVS per processor
  - Cluster 1 NVS in Cluster 2 frame
  - Cluster 2 NVS in Cluster 1 frame
  - Non-volatile battery backup for 7 days

Figure 15. ESS RISC processors

2.7.1 Processors

The Enterprise Storage Server is a Seascape architecture subsystem and uses high performance IBM RISC processors to manage its operations. Each cluster of the models F10 and F20 have a 4-way 64-bit SMP processor running at 255MHz (the clusters of the earlier E models were 4-way 32-bit SMP processors running at 332 MHz).

2.7.2 Cache

Cache is used to store both read and write data to improve ESS performance to the attached host systems. The F-models give you a choice of 8, 16, 24 or 32GB of cache. This cache is divided between the two clusters of the ESS, giving the clusters their own non-shared cache. Cache operation is described in Section 3.10, “Cache and read operations” on page 65.

2.7.3 Non-volatile storage (NVS)

NVS is used to store a second copy of write data to ensure data integrity, should we get a power failure or a cluster failure and we lose the cache copy. The NVS of cluster 1 is located in cluster 2 and the NVS of cluster 2 is located in cluster 1. In the event of a cluster failure, the remaining cluster can access the NVS of the failed cluster and destage all the unwritten data to the disk drives. This ensures that no data is lost even in the event of a component failure.

Each cluster has 192MB of NVS and is protected by a battery. The battery protects the data in the case of a total power outage for up to 7 days.

A more detailed description of the NVS use is described in Section 3.11, “NVS and write operations” on page 67.

2.8 Device adapters

SSA 160 Device Adapters

- 4 DA pairs per subsystem
- 4 x 40 MB/sec loop data rate
- 2 loops per device adapter pair

Up to 48 disk drives per loop

- No mix of 8-packs and 7133 drawers
- Each group of 8 disk drives is
  - RAID-5 array, 6+P+S or 7+P
  - or 8 JBOD
- 2 spares per loop
- No mix of different capacity disk drives

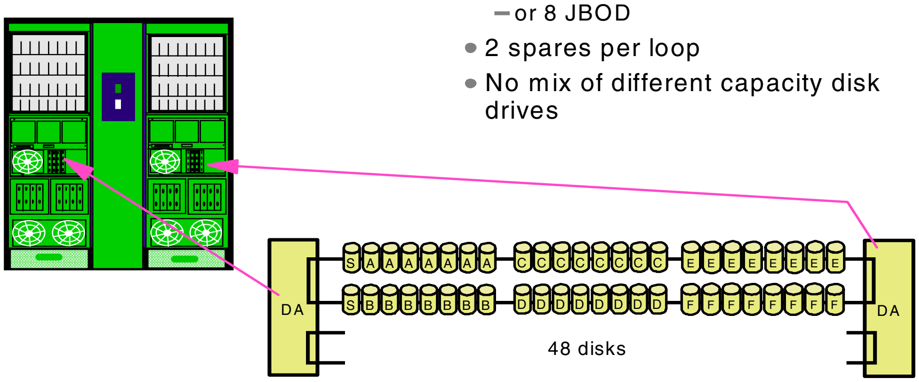


Figure 16. ESS device adapters

2.8.1 SSA 160 device adapters

The Enterprise Storage Server uses the latest SSA160 technology in its device adapters. With SSA 160, each of the four links operates at 40MB/sec., giving a total bandwidth of 160 MB/sec. for each of the two connections to the loop. This amounts for a total of 320 MB/sec. across each loop (see “Spatial reuse” on page 32). Also each device adapter card supports two independent SSA loops, giving a total bandwidth of 320 MB/sec. per adapter card.There are eight adapter cards, giving a total bandwidth capability of 2560 MB/sec. See Section 2.9, “SSA loops” on page 31 for more on the SSA characteristics.

One adapter from each pair of adapters is installed in each cluster as shown in Figure 16., “ESS device adapters” on page 29. The SSA loops are between adapter pairs, which means that all the disks can be accessed by both clusters. During the configuration process, each rank (RAID array or JBOD) is configured by the IBM TotalStorage ESS Specialist to be normally accessed by only one of the clusters. Should a cluster failure occur, the remaining cluster can takeover all the disk drives on the loop.

RAID-5 is managed by the SSA device adapters.

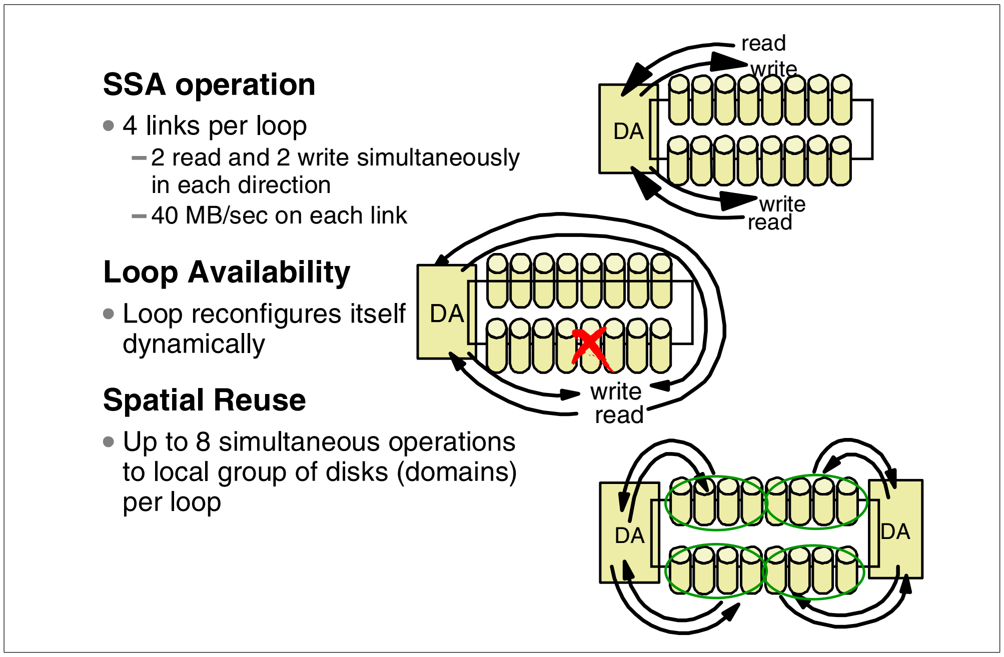
2.8.2 Disk drives per loop

Each loop supports up to 48 disk drives, and each adapter pair supports up to 96 disk drives. There are four adapter pairs supporting up to 384 disk drives in total.

The diagram in Figure 16 shows a logical representation of a single loop with 48 disk drives (RAID ranks are actually split across two 8-packs for optimum performance). In the figure you can see there are 6 RAID ranks (designated A to

F for this representation). Ranks A and B both have spare disks that are used across the loop in case of a disk failure. When the failed disk is replaced it becomes the new spare. Over time the disks (data and spare) in the RAID ranks on the loop become mixed. So it is not possible to remove an 8-pack without affecting the rest of the data in the loop. See Section 3.23, “Sparing” on page 82.

2.9 SSA loops



SSA operation

- 4 links per loop
  - 2 read and 2 write simultaneously in each direction
  - 40 MB/sec on each link

Loop Availability

- Loop reconfigures itself dynamically

Spatial Reuse

- Up to 8 simultaneous operations to local group of disks (domains) per loop

Figure 17. SSA loops

2.9.1 SSA operation

Serial Storage Architecture (SSA) is a high performance, serial connection technology for disk drives. SSA is a full duplex loop based architecture, with two physical read paths and two physical write paths to every disk drive attached to the loop. Data is sent from the adapter card to the first disk drive on the loop and then passed around the loop by the disk drives until it arrives at the target disk. Unlike bus based designs, which reserve the whole bus for data transfer, SSA only uses the part of the loop between adjacent disk drives for data transfer. This means that many simultaneous data transfers can take place on an SSA loop, and it is one of the main reasons that SSA performs so much better than SCSI. This simultaneous transfer capability is known as *spatial reuse*.

Each read or write path on the loop operates at 40MB/s, providing a total loop bandwidth of 160MB/s.

2.9.2 Loop availability

The loop is a self-configuring, self repairing design which allows genuine hot-plugging. If the loop breaks for any reason, then the adapter card will automatically reconfigure the loop into two single loops. In the ESS, the most likely scenario for a broken loop is if the actual disk drive interface electronics should fail. If this should happen, the adapter card will dynamically reconfigure the loop into two single loops, effectively isolating the failed disk drive. If the disk drive were part of a RAID array, the adapter card would automatically regenerate the missing disk drive (using the remaining data and parity disk drives) to the spare. Once the failed disk drive has been replaced, the loop will automatically be re-configured into full duplex operation, and the replaced disk drive will become a new spare.

2.9.3 Spatial reuse

Spatial reuse allows domains to be set up on the loop. A domain means that one or more groups of disk drives “belong” to one of the two adapter cards, as is the case during normal operation. The benefit of this is that each adapter card can talk to its domains (or disk groups) using only part of the loop. The use of domains allows each adapter card to operate at maximum capability because it is not limited by I/O operations from the other adapter. Theoretically, each adapter card could drive its domains at 160MB/s, giving 320MB/s throughput on a single loop! The benefit of domains may diminish slightly over time, due to disk drive failures causing the groups to become intermixed, but the main benefits of spatial reuse will still apply.

The spatial reuse feature sets SSA apart from other serial link technologies and from SCSI. With spatial reuse, SSA loops can achieve effective bandwidths well in excess of any individual link bandwidth. This is particularly true for RAID arrays where much data traffic transfer activity can occur within the disk array, independently of communication to the cluster controller.

If a cluster or device adapter should fail, the remaining cluster device adapter will own all the domains on the loop, thus allowing full data access to continue.

2.10 Host adapters

Host adapter bays

- 4 bays
- 4 host adapters per bay

ESCON host adapters

- up to 32 ESCON links
- 2 ESCON links per host adapter

FICON host adapters

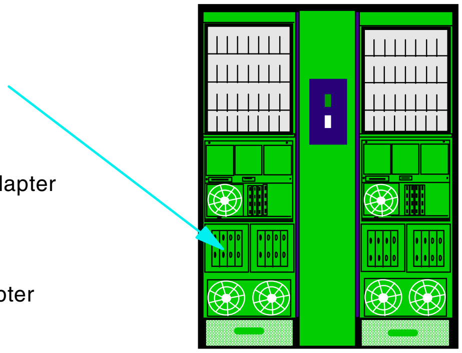
- up to 16 FICON links
- 1 FICON link per host adapter

SCSI host adapters

- up to 32 SCSI bus connections
- 2 SCSI ports per host adapter

Fibre Channel host adapters

- up to 16 Fibre Channel links
- 1 Fibre Channel port per host adapter



Adapters can be intermixed

- Any combination of host adapter cards up to a maximum of 16

Figure 18. ESS host adapters

2.10.1 Host adapter bays

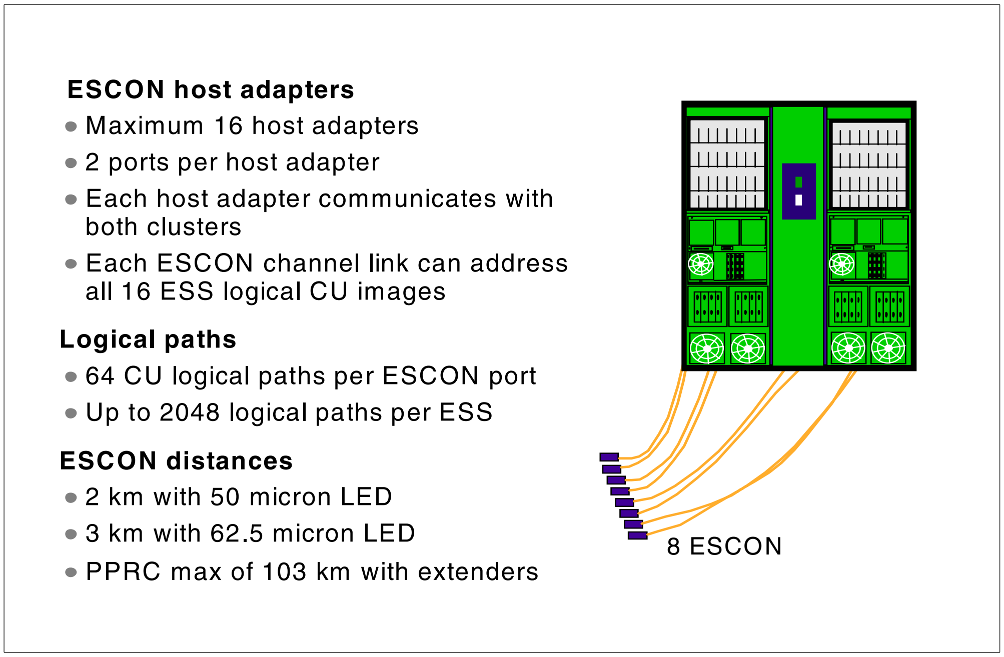
The Enterprise Storage Server has four host adapter (HA) bays, two in each cluster. Each bay supports up to four host adapters cards. Each of these host adapter cards can be either for FICON, or ESCON, or SCSI, or Fibre Channel server connection.

Each host adapter can communicate with either cluster, so there is no requirement to install a host adapter in a cluster 1 bay just to attach to cluster 1.

To install a new host adapter card, the bay must be powered off. For this reason, it is important to spread the host connections across all the adapter bays; this will minimize the impact. For example, if you have four ESCON links to a host, each connected to a different bay, then the loss of a bay for upgrade would only impact one out of four of the connections to the server. The same would be valid for a host with FICON connections to the ESS.

Similar considerations apply for servers connecting to the ESS by means of SCSI or Fibre Channel links. The Subsystem Device Driver (SDD) program that comes standard with the ESS to be installed in the connecting servers, handles errors (path failover) and balances the I/O over the paths to the ESS. See Appendix B, “Subsystem Device Driver (SSD)” on page 283.

2.11 ESCON host adapters



ESCON host adapters

- Maximum 16 host adapters
- 2 ports per host adapter
- Each host adapter communicates with both clusters
- Each ESCON channel link can address all 16 ESS logical CU images

Logical paths

- 64 CU logical paths per ESCON port
- Up to 2048 logical paths per ESS

ESCON distances

- 2 km with 50 micron LED
- 3 km with 62.5 micron LED
- PPRC max of 103 km with extenders

Figure 19. Host adapters — ESCON

2.11.1 ESCON host adapters

The Enterprise Storage Server can connect up to 32 ESCON channels, two per ESCON host adapter. Each ESCON host adapter is connected to both clusters. The ESS emulates up to 16 of the 3990 logical control units (LCUs). Half of the LCUs (even numbered) are in cluster 1, and the other half (odd-numbered) are in cluster 2. Because the ESCON host adapters are connected to both clusters, each adapter can address all 16 LCUs. More details on the CKD logical structure are presented in Section 3.6.6, “CKD implementation” on page 58.

2.11.2 Logical paths

An ESCON link consists of two fibers — one for each direction — connected at each end by an ESCON connector to an ESCON port. Each ESCON adapter card supports two ESCON ports or links, and each link supports 64 logical paths. With the maximum of 32 ESCON ports, the maximum number of logical paths is 2048.

2.11.3 ESCON distances

Apart from the standard 2 Km. with 50 micron multimode fiber, and the 3 Km. with 62.5 micron multimode fiber, you can extend the distance at which you can operate the ESS up to 103 Km. for PPRC implementations —control unit to control unit. This distance requires at least two pairs of IBM Fiber Savers 2029 Models 001 and RS1. These are a Dense Wavelength Division Multiplexer (DWDM) that can transport multiple protocols over the same fiber optic (this support was formerly provided by the IBM 9729 Optical Wavelenght Division Multiplexer). Each pair can be separated up to 50 Km. (31 miles) of fiber. The fiber can be dark-fiber (i.e. leased from a common carrier like a telephone company or cable TV operator) as long as it meets the 2029 attachment criteria.

2.12 SCSI host adapters

Ultra SCSI host adapters

- Differential Fast Wide
- 40 MB/sec
- 16 SCSI host adapters
- 2 ports per host adapter
- Each host adapter communicates with both clusters

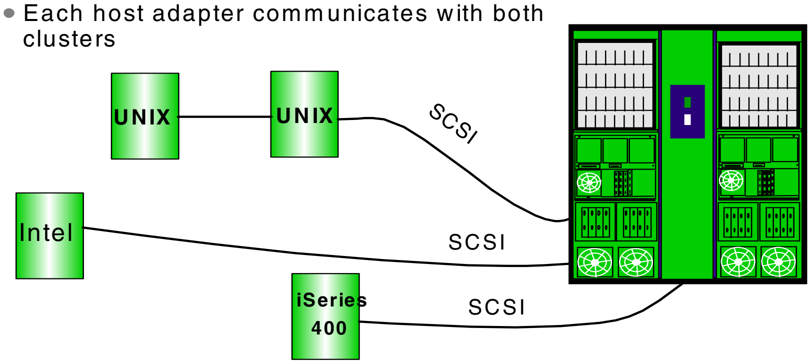


Figure 20. Host adapters — SCSI

2.12.1 SCSI host adapters

The Enterprise Storage Server provides Ultra SCSI interface with SCSI-3 protocol and command set for attachment to open systems This interface also supports SCSI-2.

Each SCSI host adapter supports two SCSI port interfaces. These interfaces are Wide Differential and use the VHDCI (Very High Density Connection Interface). These SCSI cables can be ordered from IBM.

2.12.2 SCSI supported servers

The current list of supported servers by the ESS SCSI interface include:

- IBM RS/6000, IBM RS/6000 SP, and the pSeries family of IBM @servers
- IBM AS/400 and the iSeries 400 family of the IBM @servers
- Compaq Alpha servers
- Data General AViiON
- HP9000 Enterprise servers
- SUN servers with Solaris
- Intel based servers: the IBM and the non-IBM supported servers

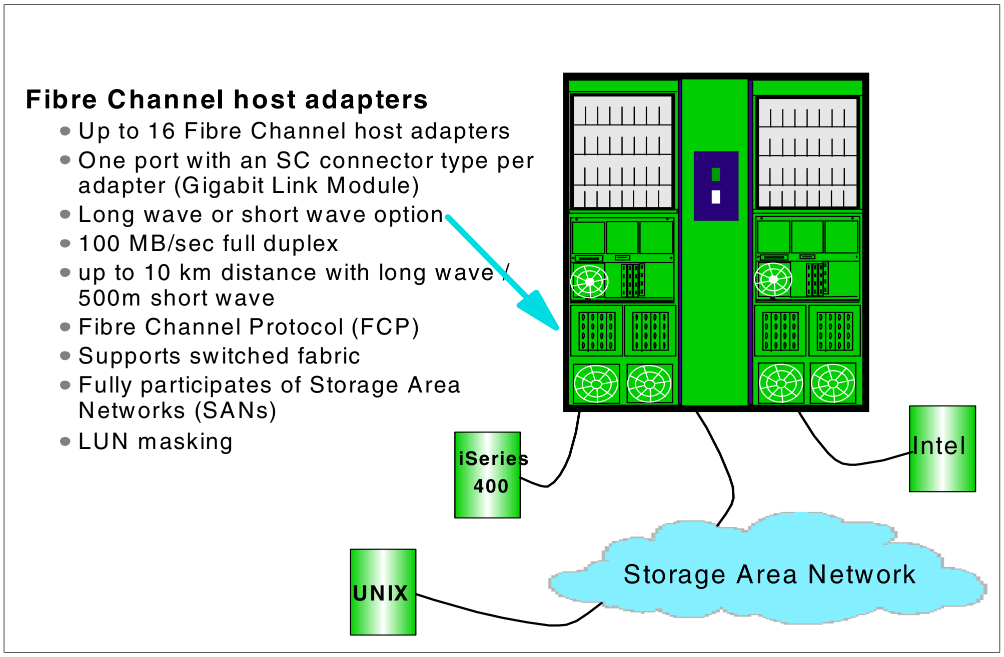
For a list of the specific server models that are supported and the most updated information when using SCSI attachment, please refer to the Web site at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

2.12.3 SCSI targets and LUNs

The ESS SCSI interface supports 16 target SCSI IDs (the host requires one initiator ID for itself, so this leaves 15 targets left for the ESS definitions) with up to 64 logical unit numbers (LUNs) per target (the SCSI-3 standard). The number of LUNs actually supported by the host systems listed above varies from 8 to 32. Check with your host server supplier on the number supported by any specific level of driver or machine.

See Section 3.6.5, “Fixed-block implementation” on page 56 for more detailed explanation on SCSI attachment characteristics.

2.13 Fibre Channel host adapters



Fibre Channel host adapters

- Up to 16 Fibre Channel host adapters
- One port with an SC connector type per adapter (Gigabit Link Module)
- Long wave or short wave option
- 100 MB/sec full duplex
- up to 10 km distance with long wave / 500m short wave
- Fibre Channel Protocol (FCP)
- Supports switched fabric
- Fully participates of Storage Area Networks (SANs)
- LUN masking

Figure 21. Host adapters — Fibre Channel

Fibre Channel is a technology standard that allows data to be transferred from one node to another at high speeds (100 MB/sec.) and greater distances (up to 10 Km.). The word *Fibre* in Fibre Channel takes the French spelling rather than the traditional spelling fiber, as in fiber optics. This is because the interconnection between nodes are not necessarily based on fiber optics, but can also be based on copper cables (*Note*: Fibre Channel on ESS is always based on fiber optics).

It is the very rich connectivity options of the Fibre Channel technology that has resulted in the Storage Area Network (SANs) implementations. The limitations seen on SCSI in terms of distance, performance, addessability and connectivity are overcame with Fibre Channel and SAN.

2.13.1 Fibre Channel host adapters

The ESS with its Fibre Channel/FICON host adapters provides FCP (Fibre Channel Protocol, which is the SCSI traffic on the serial fibre implementation) interface, for attachment to open systems that use Fibre Channel adapters for their connectivity.

Each Fibre Channel/FICON host adapter provides one port with an SC connector type, and the Gigabit Link Module (GLM) on the adapter provides an interface that supports 100 MB/second full duplex data transfer. The ESS supports up to 16 host adapters which allows for a maximum of 16 Fibre Channel/FICON ports per ESS.

There are two types of host adapter cards you can select: long wave (feature 3021) and short wave (feature 3023). With long wave laser you can connect nodes up to 10 Km. of distance. With short wave laser you can connect distances of up to 500m. See 2.15, “Fibre distances” on page 41 for more details.

Note: The Fibre Channel/FICON host adapter supports FICON and SCSI-FCP, but not simultaneously; the protocol to be used is configurable on an adapter by adapter basis.

When equipped with the Fibre Channel/FICON host adapters, configured for Fibre Channel interface, the ESS can participate in all the three topology implementations of Fibre Channel:

- Point-to-point
- Switched fabric
- Arbitrated loop

For detailed information on Fibre Channel implementations using the ESS refer to *Implementing Fibre Channel Attachment on the ESS,* SG24-6113

2.13.2 Fibre Channel supported servers

The current list of supported servers by the ESS using Fibre Channel interface include:

- IBM RS/6000, IBM RS/6000 SP, and the pSeries family of IBM @servers
- IBM @server iSeries 400
- Compaq Alpha servers
- HP9000 Enterprise servers
- Sun servers
- Intel based servers: the IBM and the non-IBM supported servers
- IBM NUMA-Q

For a list of the specific server models that are supported and the most updated information when using Fibre Channel attachment, please refer to the Web site at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

iSeries 400

The IBM @servers iSeries 400 models 820, 830, 840 and 270, with the Fibre Channel Disk adapter card #2766, and running OS/400 Version 5.1 or higher, now attach via fibre channel to the ESS.

2.13.3 Fibre Channel distances

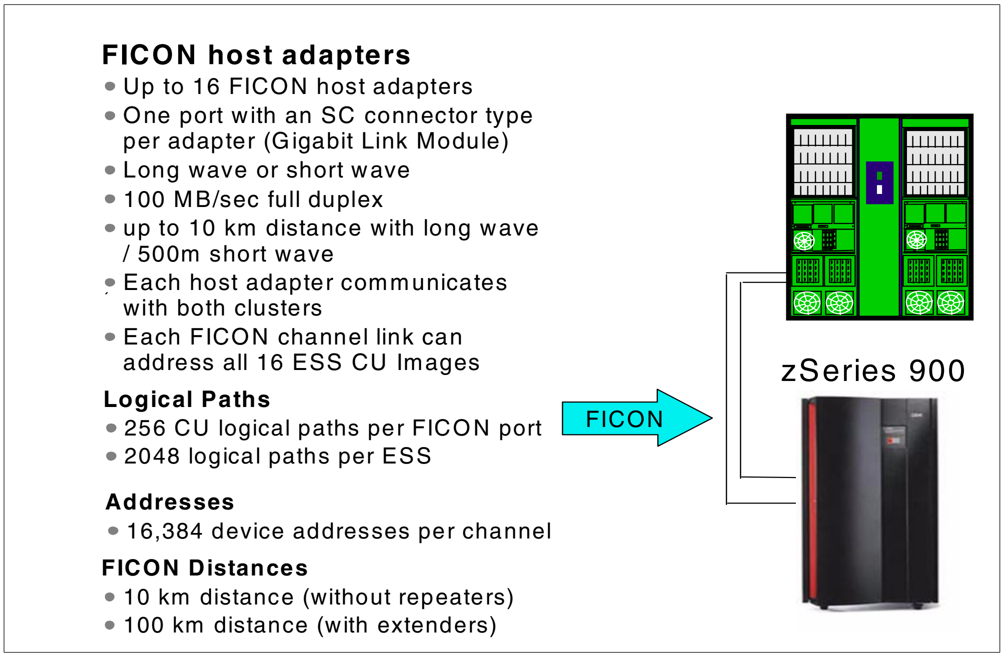
The ESS Fibre Channel host adapter you order, whether short wave or long wave, and the physical characteristics of the fiber you use to establish the link, will determine the maximum distances the nodes can connect to the ESS. See 2.15, “Fibre distances” on page 41 for additional information.

2.13.4 Storage Area Network

It is in the more complex switched fabric topology implementations, where various switches, server nodes, and storage nodes intelligently interconnect, that we see SAN implementations at its greatest capability.

Being SAN enabled, the Enterprise Storage Server can fully participate in current and future SAN implementations, like LAN-less or server-free backup solutions. For a more detailed description of these implementations please refer to *Introduction to Storage Area Network, SAN,* SG24-5470

2.14 FICON host adapters



FICON host adapters

- Up to 16 FICON host adapters
- One port with an SC connector type per adapter (Gigabit Link Module)
- Long wave or short wave
- 100 MB/sec full duplex
- up to 10 km distance with long wave / 500m short wave
- Each host adapter communicates with both clusters
- Each FICON channel link can address all 16 ESS CU Images

Logical Paths

- 256 CU logical paths per FICON port
- 2048 logical paths per ESS

Addresses

- 16,384 device addresses per channel

FICON Distances

- 10 km distance (without repeaters)
- 100 km distance (with extenders)

Figure 22. Host adapters — FICON

FICON (Fiber Connection) is based on the standard Fibre Channel architecture, and therefore shares the attributes associated with Fibre Channel. This includes the common FC-0, FC-1, and FC-2 architectural layers, the 100 MB/s bidirectional (full duplex) data transfer rate, and the point-to-point distance capability of 10 kilometers. The ESCON protocols have been mapped to the FC-4 layer, the Upper Level Protocol (ULP) layer, of the Fibre Channel architecture. All this provides full compatibility interface with previous S/390 software and puts the zSeries servers in the Fibre Channel industry standard. FICON goes beyond ESCON limits:

- Addressing limit, from 1024 device addresses per channel to up to 16,384
- Control unit logical paths per port, from 64 to up to 256
- FICON channel to ESS multiple concurrent I/O connections capability (ESCON supports only one I/O connection at one time)
- Greater channel and link bandwidth: FICON has up to 10 times the link bandwidth of ESCON (1 Gigabit/second full duplex, compared to 200 megabits/second half duplex). FICON has up to more than 4 times the effective channel bandwidth for the initial implementation (70 MB/sec. compared to 17 MB/sec.)
- FICON path consolidation using switched point-to-point topology.
- Greater unreported fiber link distances (from 3 Km. for ESCON to up to 10 Km., or 20 Km. with an RPQ, for FICON).

These characteristics allow more powerful and simpler configurations.

2.14.1 FICON host adapter

Each Fibre Channel/FICON host adapter provides one port with an SC connector type, and the Gigabit Link Module (GLM) on the adapter provides an interface that supports 100 MB/second full duplex data transfer. The ESS supports up to 16 host adapters which allows for a maximum of 16 Fibre Channel/FICON ports per machine.

There are two types of host adapter cards you can select: long wave (feature 3021) and short wave (feature 3023). With long wave laser you can connect nodes up to 10 Km. of distance. With short wave laser you can connect distances of up to 500m. See 2.15, “Fibre distances” on page 41 for more details.

Note: The Fibre Channel/FICON host adapter supports FICON and SCSI-FCP, but not simultaneously; the protocol to be used is configurable on an adapter by adapter basis.

When configured with the FICON host adapters the ESS can participate in point-to-point and in switched topologies.

2.14.2 FICON supported servers

FICON is used for attaching the Enterprise Storage Server to the IBM *e*™series zSeries 900 family of servers, and to the IBM S/390 processors 9672 G5 and G6 family of servers. These environments benefit from FICON exceptional characteristics.

2.15 Fibre distances

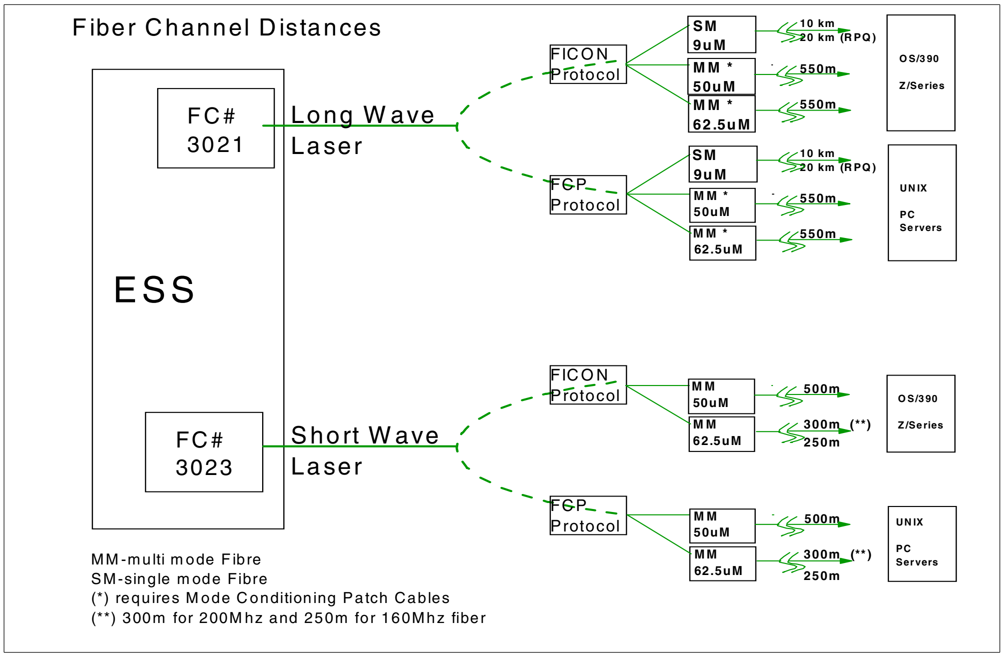


Figure 23. Fiber distances

The distance characteristics when using fiber on FICON or Fibre Channel implementations are common to both and are a function of the ESS host adapter cards been used (long wave vs. short wave) and of the fiber being used (9 vs. 50 vs. 62.5 micro-meters)

Figure 23 shows the supported maximum distances for the Fibre Channel/FICON host adapters of the ESS. As clearly depicted in Figure 23 the distance is dependent on host adapter card and the type of fiber. The quality of the fiber also determines the distance. Remember that distances up to 100 km. can be supported with the appropriate fabric components. You may refer to the publication *Fiber Optic Link Planning,* GA23-0367 for detailed considerations on fiber distances.

Note: the Fibre Channel / FICON (long wave) ESS host adapter (fc 3021) can be used with existing 50 and 62.5 micron ESCON cables, when the Mode Conditioning Patch cables are used.

2.15.1 Long wave host adapter

ESS feature number 3021 is the Fibre Channel / FICON (long wave) host adapter. The adapter supports FICON and SCSI-FCP, but not simultaneously; the protocol to be used is configurable on an adapter by adapter basis. When configured for a FICON interface, the host adapter connects to the zSeries servers. When it is configured for a Fibre Channel Protocol (FCP) interface it connects to the open systems servers that use Fibre Channel adapters for their connectivity.

2.15.2 Short wave host adapter

ESS feature number 3023 is the Fibre Channel / FICON (short wave) host adapter. The adapter supports FICON and SCSI-FCP, but not simultaneously; the protocol to be used is configurable on an adapter by adapter basis. When configured for a FICON interface, the host adapter connects to the zSeries servers. When it is configured for a Fibre Channel Protocol (FCP) interface it connects to the open systems servers that use Fibre Channel adapters for their connectivity.

2.16 Other interfaces

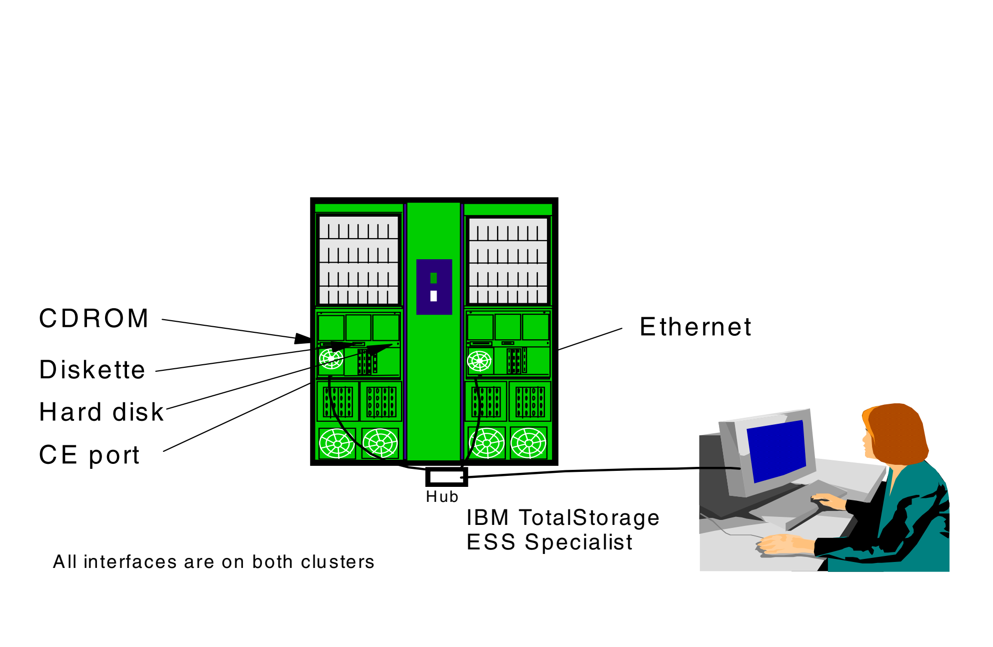


Figure 24. ESS — other interfaces

Each cluster of the Enterprise Storage Server has external interfaces that allow for licensed Internal Code (LIC) installation/update and the off-load of information.

The CDROM drive is used to load the Licensed Internal Code when LIC levels need to be upgraded. Both clusters have a CDROM drive, diskette drive, and a hard disk drive that is used to store both the current level of LIC and a new level, or the current level and the old level.

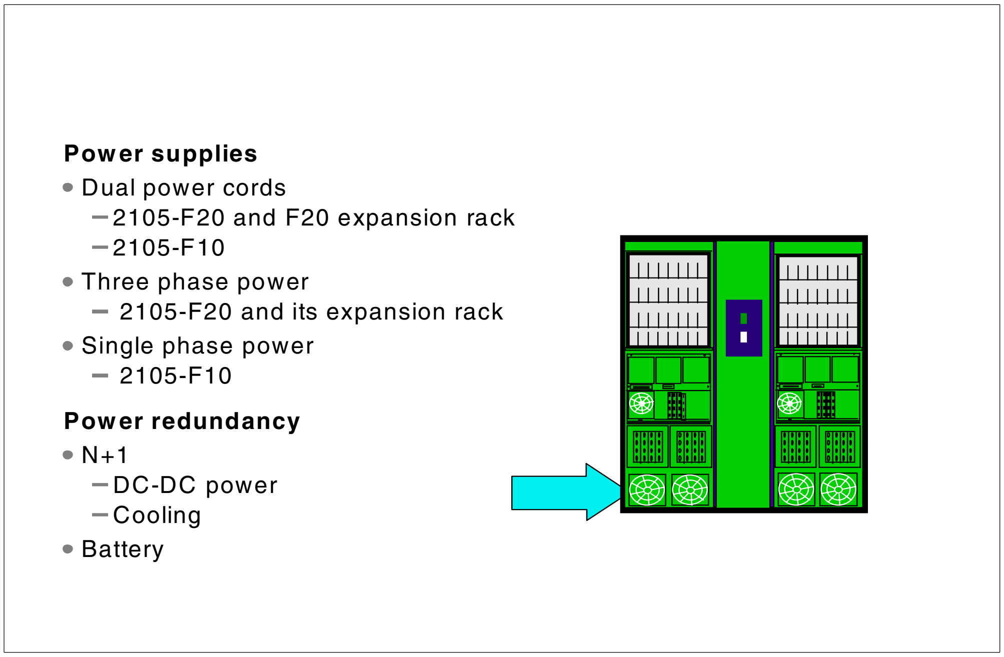
The CE port is used by the IBM System Support Representative (SSR) to connect the CE Mobile Solution Terminal (MOST). This allows the SSR to set up and test the ESS, and to perform upgrades and repair operations.

The customer interface is through an Ethernet connection (10/100BaseT) from the IBM TotalStorage ESS Specialist (ESS Specialist) running on the ESS to a customer supplied Web browser (Netscape or Internet Explorer). This interface allows you to configure the RAID ranks, and assign capacity to the various hosts. Details of how to configure the ESS are in Chapter 4, “Configuration” on page 91.

The two clusters come from the factory connected together by a simple Ethernet cable. During installation, the IBM SSR representative connects both clusters to the ESSNet hub. This way, the two clusters can communicate with each other as well as with the ESS Specialist. The ESSNet (feature code # 2715) hub has two ports available for customer use (one crossover, and one non-crossover), so you can connect your network to access the Web server from an external browser.

The ESS contains two service processors, one in each cluster, that monitor each cluster and handle power on and re-IML of the RISC processors.

2.17 Power supplies



Power supplies

- Dual power cords
  - 2105-F20 and F20 expansion rack
  - 2105-F10
- Three phase power
  - 2105-F20 and its expansion rack
- Single phase power
  - 2105-F10

Power redundancy

- N+1
  - DC-DC power
  - Cooling
- Battery

Figure 25. ESS — power supplies

The Enterprise Storage Server is a fault tolerant subsystem and has dual power cords to protect against power loss. The power units are N+1—so a failure in a single power unit has no effect and the failing unit can be replaced non-disruptively. Likewise, the fans can be replaced non-disruptively should they fail.

2.17.1 Power requirements

The two models of the ESS have different power connections, the 2105-F20 requiring 3-phase supply, and the 2105-F10 using a single phase supply. Note that you cannot just plug the F10 power connector into a wall socket; the single phase requires a 50-60 Amp supply.

2.17.2 Power redundancy

Each external power source can power the entire ESS in the event of the failure of one power input (each power cord should be attached to an independent power source). The ESS has a battery with sufficient capacity to provide power for a fully configured ESS for a minimum of 5 minutes if power from both line cords is lost. The ESS remains operational for the first 50 seconds. If power failure persists beyond the first 50 seconds, the ESS will begin an orderly shutdown. Once the orderly shutdown has started it cannot be interrupted.

When power is lost, the ESS initially destages all modified data in cache to disk, in readiness for a controlled shutdown. If the power loss is transient, then the ESS will continue with normal operations. If the power loss is permanent, then the ESS will shut down.

Chapter 3. Architecture

In this chapter we look at the logical structure of the Enterprise Storage Server (ESS) and the concept of Logical Storage Subsystems (LSS). We also look at the data flow, both read and write operations. Finally, we present the availability features of the ESS, from failure handling to maintenance procedures.

This chapter does not discuss the principles of RAID-5, as there are many redbooks that already cover this in detail, for example, *Enterprise Storage Solutions Handbook*, SG24-5250.

3.1 Overview

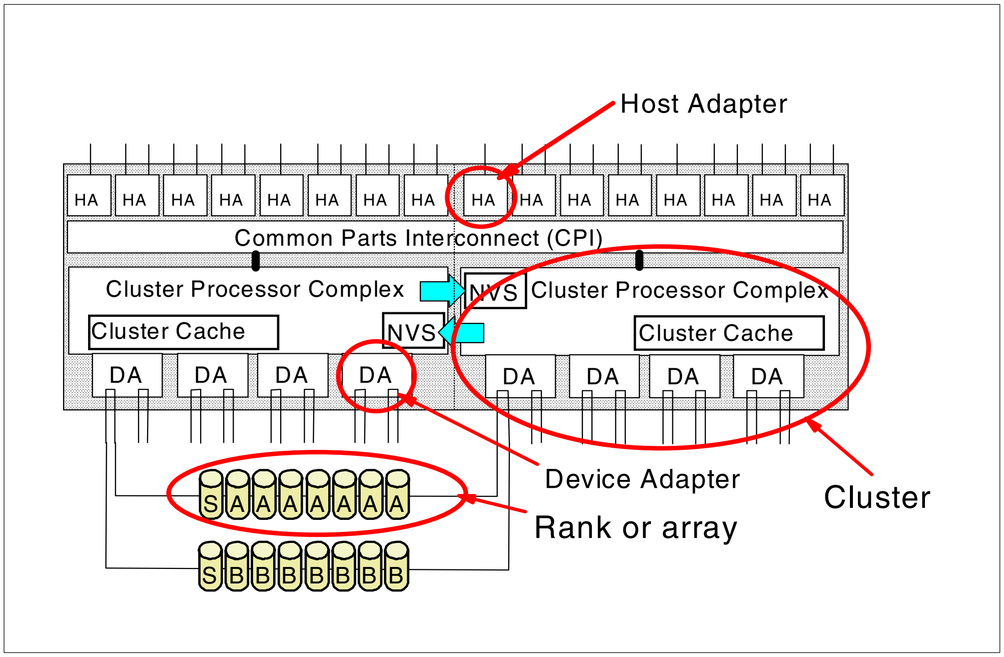


Figure 26. Overview

Figure 26 shows a schematic of the Enterprise Storage Server. At the top we have up to 16 *host adapters* (HAs), each HA supporting one Fiber Channel / FICON port or two ESCON or SCSI ports. Each HA is connected to both clusters through the *Common Parts Interconnect* (CPI) buses, so that either cluster can handle I/O from any host adapter. You can see the two clusters that contain the RISC processors, the cache, and the NVS for the opposite cluster. We shall discuss the NVS structure later. Within each cluster we have four *device adapters* (DAs). They always work in pairs, and the *disk arrays* (or *ranks*, as they are called in the ESS) are attached through an SSA loop to both DAs in a pair. The ranks can be configured as RAID *ranks*, or as *Just a Bunch Of Disks* (JBOD), as we shall later in this chapter. We shall look at each of these components in more detail on the following pages.

3.2 Rank (array) types

RAID rank (array)

- Formatted as multiple logical volumes (LV)
- LVs striped across the RAID array
- Whole rank is CKD or FB only
- Owned by one logical subsystem (LSS)

Non-RAID rank

- Group of 8 non-RAID disk drives
- Each disk drive in the group is a rank
- Each disk drive formatted as one or more logical volumes
- Mixed CKD and FB within the group
- Individual JBOD ranks owned by CKD or FB LSS

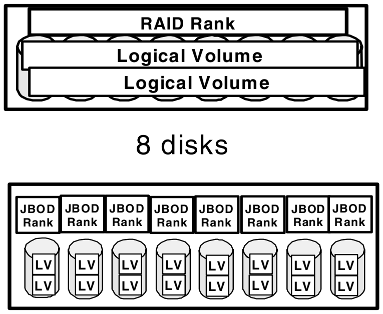


Figure 27. Rank (array) types

3.2.1 RAID rank

As we have already seen the basic unit of capacity in the ESS is the 8-pack that is a set of eight disk drives packed together and installed as a unit. These 8-packs are installed always in pairs and, initially, four disks from each of these two 8-packs make a *disk group*. You will find more detailed explanation on *disk groups* later in Section 4.11.4, “Disk groups — ranks” on page 112.

A RAID *rank* (or RAID array) is owned by one Logical Subsystem (LSS) only, either an FB LSS or a CKD (S/390) LSS.

Each rank is formatted as a set of *logical volumes* (LV). The number of LVs in a rank depends on the capacity of the disks drives in the array (9.1 GB, 18.2 GB, or 36.4 GB), and the capacity of *logical disks* being emulated (for example, 3390-3 for CKD) that were selected during the configuration steps. The LVs are striped across all the data and parity disks in the array.

As an example, a group of 8 disk drives (with a 6+P+S arrangement) has a capacity of about 53 GB (assuming 9 GB disks). This can then be formatted into 18 of the 3390-3 disks, or even one 53 GB LUN, or also in 3 LUNs of 16 GB each (plus spare) for use by a UNIX system.

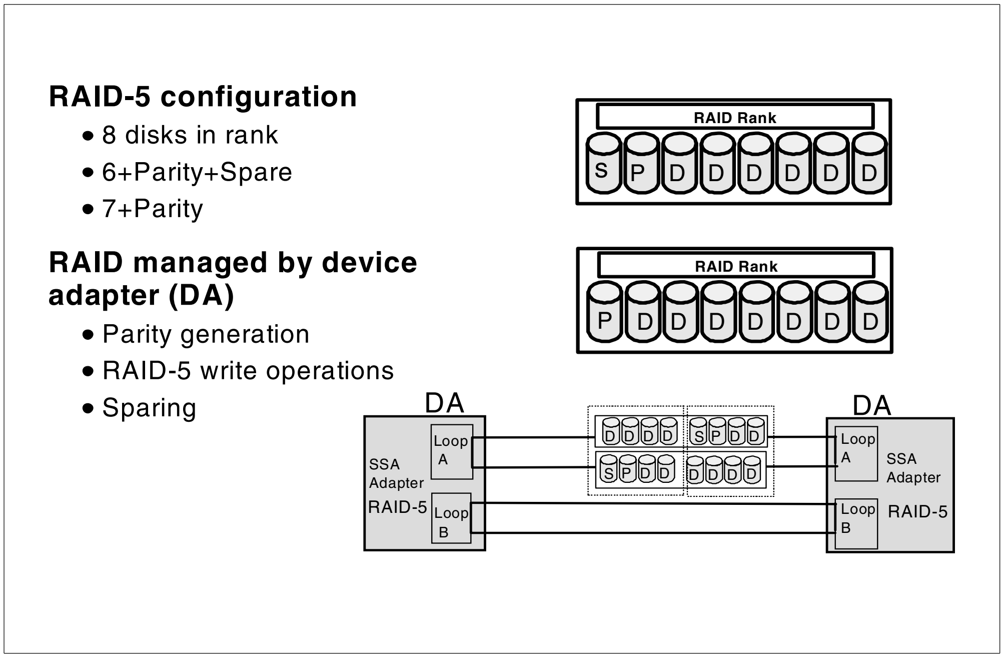
3.2.2 Non-RAID rank

A non-RAID rank, also called a JBOD rank, is very different; each disk drive in the group of 8 is a rank in itself. So there are 8 ranks in a JBOD group. Each of the 8 ranks is attached to only one of the four logical subsystems that support the loop. Each JBOD disk drive can be defined and formatted as one or more logical volumes, either as multiple CKD volumes or as multiple FB disks.

A JBOD rank is not RAID protected and, should a disk fail, all data on it will be lost.

Note that if you create a JBOD disk group in the first 16 disk drives on a loop, the ESS Specialist will leave 1 disk out of the 8 as spare, in case you create a RAID rank on the next 8 disk drives (you must always have 2 spares on a loop with RAID ranks).

3.3 Rank RAID operation



RAID-5 configuration

- 8 disks in rank
- 6+Parity+Spare
- 7+Parity

RAID managed by device adapter (DA)

- Parity generation
- RAID-5 write operations
- Sparing

Figure 28. Rank RAID operation

3.3.1 RAID-5 configuration

Protection against disk failures is provided by operating the ranks in RAID-5 mode. The Enterprise Storage Server only supports RAID ranks consisting of 8 disk drives. These are arranged in one of the two array modes shown in Figure 28. The first array type (6+P+S) contains one spare, one parity and 6 data disks. The second array type (7+P) contains a parity disk plus seven data disks. When there are disk drives in a loop, there has to be a minimum of two 8-packs in the loop, and these initial two *disk groups* are always configured as 6+P+S arrays. Additional arrays are configured as 7+P. As always with RAID 5, parity is distributed over all the disks. The 8 disk drives that make up the RAID rank are not in the same 8-pack.

3.3.2 RAID managed by device adapter

Each *device adapter* (DA) contains an SSA adapter that manages the two loops (A and B). The RAID operation is managed by the SSA adapter. Parity generation and the RAID-5 write operation is handled entirely within the SSA adapter for each loop, the SSA adapter containing on board memory to hold the data. No parity is seen by the RISC processors or is held in the ESS cache.

Sparing—the recovery of a failed disk drive onto one of the spare disk drives—is also handled automatically by the SSA adapter. The sparing process takes place in the background over a period of time, thus minimizing its impact on normal I/O operations. A failed disk drive can immediately be replaced and automatically becomes the new spare. Should a second disk drive fail in the same RAID rank before the sparing is complete, then we lose all the data in the rank and all write operations to that rank are suspended. If the second disk drive failure is on another RAID rank, or if it even occurs on the same rank but after completion of the previous rebuild process, then the rebuild takes place on the second spare.

Throughout the rest of this chapter, we show RAID ranks as if all the disk drives were grouped together on the same part of the loop. In practice, the disk drives in a RAID rank are organized for performance by grouping four DDMs from two different 8-packs into a *disk group* (you can find a more detailed discussion on disk groups in Section 4.11.4, “Disk groups — ranks” on page 112). This allows the SSA adapter to achieve maximum throughput for each RANK by having a path to each half of the rank down each leg of the loop (see the discussion on Section 2.9.3, “Spatial reuse” on page 32 for more detail).

3.4 Device adapters and logical subsystems

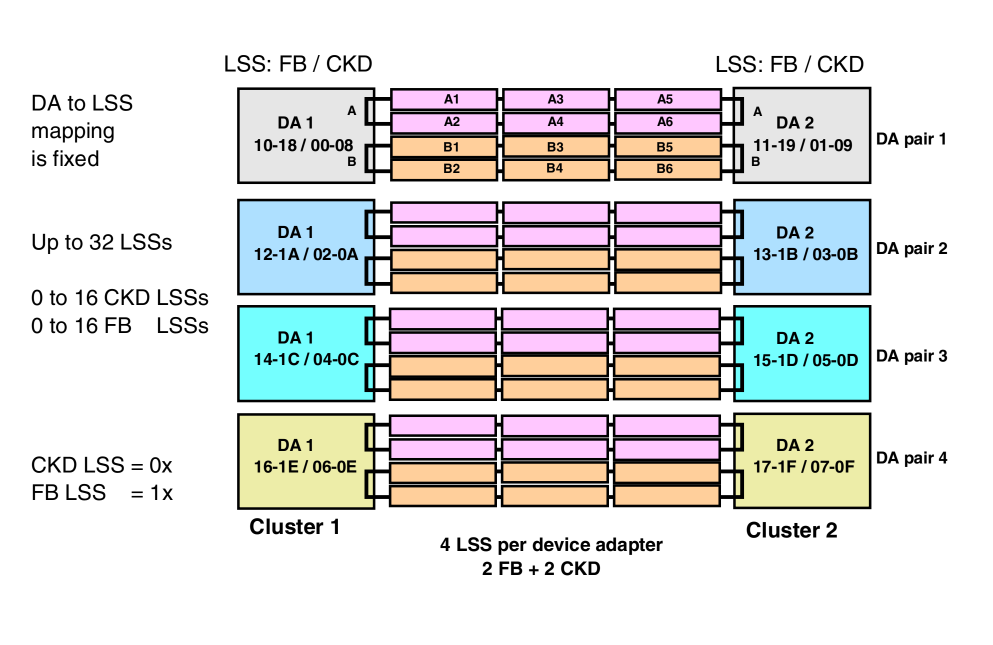


Figure 29. Device adapters and logical subsystems

The *logical subsystem* (LSS) is a logical structure that is internal to the Enterprise Storage Server and is used for configuration of the ESS. Although it relates directly to the *logical control unit* (LCU) concept of the ESCON and FICON architectures, it does not directly relate to SCSI and FCP addressing.

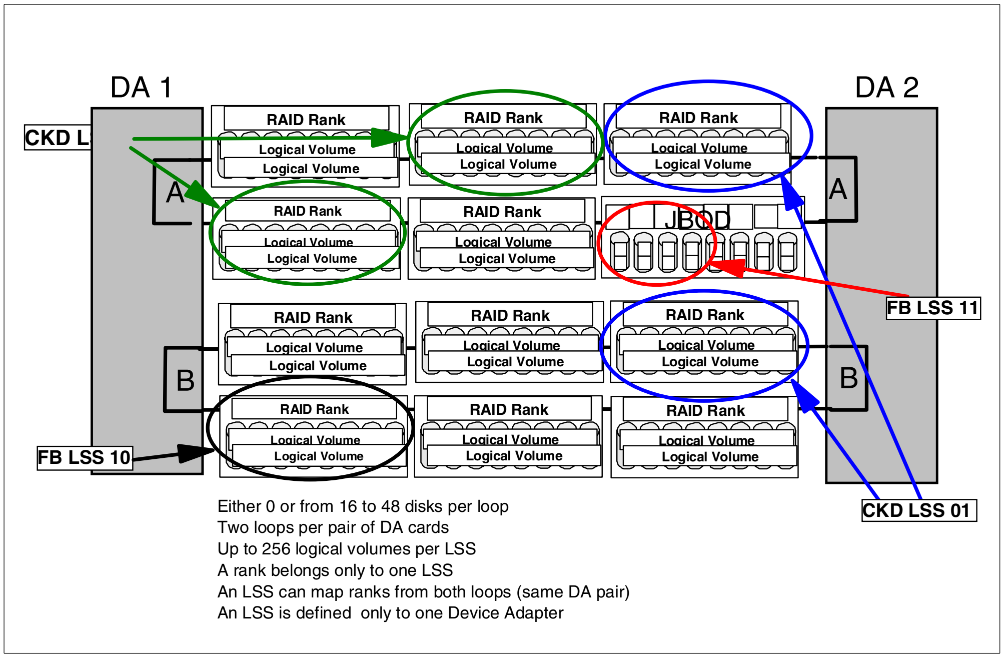
The *device adapter* (DA) to LSS mapping is a fixed relationship. Each DA supports two loops, and each loop supports two CKD logical subsystems and two FB logical subsystems (one from each cluster). So a DA pair supports four CKD LSSs and four FB LSSs.

When all the eight loops have capacity installed, then there are up to 16 CKD LSSs and up to 16 FB LSSs available to support the maximum of 48 RAID ranks (or groups of 8 JBOD ranks). Each LSS supports up to 256 *logical devices* (each logical device is mapped to a logical volume in the RAID ranks or JBOD ranks).

The numbering of the logical subsystems indicates the type of LSS. CKD logical subsystems are numbered x’00’ to x’0F’ and the FB logical subsystems are numbered x’10’ to x’1F’. For the CKD host view (i.e., zSeries server), a logical subsystem is also mapped one-to-one to a logical control unit.

As part of the configuration process, you can define the maximum number of logical subsystems of each type you plan to use. If you plan to use the ESS only for zSeries data, then you can set the number of FB LSSs to 0. This releases the definition space for use as cache storage (up to 2MB/LSS). But you must also remember that going from 8 to 16 LSSs is disruptive, so you should decide in advance how many you will need.

3.5 Logical subsystem (LSS)



Either 0 or from 16 to 48 disks per loop
Two loops per pair of DA cards
Up to 256 logical volumes per LSS
A rank belongs only to one LSS
An LSS can map ranks from both loops (same DA pair)
An LSS is defined only to one Device Adapter

Figure 30. Logical subsystem (LSS)

Let us now look at how the RAID and JBOD ranks are used within a *logical subsystem* (LSS). Logical subsystems are related to the device adapter SSA loops, and are thus managed by only one cluster. An LSS belongs to one DA.

Each loop supports from 16 to 48 disk drives. The minimum is 16 for RAID ranks, because we always need two spare disk drives on any loop with RAID ranks, so the minimum is two 6+P+S ranks. The disk drives are configured into 2 to 6 RAID ranks—either CKD or FB, or 2 to 6 JBOD groups, or a combination of the two types. If, for example, all 48 disk drives on the loop were JBOD, we would have 48 JBOD ranks, each of which would be CKD or FB.

As part of the configuration process, each rank is assigned to one logical subsystem. This LSS is either CKD or FB. Two ranks, from two different loops of the same DA pair can be associated to build up a LSS.

Example

In the example shown in Figure 30, we have the maximum of 48 disk drives installed on both loops. We plan to map six groups of disk drives onto four LSSs.

Five RAID ranks are defined and one group of 8 JBODs. We have four logical subsystems available on the DA. If we assume that this is the first DA pair in an ESS, then we can also associate the logical subsystem numbers. Please note that the order into which you define your LSS determines the LSS number: here we have created the JBOD FB LSS after the RAID FB LSS.

As a consequence of the LSS mapping algorithms, here is the LSS definition:

1. DA1 Loop A LSS(00)—CKD: Two RAID ranks with a total 16 disk drives.
2. DA1 Loop B LSS(10)—FB: One RAID rank with 8 disk drives.

3. DA 2 Loop A LSS(11)—FB: Four JBOD disk drives (ranks) formatted for FB use.
4. DA 2 *Loops A and B* LSS(01)—CKD: Two RAID ranks from two different loops are associated, with a total of 16 disk drives.

Note there are still six groups of disk drives to map onto either existing or new LSSs.

LSS mapping of ranks on more than one loop (of the same DA pair)
For an LSS, allocating ranks from different loops (from the same DA pair) could be useful, especially for FlashCopy. However you must remember that all the capacity that an LSS is able to see, must be mapped with no more than 256 logical volumes. 256 is the maximum number of logical volumes that can be defined for an LSS. Moreover, for CKD servers you should take in consideration the PAV addresses to be defined.

3.6 Host mapping to logical subsystem

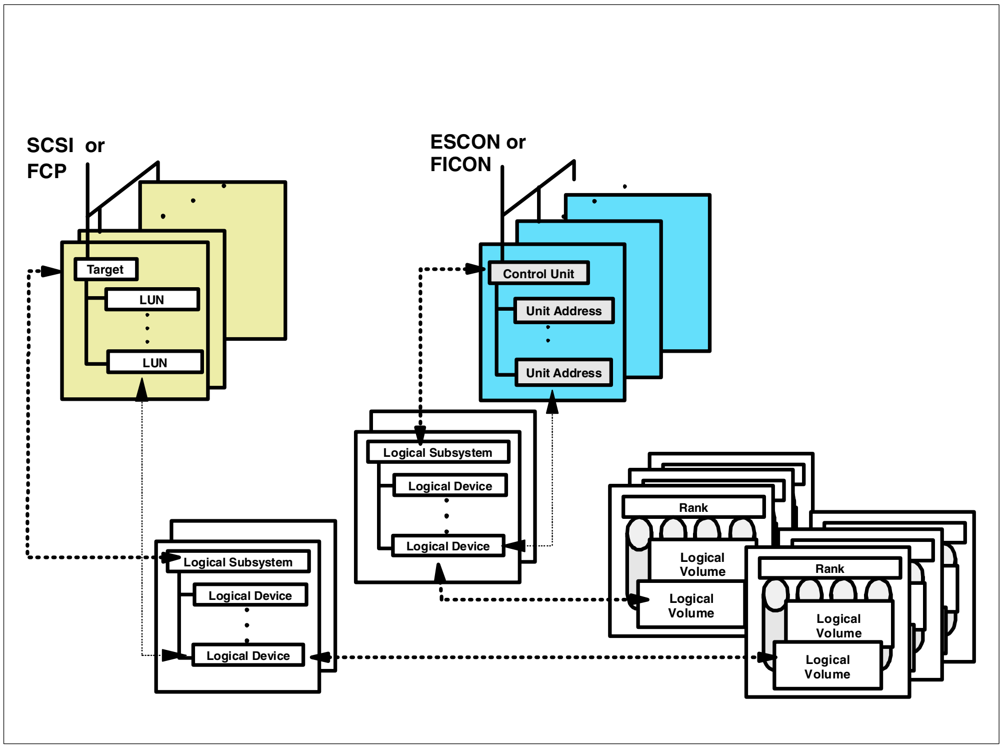


Figure 31. Host mapping to logical subsystem

3.6.1 Hosts mapping concepts

For the zSeries servers, the data stored in the ESS is arranged in a count-key-data (CKD) format, and this data is retrieved with the I/O operations that the host servers do. These I/O operations, for the zSeries servers, are done according to the ESCON or FICON architectures.

For the open system servers, the data stored in the ESS is arranged in a fixed-block (FB) format, and this data is retrieved with the I/O operations that the host servers do. These I/O operations, for the open system servers, are done according to the SCSI or Fibre Channel architectures.

Additionally, the ESS has its own logical view of the stored data, based on the *logical storage subsystem* (LSS) definitions.

3.6.2 SCSI and Fibre Channel mapping

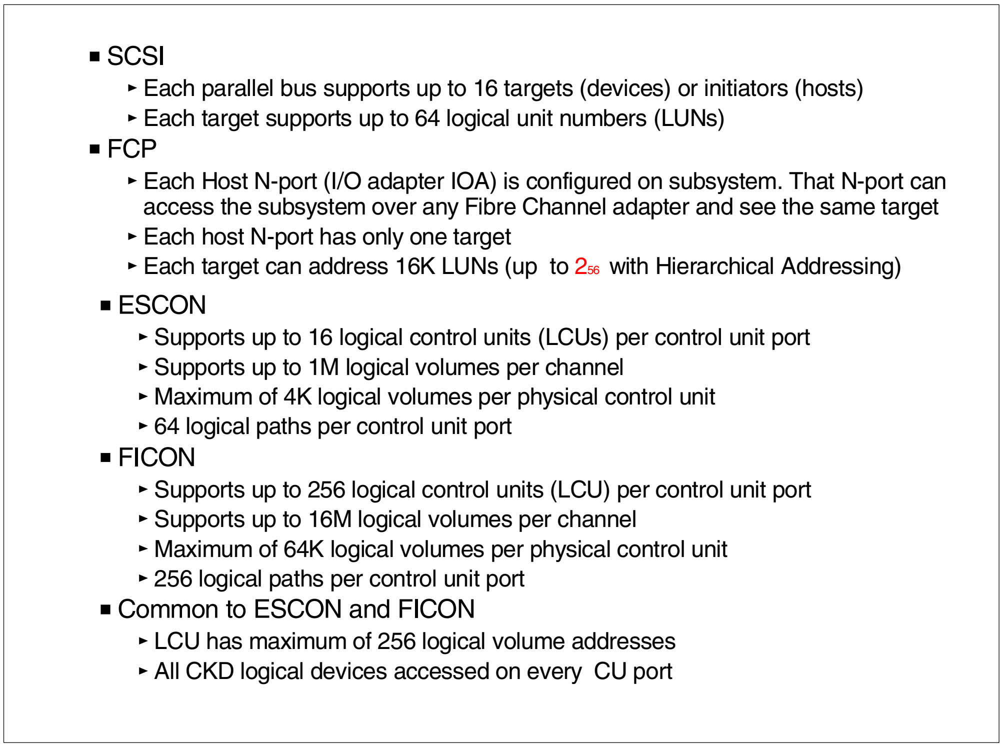
Each SCSI bus/target/LUN combination or each Fibre Channel device adapter/LUN combination is associated with one *logical device* (LD), each of which can be in only one logical subsystem. Another LUN can also be associated with the same logical device, providing the ability to share devices within systems or between systems.

3.6.3 CKD mapping

For zSeries servers every logical subsystem (LSS) relates directly to a z/ Architecture *logical control unit* (LCU), and each *logical device* (LD) relates to a z/ Architecture *unit address*.

Every ESCON or FICON port can address all 16 logical control units in the Enterprise Storage Server.

3.6.4 Addressing characteristics



- SCSI
  - Each parallel bus supports up to 16 targets (devices) or initiators (hosts)
  - Each target supports up to 64 logical unit numbers (LUNs)
- FCP
  - Each Host N-port (I/O adapter IOA) is configured on subsystem. That N-port can access the subsystem over any Fibre Channel adapter and see the same target
  - Each host N-port has only one target
  - Each target can address 16K LUNs (up to 2_{56} with Hierarchical Addressing)
- ESCON
  - Supports up to 16 logical control units (LCUs) per control unit port
  - Supports up to 1M logical volumes per channel
  - Maximum of 4K logical volumes per physical control unit
  - 64 logical paths per control unit port
- FICON
  - Supports up to 256 logical control units (LCU) per control unit port
  - Supports up to 16M logical volumes per channel
  - Maximum of 64K logical volumes per physical control unit
  - 256 logical paths per control unit port
- Common to ESCON and FICON
  - LCU has maximum of 256 logical volume addresses
  - All CKD logical devices accessed on every CU port

Figure 32. Architecture addressing characteristics

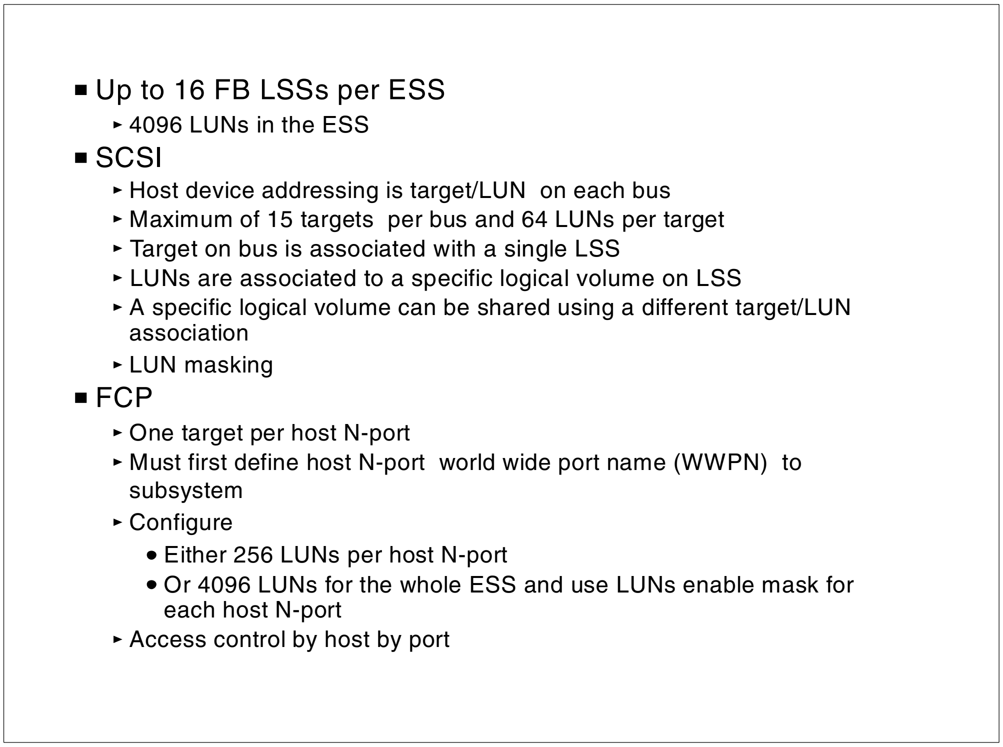
ESCON and SCSI architectures have respectively evolved to FICON and FCP architectures. The Figure 32 lists the characteristics of these architectures.

An architecture is a formal definition that, among other things, allows related components to interface to each other. The hardware and software components that make the systems, implement architectures by following the definitions described in the official documentation. Some extensions to an architecture may not be publicly available, because they are patented, and companies wanting to use them must be licensed and perhaps may have to pay a fee to access them.

Note: When looking at Figure 32 have in mind that it is showing *architectural* characteristics. Remember that the product implementation of an architecture often limits its application to only a subset of that architecture. This is the reason we distinguish between architecture characteristics and implementation. characteristics.

For detailed description of the characteristics of the FICON and the Fibre Channel architectures, and their ESS implementation you may refer to the following documents at the Web site: `http://www.storage.ibm.com/hardsoft/products/ess/support/essficonwp.pdf`, and `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf`.

3.6.5 Fixed-block implementation



- Up to 16 FB LSSs per ESS
  - 4096 LUNs in the ESS
- SCSI
  - Host device addressing is target/LUN on each bus
  - Maximum of 15 targets per bus and 64 LUNs per target
  - Target on bus is associated with a single LSS
  - LUNs are associated to a specific logical volume on LSS
  - A specific logical volume can be shared using a different target/LUN association
  - LUN masking
- FCP
  - One target per host N-port
  - Must first define host N-port world wide port name (WWPN) to subsystem
  - Configure
    - Either 256 LUNs per host N-port
    - Or 4096 LUNs for the whole ESS and use LUNs enable mask for each host N-port
  - Access control by host by port

Figure 33. ESS — FB implementation for SCSI and FCP

3.6.5.1 SCSI mapping

In SCSI attachment, each SCSI bus can attach a combined total of 16 initiators and targets. Since at least one of these attachments must be a host initiator, that leaves a maximum of 15 that can be targets. The ESS presents all 15 targets to its SCSI ports. Also, in SCSI attachment, each target can support up to 64 LUNs. The software in many hosts is only capable of supporting 8 or 32 LUNs per target, no matter the architecture allows for 64. Since the ESS supports 64 LUNs per target, it can support 15 x 64 = 960 LUNs per SCSI port. Each bus/target/ LUN combination is associated with one logical volume, each of which will be mapped in only one LSS. Another target/LUN can also be associated with the same logical volume, providing the ability to share devices within systems or between systems.

3.6.5.2 SCSI LUNs

For SCSI a target ID and all of its LUNs are assigned to one LSS in one cluster. Other target IDs from the same host can be assigned to the same or different FB LSSs. The host adapter directs the I/O to the cluster with the LSS that has the SCSI target defined during the configuration process.

LUN affinity

With SCSI attachment, ESS LUNs have an affinity to the ESS SCSI ports, independent of which hosts may be attached to the ports. Therefore, if multiple hosts are attached to a single SCSI port (The ESS supports up to four hosts per port), then each host will have exactly the same access to all the LUNs available on that port. When the intent is to configure some LUNs to some hosts and other LUNs to other hosts, so that each host is able to access only the LUNs that have been configured to it, then the hosts must be attached to separate SCSI ports. Those LUNs configured to a particular SCSI port are seen by all the hosts attached to that port. The remaining LUNs are “masked” from that port. This is referred to as “LUN Masking”.

3.6.5.3 FCP mapping

In Fibre Channel attachment, each Fibre Channel host adapter can architecturally attach up to 2^{56} LUNs in hierarchical mode. However, the software in some hosts is only capable of supporting only 256 LUNs per adapter.

The maximum number of logical volumes you can associate with any LSS is still 256, and the ESS has 16 FB LSSs. Therefore, the maximum LUNs supported by ESS, across all of its Fibre Channel and SCSI ports, is 16 x 256 = 4096 LUNs.

If the software in the Fibre Channel host supports the SCSI command “Report LUNs”, then it will support 4096 LUNs per adapter; otherwise it only supports 256 LUNs per adapter. The hosts that support “Report LUNs” are the: AIX, OS/400, HP, and NUMA-Q DYNIX/ptx based servers; hence they will support up to 4096 LUNs per adapter. All other host types, including NUMA-Q NT based, will only support 256 LUNs per adapter.

3.6.5.4 FCP LUNs

LUN affinity

In Fibre Channel attachment, LUNs have an affinity to the host's Fibre Channel adapter (via the adapter's world wide unique identifier, a.k.a. the world wide port name), independent of which ESS Fibre Channel port the host is attached to. Therefore, in a switched fabric configuration where a single Fibre Channel host can have access to multiple Fibre Channel ports on the ESS, the set of LUNs which may be accessed by the Fibre Channel host are the same on each of the ESS ports.

A new function introduced in ESS code load EC F25863, available 12/15/00, called “Access Control by Host by Port”, enables the user to restrict a host's access to one or more ports rather than always allowing access to all ports. This new capability significantly increases the configurability of the ESS.

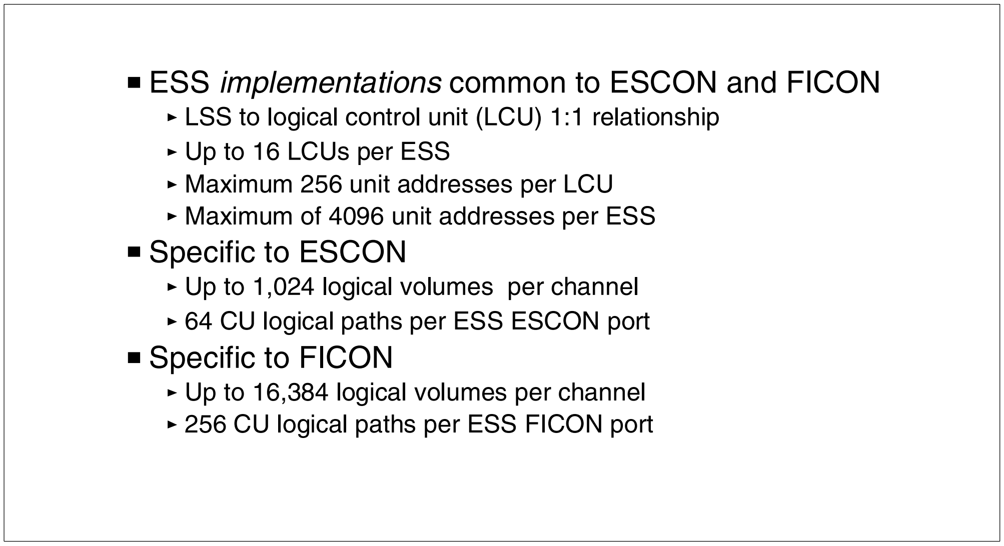
LUN access modes

In Fibre Channel attachment, ESS provides an additional level of access security via either the “Access_Any” mode, or the “Access_Restricted” mode. This is set by the IBM System Support Representative and applies to all Fibre Channel host attachments on the ESS.

In “Access_Any” mode, any host's Fibre Channel adapter for which there has been no access-profile defined can access all LUNs in the ESS (or the first 256 LUNs in the ESS if the host does not have the “Report LUNs” capability). In “Access_Restricted” mode, any host's Fibre Channel adapter for which there has been no access-profile defined, can access none of the LUNs in the ESS. In either access mode, a host's Fibre Channel adapter with an access-profile can see exactly those LUNs defined in the profile, and no others.

You can find a detailed description of LUN affinity and LUN Access modes in the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf`

3.6.6 CKD implementation



- ESS *implementations* common to ESCON and FICON
  - LSS to logical control unit (LCU) 1:1 relationship
  - Up to 16 LCUs per ESS
  - Maximum 256 unit addresses per LCU
  - Maximum of 4096 unit addresses per ESS
- Specific to ESCON
  - Up to 1,024 logical volumes per channel
  - 64 CU logical paths per ESS ESCON port
- Specific to FICON
  - Up to 16,384 logical volumes per channel
  - 256 CU logical paths per ESS FICON port

Figure 34. ESS — CKD implementation for ESCON and FICON

For CKD based servers like the zSeries servers, every LSS relates directly to one *logical control unit* (LCU), and each *logical volume* (LV) to one *unit address* (UA).

Every host channel adapter port effectively addresses all 16 logical control units, similarly for both ESCON and FICON.

The maximum number of LVs you can associate with any LSS is 256, and the ESS supports up to 16 CKD type LSSs. Therefore, the maximum number of unit addresses supported by the ESS, both for ESCON or FICON attachment, is 16 x 256 = 4096 addresses.

FICON specifics

FICON architecture allows 256 LCUs per control unit, but the ESS implementation is16 LCUs per control unit (similar to ESCON). FICON implementation in the ESS also allows 16,384 unit address per channel, and 256 control unit logical paths per port. The ESS FICON implementation supports a total maximum of 2048 logical paths per ESS (128 per LSS).

For more detailed description of the FICON architectural characteristics and its ESS implementation, you may refer to the document in the Web at `http://www.storage.ibm.com/hardsoft/products/ess/support/essficonwp.pdf`

You can define mixed channel path groups with both FICON and ESCON paths accessing the same LCUs, which allows an easy migration from ESCON to FICON.

Note

This path groups mixed configuration should only be considered for migration purposes during a short period of time. It should not be used as a permanent configuration. As soon as possible define homogeneous FICON or ESCON independent channel path groups to your ESS. An ESCON initiated I/O will never reconnect on a FICON link and conversely, RMF merges the performance information, and averages the results. Continued use of both FICON and ESCON paths to an LCU may result in inconsistent performance information being reported. Mixed ESCON/FICON path group capability is intended as a migration convenience only, not as a permanent production solution. See Section 9.7, “Migrating from ESCON to FICON” on page 276.

3.7 Data flow — host adapters

Each host adapter is connected to both clusters

- For SCSI Parallel - connect any port to any cluster
- For FCP - connect any port to any cluster according to access mode
- For CKD - all LCUs are available to every port
- Enables failover should a cluster fail

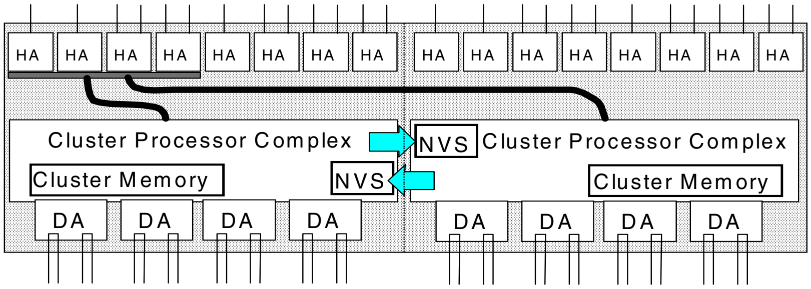


Figure 35. Data flow — host adapters

3.7.1 Each HA is connected to both clusters

The *host adapters* (HA) are the external interfaces to the Enterprise Storage Server.They provide two ESCON or SCSI link ports, or one FICON or FCP port. Each HA plugs into a bus in its bay and the bus is connected to both clusters (black lines in Figure 35).

The host adapters direct the I/O to the correct cluster, based upon the defined configuration for that adapter port.

For an ESCON or FICON port, the connection to both clusters is an active one allowing I/O operations to logical devices in the CKD LCUs in either cluster. The LCUs map directly to the ESS logical subsystems, each LSS being related to a specific LSS loop and cluster.

For SCSI, a target ID and all of its LUNs are assigned to one LSS in one cluster. Other target ID s from the same host can be assigned to the same or different FB LSSs. The host adapter in the ESS directs the I/O to the cluster with the LSS that has the SCSI target defined during the configuration process.

For FCP, any Fibre Channel initiator can access any LUN in the ESS. The LUNs can be associated in a LUN class that is related to the initiator WWPN (for access-restricted mode). Then the HA in the ESS directs the I/O to the cluster that owns the LSS that maps the LUN class that has been accessed. These LUN classes can span LSSs. There are alternatives in the relationship between the ESS host adapter and the LUNs it will finally access. You can find a detailed description of this characteristics in the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf`

3.7.2 Failover

The advantage of having both clusters actively connected to each HA, is that in the case of a failure in one cluster, all I/Os will automatically be directed to the remaining cluster. See 3.24.2, “Failover” on page 85.

3.8 Data flow — read

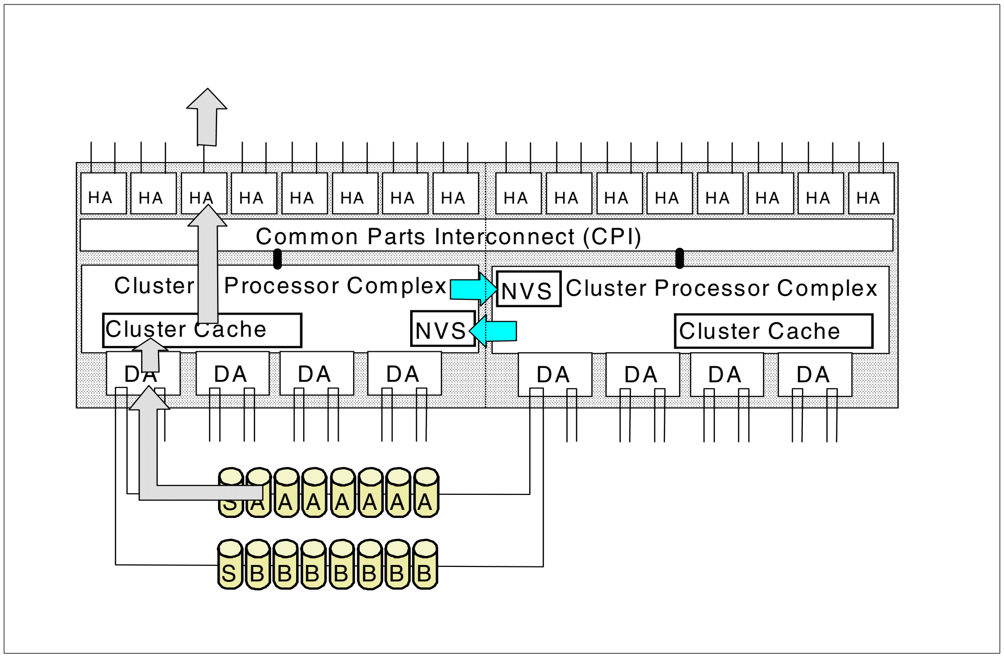

HA
Common Parts Interconnect (CPI)
Cluster Processor Complex
NVS
Cluster Processor Complex
Cluster Cache
NVS
Cluster Cache
DA
S A A A A A A A
S B B B B B B B

Figure 36. Data flow — read

The schematics of Figure 36 shows the structure of the ESS with its two clusters, each with their own cache and NVS. In the following pages on data flow, assume that the description applies to both CKD and FB servers unless mentioned otherwise.

3.8.1 Host adapter

The *host adapter* (HA) accepts the commands from the host and directs them to the appropriate cluster. For ESCON and FICON, each LCU is mapped to one LSS in one cluster, so the command can be directed to the correct cluster. For SCSI, each target is mapped to an LSS in either cluster, so part of the configuration process is to provide the HA with the SCSI target to LSS mapping. For FCP, any Fibre Channel adapter initiator can access any open systems device, so part of the configuration process is to provide the HA with the LUN class association to the WWPN of the server adapter (access-restricted mode). There are alternatives in the relationship between the ESS host adapter and the LUNs it will finally access (and consequently the cluster that will be used). See 3.6.5.2, “SCSI LUNs” on page 56, and 3.6.5.4, “FCP LUNs” on page 57.

3.8.2 Cluster Processor Complex (CPC)

The CPC processes the commands. If the data is in cache then the cache-to-host transfer takes place. If the data is not in the cache, then a staging request is sent to the device adapter (DA) to fetch the requested data.

3.8.3 Device adapter

The *device adapter* (DA) is the SSA adapter for the loop that requests the data blocks from the disk drives in the rank. SSA can multiplex multiple request so that the disk drives can start searching and reading the requested data at the same time. The SSA adapter has buffers that it uses for holding recently used data, primary for RAID-5 operations.

3.8.4 Disk drives

The disk drive(s) in the rank will read the requested data into their buffers and continue to read the rest of a 64K buffer contained in the individual adapter. Once in the buffer, data can be transferred at 40MB/sec. to the DA and the cache. Subsequent reads of data from the same track will find it already in the disk drive buffer, and it will be transferred without seek or latency delays.

3.9 Data flow — write

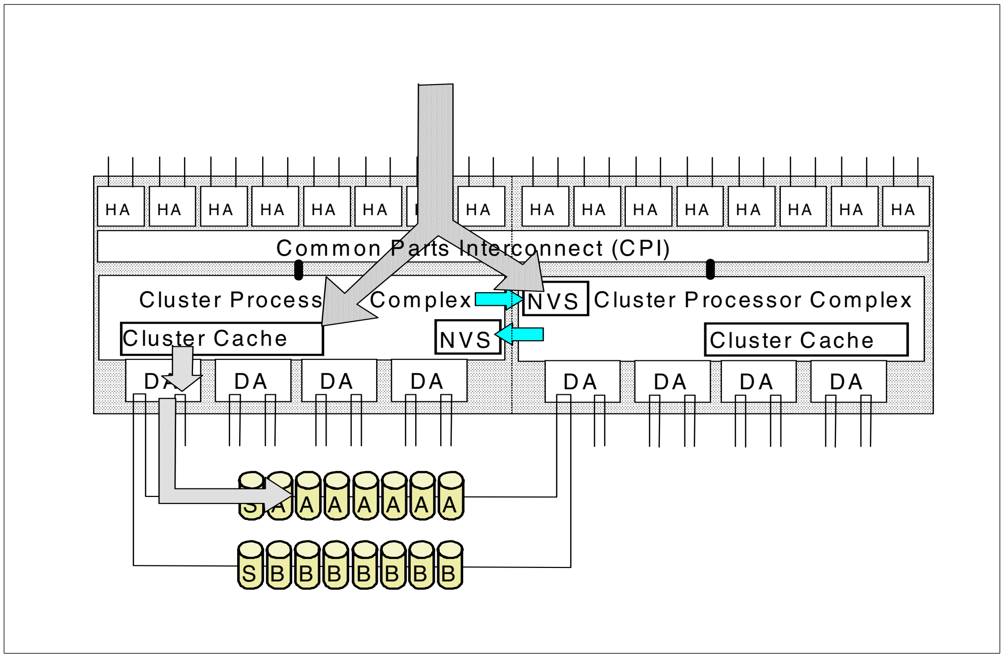


Figure 37. Data flow — write

3.9.1 Host adapter

The *host adapter* (HA) accepts the commands from the host and routes them to the correct cluster. For most write operations, data is already resident in cache from a previous operation, so the update is written to the NVS and cache. The I/O completes once the data is in NVS and cache.

3.9.2 Cluster processor complex (CPC)

The cache copy of data will remain in cache of the same *cluster processor complex* (CPC) until the LRU algorithm of the cache (of this CPC) or NVS (of the other CPC) determines that space is needed, and the data is scheduled to be destaged. All modified data for the same track is sent to the device adapter at the same time to maximize the destage efficiency.

3.9.3 Device adapter

The *device adapters* (DA) contain an SSA adapter which manages the two loops. The SSA adapter also manages the RAID-5 operation. If we are performing an update write to several blocks on a track, the data track and the parity must first be read into the SSA adapter RAM, and the updates made, the parity re-calculated and the data and new parity written back to the two disks. In the Enterprise Storage Server all the RAID-5 parity handling is done by the SSA adapter.

3.10 Cache and read operations

Cache

- 4 to 16 GB / cluster
- Managed in 4 KB segments
  - 3380= 12 segments/track
  - 3390=14 segments/track
  - SCSI = 9 segments/track
- Same adaptive staging for FB and CKD
  - Record (block)
  - Partial track
  - Full track staging

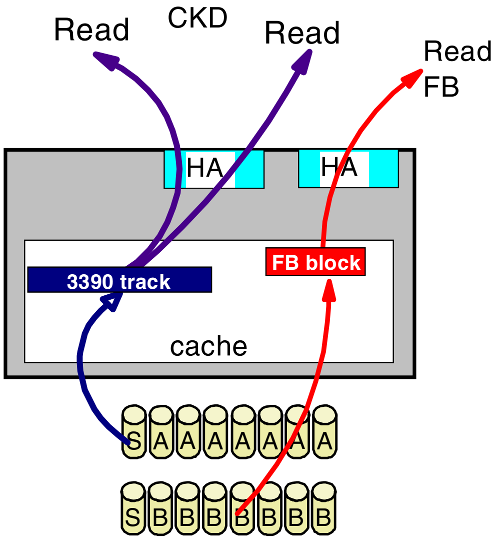


Figure 38. Cache and read operations

3.10.1 Cache

The cache in the ESS is split between the clusters and is not shared. As Figure 38 shows, each cluster has from 4 to 16 GB of read (and write) cache. The cache is managed in 4 KB segments, a full track of data in 3380 track format taking 12 segments, and a full track in 3390 track format taking 14 segments. The small size allows efficient utilization of the cache, even with small records and blocks operating in record mode. For FB mode, a track is up to 9 segments.

3.10.1.1 Read operations

Read operations on the ESS in both CKD and FB mode are similar to the operations of the IBM 3990 Storage Control.

A read operation sent to the cluster processor complex (CPC) will result in:

- A cache hit if the requested data resides in the cache. In this case the I/O operation will not disconnect from the channel/bus until the read is complete. Highest performance is achieved from read hits.
- A cache miss occurs if the data is not in the cache. The ESS I/O is logically disconnected from the host, allowing other I/Os to take place over the same interface,and a stage operation from the RAID rank takes place. The stage operation can be one of three types:
  - Record or block staging

    Only the requested record or blocks are staged into the cache.
  - Partial track staging

    All the records or blocks on the track from the requested record until the end of the track are staged.

- Full track stage

  The entire track is staged into the cache.

The method selected by the ESS to stage data is determined by the data access patterns. Statistics are held in the ESS on each *zone*. A zone is a contiguous area of 128 cylinders or 1920 32-KB tracks. The statistics gathered on each zone determine which of the three cache operations is used for a specific track.

- Data accessed randomly will tend to use the record access or block mode of staging.
- Data that is accessed normally with some locality of reference will use partial track mode staging. This is the default mode.
- Data that is not a regular format, or where the history of access indicates that a full stage is required, will set the full track mode.

The adaptive caching mode data is stored on disk and is reloaded at IML

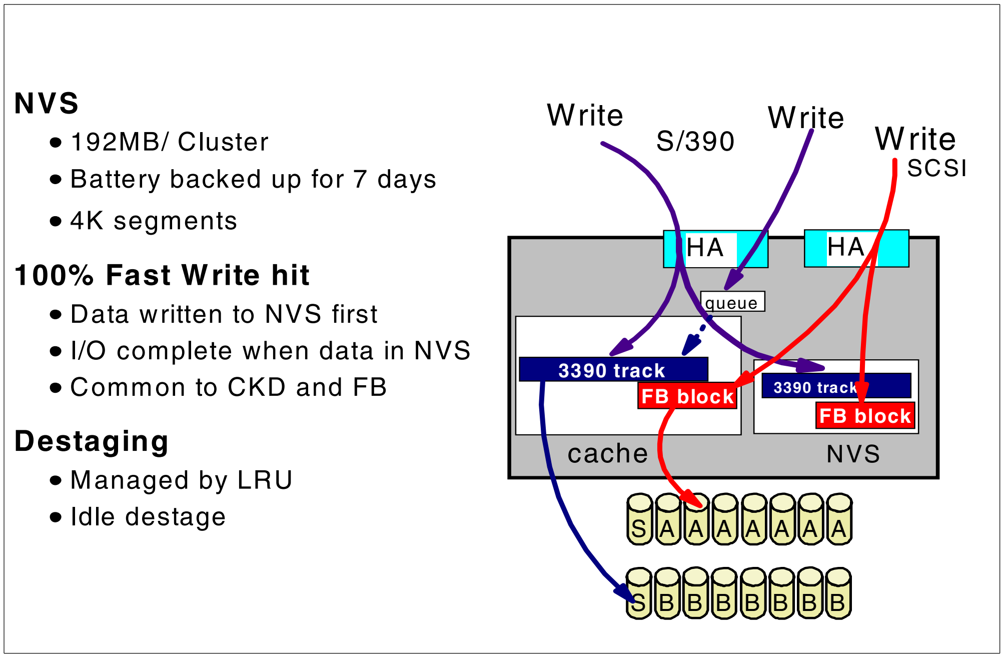
3.10.1.2 Staging

Cache space is released through the use of a Least-Recently-Used (LRU) algorithm. Space in the cache used by sequential data is freed up quicker than normal cache or record data. Use of Inhibit Cache Load and Bypass Cache, will also cause the tracks/records to be freed quickly.

The ESS will continue to pre-stage sequential tracks when the last few tracks in a sequential staging group are accessed.

Stage requests can be performed by the RAID array in parallel for sequential operations, giving the ESS its high sequential throughput characteristic. Parallel operations can take place because the logical data tracks are striped across the physical data disks in the RAID array.

3.11 NVS and write operations



NVS

- 192MB/ Cluster
- Battery backed up for 7 days
- 4K segments

100% Fast Write hit

- Data written to NVS first
- I/O complete when data in NVS
- Common to CKD and FB

Destaging

- Managed by LRU
- Idle destage

Figure 39. NVS and write operations

At any moment there are always two secured copies of any update into the ESS. See Figure 39.

3.11.1 NVS

The NVS size is 192MB per cluster. The NVS is protected by a battery that must be operational and charged for the NVS to be used. The battery will power the NVS for up to 7 days following a total power failure.

3.11.2 Write operations

Data written to an ESS is almost 100% *fast write* hits. A fast write hit occurs when the write I/O operation completes as soon as the data is in the ESS cache or NVS. The benefit of this is very fast write operations. This applies to both CKD and FB I/O operations.

3.11.3 Fast write

Data received by the host adapter is transferred first to the NVS and a copy held in the host adapter buffer. The host is notified that the I/O operation is complete as soon as the data is in NVS. The host adapter, once the NVS transfer is complete, then transfers the data to the cache.

The data remains in the cache and NVS until it is destaged. Destage is triggered by cache and NVS usage thresholds.

This requires that the zSeries host indicate DFW or CFW in the Define Extent command.

3.11.4 NVS LRU

NVS is managed by a *Least Recently Used* (LRU) algorithm. The Enterprise Storage Server attempts to keep free space in the NVS by anticipatory destaging of tracks when the space used in NVS exceeds a threshold. In addition, if the ESS is idle for any period of time, an idle destage function will destage tracks until, after about 5 minutes, all tracks will be destaged.

Both cache and NVS operate on LRU lists. Typically space in the cache and NVS occupied by sequential data and Inhibit Cache Write or Bypass Cache data is freed faster than space occupied by data that is likely to be re referenced.

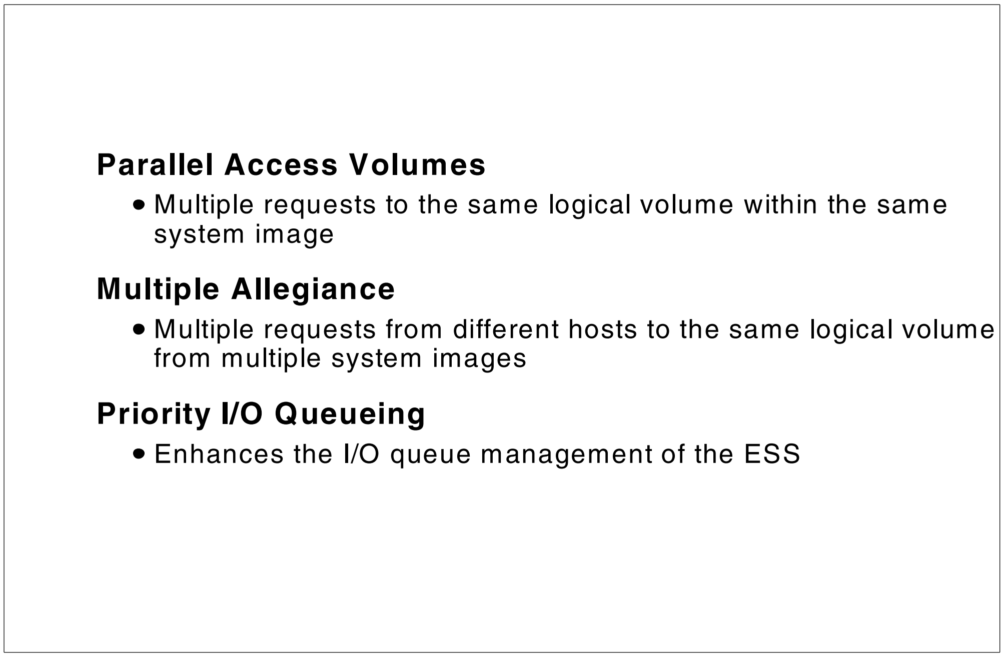
When destaging tracks, the ESS attempts to destage all the tracks that would make up a RAID stripe, minimizing the RAID-5 parity handling operation in the SSA adapter.

3.11.5 NVS location

NVS for cluster 1 is located in cluster 2, and the NVS for cluster 2 is located in cluster 1. This ensures that we always have one good copy of data, should we have a failure in one cluster.

Section 3.24, “Cluster operation: failover/failback” on page 84 discusses failover in more detail.

3.12 S/390 and z/OS I/O accelerators



Parallel Access Volumes

- Multiple requests to the same logical volume within the same system image

Multiple Allegiance

- Multiple requests from different hosts to the same logical volume from multiple system images

Priority I/O Queueing

- Enhances the I/O queue management of the ESS

Figure 40. S/390 and z/OS I/O accelerators

For the zSeries servers, the Enterprise Storage Server provides some specific performance features that we briefly present here. These features are explained in detail in Chapter 5, “Performance” on page 147.

3.12.1 Parallel Access Volumes (PAV)

Parallel Access Volumes (PAV) allows the host system to access the same logical volume using alternative device address UCBs (operating system unit control blocks). There are two types of PAVs, *base* unit addresses and *alias* unit addresses. The base represents the real device and the aliases represent an alternate access. Multiple read requests for the same track in cache will be read hits and will provide excellent performance. Write operations will serialize on the write extents and prevent any other PAV address from accessing these extents until the write I/O completes. As almost all writes are cache hits, there will be only a short delay. Other read requests to different extents can carry on in parallel.

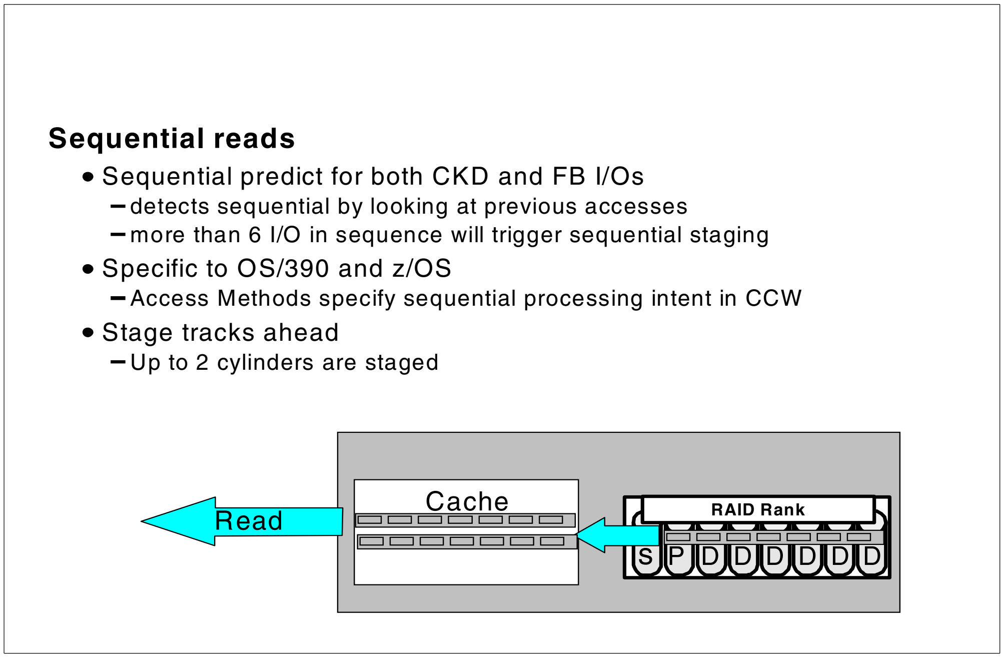
3.12.2 Multiple Allegiance

Multiple Allegiance (MA) allows multiple requests, each from multiple hosts to the same logical volume. Each read request can operate concurrently if data is in the cache, but may queue if access is required to the same physical disk in the array. If you try to access an extent that is part of a write operation, then the request will be queued until the write operation is complete.

3.12.3 I/O priority queuing

I/Os from different z/OS system images can be queued in a priority order. It is z/OS’s Workload Manager with the ESS that can utilize this priority to favor I/Os from one system against the others.

3.13 Sequential operations — read



Sequential reads

- Sequential predict for both CKD and FB I/Os
  - detects sequential by looking at previous accesses
  - more than 6 I/O in sequence will trigger sequential staging
- Specific to OS/390 and z/OS
  - Access Methods specify sequential processing intent in CCW
- Stage tracks ahead
  - Up to 2 cylinders are staged

Figure 41. Sequential operations — read

3.13.1 Sequential detection

The Enterprise Storage Server sequential detection algorithm analyzes sequences of I/Os to determine if data is being accessed sequentially. This algorithm applies equally when accessing CKD data or FB data, although the zSeries servers generally benefit more because of the way they manage the data.

As soon as the algorithm detects that 6 or more tracks have been read in succession, the algorithm triggers a sequential staging process. One area where the new ESS algorithms will detect sequential operations is for the z/Architecture VSAM files. VSAM does not set any sequential mode through software, and its sequential processing often skips areas of the dataset—because, for example, it has imbedded free space on each cylinder.

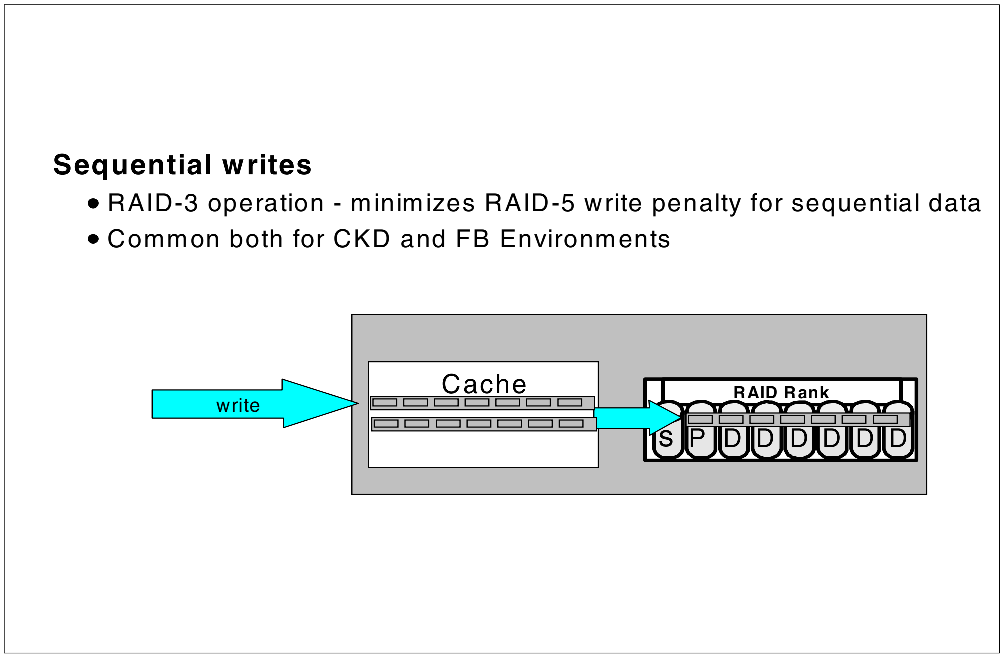
3.13.2 Software setting

The second method of triggering sequential staging in the zSeries environments is specifying through the software sequential access in the channel program.

The sequential staging reads ahead up to 2 cylinders; the actual amount depends on the array configuration, for a 6+P it is 30 tracks and for 7+P it is 28 tracks. As the tracks are read, when we get to about the middle of a staging group, we start staging the next. This delivers maximum sequential throughput with no delays waiting for data to be read from disk.

Tracks that have been read sequentially are eligible to be freed quickly to release the used cache space. This is because sequential data is rarely reread within a short period.

3.14 Sequential operations — write



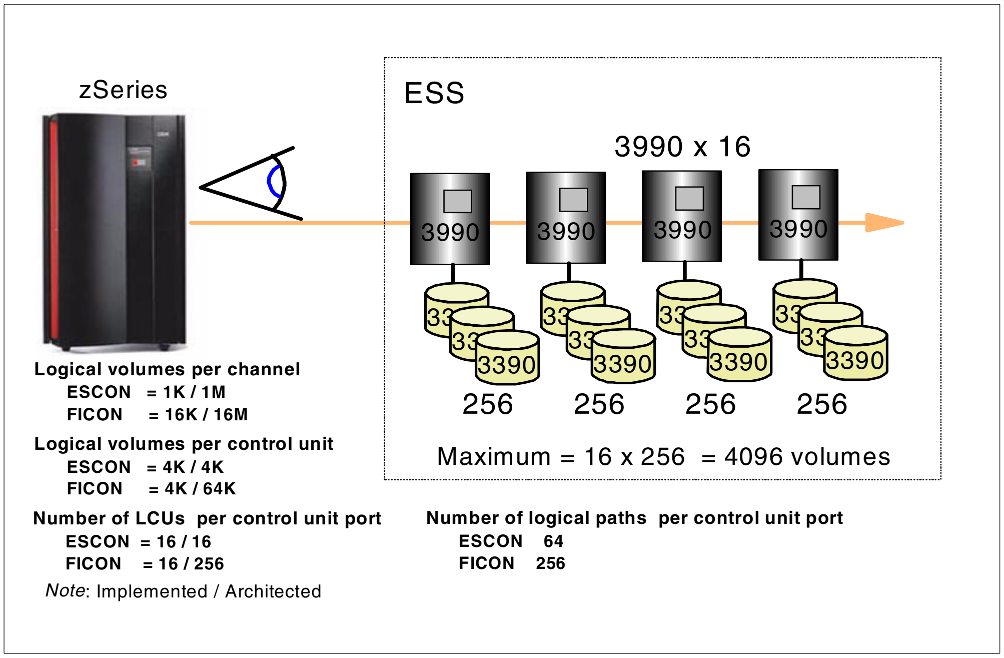
Sequential writes

- RAID-3 operation - minimizes RAID-5 write penalty for sequential data
- Common both for CKD and FB Environments

Figure 42. Sequential operations — write

Sequential write operations on the ESS minimize the RAID-5 write penalty; this is sometimes called RAID-3 mode. An entire stripe of data is written across all the disks in the RAID array, and the parity is generated once for all the data simultaneously and written to the parity disk, that is a rotating parity disk.

3.15 CKD server view of ESS



Logical volumes per channel
ESCON = 1K / 1M
FICON = 16K / 16M

Logical volumes per control unit
ESCON = 4K / 4K
FICON = 4K / 64K

Number of LCUs per control unit port
ESCON = 16 / 16
FICON = 16 / 256

Note: Implemented / Architected

Number of logical paths per control unit port
ESCON 64
FICON 256

Figure 43. CKD servers view of ESS

From the zSeries servers view, both ESCON and FICON attached, an ESS looks like multiple 3990-6 Storage Controls, each with up to 256 volumes. Up to 16 of the 3990s may be defined through HCD using the CUADD parameter to address each LCU. Each LCU is mapped directly to the CKD LSS number. So LSS 0 is mapped to LCU 0 and CUADD 0, and so on for all 16 CKD LSSs.

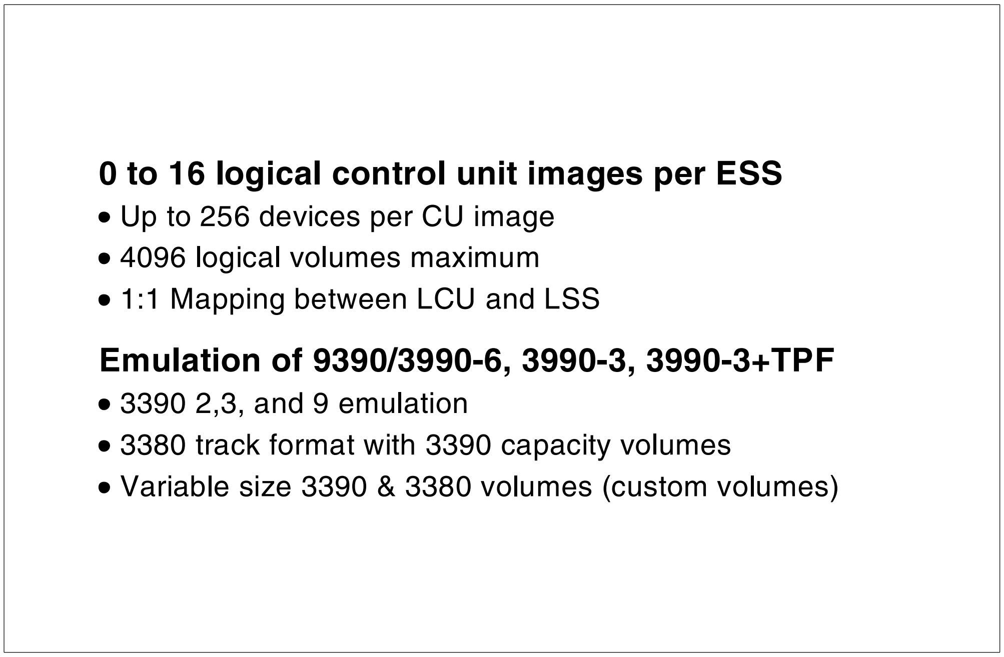
3.15.1 ESCON attached CKD server view of ESS

The zSeries servers support 256 devices per LCU. Every LSS (and therefore LCU) can be addressed by every ESCON link. This means that, in theory, an ESCON channel could see all 16 LCUs, each with 256 devices (a maximum of 4096 devices). However, the ESCON channel hardware implementation limits the number of devices that can be addressed per channel to 1024. This is unlikely to be a restriction for most customers.

3.15.2 FICON attached CKD server view of ESS

For FICON with ESS, each FICON channel also sees all 16 LCUs of the ESS, each with 256 devices (a maximum of 4096 devices). But the FICON hardware implementation extends the number of devices that can be addressed over a channel to 16,384.

3.16 CKD logical subsystem



0 to 16 logical control unit images per ESS

- Up to 256 devices per CU image
- 4096 logical volumes maximum
- 1:1 Mapping between LCU and LSS

Emulation of 9390/3990-6, 3990-3, 3990-3+TPF

- 3390 2,3, and 9 emulation
- 3380 track format with 3390 capacity volumes
- Variable size 3390 & 3380 volumes (custom volumes)

Figure 44. Logical subsystem — CKD

3.16.1 0/8/16 logical control unit images per ESS

When configuring an Enterprise Storage Server, you can specify whether you want 0, 8 or, 16 LCU to be supported. If, for example, you plan to use an ESS for FB type data only, setting the CKD LSS number to zero frees up storage for use as cache.

3.16.2 Emulation of 9390/3990-6, 3990-3, 3990-3+TPF

The Enterprise Storage Server emulates the 9390/3990-6, the 3990-3 and the 3990-3 with TPF LIC. z/OS will recognize the ESS as a 2105 device type when you have the appropriate maintenance applied to your system.

Devices emulated include 3390 Model 2, 3 and 9. You can also define *custom volumes*, volumes whose size varies from a few cylinders to as large as a 3390 Model 9 (or as large as z/OS can support). The selection of the model to be emulated is part of the ESS Specialist configuration process.

The ESS also supports 3380 track format, in a similar way to 3390 Track Compatibility Mode. A 3380 is mapped onto a 3390 volume capacity. So the 3380 track mode devices will have 2226 cylinders on a volume defined with the capacity of a 3390-2, or 3339 cylinders on a volume of 3390-3 size. If you wanted to have volumes that were exactly the same, for example, as a 3380-K, then you could use the custom volume function and define your logical volumes with exactly the same number of cylinders as a 3380-K.

3.17 SCSI server view of ESS

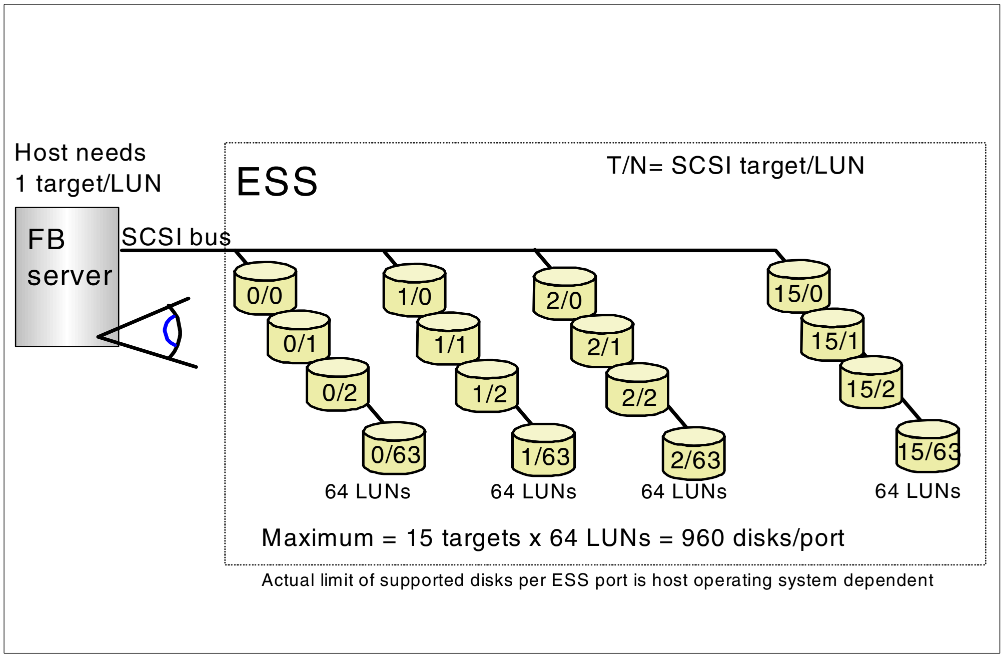
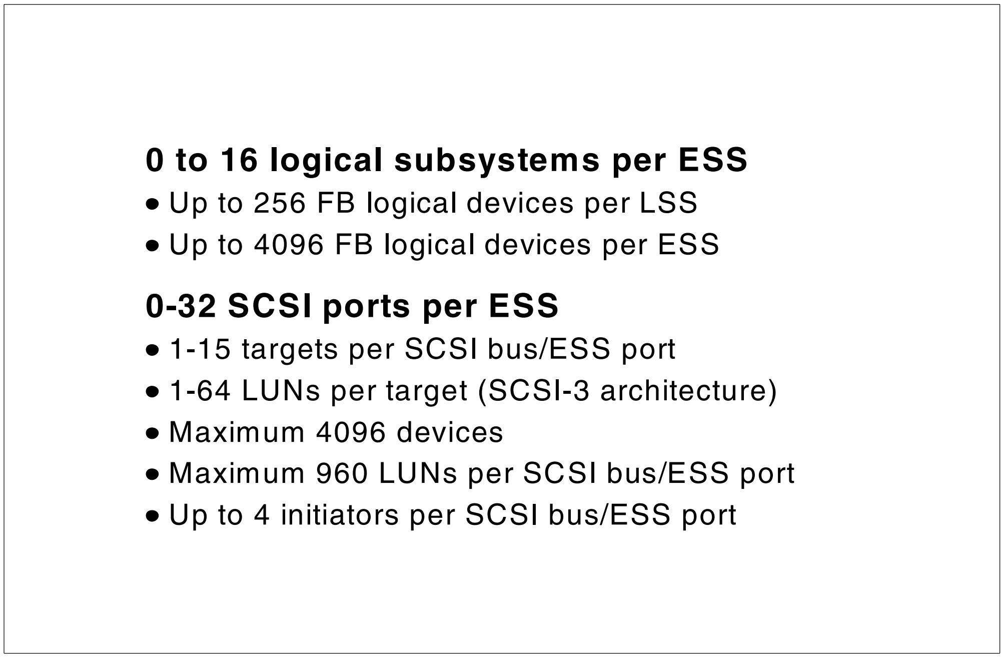


Figure 45. SCSI-parallel view of ESS

If you imagine a SCSI host's view of an Enterprise Storage Server, it looks like a bunch of SCSI disks attached to a SCSI bus. The actual number that any Fixed-Block SCSI attached server system can support is considerably less than the maximum shown here.

One target/LUN is used for each host attached to the ESS SCSI bus, and is commonly designated as initiator ID. Typically, you will only have one host per SCSI bus that is attached to one ESS port, leaving you with 15 target IDs and a number of LUNs per target that varies, depending on the host system's LUN support. Today, this operating system support can range from four to 32 LUNs per target ID.

3.18 FB logical subsystem — SCSI attachment



0 to 16 logical subsystems per ESS

- Up to 256 FB logical devices per LSS
- Up to 4096 FB logical devices per ESS

0-32 SCSI ports per ESS

- 1-15 targets per SCSI bus/ESS port
- 1-64 LUNs per target (SCSI-3 architecture)
- Maximum 4096 devices
- Maximum 960 LUNs per SCSI bus/ESS port
- Up to 4 initiators per SCSI bus/ESS port

Figure 46. Logical subsystem — SCSI

3.18.1 0/8/16 logical subsystems per ESS

When configuring a Enterprise Storage Server, you can specify the maximum number of FB type LSSs you plan to use. If you only have FB type servers connected to your ESS, then you can set the number of CKD type LSSs to zero.

Each FB type LSS supports up to 256 logical volumes. The size of the logical volume within an LSS varying from 100 MB to the maximum size of the rank (for instance 245.5 GB, the size of a RAID rank when having 36.4 GB disks and a 7+P array configuration). A single FB logical subsystem can contain logical volumes from multiple SCSI hosts.

Either 8 or 16 LSSs can be specified. The choice of whether to use 8 or 16 LSSs will be influenced by whether or not you intend to use FlashCopy, and by how many volumes you wish to attach. If 16 LSSs are specified, then it will be possible to create up to 4096 logical volumes. But in most cases 2048 logical volumes in the ESS is more than sufficient for the open systems server you may plan to attach. This is because it is possible (and often desirable) to create a smaller number of large or very large volumes. For example, it is possible to create a single 245GB volume for NT (with 36.4GB disks on a 7+P array) which can also help overcome the limited amount of drive letters available to NT.

The more important consideration then is whether or not you intend to use FlashCopy. FlashCopy is only possible within an LSS and if you choose eight LSSs, it will be much easier to manage existing and/or future flash copy requirements than with sixteen LSSs. For example, if your ESS had eight LSSs specified, with disk capacity in each LSS, then any additional arrays will be added to an existing LSS. If the existing LSS capacity was fully utilized, then the additional arrays could be used for FlashCopy straight away. If your ESS had sixteen LSSs, then additional arrays might cause a new LSS to be used. This is because although you specified sixteen LSSs, the number of LSSs actually in use is dependent upon the installed capacity of the ESS and the number of logical volumes created within the ESS. If new LSSs are used because new arrays are installed, then it will be necessary to move some of your data to those LSSs in order to FlashCopy it. For this reason, it is generally best to configure eight LSSs in the ESS.

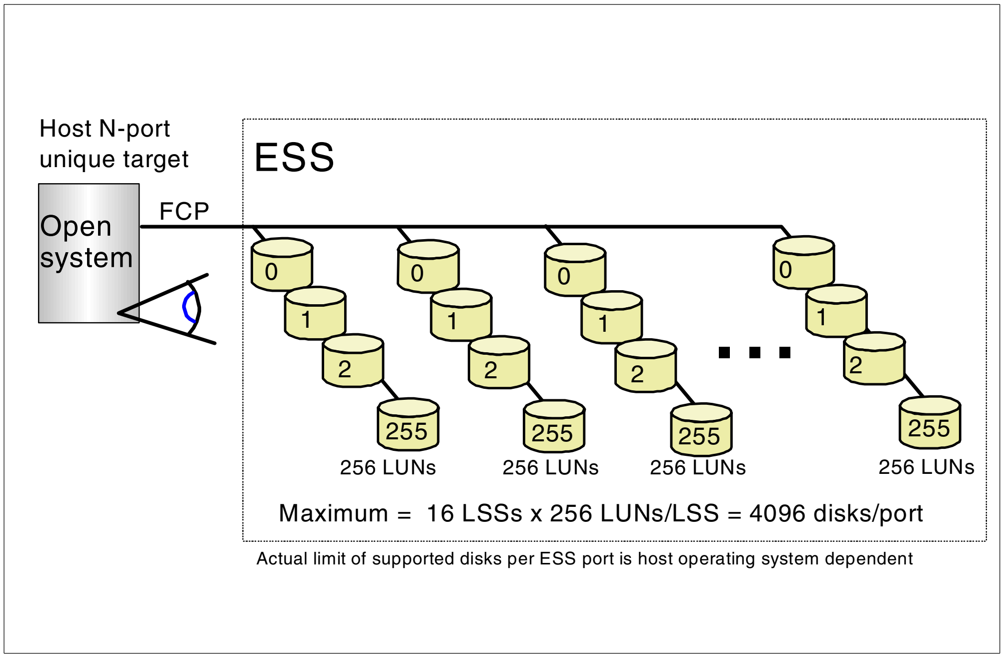
3.18.2 0-32 SCSI ports per ESS

You can install the SCSI host adapters into any of the ESS host adapter bays. Each SCSI card contains two SCSI ports. For a SCSI only ESS, you can fill all the host adapter bays with SCSI cards, giving you a maximum of 16 cards and 32 SCSI ports.

Each SCSI port supports the SCSI-3 standard—16 target SCSI IDs with 64 LUNs per target. This gives a total of 15 x 64 = 960 logical volumes on one SCSI port (only 15 because the host uses one SCSI ID).

You can attach up to four hosts to each ESS SCSI port. See Section 4.11.14, “Defining FB logical devices” on page 126 for details on the number of LUNs supported by different systems.

3.19 Fibre Channel server view of ESS



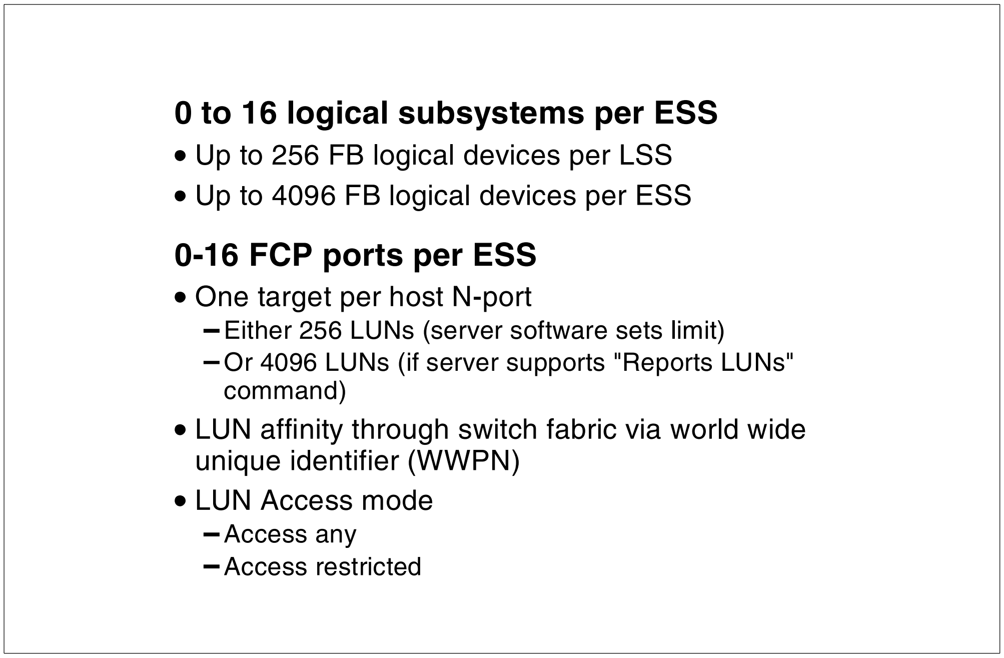
Actual limit of supported disks per ESS port is host operating system dependent

Figure 47. FCP view of ESS

In Fibre Channel attachment, each Fibre Channel host adapter can architecturally attach up to 2^{56} LUNs. However, the software in some hosts is only capable of supporting 256 LUNs per adapter. Also, in an ESS, each LSS supports a maximum of 256 LUNs, and there are 16 LSSs in an ESS. Therefore, the maximum LUNs supported by ESS, across all its SCSI and FCP ports, is 16 x 256 = 4096 LUNs.

For Fibre Channel architecture, where LUNs have affinity to the host's Fibre Channel adapter (via the world wide port name, WWPN), any fibre channel initiator can access any LUN in the ESS. This way each port in the ESS is capable of addressing all the 4096 LUNs in the ESS (should the host software not set a lower limit addressing for the adapter). This 4096 addressing may not be always desired, so there are ways to limit it. You can find more detailed description of this characteristics in the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf`.

3.20 FB logical subsystem — Fibre Channel attachment



0 to 16 logical subsystems per ESS

- Up to 256 FB logical devices per LSS
- Up to 4096 FB logical devices per ESS

0-16 FCP ports per ESS

- One target per host N-port
  - Either 256 LUNs (server software sets limit)
  - Or 4096 LUNs (if server supports "Reports LUNs" command)
- LUN affinity through switch fabric via world wide unique identifier (WWPN)
- LUN Access mode
  - Access any
  - Access restricted

Figure 48. Logical subsystem — FCP

3.20.1 0/8/16 logical subsystems per LSS

The logical subsystems used by the servers that attach using the Fibre Channel architecture are also the FB LSSs. So the implementation characteristics are the same as already described for SCSI attachment (see Section 3.18, “FB logical subsystem — SCSI attachment” on page 75).

3.20.2 0-16 Fibre Channel ports per ESS

Each Fibre Channel / FICON host adapter card has one port. For a Fibre Channel only ESS, you can fill all the host adapter bays with Fibre Channel / FICON host adapter cards, giving you a maximum of 16 cards and also 16 FCP ports.

Note: The host adapters support FICON and FCP, but not both simultaneously. The protocol to be used is configurable on an adapter by adapter basis.

Each Fibre Channel port can address all the maximum 4096 LUNs that can be defined in the ESS. For hosts that support the “Report LUNs” command, this ESS maximum can be addressed by one Fibre Channel adapter (other hosts have a software imposed limit of 256 per adapter). When you configure the LSS you have the option to control the LUNs that a server can access thru a specified port by means of the access-any or access-restricted modes of configuring it.

For more detailed explanation on Fibre Channel attachment please see the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf`.

3.21 iSeries 400

All the architectural considerations presented so far for the open systems servers, whether SCSI or Fibre Channel connected, involve also the iSeries 400 family of IBM @servers. With the Fibre Channel Disk adapter #2766, the iSeries 400 models 820, 830, 840 and 270 running OS/400 Version 5.1 attach via fibre channel to the ESS. Also with the #6501 Disk adapter, the other models of AS/ 400 running OS/400 Version 4.5 or earlier, can connect SCSI to the ESS.

All the considerations about host mapping of the LSS,SCSI addressing, Fibre Channel addressing, FB implementation, SCSI server view of the ESS, and FCP server view of the ESS, include the iSeries 400 servers.

When attaching an iSeries 400 to an ESS, the iSeries will be like any other FB server whether SCSI or Fibre Channel connected. Basically the ESS presents a host of LUNs to the iSeries 400, like it does for any other open systems server. The thing that distinguishes the iSeries 400 is the LUN sizes it will be using: 4.19, 8.59, 17.54, 35.16, 36.00, and 70.56 GB. This way, with a SCSI adapter (#6501) attachment, the LUNs will report into the iSeries as the different models of device type 9337. And with a Fibre Channel Disk adapter (#2766) connection, the LUNs will report into the iSeries as the different models of the 2105 device type. The models will depend on the size of LUNs that have been configured (see Section 4.11.11.3, “Assigning iSeries 400 logical volumes” on page 123).

The other distinguishing characteristics of the iSeries 400 come on an upper layer on top of the preceding architectural characteristics described so far for the FB servers. This is the *Single Level Storage* concept that the iSeries 400 uses. This is a powerful characteristic that makes the iSeries 400 a unique server.

To learn more about the iSeries 400 storage architecture characteristics and how they complement to the ESS own characteristics, you may refer to the redbook *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220.

3.21.1 Single level storage

Both the main memory of the iSeries 400 and the physical disk units, are treated as a very large virtual address space, known as Single Level Storage. This is probably the most significant differentiation of the iSeries 400 when compared to other open systems. As far as applications on the iSeries are concerned, there is really no such thing as a disk unit.

3.21.2 iSeries storage management

iSeries is able to add physical disk storage, and the new disk storage is automatically treated as an extension of the virtual address space. Thus the data is automatically spread across the entire virtual address space. This means that data stripping and load balancing is already automated by the iSeries 400.

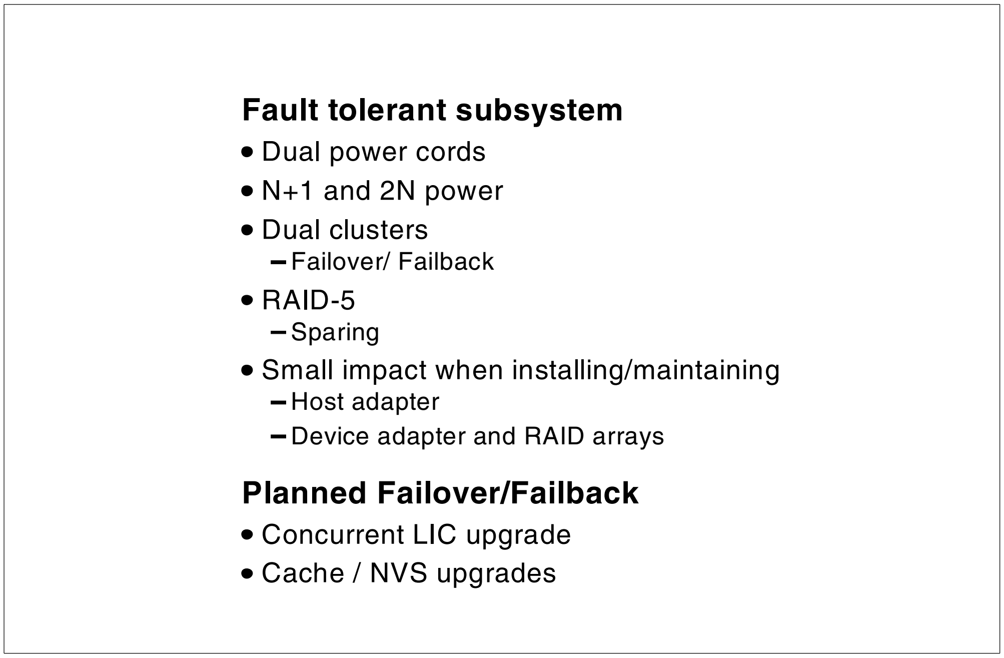
The iSeries keeps the objects in a single address space. The operating system maps portions of this address space as need arises, either to disk units for permanent storage, or to main memory for manipulation by the applications.

Storage management on the iSeries 400 system is automated. The iSeries system takes care of selecting the physical disk (DASD - Direct Access Storage

Device) to store data, spreads the data across the DASDs, and continues to add records to files until specified threshold levels are reached.

All these iSeries 400 storage management characteristics come on top of the ESS own performance characteristics.

3.22 ESS availability features



Fault tolerant subsystem

- Dual power cords
- N+1 and 2N power
- Dual clusters
  - Failover/ Failback
- RAID-5
  - Sparing
- Small impact when installing/maintaining
  - Host adapter
  - Device adapter and RAID arrays

Planned Failover/Failback

- Concurrent LIC upgrade
- Cache / NVS upgrades

Figure 49. ESS availability features

3.22.1 Fault tolerant subsystem

The Enterprise Storage Server has a number of features that make it a fault tolerant subsystem.

It has dual power supplies (2N), each with their own power cord. Each of the two power supplies is capable of powering the whole subsystem.

The DC power supplies are N+1—we have three DC power supplies, two of which are capable of supplying all the DC power.

3.22.2 Planned failover/failback

The ESS consists of two clusters; each cluster is independent and can operate all host connections and access all the disks should the other cluster fail. The failover/failback function is used to handle both unplanned failures and planned upgrades or configuration changes, eliminating most planned outages and providing you with continuous availability.

Within the RAID subsystems, we have spare disk drives, so that in the event of a disk failure, data is rebuilt onto a spare disk drive with no loss in availability.

3.23 Sparing

- Spare capacity must be greater than or equal to array DDM size
- Replacement DDM is new spare

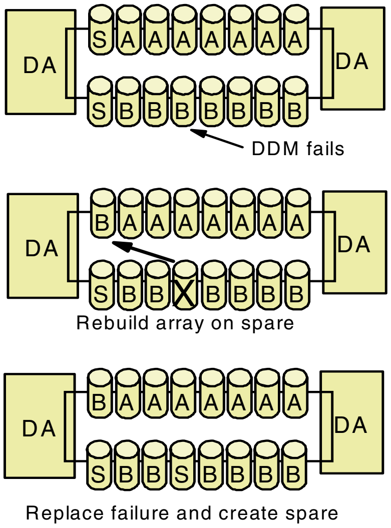


Figure 50. Sparing

3.23.1 Sparing in a RAID rank

In Figure 50, we show how sparing is handled within a RAID rank (or array).

The top diagram illustrates an SSA loop with two RAID ranks (each with a spare disk drive). When a disk drive (DDM) fails, the SSA adapter recreates the missing data by reading the corresponding track on each of the other data disk drives and the parity disk drives, and recalculating the missing data.

The SSA Adapter in the DA will—at the same time as normal I/O access—read the tracks from the data and parity disk drives and rebuild the data from the failed disk drive on one of the spares on the loop.

Once the rebuild has completed, the original spare is now part of the RAID rank, and the failed disk drive becomes the new spare, once it has been replaced.

3.23.2 Spare capacity

As you may have to rebuild any of the disk drives on the loop onto the spare, the spare must be the same size as, or larger than, all the other disk drives in the array. Presently all the configurations available, and any configuration you may build with installed 7133, only use similar disk drives in the loop.

3.23.3 Array spare considerations

The spare in an array can be used by any of the other arrays in the loop. This means that over a period of time the initial correspondence of arrays and the 8-packs that hold them will change and some of the disk drives that conform an array will be on different 8-packs than the ones they were initially.

For this reason, individual 8-packs cannot be removed from a loop without a significant disruption to all the arrays on the loop. You would have to backup all the data on the loop, then delete and re-define all the arrays once the 8-pack had been removed.

3.23.4 Replacement DDM is new spare

Once data has been rebuilt on a spare, it remains there. So the replacement disk drive always becomes the new spare; this minimizes data movement overheads, because there is no requirement to move data back to an original location. Because the spare disk drive “floats” across the arrays, the RAID array will not always map to the same 8 disk drives on which it was initially defined.

3.24 Cluster operation: failover/failback

This section discusses cluster operation concerning failover and failback.

3.24.1 Normal operation before failover

Normal operations of cluster/NVS

- Subsystem A runs in CPC 1 and NVS 1
- Subsystem B runs in CPC 2 and NVS 2
- NVS 1 is physically in Cluster 2
- NVS 2 is physically in Cluster 1

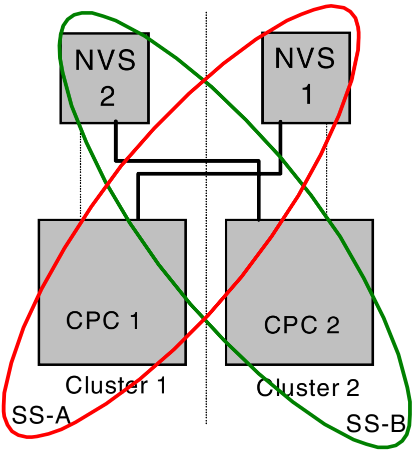


Figure 51. Normal operation of cluster — before failover

The normal setup of the clusters is shown in Figure 51. For the purposes of showing how a cluster failover is handled, we use the following terminology:

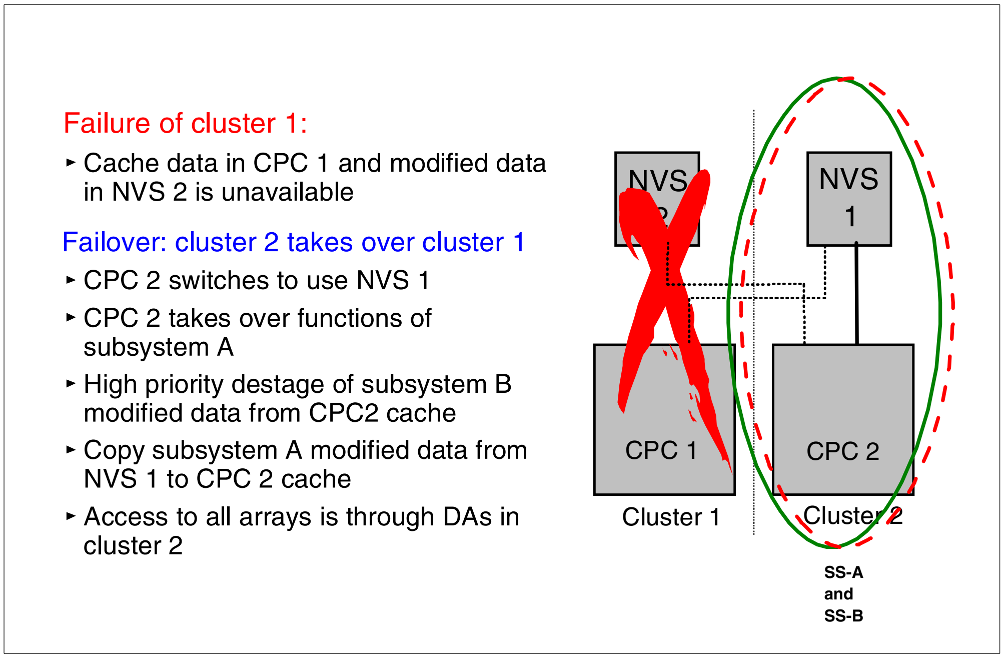
Subsystem A (SS-A): these are functions that normally run in CPC 1 and use NVS 1.

Subsystem B (SS-B): these are functions that normally run in CPC 2 and use NVS 2.

Within an ESS, the two subsystems will be handling different RAID ranks and talking to different host adapters and device adapters. During a failover, the remaining cluster must run both subsystems within the one CPC and NVS.

The host adapters are connected to both clusters, and the device adapters in each cluster can access all the RAID ranks.

3.24.2 Failover



Failure of cluster 1:

- Cache data in CPC 1 and modified data in NVS 2 is unavailable

Failover: cluster 2 takes over cluster 1

- CPC 2 switches to use NVS 1
- CPC 2 takes over functions of subsystem A
- High priority destage of subsystem B modified data from CPC2 cache
- Copy subsystem A modified data from NVS 1 to CPC 2 cache
- Access to all arrays is through DAs in cluster 2

Figure 52. Failover of cluster 1 onto cluster 2

Should the ESS have a failure in one of the clusters, as shown into Figure 52, then the remaining cluster takes over all of its functions. The RAID arrays, because they are connected to both clusters, can be recovered on the remaining device adapters. As we only have one copy of data, any modified data that was in cluster 2 in the diagram is destaged, and any updated data in NVS 1 is also copied into the cluster 2 cache. Cluster 2 can now continue operating using NVS-1.

3.24.3 Failback

- Re-activate failed cluster
- CPC 2 starts using NVS 2
- Transfer subsystem A to CPC 1
- Copy subsystem A modified data from NVS 1 to CPC 1 cache
- Return to normal processing

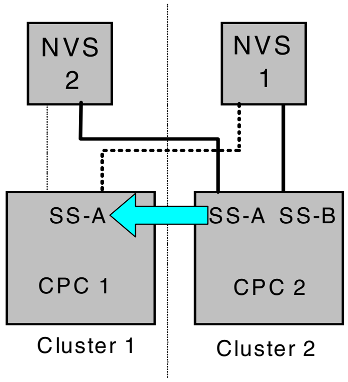


Figure 53. Failback

When the failed cluster has been repaired and restarted, the failback process is activated. CPC2 starts using its own NVS, and the subsystem function SS-A is transferred back to CPC1. Normal operations then resume.

3.25 Maintenance strategy

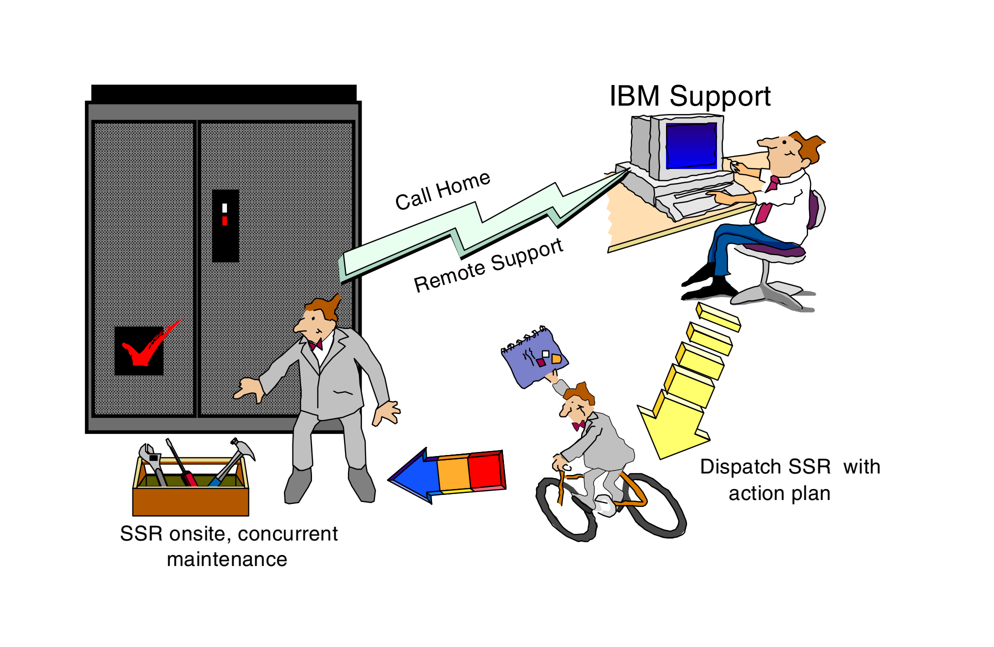

IBM Support
Call Home
Remote Support
Dispatch SSR with action plan
SSR onsite, concurrent maintenance

Figure 54. ESS maintenance strategy

Figure 54 shows the maintenance strategy of an Enterprise Storage Server. You can see that an important part of the maintenance strategy is the capability of the ESS to place a call home in case of failures as well as the possibility of remote support. These two basic aspects are essential for successful, quick, and accurate maintenance.

3.25.1 Call Home and Remote Support

The Call Home feature of the ESS enables it to contact the IBM Support Center directly in case of a failure. The advantage of the Call Home feature is that users will have a 7-day / 24-hour watch-dog of the ESS environment. The IBM Support Center will receive a short report about a failure. With that information, the IBM Support Center will already be able to start analyzing the situation by using several databases for more detailed error information. If required, the IBM Support Center will be able to dial the ESS, in case additional error logs, traces, or configuration information are needed for a more accurate failure analysis.

The capability of dialing the machine also allows the IBM Support Center to help the user with configuration problems, or the restart of a cluster after a failover. The Remote Support capability is a password-protected procedure, which is defined by the user and entered by the IBM System Support Representative (SSR) at installation time.

Note: The routines that are used to support these maintenance procedures will not allow any access to the data residing in the disk drives.

3.25.2 SSR dispatch

After failure analysis using remote support, the IBM Support Center will be able to start an immediate SSR dispatch if the reported problem requires it. The IBM SSR will get an action plan that will most likely solve the situation on-site. That action plan is based on the analysis of the collected error data, additional database searches, and if required, laboratory input. All this occurs without any intervention by the user and helps to solve the raised problem without any big delays, such as phone calls to get support, discussions with on-site people, and so on. Often, the problem diagnosis and fix occurs before the user is aware that a problem has occurred.

3.25.3 Concurrent maintenance

An SSR who is on-site running maintenance at the Enterprise Storage Server will be able to run all maintenance actions concurrent with the customer's operation. This is possible due to the fault tolerant architecture of the ESS. Procedures like Cluster Failover / Failback (see 3.24.2, "Failover" on page 85 and 3.24.3, "Failback" on page 86) will allow a service representative to run service, maintenance, and upgrades concurrently, if configured properly.

3.26 Concurrent logic maintenance

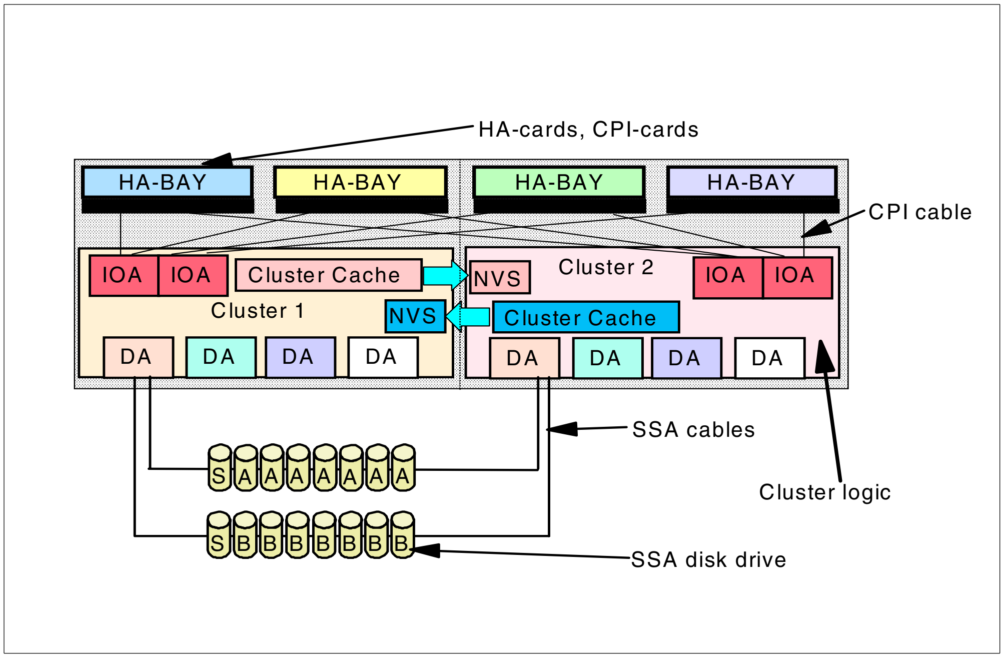


Figure 55. Concurrent logic maintenance

Figure 55 provides details about the logic maintenance boundaries of the ESS. These boundaries allow an IBM System Support Representative to do repairs, maintenance, and upgrades concurrently, without the need to take away the ESS from customer operations.

Concurrent maintenance actions

All logic components are concurrently replaceable. Some of them will even allow hot plugging. The following list indicates the parts that are concurrently replaceable and upgradable:

Cluster logic

All components belonging to the cluster, such as DA cards, IOA cards, cache memory, NVS and others, can be maintained concurrently using the failover/failback procedures. The cluster logic also manages the concurrent LIC load.

SSA disk drives

The SSA disk drives can be maintained concurrently, and because of their design, a replacement is hot-pluggable. This is also valid for SSA cables.

LIC load

The Licensed Internal Code, the control program of the ESS, is designed in such a way that an update to a newer level will take place while the machine is operational using the failover/failback procedure.

Concurrent upgrades

The ESS is upgradable with HA cards, cache size, DA cards, and disk drives. Whenever these upgrades are performed, they will run concurrently. In some cases the failover/failback procedure is used. The upgrade of an HA card will impact other cards on the same HA bay.

3.27 Concurrent power maintenance

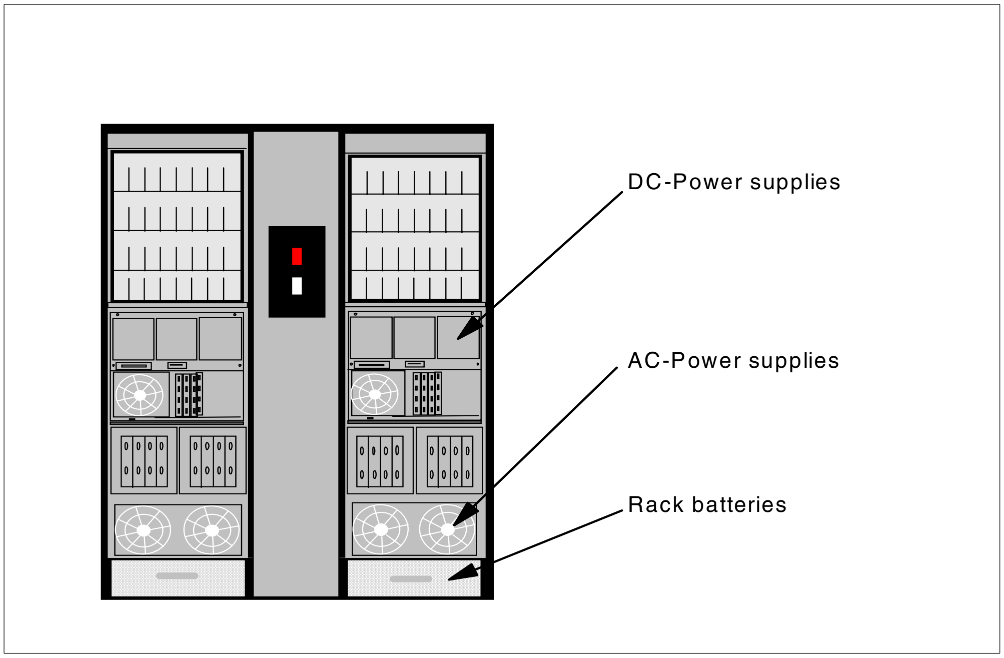


Figure 56. Concurrent power maintenance

Figure 56 shows the main power units. All maintenance actions required in the power area are concurrent, both replacement of failed units, as well as any upgrades. The three power areas in the Enterprise Storage Server are:

DC power supplies
All DC power required in the ESS is provided by an N+1 concept. This will ensure, in case of outage of one of the DC power supplies, that an IBM System Support Representative (SSR) is able to replace the failed part concurrently.

Dual AC distribution
The ESS is a dual AC cord machine, and because of that, it has a 2N concept in AC power. This allows an IBM SSR to replace or upgrade either of the AC supplies.

Rack batteries
Two rack batteries have been integrated in the racks to allow a controlled destage of cache data to the disk drives and a controlled power-down of the rack in case of power loss. The IBM SSR will be able to replace them concurrently.

Chapter 4. Configuration

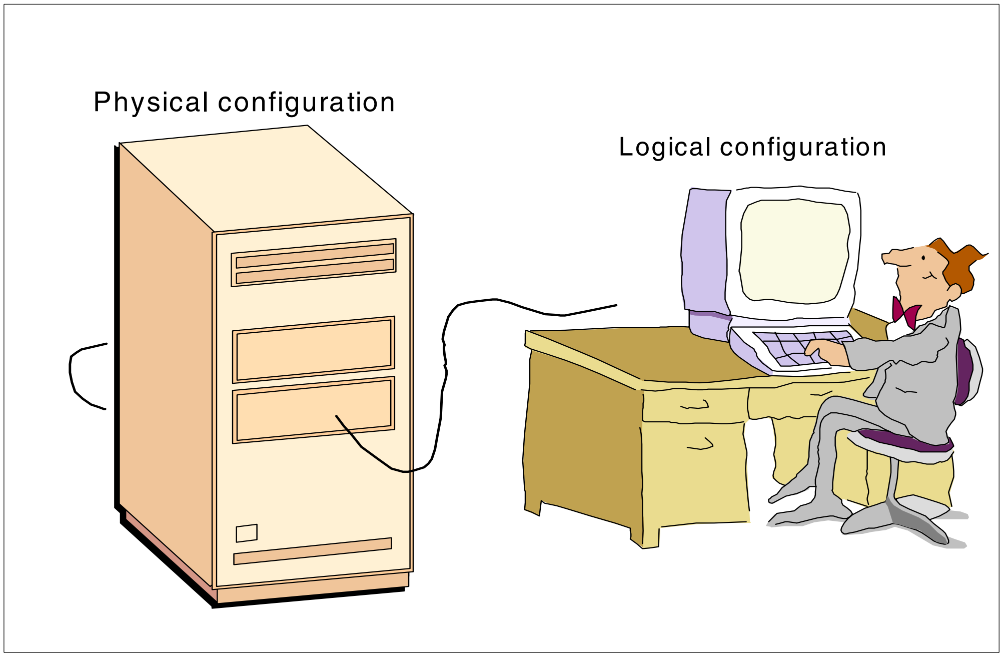


Figure 57. Configuration process

In this chapter we cover the configuration of the Enterprise Storage Server. The configuration is a two step procedure:

1. Physical configuration
2. Logical configuration

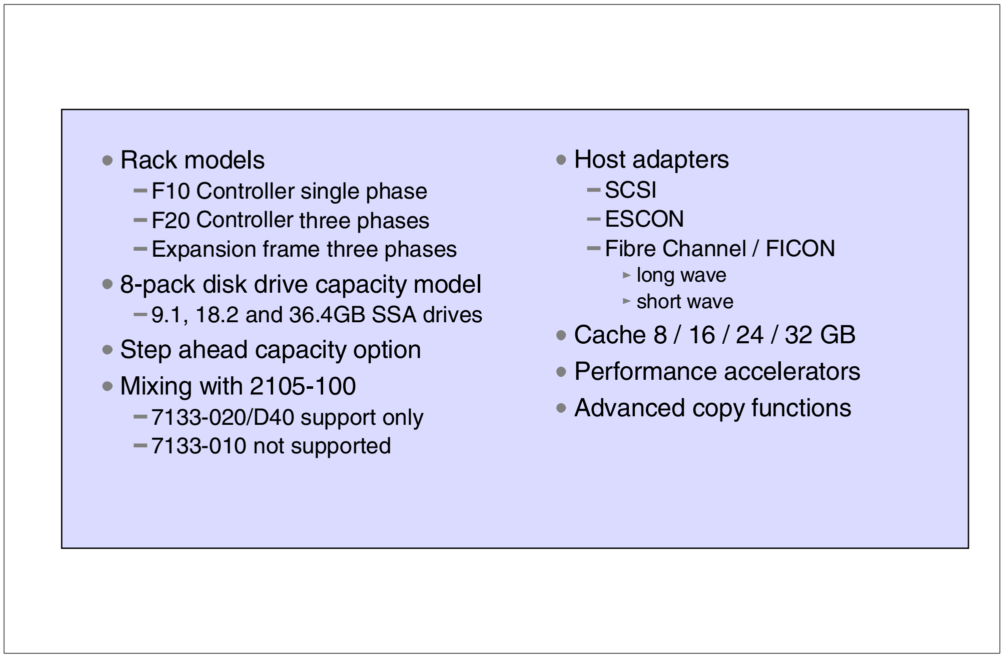
During the physical configuration you select the basic hardware configuration of the Enterprise Storage Server. This is the step where you plan for what features and functions of the ESS will best suit your needs.

During the logical configuration, you do the required definitions within the Enterprise Storage Server for it to work in the particular environment where it is been connected. This is the step where you customize the ESS for your particular servers needs.

The first part of this chapter describes the options you have when doing the physical configuration. These features and functions are also described in great detail in their corresponding chapters in the rest of this book. When dealing with the physical configuration options you may find useful to refer to Appendix E, “Feature codes” on page 305.

Then the remainder of this chapter presents in detail the logical configuration topics. The following sections provide you with the necessary information to be able to configure the Enterprise Storage Server.

4.1 Physical configuration



- Rack models
  - F10 Controller single phase
  - F20 Controller three phases
  - Expansion frame three phases
- 8-pack disk drive capacity model
  - 9.1, 18.2 and 36.4GB SSA drives
- Step ahead capacity option
- Mixing with 2105-100
  - 7133-020/D40 support only
  - 7133-010 not supported
- Host adapters
  - SCSI
  - ESCON
  - Fibre Channel / FICON
    - long wave
    - short wave
- Cache 8 / 16 / 24 / 32 GB
- Performance accelerators
- Advanced copy functions

Figure 58. Physical configuration options

During the physical configuration, you determine the basic setup of the Enterprise Storage Server — by determining the amount of storage capacity required, the rack configuration, power options, and usage of already existing 2105-100 racks and 7133 drawers if any. You also determine the type and quantity of host attachments, which can be ESCON, SCSI, and Fibre Channel / FICON. Also as part of the physical configuration process you determine the desired cache and the advanced functions.

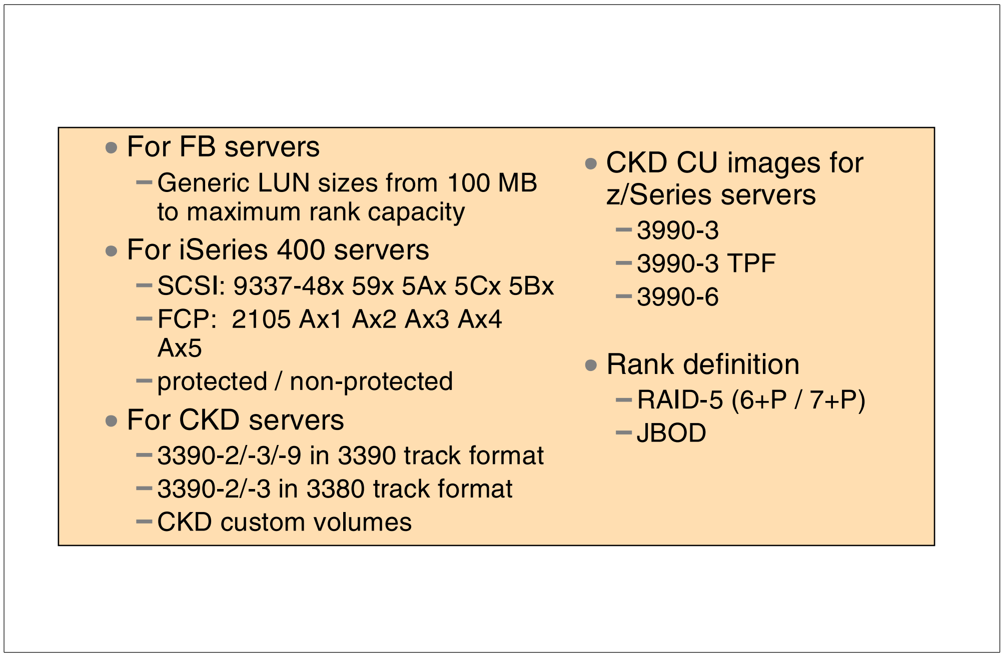
4.1.1 Standard capacity physical configurations

The ESS capacity ranges from 420 GB to 11.2 TB (RAID-5 protected user data). This ample range is available with a set of standard capacity configurations that can be ordered. You can scale easily from one capacity to the other by adding more disks drives.

In the physical configuration process you select the capacity of the disk drives that will come in the 8-packs. You do so when selecting from one of the standard capacity configurations that are available. Currently there are three disk drive models that can be configured with the ESS, with capacities of 9.1 GB, 18.2 GB, or 36.4 GB respectively. You also have the option to order an ESS *step ahead* standard configuration so you will be shipped from the plant the next larger standard configuration of the ESS. This way you will be able to respond immediately to unexpected and urgent increases of demand for additional storage capacity. Please refer to Appendix E, “Feature codes” on page 305 for information on the features and configurations that can be ordered.

Figure 58 shows all the options most of which you should consider when determining the physical configuration.

4.2 Logical configuration



- For FB servers
  - Generic LUN sizes from 100 MB to maximum rank capacity
- For iSeries 400 servers
  - SCSI: 9337-48x 59x 5Ax 5Cx 5Bx
  - FCP: 2105 Ax1 Ax2 Ax3 Ax4 Ax5
  - protected / non-protected
- For CKD servers
  - 3390-2/-3/-9 in 3390 track format
  - 3390-2/-3 in 3380 track format
  - CKD custom volumes
- CKD CU images for z/Series servers
  - 3990-3
  - 3990-3 TPF
  - 3990-6
- Rank definition
  - RAID-5 (6+P / 7+P)
  - JBOD

Figure 59. Logical configuration options

When doing the logical configuration, you make definitions in the Enterprise Storage Server that customize the relationship between the ESS and the attached hosts. You can configure the ESS for the open systems FB servers, either with SCSI or Fibre Channel attachment. You also configure the ESS for the zSeries servers, either with ESCON or FICON attachment. For FB architecture servers the ESS is viewed as generic LUN devices, except for the iSeries 400 servers for which the devices are 9337 or 2105 volumes. For the zSeries servers, the ESS is seen as up to sixteen 3990 Storage Controllers with 3390-2, 3390-3 and 3390-9 DASDs. For these servers, the ESS can also emulate 3380 track formats to be compatible with 3380 devices.

Figure 59 shows some initial options you will be considering when starting with the logical configuration of the ESS. The logical configuration process is done primary using the IBM TotalStorage ESS Specialist. During this procedure you present to the ESS all the required definitions needed to logically set up the ESS and put it operative.

4.2.1 Standard logical configurations

To assist with the configuration process, there is the alternative of specifying some standard formatting options. Once the ESS has been installed, the IBM System Support Representative will format each loop according to the standard configurations you may have selected. This alternative eliminates some of the ESS Specialist steps needed for the logical configuration, shortening the installation procedure. These standard formatting options are available for all, CKD and FB servers. See Section 4.18, “Standard logical configurations” on page 145 for description of these configuration options.

4.3 The 2105 — F10/F20 racks

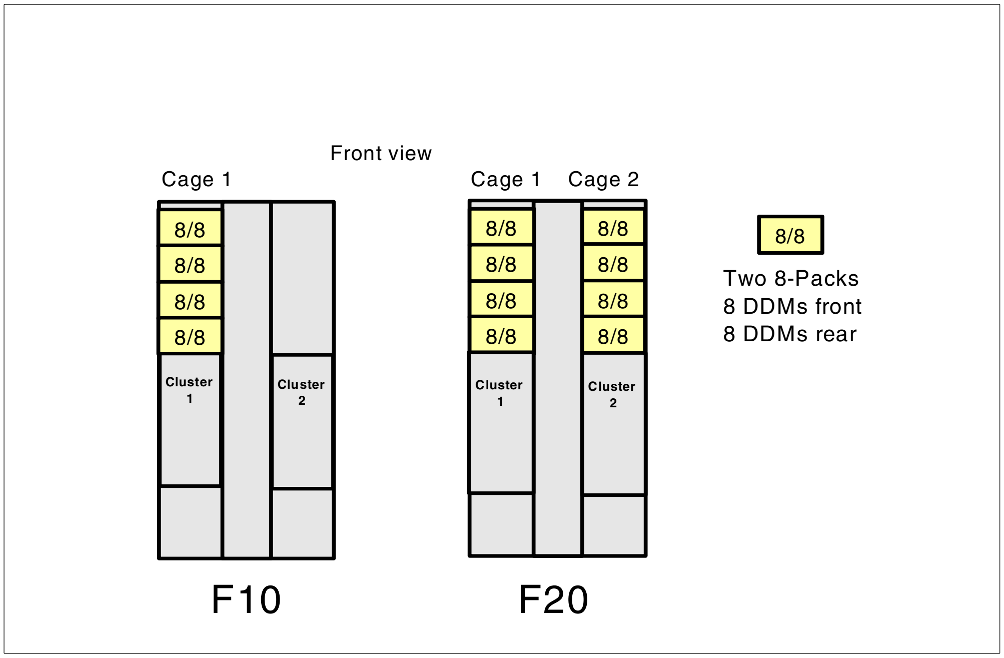


Figure 60. 2105 F10 and F20 racks

Figure 60 shows the Enterprise Storage Server racks. These can be either the 2105-F10 model or the 2105-F20 model. Notice that the F10 model can hold fewer disk drives in the rack because it is a single-phase power machine. The F20 model is a three-phase power machine, and because of that, it can provide the maximum disk drive capacity in the rack. The F10 and the F20 models both carry the disk drives in 8-packs only. The F10 can hold up to 64 disk drives, while the F20 can hold up to 128 disk drives.

The 8-packs are installed in cages. The cages provide the 8-packs with power. The F10 can hold 8-packs only in cage 1, while the F20 can hold 8-packs in cage 1 and in cage 2. The minimum configuration accounts for four 8-packs in cage 1. When upgrading the machines with 8-packs, you will do this from the bottom front of the cage to the top rear. Cage 1 will be filled out first, and then you will continue with Cage 2 in the same order. The upgrades will be done adding 8-pack pairs.

4.4 The 2105 Expansion Enclosure

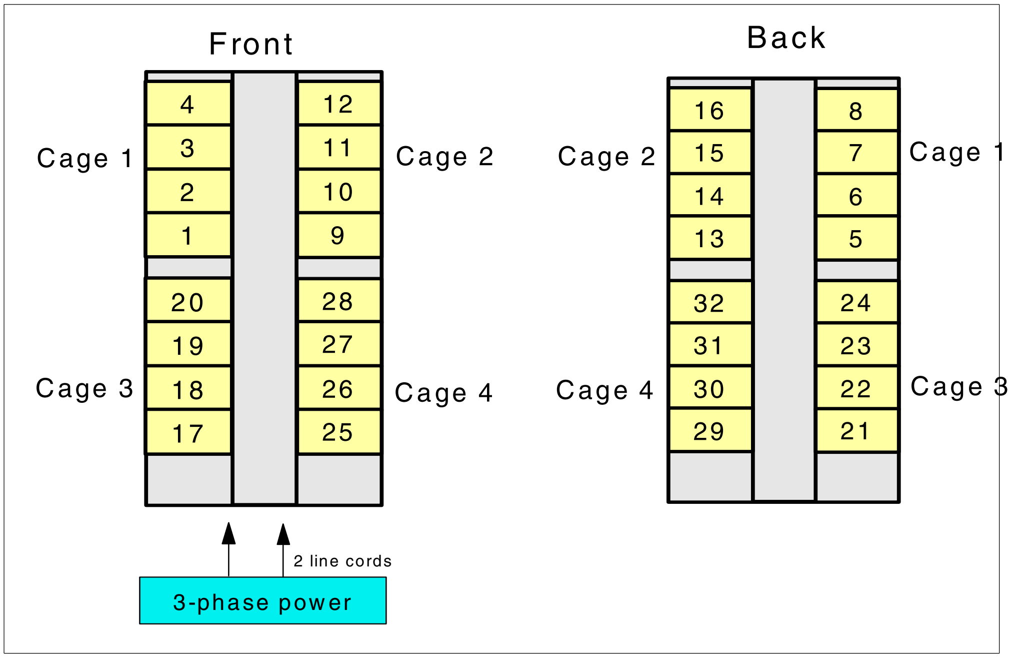


Figure 61. ESS Expansion Enclosure rack

The ESS Expansion Enclosure rack (feature 2100 of the 2105 model F20) can accommodate 4 cages. The same considerations presented for the 8-packs of the base ESS rack apply to the expansion rack. When adding disk drives to the Expansion Enclosure rack, it will be done by installing 8-pack quartets, starting with cage 1. The numbers 1 to 32 in Figure 61 indicate the sequence in which the 8-packs are installed in the cages of the Expansion Enclosure. The first two 8-packs are installed in the lowest positions of the front of cage 1. Then cages are filled-up from bottom to top.

When the 2105-F20 Expansion Enclosure is fully populated with 8-packs, it holds 256 disk drives. The Expansion Enclosure rack is powered by two 3-phase power supplies (two power line cords are used by the Expansion Enclosure rack, additional to the ESS base rack power line cords).

4.5 Mixing with 2105-100 racks

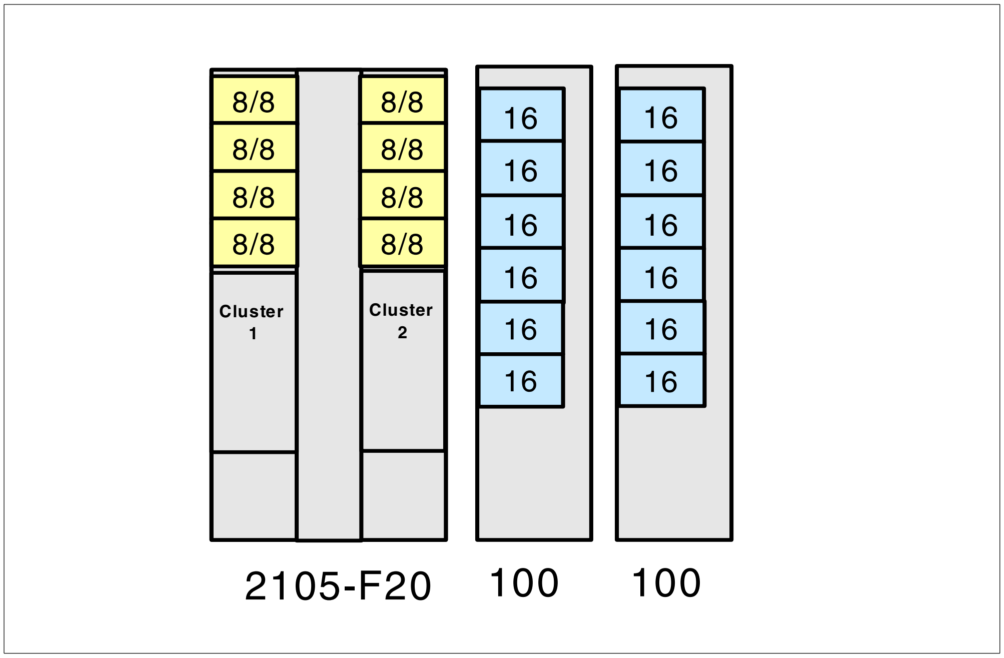


Figure 62. ESS with 2105-100 racks

The Enterprise Storage Server provides investment protection by supporting the attachment of your existing 7133 Serial DIsks. To hold the 7133 drawers, you use the 2105-100 racks. The 7133 drawers must contain a full complement of 16 disks drives of the same capacity to be supported by an ESS. Configurations with less than full drawers must first be upgraded. Note that the capacity of the 7133 disk drives does not have to match the capacity of the disk drives in the base ESS.

Also if you plan to attach 7133 drawers to the ESS, then you must reserve ESS loops for the 7133s. You do so using feature 9904 in the ESS. With feature 9904, loops 3B and 4B are reserved with the first feature 9904 and loops 1B and 2B with the second feature 9904 ordered. The first two 7133 drawers in the loop are configured as 6+P+S arrays while additional 7133s are configured as two 7+P arrays.

See Section 2.6.3, “7133-D40 drawers” on page 25 for details on 7133 drawer configuration and ESS loop considerations.

4.5.1 Maximum configuration support

Up to two 2105-100 racks can be attached to the Enterprise Storage Server. Each 2105-100 rack must include the battery backup power feature 1000. The 2105-100 that attaches directly to the ESS (first in string) must include feature 1121. This feature provides the installation, formatting and checkout of up to six 7133-020 or D40.

A second 2105-100 rack can be attached to the first 2105-100 and must include feature 1122 (similar to 1121, but for second rack). This way, up to twelve 7133 drawers can be attached to the ESS, six in each rack. The ESS cannot have the

Expansion Enclosure rack attached (ESS feature 2100) when 7133 drawers are attached.

Whenever you plan to attach the ESS with the 2105-100 racks, the maximum number of DDMs that can be in the subsystem is 320 - including the 192 DDMs in the 2105-100 racks. The 7133 drawers must not be mixed with 8-packs in a loop. And any SSA loop can hold up to a maximum of 48 disks.

For a full configuration using twelve 7133 drawers, you will be using two 2105-100 racks, each with six fully populated 7133 drawers. You must also reserve four ESS loops (two features 9904). This way each reserved ESS loop will hold the maximum 48 disk drives that an SSA loop can hold. You must remember that feature 9904 is not installed in the field. It can be ordered only when you order a new machine from the manufacturing plant, so you must plan ahead whether you will at sometime use the 7133 Serial DIsks in your ESS configuration.

Note that for an ESS configuration with two 9904 features (loops 1B, 2B, 3B and 4B reserved for 7133 attachment) you can still have up to sixteen 8-packs in the 2 cages of the base ESS. This time the sixteen 8-packs in the base ESS will not be using the 8 loops but only four loops as a consequence of the 9904 feature. This alternative configuration also has the uniqueness of allowing 7+P arrays in the base ESS.

4.5.2 Migration considerations

When attaching existing 7133 Serial Disks to an Enterprise Storage Server, the disk drives of the 7133 drawers must be reformatted to be supported by the ESS. The ESS uses a special 524-byte sector which includes a cyclic redundancy field used to ensure data integrity within the ESS. Disk drives must be reformatted from 512-byte sectors to 524-byte sectors.

For this reason, when attaching your 7133 Serial Disks to the ESS, the existing data in those disk drives must be removed before they are reformatted. Then you can restore the data after the 7133s are installed in the ESS.

4.6 Physical configuration

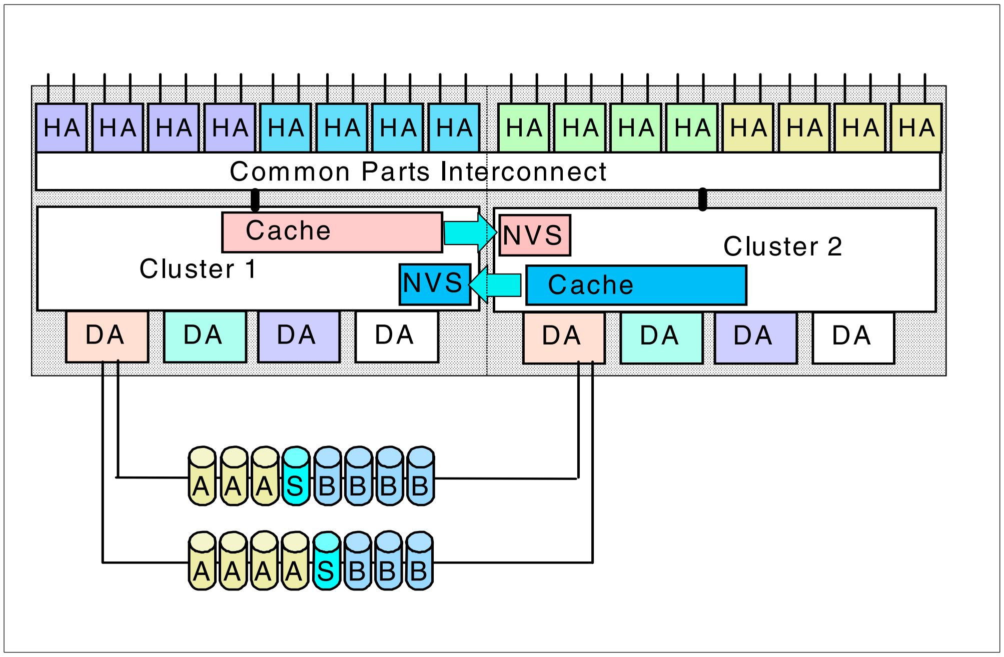


Figure 63. Block diagram of an ESS

Figure 63 illustrates the basic layout of the Enterprise Storage Server architecture showing only one SSA loop. At this time, you will start with the configuration of the device adapter loops and the host adapters.

4.6.1 Host adapters

The host adapters (HAs) are mounted in bays. Each of the 4 bays is able to hold up to 4 host adapters, making a maximum of 16 host adapters for the ESS. Each cluster has 2 host adapter bays installed. The host adapter cards can either be ESCON, SCSI or Fibre Channel / FICON. The Fibre Channel / FICON host adapters can be configured as either Fibre Channel or FICON (one or the other but not simultaneously) on an adapter by adapter basis.

All of the host adapter cards are connected to the clusters by the CPI or Common Parts Interconnect. This allows any of the cards to communicate with either cluster. You can mix ESCON, SCSI, and FICON / Fibre Channel host adapter cards in the ESS. Remember that the total number of host adapter cards, be it ESCON, SCSI or Fibre Channel / FICON, cannot exceed 16. The Fibre Channel / FICON card is a single port host adapter, whereas SCSI and ESCON have two ports for connection.The upgrade of host adapters is done by installing additional HA cards to the bays.

You must specify the type and number of host adapters for the machine you are ordering. Later when installed, you can also add or replace HAs (See Appendix E, “Feature codes” on page 305). Note that the four bays are a standard part of the ESS and come always with the machine independently of the configuration ordered.

The order in which the ESS host adapter cards are installed in the machine during the manufacturing process is:

- Cluster 1- Bay 1 - Adapter 1
- Cluster 2 - Bay 3 - Adapter 1
- Cluster 1 - Bay 2 - Adapter 1
- Cluster 2 - Bay 4 - Adapter 1
- Cluster 1 - Bay 1 - Adapter 2
- Cluster 2 - Bay 3 - Adapter 2
- Cluster 1 - Bay 2 - Adapter 2
- Cluster 2 - Bay 4 - Adapter 2
- Cluster 1 - Bay 1 - Adapter 3
- And so on, until filling the 16 host adapter card positions across the 4 bays

In addition, the ESCON host adapters are installed first, then the SCSI adapters, and finally the Fibre Channel / FICON adapters.

4.6.2 Cache

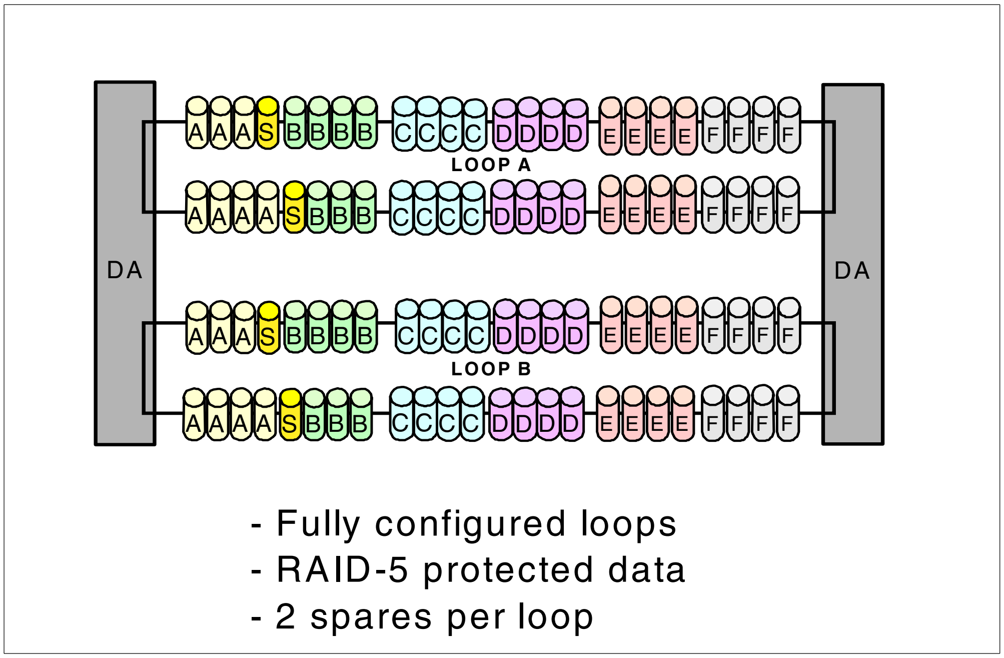
You have the option of selecting any of four cache sizes to accommodate your performance needs: 8, 16, 24, or 32 GB. This total cache capacity, that is split among both clusters, is selected when you order the Enterprise Storage Server from the plant by specifying one of the following ESS features: 4002 (8 GB), 4004 (16 GB), 4005 (24 GB), or 4006 (32 GB).

4.6.3 Device adapters

The device adapter (DA) cards are installed into the cluster logic. There are no bays for the DA cards. The device adapter cards connect in pairs and support two SSA loops. This pairing of DAs not only adds performance, but also provides redundancy. There are four pairs of DAs in the Enterprise Storage Server, each supporting two SSA loops. When a loop holds disk drive capacity, the minimum it can have installed is 16 disk drives (two 8-packs). For capacity upgrades, additional disk drives in groups of 16 (two 8-packs) are installed in a predetermined sequence (As explained in 4.7.1, “2105-F20 upgrade with 8-packs” on page 102). A maximum of 48 disk drives are supported in each SSA loop. A pair of device adapters will always have access to all the disk drives that belong to the SSA loop.

There is no feature specification needed for the device adapters, they are a standard component of the ESS. The eight DAs will come installed four in each cluster, whether you order a small capacity configuration or a large capacity configuration.

4.6.4 DA SSA loop configuration



- Fully configured loops
- RAID-5 protected data
- 2 spares per loop

Figure 64. DA-Pair ESS loop configuration

Figure 64 provides details about the SSA loop configuration in an ESS. Each DA-Pair supports two loops: one is loop A and the other is loop B. The example shows the maximum drive configuration. If you plan to run in RAID-5 mode, then the logical configuration procedure that you run, will leave two spare drives assigned to each loop. These drives are shown in the figure with an "S" and are globally available for any array in that loop. Since every loop running RAID-5 must have 2 spare disks, the first two configured arrays for these loops will contain the 2 spare disks. After this, subsequent arrays added to the loop will not be configured with spares. A loop supports a maximum of 48 disk drives. The disks are added to the ESS loops in groups of sixteen, two disk 8-packs. All drives in the loops are of the same size and the same loop performance: 9.1 GB, 18.2 GB, 36.4 GB, capacity and 40 MB/s loop speed respectively. See Section 2.6.1, “2105 8-pack” on page 24 and Section 2.6.2, “Disk intermixing” on page 25 for additional details on configuring the ESS loop with disk 8-packs.

A loop consists either of 8-packs or 7133 drawers only. The 7133-020 and D40 drawers are supported to protect your investment in existing technology. See “ESS disks” on page 24, and “Mixing with 2105-100 racks” on page 96 for considerations when attaching 7133 disks to your ESS configuration.

4.6.5 Standard configurations

All Enterprise Storage Server configurations you can order, meet the configuration recommendations discussed so far in this section. See Appendix E, “Feature codes” on page 305 for the standard capacity configurations you can order.

4.6.6 Step ahead configurations

The Enterprise Storage Server provides a wide range of step ahead configurations scaling from 420 GB (and the 210 GB Step Ahead) to 10,220 GB (and the 980 GB Step Ahead). Field installed capacity upgrades are enabled for configurations of the same disk drive size (18.2 GB, and 36.4 GB). This capacity upgrade between any configuration of like disk drives can be performed concurrently with normal I/O operations on the ESS. This option allows you to be prepared to immediately respond to unexpected demands of storage capacity. See Appendix E, “Feature codes” on page 305 for the standard configurations.

4.6.7 RAID-5 implementation

For disk drive redundancy, the SSA device adapters support RAID-5 arrays. This option is recommended because of its high availability characteristic. When configuring RAID-5 ranks then groups of up to 8 disk drives are selected by the microcode for that purpose, and a total of 2 disk drives per loop will become spares. Figure 64 illustrates how, initially, the RAID-5 arrays are grouped. Disk drives of the same letter belong to a RAID rank (array).

4.6.8 JBOD implementation

The Enterprise Storage Server also supports JBOD (Just a Bunch Of Disks) configurations. JBODs do not provide disk drive redundancy. If this option is used, the attached host must provide the disk drive redundancy options if the user wishes redundancy for availability. This may mean that the operating system will need to provide that kind of solution, for example with software disk mirroring. JBODs may be of interest also for customers using TPF on z/Series systems. TPF has routines that ensure volume redundancy from the operating system.

4.6.9 JBOD versus RAID-5 considerations

Very often you will hear that open systems users like to use JBODs for performance reasons. This is correct when disk are grouped to logical volumes at operating system level, because data striping may be used for the LVs. Data striping, on the other hand, must be managed by host system software, and once users add mirroring for availability, you must analyze how big the impact to the host system performance will be, because mirroring is done by software tools. In such a case you may end up with a host system that is very busy managing the striping and mirroring. The Enterprise Storage Server, because of its design, will do both tasks at hardware level when using RAID-5, because it stripes the data across several disk drives and provides disk redundancy. The result of using RAID-5 in the ESS, instead of letting it be done by the host system, will be that the host will experience a dramatic reduction in processor load and a gain in processor performance. And on the other hand, this means no penalty at all in terms of performance for the ESS because of its sophisticated design.

4.7 Base and Expansion Enclosure loop configuration

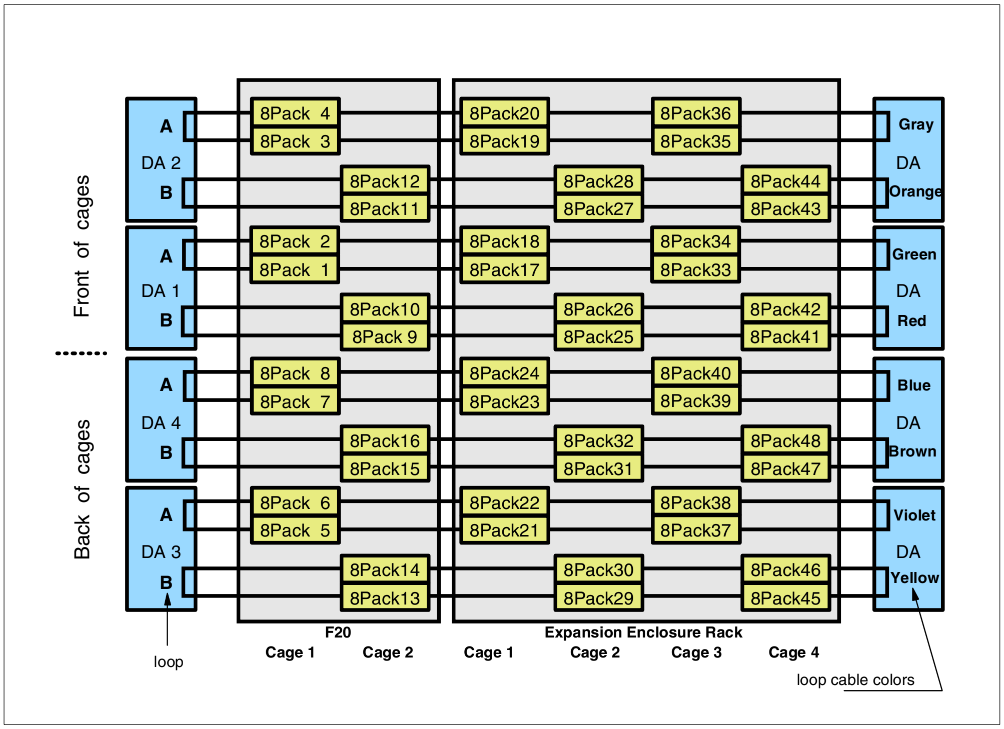


Figure 65. 8-pack configuration — 2105-F20 with Expansion Enclosure

Figure 65 shows a fully configured Enterprise Storage Server 2105-F20 with the Expansion Enclosure rack and the maximum number of 8-packs. As you can see, the 8-packs on the expansion rack are on the same loops that hold the 8-packs on the base ESS.

The minimum standard capacity configuration available has four 8-packs in the front of cage 1 of the base ESS —identified as 1, 2, 3 and 4 in Figure 65.

4.7.1 2105-F20 upgrade with 8-packs

If you plan to upgrade the capacity of the 2105-F20 with more disk drives, you do it by adding pairs of 8-packs. The available standard capacity configurations respond to this criteria (see Appendix E, “Feature codes” on page 305). If the cages are not already present, they will be automatically shipped.

The cages will be filled in sequence. Cage 1 is filled first, bottom up, starting on the front then following to the back. Then cage 2 until completing the base ESS. In Figure 65 you can follow the order of this sequence by the numbers shown on the 8-packs.

4.7.2 Expansion Enclosure upgrade with 8-Packs

In the ESS Expansion Enclosure the 8-packs are also added in pairs to each loop, once the base ESS is complete. In fact, the upgrades for the expansion rack consist of 8-pack quartets, so two pairs of 8-packs are added with each capacity upgrade.

For the ESS expansion rack the sequence followed to add the capacity upgrade is the same as for the base ESS, except that the expansion rack has cages 3 and 4 in addition to 1 and 2. You can also follow the sequence order for the capacity upgrade of the Expansion Enclosure by the numbers on the 8-packs in Figure 65.

4.8 The IBM TotalStorage ESS Specialist

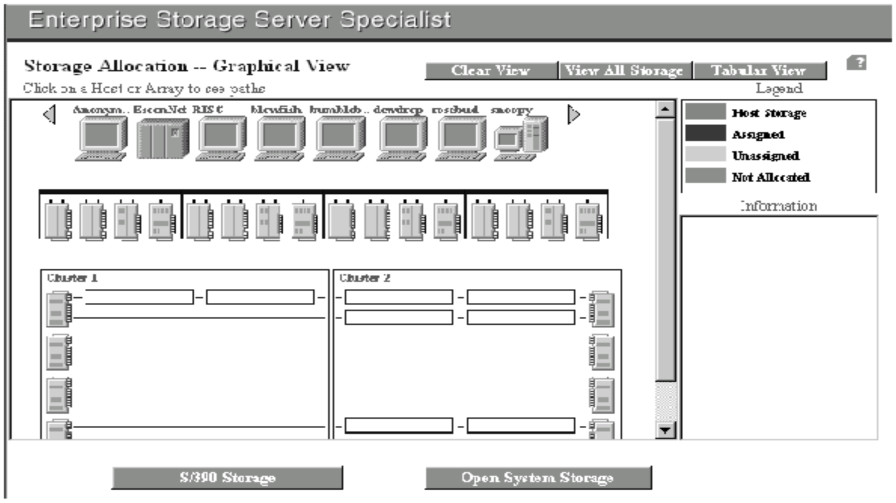


Figure 66. ESS Specialist — Storage Allocation panel

Before explaining the logical configuration, let us see the interface you will be using to do it. The IBM TotalStorage ESS Specialist (ESS Specialist), is the interface that will assist you in most of the logical configuration definitions.

In this section we do a very brief presentation of the ESS Specialist. To learn more about it and how to use it refer to the *IBM Enterprise Storage Server Web Interface Users Guide for the ESS Specialist and ESS Copy Services,* SC26-7346.

4.8.1 ESS Specialist configuration panels

Figure 66 shows the *Storage Allocation* panel of the IBM TotalStorage ESS Specialist. You arrive at this panel by clicking the Storage Allocation button from the initial ESS Specialist welcome panel, that you get once you access to the ESS Specialist from your Web browser.

The *Storage Allocation* panel shows the ESS logical view for the host systems, host adapter ports, device adapters, arrays, and volume assignments. You start the logical configuration process starting from the *Storage Allocation* panel by selecting either the *Open Systems Storage* button or the *S/390 Storage* button, that will take you to their associated panels.

You use the *Open Systems Storage* panel and its buttons, to configure new host system attachments and storage, or to modify existing configurations, for FB type of servers that attach using SCSI or Fibre Channel. From the *S/390 Storage* panel and its buttons, you go to configure the LCUs, their associated volumes, PAVs, and FICON host adapters for the CKD type of servers that attach using ESCON or FICON (ESCON ports do not need to be configured).

4.8.2 Standard logical configurations

Remember that you have the recommended alternative of specifying some standard formatting options for the logical configuration of the Enterprise Storage Server. Choosing these standard options allows the IBM System Support Representative to use his Batch Configuration Tool for making some of the definitions in an automatic way instead of doing them manually with the ESS Specialist panels. No matter if the Batch Configuration Tool is used for some of the configuration steps, the ESS Specialist will still have to be used for many other steps of the configuration procedure.

This Batch Configuration Tool is used only during the initial installation of the ESS. Section 4.18, “Standard logical configurations” on page 145 for a description of this formatting options.

4.9 IBM TotalStorage ESS Specialist setup

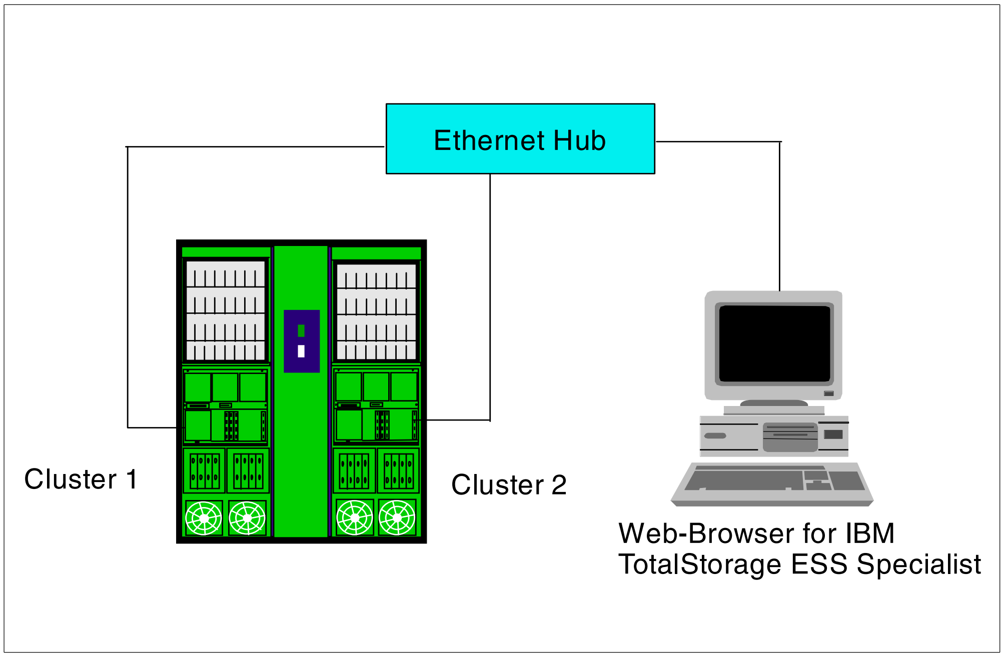


Figure 67. ESS Specialist setup

In order to start using the ESS Specialist for the logical configuration process, you first set up a local-area network connecting to the Enterprise Storage Server, and a Web browser interface to access the ESS Specialist panels.

4.9.1 Web browser

The IBM TotalStorage ESS Specialist provides a Web-based interface connection. This is an easy to use interface. Using an Internet browser, such as Netscape Navigator Professional Edition or Microsoft Internet Explorer, you can access the ESS Specialist from a desktop or mobile computer as supported by the network. The key is that the browser must support Java 1.1.4. (Netscape Navigator 4.04, or later provides this support. Microsoft Internet Explorer 4.00 does too.)

Access to the ESS Specialist requires a valid user name and password. When started for the first time on a newly installed ESS, the ESS Specialist has one administrator user name predefined. Logon with this user name and immediately create one or more administrative users with secure passwords. This is required because the provided user name will automatically be deleted by the ESS Specialist after the first new administrator id has been added to the list and this user name cannot be redefined.

4.9.2 ESSNet

The IBM ESSNet (Enterprise Storage Server Network) is a dedicated local-area network connecting up to seven Enterprise Storage Servers. Included with ESSNet is the ESSNet console, a PC running Windows NT, which provides a management interface to the Enterprise Storage Servers through a Web browser.

Where more than seven Enterprise Storage Servers are used, their ESSNets may be interconnected creating the effect of a single ESSNet and a single point of control for many Enterprise Storage Servers. Alternatively, one or more ESSNets may be connected to an enterprise wide-area network enabling control of Enterprise Storage Servers from a central location, regardless of where they are may be located.

The ESSNet is a self-contained Ethernet LAN. An Ethernet hub is provided as part of ESSNet (two 10BaseT cables, connected to a 10BaseT Ethernet LAN hub). The IBM System Support Representative will attach the LAN cables (one per cluster) to the Enterprise Storage Server.

4.9.3 User local-area network

Attachment to the customers local-area network permits access to the ESS Specialist outside the immediate area of the ESS. But it also has the potential of increasing the number of people who can access, view, administer and configure the ESS.

If you want to attach your LAN to the ESSNet hub, you will need to provide the required TCP/IP information to the IBM System Support Representative (the ESSNet needs to be configured with TCP/IP addresses which are recognized as part of your IP network). The IBM SSR will connect your LAN cable to the ESSNet hub and enter the required TCP/IP information.

The ESS can be installed in a secured LAN, limited access environment. IBM recommends that the ESS network be established so that it is limited to those requiring access to the ESS. It is not recommended that it be installed on the enterprise intranet nor the world wide Internet. Installing ESS behind a *firewall* is one method of insuring ESS security.

4.9.4 Physical setup

The ESSNet comes with feature 2715 of the Enterprise Storage Server. This feature, besides providing the ESSNet PC server and monitor, it also provides the modem, switch, cables and connectors required to attach the first ESS to the telephone system for the remote support function. For two to seven additional ESSs, feature 2716 provides the additional Ethernet cables to attach to the ESSNet server.

Remember that besides the power outlet needed for the IBM System Support Representative MosT terminal to service the Enterprise Storage Server, and the two power outlets for the remote services modem and the modem expander, for the ESSNet setup you will also require:

- Two power outlets for the ESSNet PC and monitor
- One power outlet for the Etherrnet hub

If the ESSNet will attach to the customers LAN, Ethernet cables need to be obtained. No cable is provided to attach the ESSNet to the customers network.

4.10 ESS logical configuration

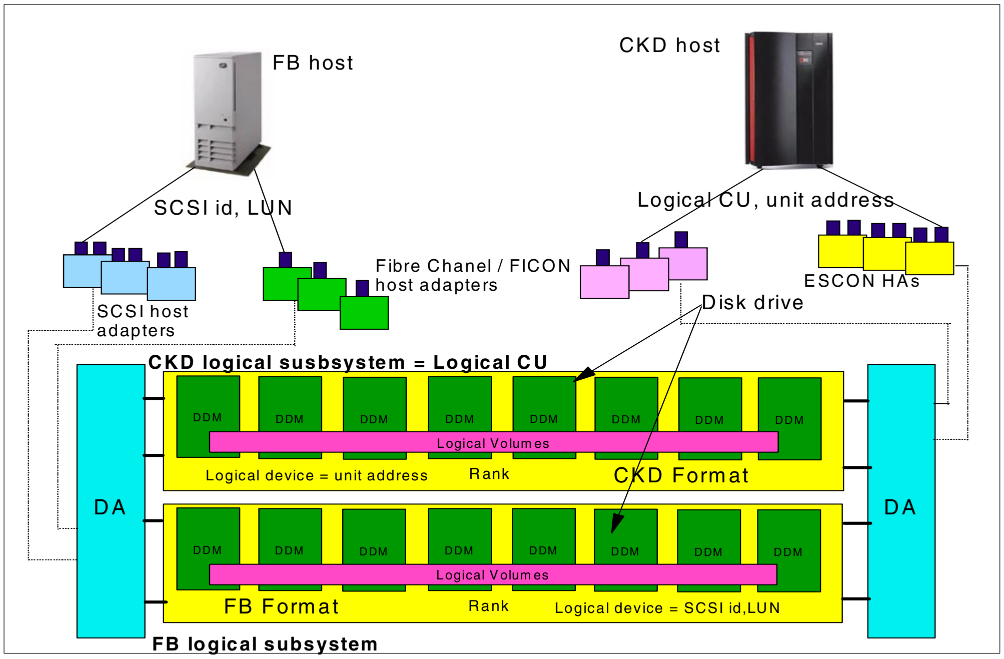


Figure 68. Logical configuration terminology

The terminology that we will be using in the following sections when describing the logical configuration procedure, has already been presented in previous chapters. You may refer to Section 1.2, “Terminology” on page 8. Also Chapter 2, “Hardware” on page 17 and Chapter 3, “Architecture” on page 45 already define the various terms when they describe in detail the ESS components that this terms designate.

Figure 68 presents most of the terms we will be using in the following sections when describing the logical configuration.

4.11 Logical configuration process

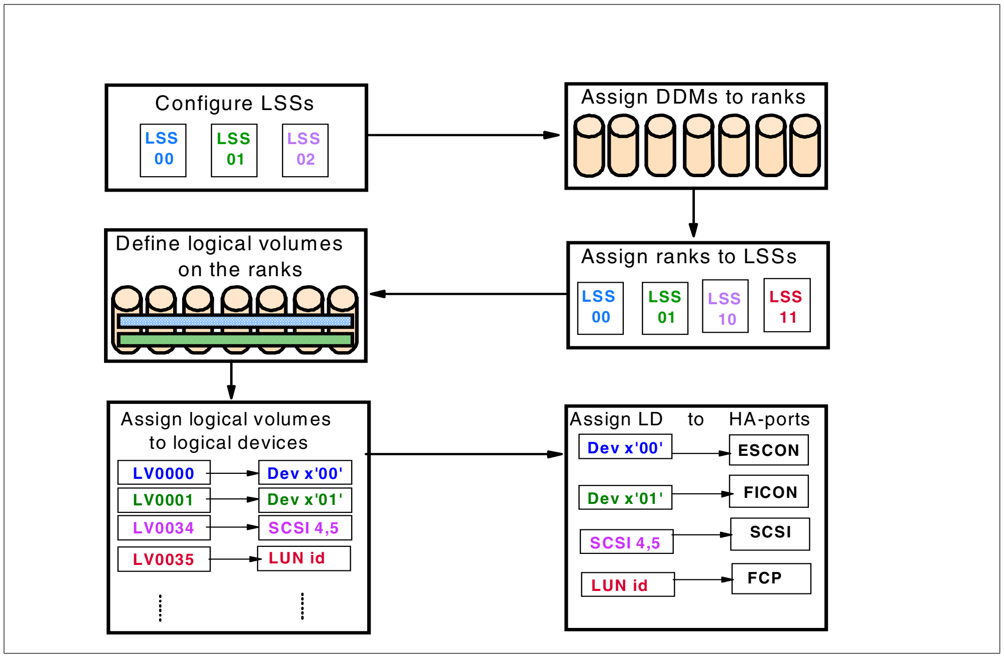


Figure 69. Logical configuration process

The diagram in Figure 69 provides a basic idea of the logical configuration process. The logical configuration requires that the physical installation of the Enterprise Storage Server has been completed by the IBM System Support Representative. Then the logical configuration is made using the ESS Specialist only, or also using the Batch Configuration Tool if choosing standard logical configurations options. The basic steps that are done during logical configuration are:

1. Identify the host systems that are attaching to the ESS.
2. Configure LCUs (for CKD LSSs only).
3. Select disk drives (DDMs) to form ranks.
4. Define the ranks as Fixed Block (FB) or Count-Key-Data (CKD).
5. Assign ranks to the corresponding logical subsystems (LSSs).
6. Define logical volumes (LVs) on the ranks.
7. Assign LVs to host logical devices (LDs).
8. Relate LDs with HAs (for SCSI only). CKD LDs will have an exposure to all ESCON and FICON host adapters in the ESS. For Fibre Channel attachment, you assign LDs to HA when setting the adapter in *Restricted Mode* (using the IBM TotalStorage ESS Specialist).

Some of these steps will not require an action from your side. For example, when assigning FB ranks to an LSS, the Enterprise Storage Server will do this. The process of logical configuration is described in the following pages.

4.11.1 Defining logical subsystems

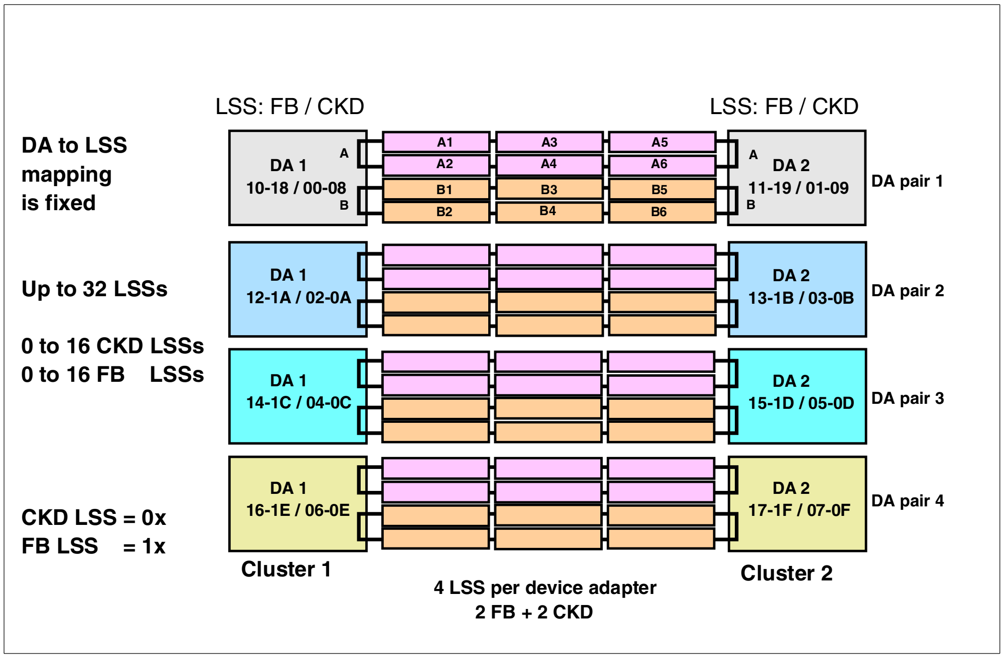


Figure 70. Logical subsystem mapping

Up to 32 LSSs can be configured in an ESS, 16 for CKD servers and 16 for FB servers. Each LSS will get an hexadecimal identifier. LSSs 0x are CKD and LSSs 1x are FB as Figure 70 shows. An LSS can have up to 256 logical devices defined to it. LSSs can map arrays from both loops of the DA pair.

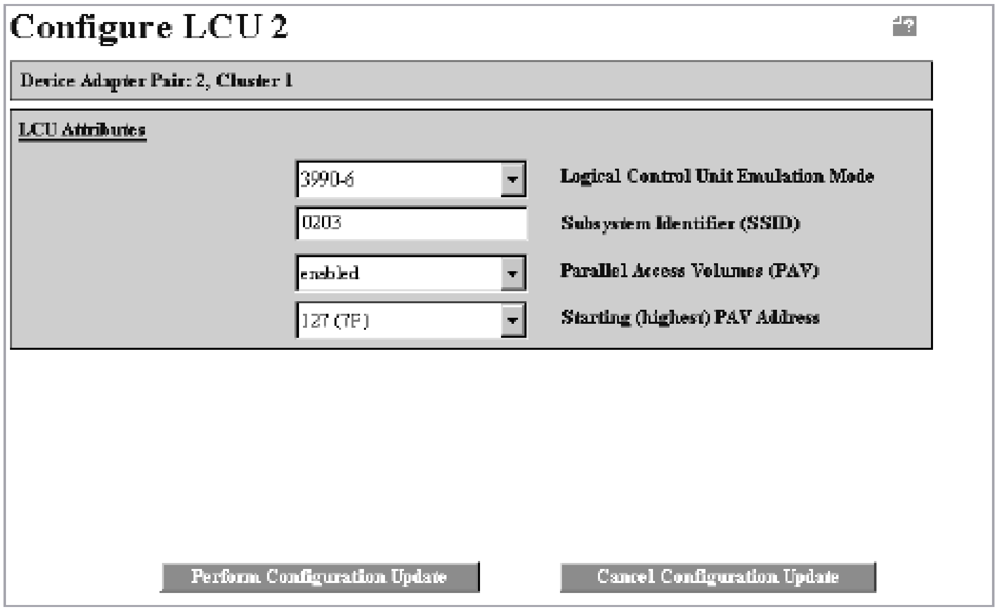
4.11.2 CKD logical subsystems

The ESCON and FICON protocol support up to 16 *logical control unit* (LCU) images from x’00’ to x’0F’ (Note: FICON has an architectural capability of supporting 256 control unit images, but its current implementation is 16 as in ESCON). In other words, any ESCON or FICON channel link arriving to the corresponding ESS port, will be able to address any of the sixteen logical control units that the ESS has available for CKD hosts: x’00’ to x’0F’. This settings are mapped directly to the LSSs IDs, which means that LSS 00 will be the logical CU 0, LSS 01 will logical CU 1, and so on. The LSS concept is very straight for the zSeries users because LSSs in the ESS map one to one the logical control units the zSeries server is viewing.

You will be using the ESS Specialist *Configure LCU* panel to make the CKD LSSs definitions (see Figure 71).

For each control unit image, you must specify its emulation mode. You can choose between the following CU emulations:

- 3990-6
- 3990-3
- 3990-3 TPF



Configure LCU 2

Device Adapter Pair: 2, Cluster 1

LCU Attributes

3990-6 Logical Control Unit Emulation Mode

0203 Subsystem Identifier (SSID)

enabled Parallel Access Volumes (PAV)

127 (7F) Starting (highest) PAV Address

Figure 71. ESS Specialist — Configure LCU panel

For each of the configured logical control units will need you to specify a 4-digit subsystem identifier (SSID). This is the usual setting done for a real 3990, and it is required to identify the CU to the host for error reporting reasons and also for functions like Peer-to-Peer Remote Copy (PPRC). If the Batch Configuration Tool has been used by the IBM SSR, then the SSIDs will already by assigned. If the tool has not been used, then the SSIDs will need to be configured.

Remember that SSIDs must be unique. The system does not allow bringing online a control unit with an already assigned SSID. Users must keep a record of the SSIDs in use, and must assign a unique SSID to each new LCU been configured. By default the Batch Configuration Tool assigns the SSIDs with an xxyy format, where xx is the two last digits of the ESS serial number, and yy can be from 01 to 16 in correspondence to LCUs 01 to 16 (this can be modified to adjust to the installation requirements).

Also in this panel you enable PAV if your ESS has this optional feature, and you set the PAV starting address. Next you proceed to define the ranks (RAID-5 or JBOD) and its logical volumes for the CKD LSS as is explained later in 4.11.5, “Configuring CKD ranks” on page 114

4.11.3 FB logical subsystems

As with CKD, the ESS allows to have either 8 or 16 FB LSSs. When your choice is to have the IBM SSR to configure 8 LSSs, then each of the four DA pairs in the ESS will have two LSSs, one LSS per DA. For each DA pair, one LSS is assigned to cluster 1 and the other to cluster 2. When your choice is 16 LSSs, then each of the four DA pairs will have four LSSs, two LSSs per DA. For each DA pair. two LSSs are allocated to cluster 1 and two to cluster 2.

The FB LSSs will implicitly be configured when you define the ranks as RAID-5 or JBOD, select the storage type as FB, and ESS Specialist assigns the ranks to the LSSs. There is no specific ESS Specialist panel for FB LSS configuration.

4.11.4 Disk groups — ranks

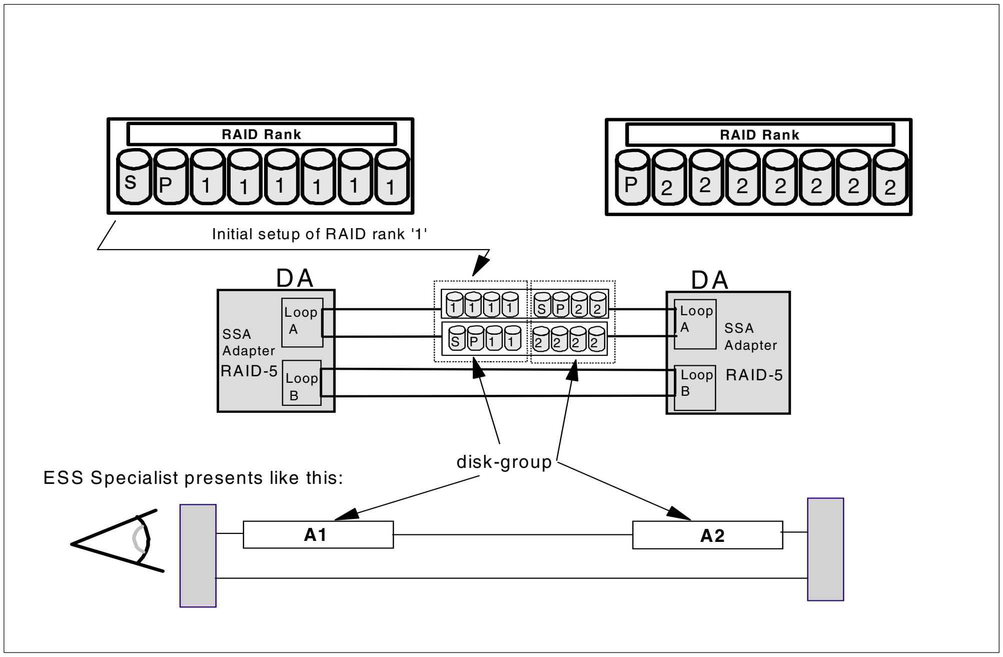


Figure 72. Disk groups and RAID ranks — initial setup

Before describing the configuration of CKD and FB ranks, let us review some definitions we already presented and let us see some additional basic concepts that are involved. Please refer to Figure 72 for the following explanations

- The basic units of non-volatile storage that are field replaceable units (FRUs) are the *disk drive modules* (DDMs) that hold the *hard disk drives* (HDDs), to which we also refer generically as *disk drives*.
- Eight DDMs are grouped into one *8-pack* assembly (8-packs). The 8 packs must be installed in pairs on the same loop. These pairs of 8-packs are the way to add capacity to the ESS.
- There are two SSA loops (A and B) per device adapter (DA) pair.
- Each SSA loop can hold from two to six 8-packs, if it has some capacity installed on it. Else it is a loop not been used, not holding any 8-pack, and is available for future capacity upgrade.
- Initially, for each loop, four DDMs from two different 8-packs make what is called a *disk group* by the ESS Specialist (see Figure 72). The ESS Specialist refers to disk groups and displays them graphically as whole rectangles. Later when the spare disks are used, the DDM set that conforms the disk group will differ from the original set. The ESS Specialist will still show the disk group with the same whole rectangle representation it used to show the initial disk group setup. So a disk group is basically eight DDMs of the same loop, and these DDMs at some point, will no more be part of the same initial 8-pack assembly.
- The up to six 8-packs that a loop may have correspond to the disk groups that are designated as A1 to A6 for loop A, and B1 to B6 for loop B.
- This means that there can be from two to twelve disk groups per DA pair (A1 thru A6 and B1 thru B6).

- Disk groups A1, A3, A5, B1, B3, B5 are associated with DA1; disk groups A2, A4, A6, B2, B4, B6 are associated with DA2. This is the default association before any ESS Specialist configuration process.
- Using the ESS Specialist, ranks are created from these disk groups by formatting the set of eight DDMs. A rank results from the ESS Specialist formatting process of a disk group: it will become either a RAID rank (RAID array) or a set of eight non-RAID ranks.
- Each rank is mapped by one od the DAs only.

Also, when working with the ESS Specialist to do the logical configuration, you will recognize the following characteristics:

- The device adapter (DA) to logical subsystem (LSS) mapping is a fixed relationship.
- Each one of the DAs of a pair, has up to four LSSs that map ranks: two are for CKD ranks and the other two are for FB ranks.
- For a pair of DAs, any of its LSSs can map ranks from any of its two loops (FB LSSs map FB ranks, and CKD LSSs map CKD ranks).
- Logical volumes (up to 256 per LSS) are created on these ranks.
- Each *logical volume* (LV) is part of a rank. The rank is part (non-RAID) or the whole (RAID-5) of a disk group. This disk group is mapped in an LSS, which is associated with a single DA in a DA pair.

The assignment of disk groups to an LSS is made in the context of the corresponding addressing architectures, either *fixed-block* architecture (FB) or *count-key-data* (CKD) architecture. Also this disk group assignment to the LSSs is made in the context of the communication architecture that will be used (SCSI, FCP, ESCON or FICON).

As already said, ranks can either operate in RAID-5 mode or as JBOD. The RAID-5 arrays will have the following setup:

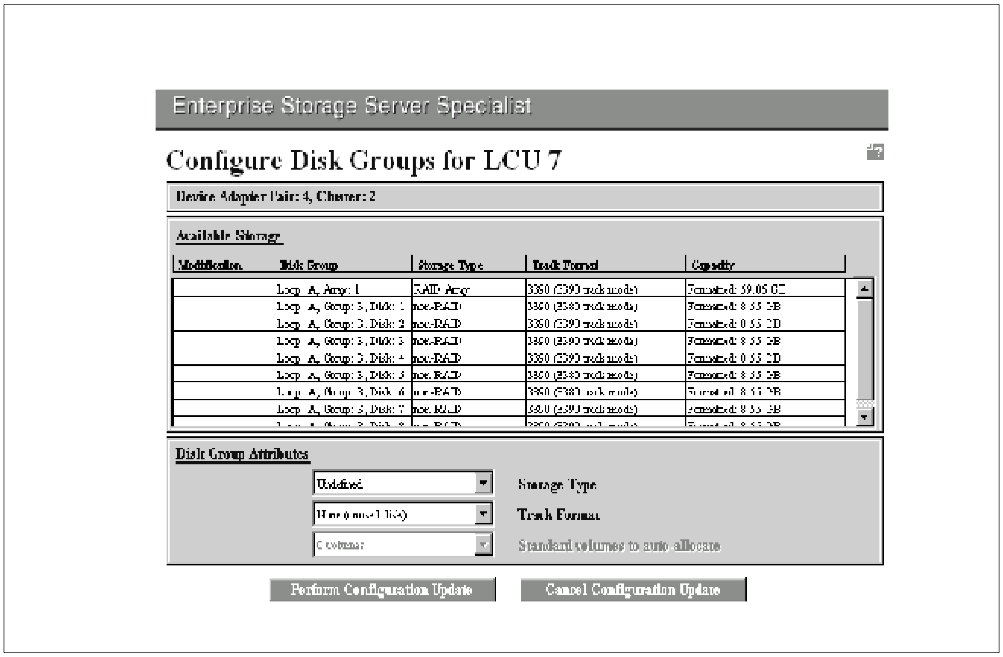
- **6+P:** This setup is for leaving spares in the loop. Because 16 drives are the minimum configuration for a loop, whenever configuring for RAID-5, the first 2 arrays in the first loop will be 6+P, leaving 2 drives in the loop as spares.
- **7+P:** This is the setup for all the other ranks in the loop you are configuring as RAID-5.

For JBOD ranks each disk drive in the disk group, become a rank by itself.

4.11.5 Configuring CKD ranks

The CKD ranks can either be RAID-5 or JBODs. You will assign these disk group characteristics with the *Configure CKD Disk Group* panel of the ESS Specialist, when doing the logical configuration (see Figure 73).

The CKD RAID ranks can be configured in interleaved mode. Ranks that are configured in interleaved mode, will have logical volumes assigned to it already during this step. For an explanation of interleaved and non-interleaved partitions see 4.11.6, “CKD interleaved partitions” on page 115 and 4.11.7, “CKD non-interleaved partitions” on page 116.



Configure Disk Groups for LCU 7

Device Adapter Pair: 4, Cluster: 2

Available Storage

Modification	Disk Group	Storage Type	Track Format	Capacity
	Loop A, Array: 1	RAID Array	3390 (3390 track mode)	Formatted: 59.05 GB
	Loop A, Group: 3, Disk: 1	non-RAID	3380 (3380 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 2	non-RAID	3390 (3390 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 3	non-RAID	3390 (3390 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 4	non-RAID	3390 (3390 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 5	non-RAID	3380 (3380 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 6	non-RAID	3380 (3380 track mode)	Formatted: 8.55 GB
	Loop A, Group: 3, Disk: 7	non-RAID	3390 (3390 track mode)	Formatted: 8.55 GB

Disk Group Attributes

Storage Type: Undefined

Track Format: None (unused disk)

Standard volumes to auto-allocate: 0 volumes

Figure 73. ESS Specialist — Configure CKD Disk Group panel

When working with the *Configure CKD Disk Group* panel you select from the list of disk groups that the panel displays, the ones you will associate to the LCU you are configuring. And for the selected disk groups you define the storage type as either RAID, non-RAID or *Undefined* (see Figure 73)

Then you select the track format as either 3380 or 3390. If you selected non-RAID, then for each disk drive in the disk group you will have to define its track format as either 3380 or 3390 or let it default to *none* (If you keep the default, you can define these disks drives at a later time as either FB or CKD).

Next you select the number of standard *logical volumes* to allocate automatically, with the *Standard volumes to auto allocate* drop down list of the *Configure CKD DIsk Group* panel. Remember that *logical volumes* are associated to *device numbers* (unit address) of the host. If you define zero, the array is created with the specified track format and zero logical volumes. Logical volumes can later be created using the *Add CKD Volumes* panel of the ESS Specialist.

You do this definitions for all the disk groups you want to format, making the selections from the list displayed by the ESS Specialist panel shown in Figure 73. Then you finish this disk group to LSS association by clicking the “Perform

Configuration Update" button. Next the last task will be to add *custom volumes* using the *Add CKD Volumes* panel of the ESS Specialist.

Once the "Perform Configuration Update" button has been clicked, the LSS definitions are executed, and the disk group and the logical volume size definitions are established. Any change (even removal) should require a global LSS re-definition, which is disruptive for the data. This data has to be moved or backed-up previously.

Note however that in case of adding ranks to an existing LSS, you can do this without affecting the existing data. Also note, that you can update non-disruptively PAV assignments after a Perform Configuration Update.

4.11.6 CKD interleaved partitions

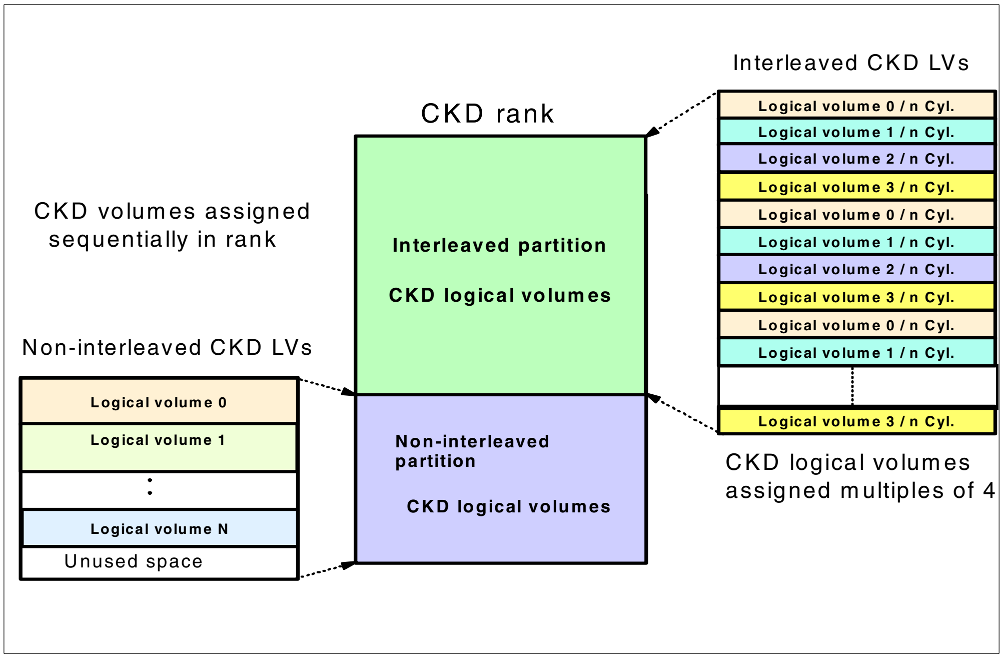

Interleaved CKD LVs
CKD rank
Logical volume 0 / n Cyl.
Logical volume 1 / n Cyl.
Logical volume 2 / n Cyl.
Logical volume 3 / n Cyl.
Logical volume 0 / n Cyl.
Logical volume 1 / n Cyl.
Logical volume 2 / n Cyl.
Logical volume 3 / n Cyl.
Logical volume 0 / n Cyl.
Logical volume 1 / n Cyl.
Logical volume 3 / n Cyl.
CKD volumes assigned sequentially in rank
Interleaved partition
CKD logical volumes
Non-interleaved CKD LVs
Logical volume 0
Logical volume 1
Logical volume N
Unused space
Non-interleaved partition
CKD logical volumes
CKD logical volumes assigned multiples of 4

Figure 74. CKD interleaved and non-interleaved partitions

Figure 74 shows CKD LVs mapped into an interleaved partition and into a non-interleaved partition. An interleaved partition is one where a number of logical volumes of identical size are striped across all the disks in the array. The mapping of CKD LVs into an interleaved partition will occur in multiples of 4 logical volumes with the same number of cylinders per volume. When you format an interleaved partition, you specify the size of volume you require (for example 3390-3) and the array is formatted with as many 3390-3s as can be fitted into the space in multiples of four. At the end of each interleaved partition there is an unformatted area with a minimum size of 5000 cylinders, plus any space that could not be filled with a multiple of four volumes.

The volumes are formatted from the beginning of the partition until all the CKD logical volumes have been fitted into the partition. This method ensures that none of the CKD LVs will have performance disadvantages related to the physical characteristics of the disk drives forming a rank, such as long seeks to reach the data of a logical volume, because it has been placed at the end of a disk drive.

All the volumes that are automatically defined when you specify an interleaved partition must be the same size. The ESS Batch Configuration Tool that the IBM System Support Representative uses at initial installation, and the *Standards volume to auto-allocate* option of the ESS Specialist *Configure CKD Disk Group* panel, define the volumes as interleaved as when configuring CKD ranks.

However, after these standard volumes allocations, there will be some unformatted space at the end of the interleaved partition. You can use this space for manually formatting more volumes of standard size, or for custom volumes. You do this using the ESS Specialist, which for this instance will do the definitions in a non-interleaved mode (see 4.11.8, “Defining interleaved volumes with ESS Specialist” on page 116).

We recommend that, wherever possible, you use interleaved partitions for ease of management and optimum performance, and use the remaining non-interleaved partition for more custom or standard volumes that you may need.

4.11.7 CKD non-interleaved partitions

In a non-interleaved partition, all CKD LVs assigned to it are defined to the rank in the order of creation. Each logical volume must be defined individually. Because of that, the first LV created gets space assigned at beginning of the partition, the second one the next available space in the partition and so on. This may cause some performance disadvantages for LVs that are placed near the end of a partition, because it may require longer seeks from the disk drive to reach the assigned space for a specific LV. On the other hand, if the rank is configured in RAID-5, the data of the LVs is striped across several disk drives, and this reduces the problem mentioned. Only the non-interleaved partitions accept CKD logical volumes as custom volumes. If you use non-interleaved partitions, you must plan how many volumes of the required size will fit into your partition.

The ESS Specialist defaults to a non-interleaved mode of defining the logical volumes when configuring from the *Add CKD Volumes* panel. In this instance each logical volume will be defined in the order of its creation.

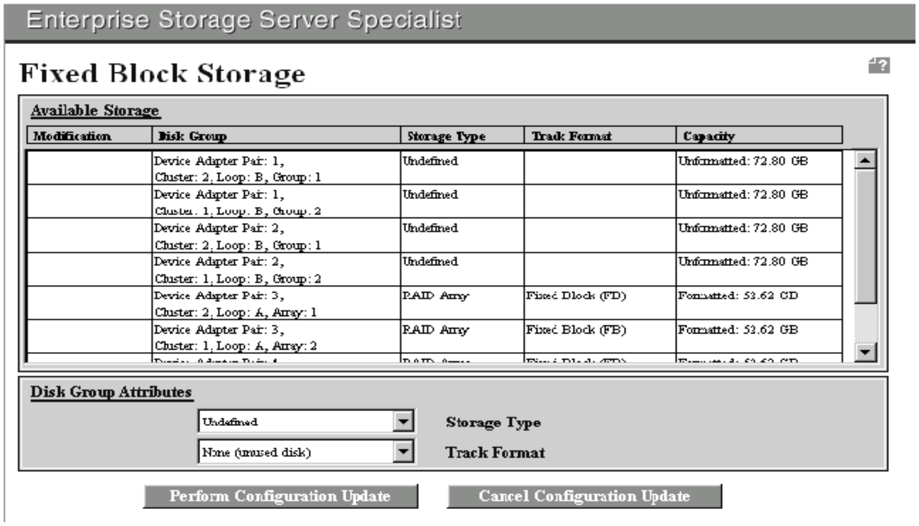
4.11.8 Defining interleaved volumes with ESS Specialist

When using the ESS Specialist to configure CKD disk groups, there is the option to select a standard volume type for auto allocation. This is done when using the *Standard volumes to auto-allocate* option in the *Configure CKD Disk Group* panel of the ESS Specialist. When using this option, the functional microcode creates auto allocated standard size volumes in interleaved mode.

On the other hand, CKD volumes created by the *Add Volumes* panel of the ESS Specialist, are non-interleaved volumes.This will happen when you have selected zero volumes when initially using the *Configure CKD Disk Groups* panel. In this case the array was created with zero volumes and the specified track format. Also after running the Batch Configuration Tool, when you go to define full or custom volumes in the remaining capacity. In both cases when you come to create the volumes, you will be using the *Add CKD Volumes* panel. In this situation, the ESS Specialist will use a non-interleaved setting.

4.11.9 Configuring FB ranks

In this particular step you configure the ranks from the disk groups available on the loop. When doing this rank definition, ESS Specialist will also automatically assign the ranks to the LSSs in the DA pair. The ranks that are been formatted upon even numbered disk groups are assigned to an LSS belonging to Cluster 1. The ranks that are been formatted upon odd numbered disk groups, are assigned to an LSS belonging to Cluster 2. The allocation o ranks to LSSs is dependent upon the number of ranks and of logical volumes that have been created. At first, ESS Specialist will assign the ranks to the first of the two LSSs belonging to a cluster. ESS Specialist will assign ranks to the second LSS when the ranks assigned to the first LSS contain a total of 192 or more logical volumes. Remember that each LSS can contain a maximum of 256 logical volumes.



Enterprise Storage Server Specialist

Fixed Block Storage

Available Storage

Modification	Disk Group	Storage Type	Track Format	Capacity
	Device Adapter Pair: 1, Cluster: 2, Loop: B, Group: 1	Undefined		Unformatted: 72.80 GB
	Device Adapter Pair: 1, Cluster: 1, Loop: B, Group: 2	Undefined		Unformatted: 72.80 GB
	Device Adapter Pair: 2, Cluster: 2, Loop: B, Group: 1	Undefined		Unformatted: 72.80 GB
	Device Adapter Pair: 2, Cluster: 1, Loop: B, Group: 2	Undefined		Unformatted: 72.80 GB
	Device Adapter Pair: 3, Cluster: 2, Loop: A, Array: 1	RAID Array	Fixed Block (FB)	Formatted: 53.62 GB
	Device Adapter Pair: 3, Cluster: 1, Loop: A, Array: 2	RAID Array	Fixed Block (FB)	Formatted: 53.62 GB

Disk Group Attributes

Storage Type: Undefined

Track Format: None (unused disk)

Figure 75. ESS Specialist - Fixed Block Storage panel

FB ranks can either be RAID-5 or JBODs. You will be choosing the disk groups and specifying for them the *storage type* as either RAID, non-RAID or *Undefined* (see Figure 75).

Then you select the *track format* as FB or *None* (undefined). If you previously selected non-RAID then for each disk drive in the disk group you will have to define its track format (Undefined disk drives can later be defined as either FB or CKD).

You do this definition for all the disk groups you want to format, using the *Fixed Block Storage* panel of the ESS Specialist shown in Figure 75. Then you finish this rank definitions by clicking the “Perform Configuration Update” button.

Next step will be to define to the LSS the *logical volumes* of these ranks. You do this using the open systems *Add Volumes* panels of the ESS Specialist. The logical volumes will be associated to one or more host’s *Logical Unit Numbers* (LUNs). This step establishes relationships with the hosts, and defines logical volume sizes and numbers.

For both these configuration steps, once the “Perform Configuration Update” button has been clicked, the LSS definition are done, and the disk group and the logical volume assignments will be completed. Changes to the logical volume assignments (e.g. attaching or detaching hosts) can be done with the ESS Specialist. Changes to the logical volume structure on an array will require the whole array to be reformatted e.g. if you have two 16GB logical volumes and you want to make them into one 32 GB logical volume, then the whole array has to be reformatted.This will require any other data on that array to be backed up first and all the volumes on that array to be detached from their host systems.

Note however that in case of adding ranks to an LSS, you can do this without affecting the existing data of that LSS.

4.11.10 Rank capacities

Table 1 provides details about the rank capacities in the ESS when they are configured as RAID-5 or JBOD. The ESS disks are available in either 9.1 GB, 18.2 GB and 36.4 GB capacities. The rank with 6+P setup is automatically configured by the ESS Specialist for the first two ranks in the loop. The 7+P setup is configured automatically by the ESS Specialist for the rest of the ranks in the loop. When in the ESS Specialist you define a disk group to be formatted as a JBOD rank, then each of the eight disks of the disk group will become a rank.

Table 1. Rank capacities

DDM capacity (is also JBOD rank capacity)	RAID-5 6+P	RAID-5 7+P
9.1 GB	53.81 GB	62.79 GB
18.2 GB	107.67 GB	125.62 GB
36.4 GB	215.38 GB	251.28 GB

4.11.11 Assigning logical volumes to a rank

Once the ranks have been set up, you can start deafening logical volumes (LV) in the ranks. For FB ranks, from the ESS Specialist *Open Systems Storage* panel you click the *Add Volumes* button and that gets you to the first *Add Volumes* panel where you select the server, the port, and the FB ranks to which the logical volumes will be available. The selections in this panel take you to the second *Add Volumes* panel where you define the volume sizes and number of volumes. Once you click the Perform Configuration Update button, you will be defining the logical volumes within the FB rank and also assigning them to the host logical devices.

For CKD ranks you will be defining the logical volumes when defining the track format as 3380 or 3390, and specifying the number of standard volumes, in the *Configure CKD Disk Group* panel. Later you can also add volumes to a rank using the (CKD) *Add Volumes* panel.

For both, FB and CKD ranks you have the recommended alternative of having the IBM System Support Representative use the Batch Configuration Tool that will define the logical volumes to the LSSs. This will happen if you choose the standard logical configuration options.

4.11.11.1 Assigning Fixed Block logical volumes

Logical volumes for open systems servers, can have an LV size from 100 MB to full rank size. This increased granularity of LUN sizes enables improved storage management efficiencies, especially for Windows NT systems which have a limited number of LUNs and therefore, require full exploitation of the rank capacity. Also supported are 16 GB or 32 GB LVs.

For the standard LUN sizes of 4, 8, or 16 GB the rank capacity is shown in Table 2.

Table 2. Open systems FB ranks — available capacity

Available capacity for open systems standard size volumes		
Volume size	DDM capacity and array	Number of LUNs
4 GB	9.1 GB (6+P+S)	13
4 GB	18.4GB (6+P+S)	26
4 GB	36.4 GB (6+P+S)	53
8 GB	9.1 GB (6+P+S)	6
8 GB	18.4GB (6+P+S)	13
8 GB	36.4 GB (6+P+S)	26
16 GB	9.1 GB (6+P+S)	3
16 GB	18.4GB (6+P+S)	6
16 GB	36.4 GB (6+P+S)	13

4.11.11.2 Assigning CKD logical volumes

The CKD logical volumes defined in the *interleaved* partition of the rank, will match full 3390-2 (1.89 GB) or 3390-3 (2.83 GB) or 3390-9 (8.51 GB). This applies if the track format selected is 3390. The 3390-2 and the 3390-3 LVs can also run in 3380 track format. This does not mean that the 3390s defined are running in track compatibility mode. What happens is that the zSeries will see a 3380 with either 2226 cylinders (3390-2) or 3339 cylinders (3390-3).

The CKD volumes defined in the *non-interleaved* partition (when using the ESS Specialist *Add Volumes* panel), may be defined as *custom volumes*. This option allows you to configure 3390s with a cylinder range from 1 cylinder up to the size of a 3390-9 (10017 cylinders). A CKD custom volume will allow you to break down a dataset to a single logical volume. The advantage of this is that the dataset will have a dedicated volume for it, which will result in less logical device contention.

Table 3 shows the capacities available when you configure the CKD ranks. The table shows, for each type of CKD logical volume, how many fit in the rank's *interleaved* an in the *non-interleaved* partitions.This table shows RAID-5 ranks. Normally you do not need to calculate these numbers, because when assigning

CKD LVs to a rank, the configuration process will give you information about the available and remaining capacities in the rank you are configuring.

Table 3. CKD ranks — available capacities

Logical device type	**Logical device physical capacity (GB)**	**Physical Capacity of Interlved partition (GB)**	**Physical capacity of non-Interlved partition (GB)**	**Logical devices in interlved partition**	**Logical devices in non-interlved partition**
RAID 5 rank, 6+P+S, 9.1 GB DDMs					
3390-2	1.96	47.04	6.77	24	3 (1)
3390-3	2.94	47.04	6.77	16	2
3390-9	8.82	35.28	18.53	4	2
3390-2 (3380)	1.82	43.68	10.13	24	3 (1)
3390-3 (3380)	2.73	43.68	10.13	16	3
RAID 5 rank, 7+P, 9.1 GB DDMs					
3390-2	1.96	54.88	7.91	28	4 (1)
3390-3	2.94	47.04	15.75	16	5
3390-9	8.82	35.28	27.51	4	3
3390-2 (3380)	1.82	58.24	4.55	32	4 (1)
3390-3 (3380)	2.73	54.60	8.19	20	5
RAID 5 rank, 6+P+S, 18.2GB DDMs					
3390-2	1.96	101.92	5.75	52	2 (1)
3390-3	2.94	94.08	13.59	32	4
3390-9	8.82	70.56	37.11	8	4
3390-2 (3380)	1.82	101.92	5.75	56	3 (1)
3390-3 (3380)	2.73	98.28	9.39	36	3
RAID 5 rank, 7+P, 18.2GB DDMs					
3390-2	1.96	117.60	8.02	60	4 (1)
3390-3	2.94	117.60	8.02	40	2
3390-9	8.82	105.84	19.78	12	2
3390-2 (3380)	1.82	116.48	9.14	64	4 (1)
3390-3 (3380)	2.73	120.12	5.50	44	2
RAID 5 rank, 6+P+S, 36.4 GB DDMs					
3390-2	1.96	203.84	11.54	104	5 (1)
3390-3	2.94	199.92	15.46	68	5
3390-9	8.82	176.40	38.98	20	4
3390-2 (3380)	1.82	211.12	4.26	116	2 (1)

Logical device type	Logical device physical capacity (GB)	Physical Capacity of Interlved partition (GB)	Physical capacity of non-Interlved partition (GB)	Logical devices in interlved partition	Logical devices in non-interlved partition
3390-3 (3380)	2.73	207.48	7.90	76	2
RAID 5 rank, 7+P, 36.4 GB DDMs					
3390-2	1.96	243.04	8.24	124	9 (1)
3390-3	2.94	235.20	16.08	80	5
3390-9	8.82	211.68	39.60	24	4
3390-2 (3380)	1.82	240.24	11.04	132	9 (1)
3390-3 (3380)	2.73	240.24	11.04	88	4
Notes: 1. These custom volumes are reported by the system as 3390-3 devices with 2226 cylinders					

Table 4 shows the CKD logical devices capacities. Both Table 3 and Table 4 also show the physical capacity that is allocated in the rank, when defining the logical volumes. As you can see, this physical capacity allocated in the rank is slightly greater than the logical device capacity. This difference reflects the algorithms the ESS uses as it allocates space for logical volumes.

Table 4. CKD logical device capacities

Logical device type	Cylinders	Bytes per cylinder	Logical device capacity (GB)	Physical capacity used (GB)	524 byte sectors per cylinder
3390-2 (1)	2,226	849,960	1.892	1.962	1,680
3390-3	3,339 (2)	849,960	2.838	2.943	1,680
3390-3 custom (3)	1 - 3,339 (2)	849,960	0.00085 - 2.838	0.00176 - 2.943	1,680
3390-9	10,017 (4)	849,960	8.514	8.828	1,680
3390-9 custom (3)	3,340 - 10,017 (4)	849,960	2.839 - 8.514	2.944 - 8.827	1,680
3390-2 (3380) (1)	2,226	712,140	1.585	1.821	1,560
3390-3 (3380)	3,339 (2)	712,140	2.377	2.731	1,560
3390-3 (3380) custom (3)	1 - 3,339 (2)	712,140	0.00071 - 2.377	0.00163 - 2.731	1,560

<table>
<tr><th>Logical device type</th><th>Cylinders</th><th>Bytes per cylinder</th><th>Logical device capacity (GB)</th><th>Physical capacity used (GB)</th><th>524 byte sectors per cylinder</th></tr>
<tr><td colspan="6">Notes:
1. Only allowed in an interleaved partition
2. In an interleaved partition, the number of cylinders is 3,339. In a non-interleaved partition, the number of cylinders may be from 1 to 3,339
3. A CKD volume that has a capacity different from that of a standard 3390 device type, is referred as custom volume
4. In an interleaved partition, the number of cylinders is 10,017. In a non-interleaved partition, the number of cylinders may be from 3,340 to 10,017</td></tr>
</table>

The following formula allows you to *approximate* the physical capacity of a logical CKD device. The formula does not completely reflect the algorithms of the ESS as it allocates space for a logical volume. It uses information from Table 3 and Table 4.

The amount of physical capacity for 3390 devices can be approximated by:

Capacity = (((Nr. of cyls. +1) * Bytes per cyl. * 524) / 512) * 1.013 * 10 $^{-9}$ GB

The amount of physical capacity for 3380 devices can be approximated by:

Capacity = (((Nr. of cyls. +1) * Bytes per cyl. * 524) / 512) * 1.122 * 10 $^{-9}$ GB

The equations compensate for any overhead in the logical device such that the result is always greater than or equal to the physical capacity required to configure the logical device.

4.11.11.3 Assigning iSeries 400 logical volumes

For the iSeries 400 servers the logical volume sizes match the 9337 or 2105 devices with sizes of 4.19 GB, 8.59 GB, 17.54 GB, 35.17 GB, 36 GB. 70.56 GB.

Volume Rank	9337-48C 9337-48A 4.19 GB	9337-59C 9337-59A 2105-A01 2105-A81 8.59 GB	9337-5AC 9337-5AA 2105-A02 2105-A82 17.54 GB	9337-5CC 9337-5CA 2105-A05 2105-A85 35.16 GB	9337-5BC 9337-5BA 2105-A03 2105-A83 36.00 GB	2105-A04 2105-A84 70.56 GB
9.1 GB DDMs 6+P	12 LUNs + 2.33 GB	6 LUNs + 1.07 GB	2 LUNs + 17.91 GB	1 LUN + 17.81 GB	1 LUN + 16.97 GB	0 LUNs
9.1 GB DDMs 7+P	14 LUNs + 2.73 GB	7 LUNs + 1.26 GB	3 LUNs + 8.94 GB	1 LUN + 26.79 GB	1 LUN + 25.95 GB	0 LUNs
18.2 GB DDMs 6+P	25 LUNs + 0.42 GB	12 LUNs + 2.19 GB	5 LUNs + 17.92 GB	2 LUNs + 35.67 GB	2 LUNs + 33.99 GB	1 LUN + 35.06 GB
18.2 GB DDMs 7+P	29 LUNs + 1.21 GB	14 LUNs + 2.56 GB	6 LUNs + 17.92 GB	3 LUNs + 17.62 GB	3 LUNs + 15.1 GB	1 LUN + 52.58 GB
36.4 GB DDMs 6+P	50 LUNs + 0.88 GB	24 LUNs + 4.42 GB	11 LUNs + 17.93 GB	5 LUNs + 35.38 GB	5 LUNs + 31.18 GB	2 LUN + 69.75 GB
36.4 GB DDMs 7+P	58 LUNs + 2.46 GB	28 LUNs + 5.16 GB	13 LUNs + 17.93 GB	6 LUNs + 35.28 GB	6 LUNs + 30.24 GB	3 LUN + 34.27 GB

Figure 76. iSeries 400 logical volumes per rank

Figure 76 shows the capacities available when you configure the ranks for use with an iSeries 400 server. Note that the residual GB's maybe configured as other iSeries logical volumes, with smaller size. For example; a residual of 17.92 GB could be defined as 2x 8.59 GB LUNs.

With SCSI attachment, the LUNs will be reported to the iSeries 400 as a device type 9337. With Fibre Channel attachment, the LUNs will be reported to the server as device type 2105. The model will depend on the LUN sizes defined.

Remember that iSeries only support RAID-5 storage type

Protected versus non-protected

With its RAID-5 architecture, the ESS emulated 9337 and 2105 are treated as *protected* logical volumes which prohibits software mirroring. To solve this, the ESS permits disk units to be defined as *non-protected* models. Software mirroring is only allowed on non-protected 9337s and 2105s. From an ESS perspective, all iSeries volumes are RAID-5 and are protected within the ESS. The ESS Specialist *Add Volumes* panel, allows you to define the volume as *Unprotected*.

For additional considerations when attaching external storage to the iSeries you may refer to *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220 for further information.

4.11.12 Configuring SCSI and Fibre Channel host adapters

With the *Modify Host Systems* panel of the ESS Specialist you define to the ESS the attached host systems and you identify SCSI or Fibre Channel hatched hosts by type and name. Once you have finished with the *Modify Host Systems* panel definitions, then you proceed to configure the host adapters. You do so from the *Open Systems Storage* panel, by clicking the Configure Host Adapter Ports button that will take you to the *Configure Host Adapter Ports* panel.

For each SCSI port you define the type of server it will handle. The ESS Specialist will provide a list of hosts that are compatible with the selected bus configuration. This is required to run the correct protocols.

Unlike SCSI, where you link the server host icon to the SCSI host adapter port attached to it, Fibre Channel requires a host icon for every Fibre Channel adapter installed in the hosts (even if the adapters are installed in the same host). This is because each Fibre Channel adapter has a unique WWPN (World Wide Port Name), and hosts are defined based on this adapter identification. Figure 77. shows the ESS Specialist *Storage Allocation* panel, where one server with two Fibre Channel adapter cards appear as two host icons in the first row of the panel.

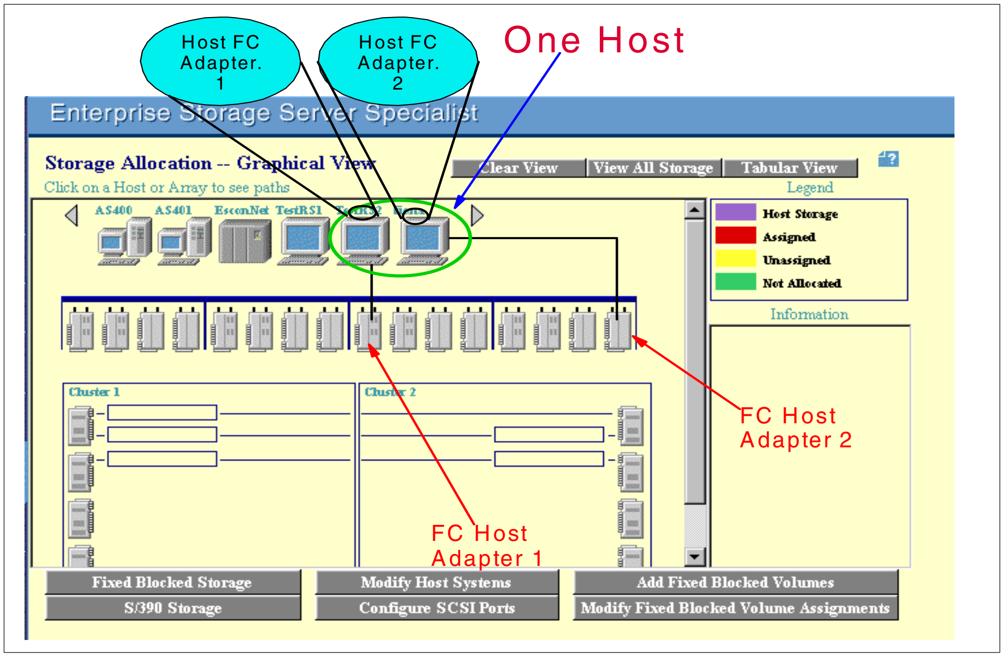


Figure 77. Fibre Channel adapter

For SCSI attachments, another important consideration when configuring the ports is the SCSI host initiator ids used by the host adapters. The default values that ESS Specialist uses are in accordance with the SCSI protocol, which defines SCSI ID 7 as having the highest priority. The SCSI ID priority order is 7-0 then 15-8. The first host system that you add is assigned to SCSI ID 7, the second is assigned to SCSI ID 6. You must verify that these assignments match the SCSI ID setting in each host system SCSI adapter card, and make adjustments to the map of SCSI IDs if necessary.

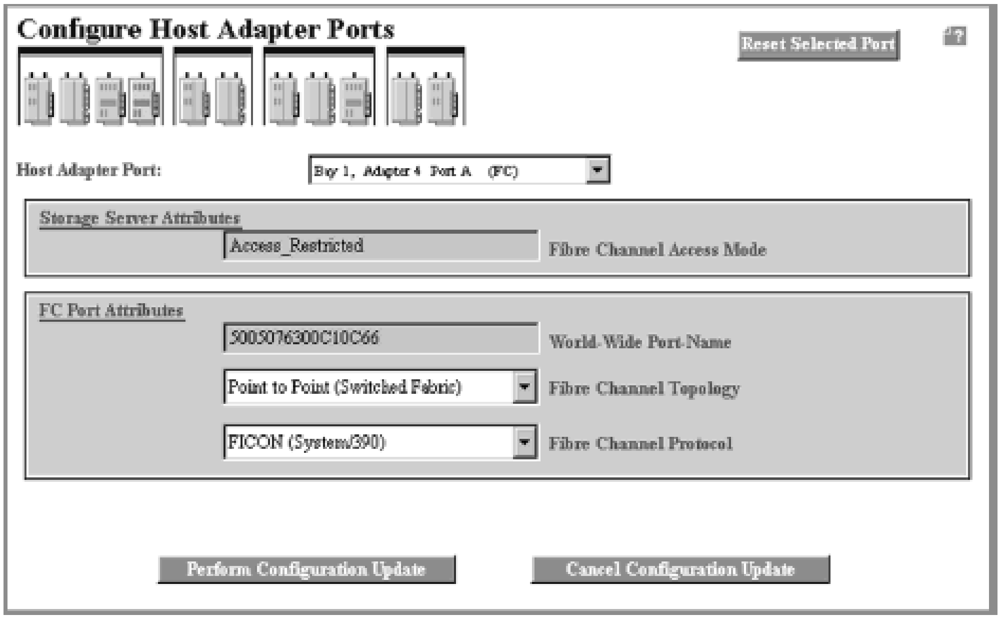
For Fibre Channel attachments, you will have some different considerations when configuring the host adapter ports. Because Fibre Channel allows any fibre channel initiator to access any logical device, without access restrictions, you will have the option to limit this characteristic, or not. The ESS Specialist *Configure Host Adapter Ports* panel, when a fibre channel host adapter has been selected, will allow you to specify the access mode as *Access any* mode or *Access restricted* mode. The *Access restricted* mode allows access to only the host system for which you have defined a profile. The profile limits the access of the host system to only those volumes assigned to the profile.The ESS adds anonymous hosts whenever you configure one or more Fibre Channel host adapter ports in the ESS and set the access mode to *Access any*.

For Fibre Channel you also can specify the fabric topology to which the port connects. If the topology is undefined you can use the menu drop-down to select either *Point-to-Point* or *Arbitrated-Loop*. If the topology is defined, you can only change the setting to *Undefined*.

4.11.13 Configuring ESCON and FICON host adapters

The ESCON and FICON protocols support up to 16 logical control unit images from x'00' to x'0F' (Note: FICON has the architectural capability of supporting 256 control unit images, but its current implementation is 16 as in ESCON). In other words, any ESCON or FICON channel link arriving at the corresponding ESS host adapter port, will be able to address any of the sixteen logical control units that the ESS has available for CKD hosts: x'00' to x'0F'.This settings are mapped directly to the LSSs IDs, which means that LSS 00 will be the logical CU 0, LSS 01 will logical CU 1, and so on. This access is restricted by means of the host IOCP and HCD definitions.

You don't need to identify the ESCON host adapter ports. The ESCON ports are identified to the ESS when the physical connection between the hosts and the ESS is made.



Configure Host Adapter Ports

Host Adapter Port: Bay 1, Adapter 4 Port A (FC)

Storage Server Attributes

Fibre Channel Access Mode: Access_Restricted

FC Port Attributes

World-Wide Port-Name: 5005076300C10C66

Fibre Channel Topology: Point to Point (Switched Fabric)

Fibre Channel Protocol: FICON (System/390)

Figure 78. Configure Host Adapter Ports panel — FICON connection

When you have FICON attachments in your configuration you will have to configure those ports using the *Configure Host Adapter Ports* panel. This panel is similar to the panel you use to configure the open systems Fibre Channel ports, but for FICON you arrive here clicking the Configure Host Adapter Ports button from the *S/390 Storage* panel (not from the *Open System Storage* panel).

Figure 78 gives an example of the fields that are enabled on the *Configure Host Adapter Ports* panel when you click a Fibre Channel host adapter.

For un-configured ports, *Point-to-point* (Switched Fabric) will be the only choice in the Fibre Channel topology field, when accessing this panel from the *S/390 Storage* panel.

The *Fibre Channel Protocol* field shows the current Fibre Channel protocol for the port you selected. If the topology is undefined the protocol can be FCP (open systems) or FICON (zSeries). FICON will be your only choice for un-configured ports, when you access this panel from the S/390 Storage panel.

If the topology is defined, you must first change the setting to *Undefined* before the ESS can make an alternate setting available for configuration.

4.11.14 Defining FB logical devices

Once the ranks have been assigned to the corresponding LSS and all host configuration information has been defined, you will start defining the logical devices (LDs) in the LSS. Remember, LDs are the way for the host to access the already defined logical volumes. Each logical volume will receive an LD identification that will be used by the host to access that volume. Remember that each LSS supports up to 256 LDs.

4.11.14.1 SCSI attached hosts

For FB logical volumes that are accessed by a SCSI host, you must set the SCSI targets of the host adapters. The SCSI target and the LUN ID of the devices are assigned by the ESS. The ESS can assign up to 64 LUNs per target, but not all hosts operating systems and host SCSI adapters are able to support these 64 LUNs per target. So the number of LUNs the ESS will automatically assign to each target will depend on the host type. These SCSI attached host types will be known to the ESS as they must have been already defined when using the *Modify Host Systems* panel.

When you first define the FB logical volumes you are simultaneously making them accessible to a host and to a SCSI port. This way you are relating the logical volumes to the logical devices view (target, ID LUN) of the host. Once each of the logical volumes has a logical device and SCSI port assignment you can map the logical devices to multiple SCSI ports using the ESS Specialist *Modify Volume Assignment* panel. Doing this will result in shared logical volumes. It is the host applications responsibility to handle shared logical volumes. These definitions may be of interest if you wish to configure for high availability. The Subsystem Device Driver (SDD) program, that comes with the ESS, allows for the handling of shared logical volumes. SDD runs on the host system side (See Appendix B, “Subsystem Device Driver (SSD)” on page 283 for further information).

Figure 79 shows the SCSI port assignment and logical device mapping that occurs when initially defining the logical volumes in the FB ranks, using the ESS Specialist *Add Fixed Block Volumes* panel.

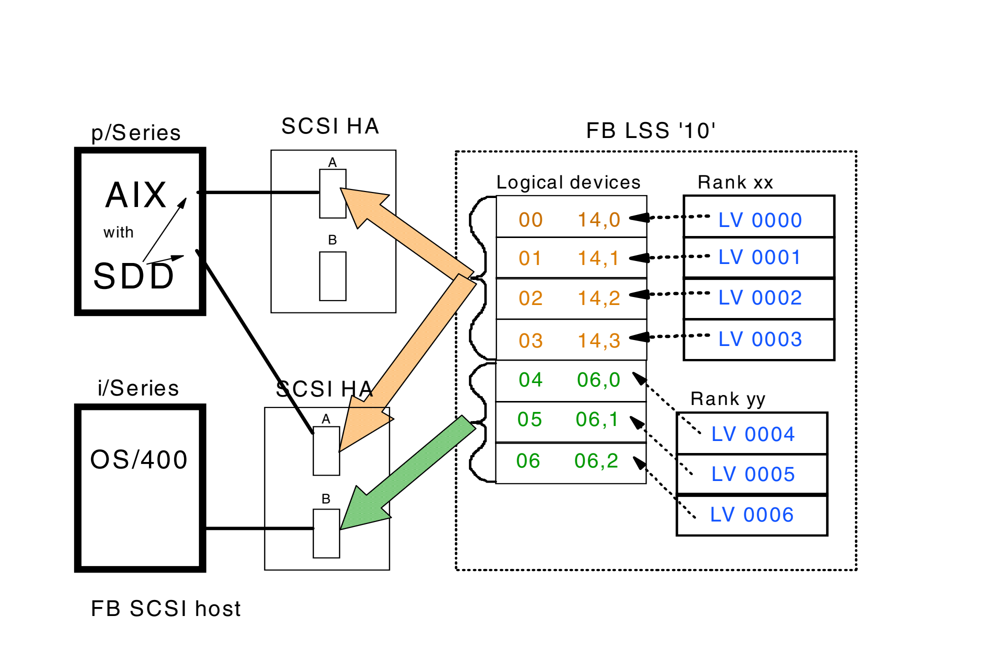


Figure 79. FB logical device mapping for SCSI attached host

The ESS supports a full SCSI-3 protocol set, and because of that, it allows the definition of up to 64 LUNs per target. Not every host operating system can support 64 LUNs per target. The following restrictions must be taken into consideration:

- IBM iSeries 400 servers accept only 6 targets and 8 LUNs ranging from 0 to 7. All LUNs on an LSS are required to be on the same target ID.
- IBM p/Series and RS6000 servers with AIX support up to 32 LUNs per target for the ultra SCSI adapters, and a maximum of 8 LUNs with the SCSI-2 fast wide differential adapters.
- IBM xSeries with Windows NT 4.0 will handle only up to 8 LUNs per target.
- Sun systems with Solaris 2.6 will allow a maximum of 8 LUNs per target (With a PTF to Solaris, the number can be 32).

You can check the most updated and completed information on host LUN support in the *IBM Enterprise Storage Server Host System Attachment Guide,* SC26-7296.

4.11.14.2 Fibre Channel attached hosts

For FB logical volumes that are given access to Fibre Channel attached host, the things are different than in SCSI. In SCSI, the LDs are assigned based on SCSI ports, independent of which hosts may be attached to those ports. So if you have multiple hosts attached to a single SCSI port (ESS supports up to four hosts per port), all of them will have access to the same LDs available on that port.

For Fibre Channel, the LD's affinity (LUN affinity) is based on the world wide port name (WWPN) of the adapter on the host, independent of to which ESS Fibre Channel host adapter port the host is attached (See Section 3.6.5.4, "FCP LUNs" on page 57 for further discussion on LUN affinity).

Figure 80, on page 128 shows the logical device (LD) assignment to the FB logical volumes (LV) when given access to Fibre Channel attached hosts.

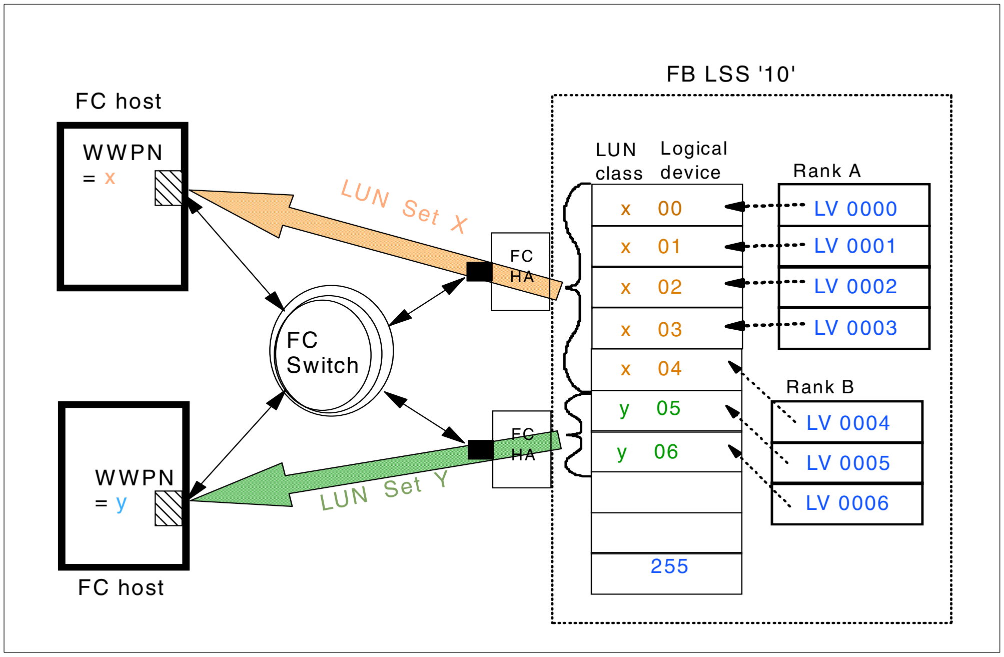


Figure 80. FB logical device mapping for Fibre Channel attached host

These FB logical volumes mapping to the host viewed logical devices (LDs) result from the definitions you do using the ESS Specialist, when initially working with *Add Fixed Block Volumes* panel. Previous to this step you already identified the host, its Fibre Channel attachment, and the Fibre Channel host adapter WWPN (World wide port name) when using the ESS Specialist *Modify Host System* panel.

Remember that Fibre Channel architecture allows any Fibre Channel initiator to access any open system logical device without access restrictions. You can restrict this access when the IBM SSR sets the access mode for your ESS during initial configuration (See Section 4.11.12, "Configuring SCSI and Fibre Channel host adapters" on page 124).

4.11.15 Defining CKD logical devices

The CKD logical volumes are mapped into a logical device map in a CKD LSS. Figure 81 shows that the logical devices in such an LSS represent the *device address* of the logical volume. It ranges from x'00' to x'FF'. Because each LSS is seen as a logical control unit, the z/Series systems will see it as a 3990-x with up to 256 devices. The logical devices need not be mapped to the ESCON or FICON host adapters, because the z/Series hosts have access to all the LDs through any of the ESCON or FICON connections available. The set of logical devices accessed by any S/390 image is defined with the HCD (Hardware Configuration Definition) in the IODF file that the operating system uses to recognize its hardware topology.

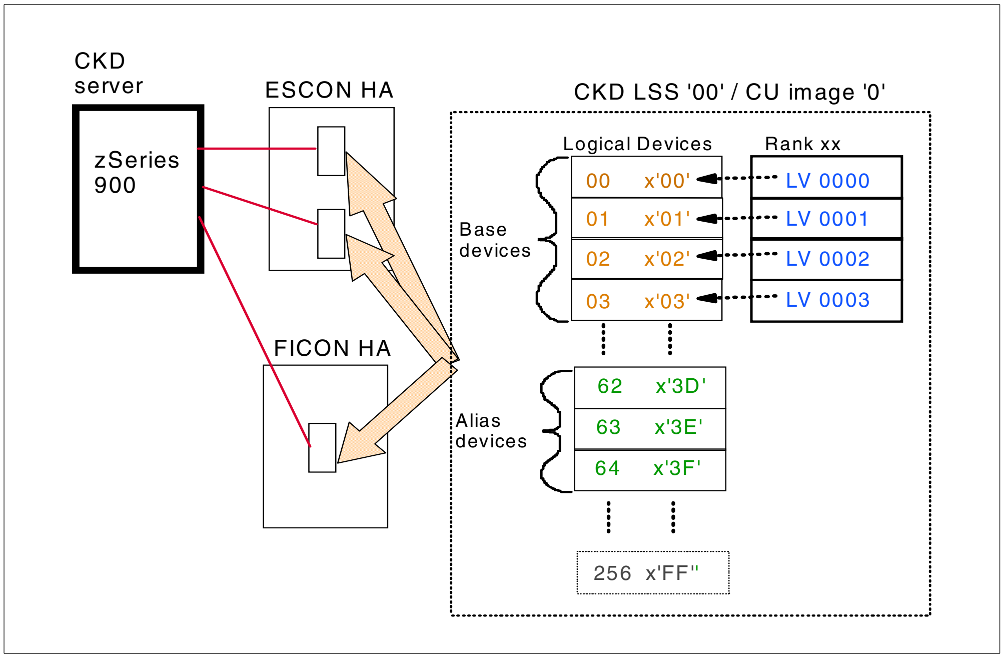


Figure 81. CKD logical device mapping

ESCON and FICON attached hosts are identified to the ESS when you make the physical connection between the hosts and the storage server. These hosts are seen as a single net in the ESS Specialist for improved graphical presentation.

4.11.16 Configuring CKD base/aliases

Since the ESS supports also alias addresses for *Parallel Access Volumes* (PAV) (see Section 5.2, "Parallel Access Volume (PAV)" on page 149 for details), you must specify two types of logical devices.

- Base Devices, for primary addressing from the host
- Alias Devices, as an alternate UCB to a base device

At least device x'00' must be a base device. The ESS is capable having up to 4096 (16 LSSs x 256 devices) devices configured. The ESCON channel can handle 1024 devices. And the FICON channel can handle 16 times 1024. Considering that the ESS gives you the capability of having larger volumes sizes (i.e., the 3390 model 9) for easier administration, without paying any penalty in response time, you may not need to address not even the 256 devices on a 3990

CU image. For this you can set an address range to 64, 128, or 256 devices for the CU images you are configuring. This will match your definition in the HCD for the operating system, and will also save space in the IODF.

Whenever you set an address range, you must understand that it will be used for both base and alias devices. The base devices are assigned from the lowest order address in the order of creation, which means that the last used device address will be assigned to the logical volume you are configuring. Alias devices are assigned from the highest device address available in the boundary. Base and alias addresses are defined both in the ESS Specialist and in the System z/ OS IOCP IODEVICE macro specifying UNIT=3390B and UNIT=3390A.

Figure 81 shows a CKD storage map and an example of how base devices and alias are mapped into a 64 address range boundary (when UNITADD=((00,64)) in the CNTLUNIT macro of the IOCP definition). The figure shows the 256 logical devices you have available in the LSS, but not been defined at the moment (*Note*: the intermixing of ESCON and FICON channels on one control unit is only supported for migration purposes. It is not a recommended configuration for the production environment).

The ESS and the ESS Specialist allow you to define from zero to 255 alias per base device address. The maximum devices (alias plus base) is 256 per LCU.

4.12 LSS/ranks configuration example

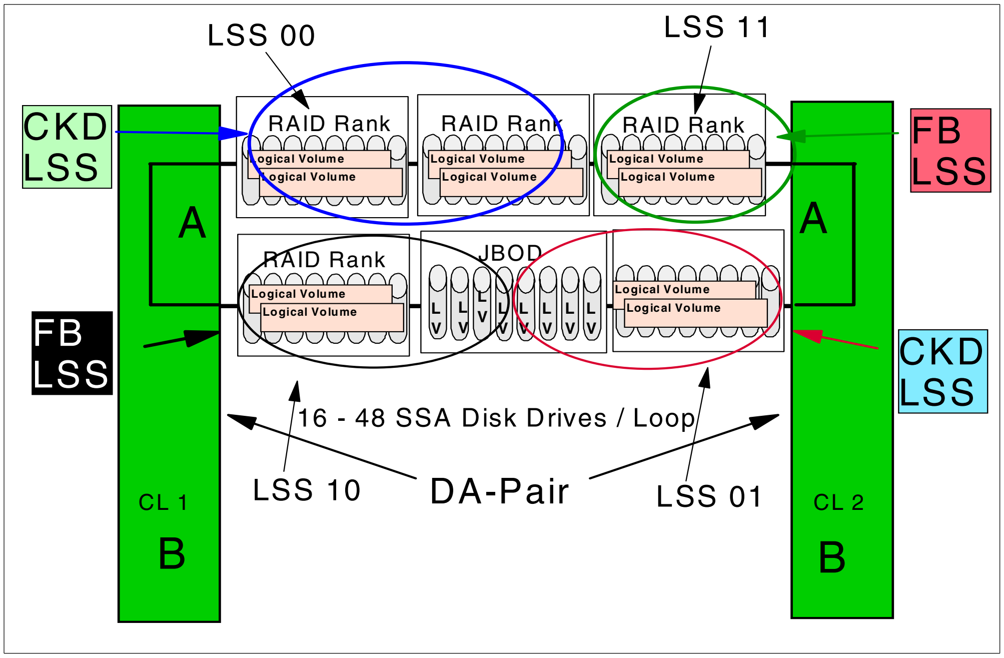


Figure 82. LSS ranks assignment

In the example in Figure 82, you can see how a final logical configuration of a loop may look. In this case, LOOP A has the maximum possible (48) drives installed. The loop has been configured with five RAID-5 ranks and eight JBOD ranks (Note: probably you will not be using JBOD ranks, but it is considered here for didactic purposes). A single DA pair loop can have up to four LSSs assigned to it, two CKD LSSs and two FB LSSs. Lets discuss this in more detail. Assuming that this example shows the first DA pair, then the LSSs defined are:

- DA CL1 Loop A: CKD LSS 00 (CU Image 0)
  - Two RAID ranks
- DA CL1 Loop A: FB LSS 10
  - One RAID rank and four JBOD ranks
- DA CL2 Loop A: CKD LSS 01 (CU Image 1)
  - One RAID rank and four JBOD ranks
- DA CL2 Loop A: FB LSS 11
  - One RAID rank

4.13 SCSI host connectivity

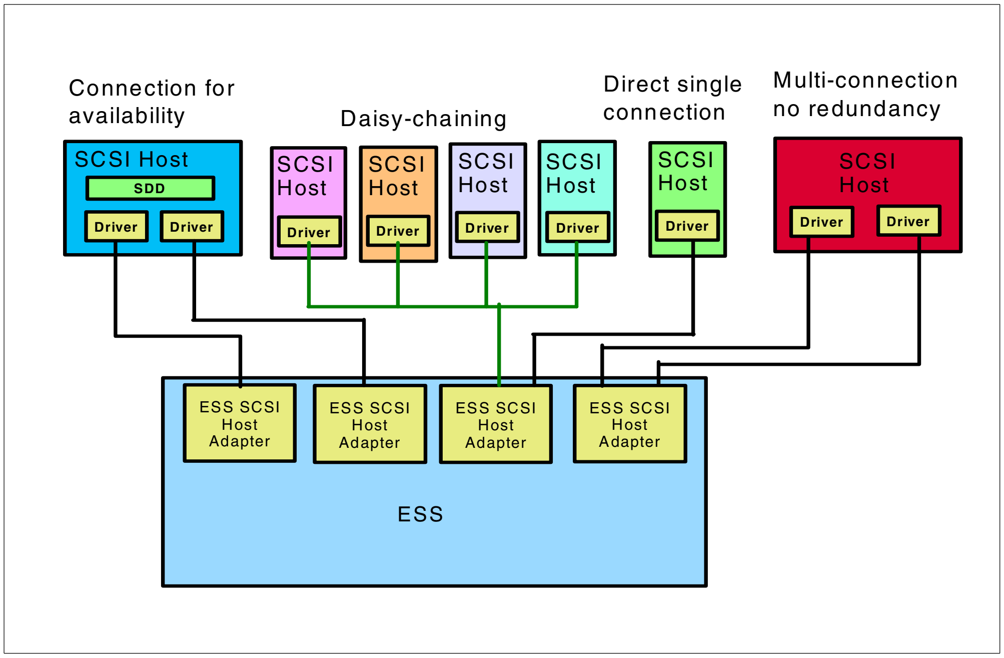


Figure 83. SCSI connectivity

4.13.1 Single host connection

Figure 83 shows different possibilities for attaching the SCSI hosts to the Enterprise Storage Server. The simplest of these possible attachments is the single host connected to only one SCSI host adapter port. In this type of connection, the server has only one path to its logical volumes in the ESS. In case the path fails, then the server loses all access to its data because no redundancy was provided.

4.13.2 SCSI connection for availability

For availability purposes, you can configure a logical device in the ESS as a shared device. To do that, you must assign it to two different SCSI ports in the ESS. This allows you to interconnect your host to two or more separate SCSI host adapter ports located on different bays, both seeing the same set of shared logical devices. You can use then IBM Subsystem Device Driver (IBM SDD, that comes standard with the ESS) to distribute the I/O activity among the SCSI adapters in the host and it will automatically recover I/Os that failed on the alternate path. This is valid for any cause of connection failures, such as SCSI interface failures, SCSI host adapter failures, or even ESS host adapter port failures. Another advantage of using the SDD is the capability of having some concurrent maintenance of the SCSI host adapter cards. In such case, SDD offers commands that allow you to deactivate the I/Os through a specific adapter and return it back to operation once the maintenance action has finished. One last consideration on the SDD benefits, is that it will automatically balance the I/O over the available paths improving the overall server I/O performance. See Appendix B, “Subsystem Device Driver (SSD)” on page 283 for more details on the Subsystem Device Driver.

4.13.3 Multi-connection without redundancy

Figure 83 also illustrates a multi-connection setup without path redundancy. This connection can be done having multiple SCSI adapters in the host and having each SCSI adapters connected to a different SCSI port in the ESS. But in this occasion, having no SDD software in the server to fail-over to the alternate path, in the event that one SCSI adapter fails then all the logical volumes associated with it become unavailable.

4.13.4 Daisy-chaining host SCSI adapters

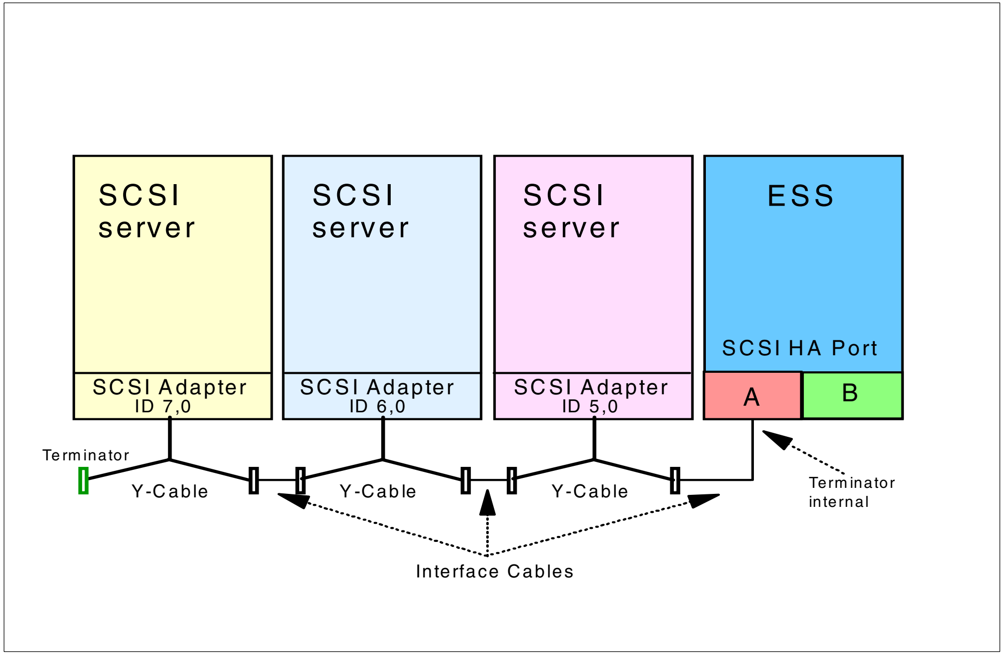


Figure 84. SCSI daisy-chain

As you can see from Figure 84, the Enterprise Storage Server allows daisy chaining of several host adapters. Although it is not the most efficient connection, whenever you need to do this, follow these rules:

- A maximum of four host initiators is recommended on a single ESS host adapter SCSI port. The SCSI ID priority order is 7 – 0 then 15 – 8. The first host system that you add is assigned to SCSI ID 7, the second is assigned to SCSI ID 6. You must verify that these assignments match the SCSI ID setting in each host system SCSI adapter card, and make adjustments to the map of SCSI IDs as necessary.
- You can daisy chain: up to four Data General Aviion; up to two HP-9000;up to four homogeneous NT hosts, when supported by the server; up to four pSeries or RS6000. The iSeries 400 do not allow daisy chaining with the adapters used to connect the external 9337 devices. This interface is not designed for that. Note: you should check with the server adapter provider for the support, since this is dependent on the server model.
- The SCSI adapters are daisy chained with Y-Cables. Both ends of the cables must be terminated. The ESS must be at one end of the interface, because it has internal terminators on the SCSI host adapter cards.

- Avoid mixing host SCSI adapters of different types in the chain. The best results are obtained when running the chain with the same type of adapter.
- The cables must be 2-byte differential SCSI cables and must match the requirements for the host SCSI adapters. For more details about the supported host SCSI adapters, see the Web site:

  `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`
- When more than one server is daisy chained from the ESS, the length of the cables in the chain must be added together and the sum must not exceed 25 meters (82 feet) This includes the length of the cable branches (Y-cables) to each server. Information on the Y-cables requirements should be referred to the provider of the server adapter to which the ESS will attach since daisy chaining support varies from vendor to vendor.
- Daisy chaining should be avoided because it creates an overhead of SCSI arbitration on the interface, which may result in performance degradation. Note that this is a SCSI limitation and not an ESS limitation.

Remember, when the ESS is daisy chained to multiple servers, all of these SCSI servers on the same bus can address any LUN (logical device), defined on that port of the ESS.

4.14 Fibre Channel host connectivity

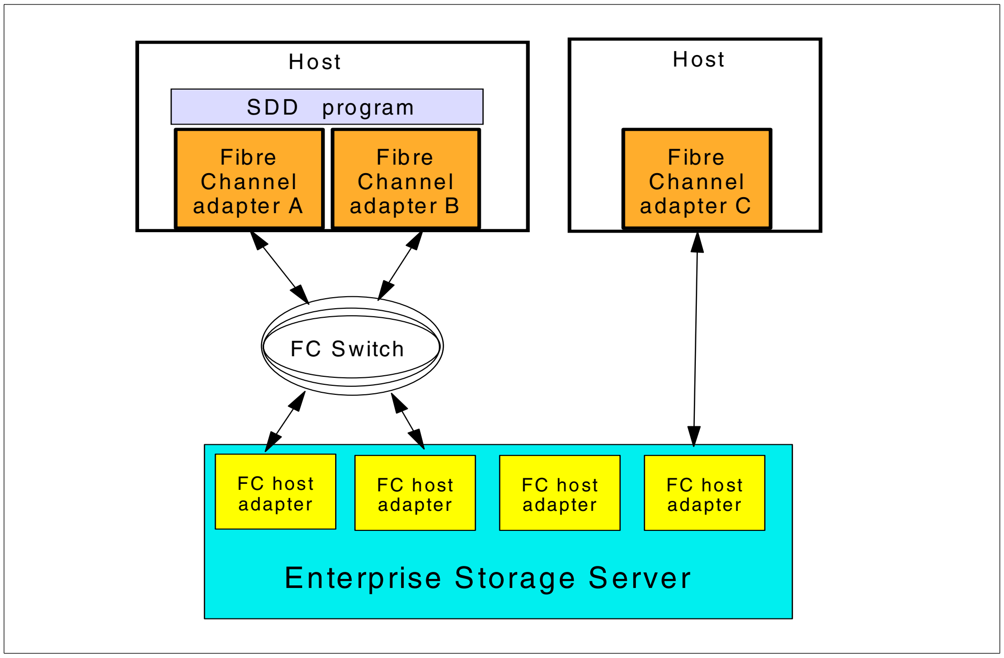


Figure 85. Fibre Channel connectivity

With Fibre Channel, the attachment limitations seen on SCSI in terms of distance, addressability and performance are now overcome. Fibre Channel is not just a replacement of the parallel SCSI by a serial based interface, it is more the ability to build Storage Area Networks (SANs) of interconnected host systems and storage servers.

4.14.1 Fibre Channel topologies

Three different topologies are defined in the Fibre Channel architecture. All of the three topologies are supported by the Enterprise Storage Server. The three topologies are discussed briefly below:

4.14.1.1 Point-to-point

This is the simplest of all the topologies. By using a fiber cable, two Fibre Channel adapters (one host and one ESS) are connected. Fibre Channel host adapter card C in Figure 85 is an example of a point-to-point connection. This topology supports the maximum bandwidth of Fibre Channel, but does not exploit all the benefits that come with SAN implementations. The distance between the host and the ESS can be up to 500 meters or 10 Km., respectively for short or for long wave host adapters.

4.14.1.2 Arbitrated Loop

Arbitrated loop is a uni-directional ring topology very much like token ring. Information is routed around the loop and repeated by intermediate ports until it arrives at its destination. If using this topology, all other Fibre Channel ports in the loop must be able to perform this routing and repeating functions in addition to all the functions required by the point-to-point ports. The acronym FC-AL refers to this topology. Up to a maximum of 127 FC ports can be interconnected via a looped interface. All ports share the FC-AL interface and therefore also share the bandwidth of the interface. Only one connection may be active at a time.

4.14.1.3 Switched fabric

Whenever a switch is used to interconnect Fibre Channel adapters, we have a switched fabric. A switched fabric is an intelligent switching infrastructure, which delivers data from any source to any destination. Figure 85, with Fibre Channel adapters A and B, shows an example of a switched fabric. This is where a SAN starts to build up. When using the IBM TotalStorage ESS Specialist to configure switched fabric, always use point-to-point in Fibre Channel port attributes. This is the protocol used in fabric.

An ESS Fibre Channel port is configured as either FC-AL or point-to-point using the ESS Specialist. The point-to-point connection may be to a server, or to a Fibre Channel switch (and to a Fibre Channel fabric). If an ESS is directly cabled to a fibre channel adapter on a RS/6000 host, the ESS port must be defined as FC-AL. It is not necessary to configure the server adapter for this. If an ESS is attached to a RS/6000 server through a Fibre Channel switch, then the ESS port must be defined as point-to-point.

Note: You may refer to the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcwp.pdf` for further information on Fibre Channel topologies.

4.14.2 Fibre Channel connection for availability

In Figure 85, the attachment for the host that holds the Fibre Channel adapters A and B does not alone provide for redundant access to the data. Besides configuring the LUNs to both Fibre Channel adapters, in order for the host to take advantage of both paths, you must also run the Subsystem Device Driver (SDD) program that is distributed with the ESS. This program runs in the host, and besides giving automatic switching between paths in the event of a path failure, it will also balances the I/O across them.

SDD supports up to 32 separate paths (See Appendix B, “Subsystem Device Driver (SSD)” on page 283 for further information).

4.15 ESCON host connectivity

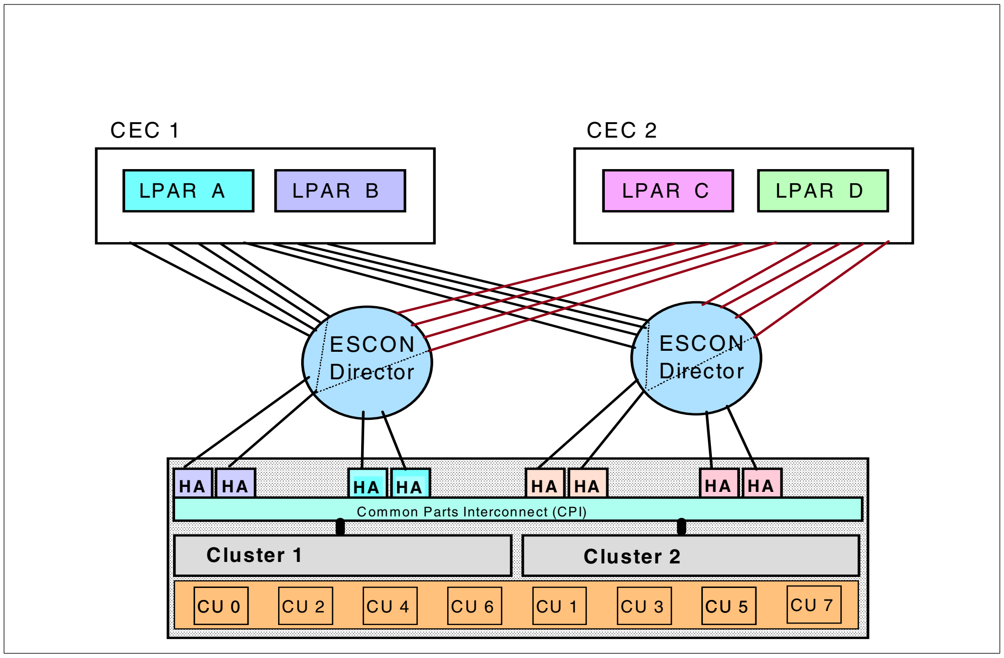


Figure 86. ESCON connectivity example

Figure 86 provides an example of how an Enterprise Storage Server can be attached through ESCON links to different CECs and LPARs. This diagram also considers availability. For the best availability, you should spread all ESCON host adapters through all available bays. These basic connectivity concepts remain similar for ESCON and FICON. What is not similar is the characteristics that they provide you, and the resulting attachment capabilities.

4.15.1 ESCON control unit images

The Enterprise Storage Server allows you to configure up to 16 LSSs that will represent up to 16 CU images (or logical control units, LCUs) in the ESS. The CU images will handle the following interconnections:

- Up to 256 devices (base and alias) per CU image.
- Up to 4096 devices (base and alias) on the 16 CU images of the ESS. This is the 16 LCUs, each capable of addressing 256 devices, makes a total of 4096 addressable devices within the ESS. But only a maximum of 1024 devices is addressable by each ESCON channel.
- Up to 128 logical paths, and up to 64 path groups, for each CU image.
- Total of 2048 logical paths in the ESS (128 logical paths per CU image, times 16 CU images, makes 2048 logical paths).

4.15.2 ESCON logical paths establishment

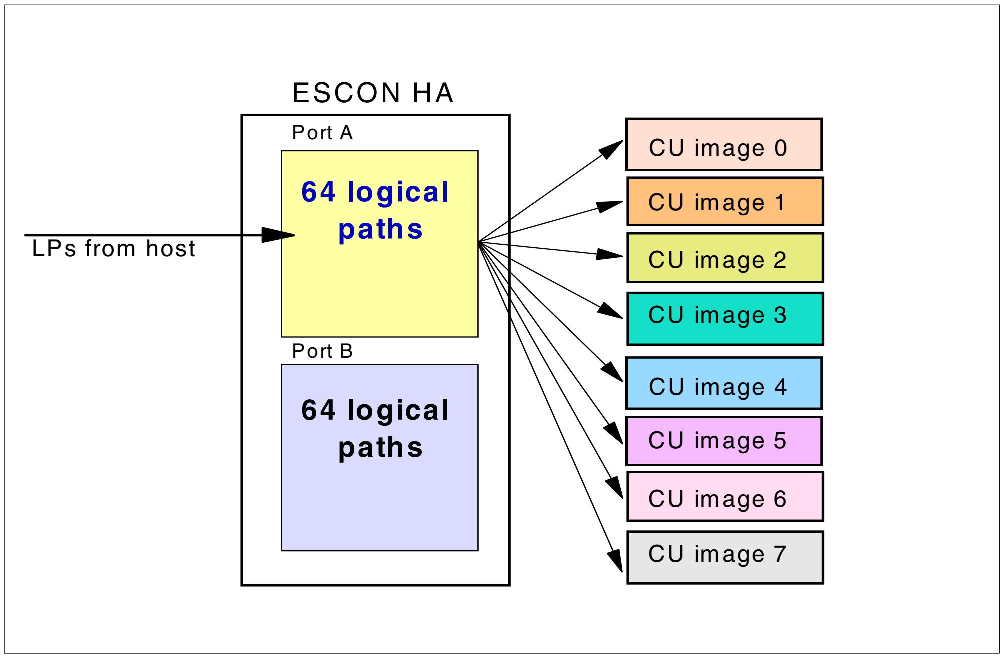


Figure 87. Establishment of logical paths for ESCON attachment

Figure 87 displays how logical paths are established in the ESS with the ESCON host adapters. An ESCON host adapter port will handle a maximum of 64 logical paths. This example shows a single port on an ESCON-HA card and an Enterprise Storage Server with 8 CU images (LSSs) configured.

4.15.3 Calculating ESCON logical paths

For the following explanation refer to Figure 86:

- The ESS is configured for 8 CU images.
- All 4 LPARs have access to all 8 CU images
- All LPARs have 8 pathing to each CU image.

This results in:

4 LPARs x 8 CU images = 32 Logical Paths

So 32 logical paths per ESCON adapter port, which does not exceed the 64 LPs per port ESCON maximum.

Under the same assumptions, each CU image must handle:

4 LPARs x 8 CHPIDs = 32 LPs

This will not exceed the 128 LPs a single ESS CU image can manage. These calculations may be needed if the user is running large sysplex environments. In such a case, it is also recommended to have many more channel paths attached to the ESS, to spread the CU images to several different channels.

4.16 FICON host connectivity

The FICON channel connectivity brings some differences and provides a list of benefits over the ESCON channel connectivity. Among the benefits you must consider:

- Addressing (from 1024 device addresses for ESCON to up to 16,384 for FICON).
- Reduced number of channels and required fibers with increased bandwidth and I/O rate per FICON channel.
- FICON channel to ESS multiple concurrent I/O connections capability (ESCON supports only one I/O connection at one time)
- Greater channel and link bandwidth: FICON has up to 10 times the link bandwidth of ESCON (1 Gigabit/second full duplex, compared to 200 megabits/second half duplex). FICON has up to more than 4 times the effective channel bandwidth for the initial implementation (70 MB/sec. compared to 17 MB/sec.)
- FICON path consolidation using switched point-to-point topology.
- Greater un-repeated fiber link distances (from 3 Km. for ESCON to up to 10 Km., or 20 Km. with an RPQ, for FICON).

The configuration shown in Figure 88 is an example of connectivity of an Enterprise Storage Server, using 8 FICON channel paths to half of the CU images of an ESS. You may refer to *FICON Native Implementation and Reference Guide,* SG24-6266 for more detailed information on ESS FICON connectivity.

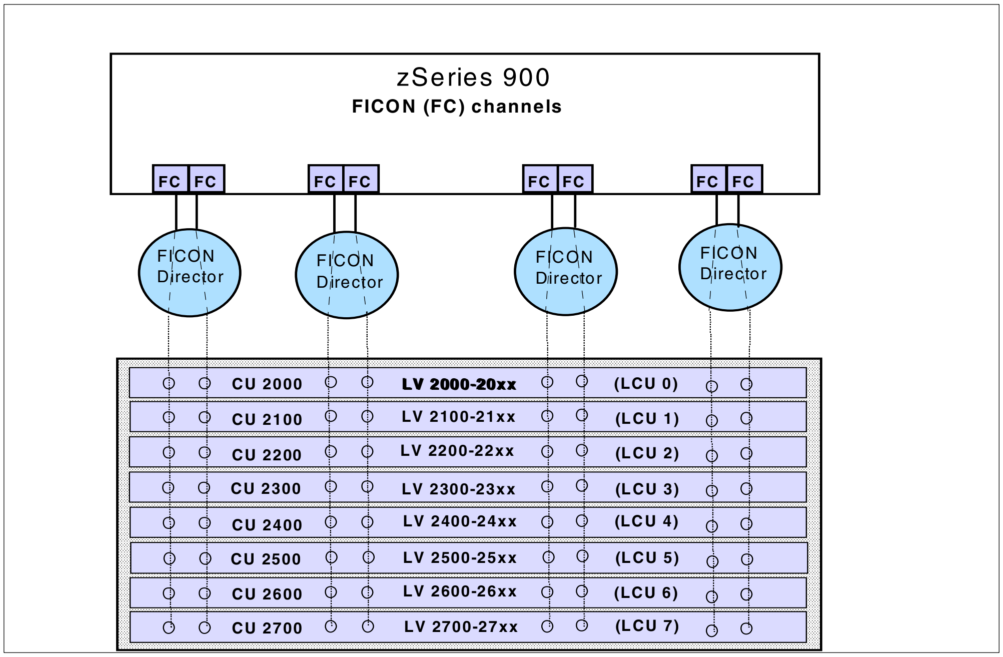

zSeries 900
FICON (FC) channels
FC FC
FC FC
FC FC
FC FC
FICON Director
FICON Director
FICON Director
FICON Director
CU 2000 LV 2000-20xx (LCU 0)
CU 2100 LV 2100-21xx (LCU 1)
CU 2200 LV 2200-22xx (LCU 2)
CU 2300 LV 2300-23xx (LCU 3)
CU 2400 LV 2400-24xx (LCU 4)
CU 2500 LV 2500-25xx (LCU 5)
CU 2600 LV 2600-26xx (LCU 6)
CU 2700 LV 2700-27xx (LCU 7)

Figure 88. FICON connectivity

4.16.1 FICON control unit images

The 16 LSSs (CU images, or LCUs) of the Enterprise Storage Server will handle the following interconnections:

- Up to 256 devices (base and alias) per CU image (same as ESCON).

- Up to 4096 devices (base and alias) on the 16 CU images of the ESS. This is the 16 LCUs, each capable of addressing 256 devices, makes a total of 4096 addressable devices within the ESS.
- Up to 128 logical paths for each CU image.
- Total of 2048 logical paths in the ESS (128 logical paths per CU image * 16 CU images = 2048).

When you plan your ESS connectivity layout you realize the dramatic benefits you get from the FICON implementation because of:

- The increased number of concurrent connections, and
- The increased channel device address supported by the FICON implementation.

Increased number of concurrent connections

FICON provides an increased number of channel-to-control unit concurrent I/O connections. ESCON supports one I/O connection at any one time while FICON channels support multiple concurrent I/O concurrent connections. While an ESCON channel can have only one I/O operation at a time, the FICON channel can have I/O operations to multiple LCUs at the same time, even to the same LCU, by using the FICON protocol frame multiplexing.

Increased channel device-address support

From 1,024 devices on an ESCON channel to 16,384 devices for a FICON channel. This makes possible to any FICON channel connected to the ESS, to address all the 4096 devices you can have within the ESS. This extra flexibility will simplify your configuration setup and management.

All these factors, plus the increased bandwidth, allow you to take advantage of FICON and allow you to layout more simple redundant configurations, accessing more data with even better performance than it is possible with ESCON. You may refer to the document at `http://www.storage.ibm.com/hardsoft/products/ess/support/essficonwp.pdf` for further considerations on FICON system attachment.

4.16.2 FICON logical path establishment

Figure 89 displays how logical paths are established in the Enterprise Storage Server with the FICON host adapters. The FICON host adapter port will handle a maximum of 256 logical paths. This example shows a single port on a FICON host adapter card and an ESS with 8 CU images (LSSs) configured.

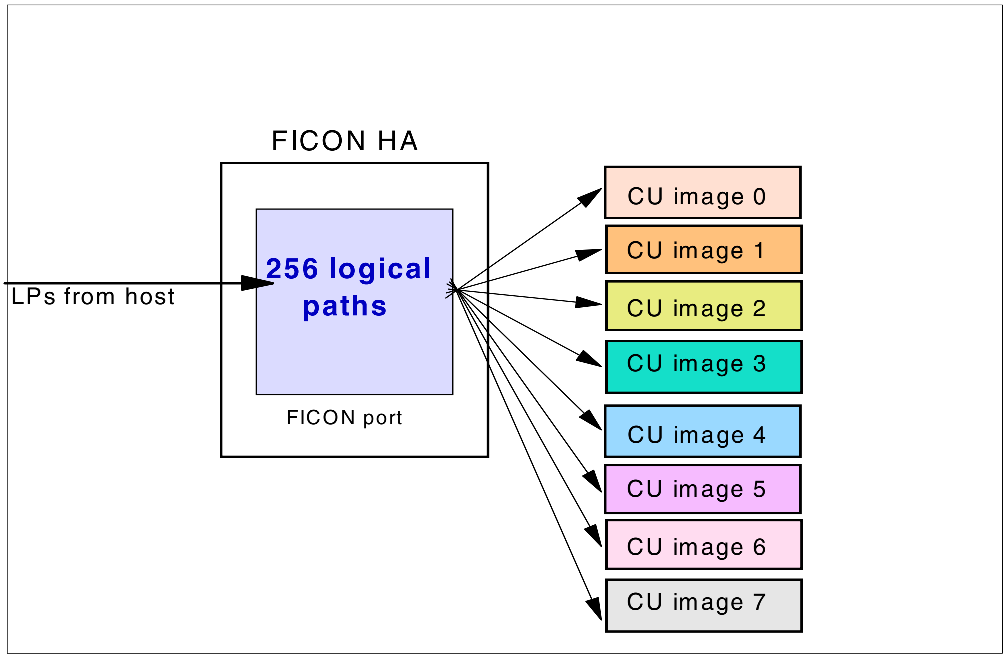


Figure 89. Establishment of logical paths for FICON attachment

4.16.3 Calculating FICON logical paths

For the following explanation refer to Figure 90. In the example there are two control units, each with eight FICON host adapters and eight logical control units configured. All adapters can be used by all logical control units within each ESS (conforming to the S/390 and z/Architecture maximum of eight paths from a processor image to a control unit image); and each logical control unit has 256 device addresses. All control unit FICON host adapters are connected through two FICON directors.

There are two FICON switches and each is connected to four FICON channels of each CEC, resulting in16 ports on each switch for channel connectivity. Each of the FICON switches is also connected to four adapters of each ESS, resulting in 24 ports in total for each switch.

The resources used by this example configuration are:

- 8 FICON channels per CEC
- 24 FICON ports per switch (16 for channels, and 8 for ESS host adapters)
- 4,096 subchannels per processor image.

  Two ESSs, each one configured with eight logical control units, and 256 devices (the maximum) per logical control unit, makes 4,096 subchannels per processor image (2 x 8 x 256 = 4,096). The maximum number of subchannels per image is CEC dependent.

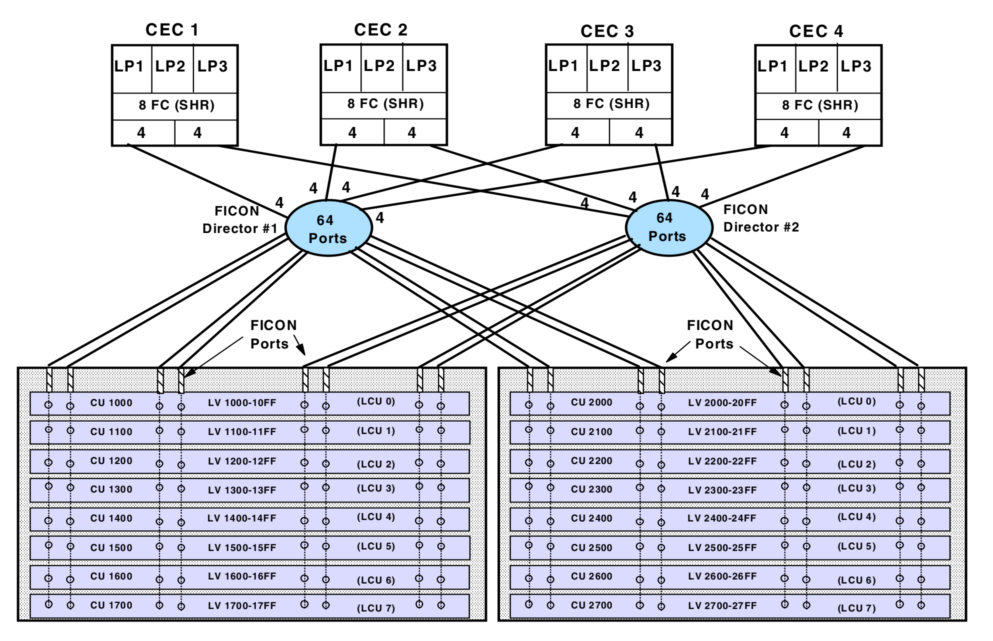

CEC 1
LP1 LP2 LP3
8 FC (SHR)
4 4
CEC 2
LP1 LP2 LP3
8 FC (SHR)
4 4
CEC 3
LP1 LP2 LP3
8 FC (SHR)
4 4
CEC 4
LP1 LP2 LP3
8 FC (SHR)
4 4
FICON Director #1
64 Ports
FICON Director #2
64 Ports
FICON Ports
FICON Ports
CU 1000 LV 1000-10FF (LCU 0)
CU 1100 LV 1100-11FF (LCU 1)
CU 1200 LV 1200-12FF (LCU 2)
CU 1300 LV 1300-13FF (LCU 3)
CU 1400 LV 1400-14FF (LCU 4)
CU 1500 LV 1500-15FF (LCU 5)
CU 1600 LV 1600-16FF (LCU 6)
CU 1700 LV 1700-17FF (LCU 7)
CU 2000 LV 2000-20FF (LCU 0)
CU 2100 LV 2100-21FF (LCU 1)
CU 2200 LV 2200-22FF (LCU 2)
CU 2300 LV 2300-23FF (LCU 3)
CU 2400 LV 2400-24FF (LCU 4)
CU 2500 LV 2500-25FF (LCU 5)
CU 2600 LV 2600-26FF (LCU 6)
CU 2700 LV 2700-27FF (LCU 7)

Figure 90. FICON connectivity example

- 12,288 subchannels per CEC.

  Three images per CEC, each with 4,096 subchannels to access all the devices in both ESSs, makes 12,288 subchannels per CEC (3 x 4,096 = 12,288). Note that this number does not take into consideration any PAV alias devices that may be configured. Each PAV alias device requires a subchannel.

  The maximum number of subchannels per CEC, is CEC dependent.

- 4,096 subchannels per FICON channel

  As each FICON channel is connected to all eight logical control units on both ESSs, and each logical control unit has 256 devices configured (the maximum), the number of subchannels per FICON channel is 4,096 (2 x 8 x 256 = 4,096).

  The maximum number of devices per FICON channel is 16,384.

- 4 Fibre Channel N_Port logins, per FICON host adapter

  There are four CECs and all control unit host adapters are accessed by a channel to all CECs, so there are four N_Port logins per ESS host adapter.

  The maximum number of N_Port logins is control unit dependent. For the ESS this is 128 N_Port logins per FICON host adapter. This means that the maximum number of FICON channels that can be attached to a FICON port (using a switch) is 128.

- 96 logical paths per FICON host adapter

  There are 12 images in total, and each image has eight logical paths through each of the FICON host adapters (one logical path per logical control unit within the ESS). This makes 96 logical paths per FICON host adapter in the configuration example (12 x 8 = 96)

The maximum number of logical paths per host adapter is control unit dependent. For the ESS this is 256 for FICON attachment (vs. 64 for ESCON).

- 96 logical paths per logical control unit

  There are eight paths per logical control unit to each processor image. In all four CECs there are 12 images, so there are 96 (8 x 12) logical paths per logical control unit.

  The maximum number of logical paths per logical control unit is control unit dependent. For the ESS this is 128.

  So the example ESS configuration shown in Figure 90 is within the FICON resources limit, and you can realize that it requires less channel and connectivity resources than an equivalent ESCON configuration would require.

FICON resources exceeded

When planning for your configuration you must take care of not over-defining the configuration. One of the most common situations is to over-define the number of logical paths per logical control unit. For the ESS you cannot have more than 128 logical paths *online* to any logical control unit. Any attempt to vary online more logical paths will fail.

The problem with the over-defined configuration may not surface until an attempt is made to vary online paths to the devices beyond the already established limit of logical paths for the logical control unit or host adapter.

The following message is issued to reject the vary path processing:

```
VARY PATH(dddd,cc), ONLINE
IEE714I PATH(dddd,cc) NOT OPERATIONAL
```

Figure 91. Over-defined paths — system message

Note that there is no additional indication of the cause of the not operational condition. For this situation you can run the ICKDSF logical path report to identify which channel images have established logical paths to the logical control unit. For more detailed information on running and interpreting the ICKDSF logical paths report, you may refer to the FICON problem determination chapter in *FICON Native Implementation and Reference Guide,* SG24-6266.

4.17 ESCON and FICON connectivity intermix

Intermixing ESCON channels and FICON native channels to the same CU from the same operating system image is supported as a transitional step for migration only.

Intermixing ESCON channels, FICON Bridge channels and FICON native channels to the same control unit from the same processor image is also supported, either using point-to-point, switched point-to-point or both. IBM recommends that FICON native channel paths only be mixed with ESCON channels and FICON Bridge channel paths to ease migration from ESCON channels to FICON channels using dynamic I/O configuration.

The coexistence is very useful during the transition period from ESCON to FICON channels. The mixture allows you to dynamically add FICON native channel paths to a control unit while keeping its devices operational. A second dynamic I/O configuration change can then remove the ESCON channels while keeping devices operational. The mixing of FICON native channel paths with native ESCON and FICON Bridge channel paths should only be for the duration of the migration to FICON.

4.18 Standard logical configurations

There are two ways that the Enterprise Storage Server can be configured when it is first installed - just using the ESS Specialist, or using the ESS Batch Configuration Tool and the ESS Specialist. The latter is a way to configure the ESS in a simplified fashion which utilizes standard configurations and standard volume sizes. This tool is designed to perform the initial configuration of the ESS. Subsequent changes to the configuration should be done with the ESS Specialist although the ESS Batch Configuration Tool can be used for previously unallocated (unassigned) 8-packs on an SSA adapter card. The ESS Batch Configuration Tool can also be used where the data on the 8-pack does not need to be preserved. The ESS Batch Configuration Tool is a tool that is run by the IBM System Support Representative. The ESS Specialist is available for anyone to use.

These standard logical configuration options will configure the partition with one standard logical volume or LUN size. The logical configuration options, which are available for all the supported platforms, are:

- For CKD servers:
  - 3390-3 in interleaved mode. If you have the PAV feature code installed, one PAV alias is defined for each 3390-3.
  - 3390-9 in interleaved mode. If you have the PAV feature code installed, three aliases are defined for each 3390-9.
- For iSeries 400:
  - 9337-590 (8.59 GB LUNs)
- For FB servers:
  - 4 GB LUNs
  - 8 GB LUNs
  - 16 GB LUNs
  - Maximum array size LUN

  The size of the logical device defined does not generally have an impact on performance of the subsystem. The ESS does not serialize I/O on the basis of logical devices.

These standard options speed the logical configuration process, and the only setup you must do is the assignment to the host adapter ports, which is a quick process. The effective capacity of each standard configuration depends on the disk array capacity.

Please refer to the *IBM Enterprise Storage Server Configuration Planner,* SC26-7353*.*

Chapter 5. Performance

The Enterprise Storage Server is an unparalleled performing storage subsystem.

The ESS design based upon the Seascape Architecture uses the latest IBM technology that includes advanced RISC microprocessors, Serial Storage Architecture disk adapters, high performance disk drives and microcode intelligent algorithms. These features gathered together under a superb architectural design, deliver the best performance you could presently expect from a storage server solution.

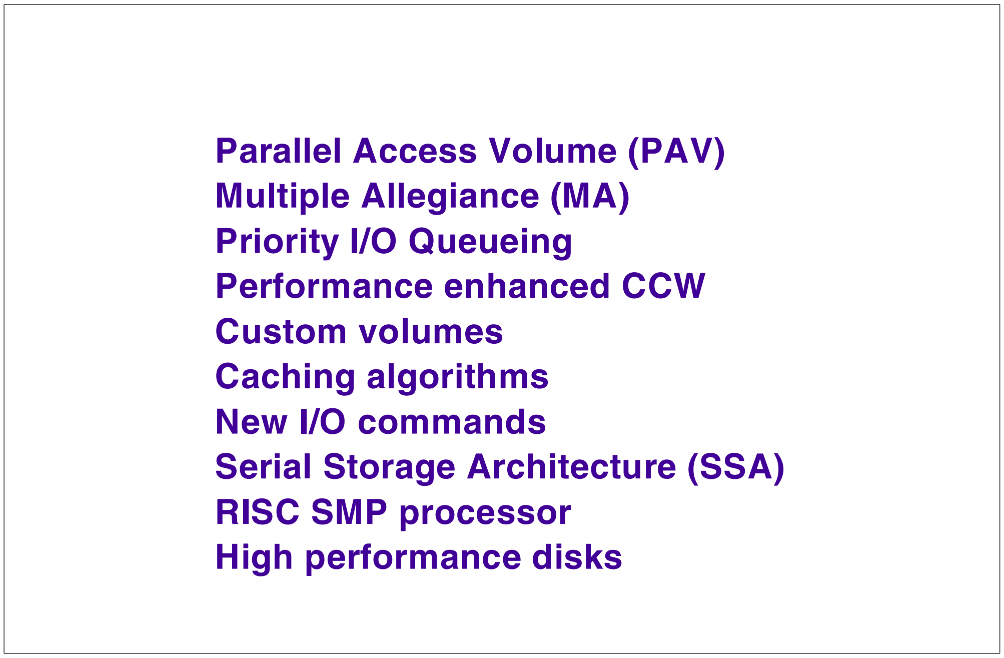
The ESS also delivers a multi-feature synergy with the zSeries server operating systems, allowing an unprecedented breakthrough in performance for those environments.

In this chapter we will look at these characteristics, that put the ESS as the performance leader for the disk storage solutions in the market, for all its attachable platforms

5.1 Performance accelerators

The Enterprise Storage Server has an ample repertoire of features and innovative performance functions. Some of these features and functions are common to all the platforms to which the ESS can attach. Others are platform specific.

Figure 92 lists the performance accelerator features and functions that, tuned-up with the sophisticated Seascape overall design, make the ESS the leader in performance for all the heterogeneous environments it attaches.



Parallel Access Volume (PAV)
Multiple Allegiance (MA)
Priority I/O Queueing
Performance enhanced CCW
Custom volumes
Caching algorithms
New I/O commands
Serial Storage Architecture (SSA)
RISC SMP processor
High performance disks

Figure 92. ESS performance accelerators

5.2 Parallel Access Volume (PAV)

Parallel Access Volume (PAV) is one of the exclusive features that the Enterprise Storage Server brings specifically for the z/OS and OS/390 operating systems, helping the zSeries running applications to concurrently share the same logical volumes.

The ability to do multiple I/O requests to the same volume nearly eliminates IOSQ time, one of the major components in z/OS response time. Traditionally, access to highly active volumes has involved manual tuning, splitting data across multiple volumes, and more. With PAV and the Workload Manager, you can almost forget about manual performance tuning. WLM manages PAVs across all the members of a sysplex too. The ESS in conjunction with z/OS has the ability to meet the performance requirements by its own.

Before we see PAV in detail, let us review how z/OS traditionally behaves when more than one application need concurrently to access a logical volume.

5.2.1 Traditional z/OS behavior

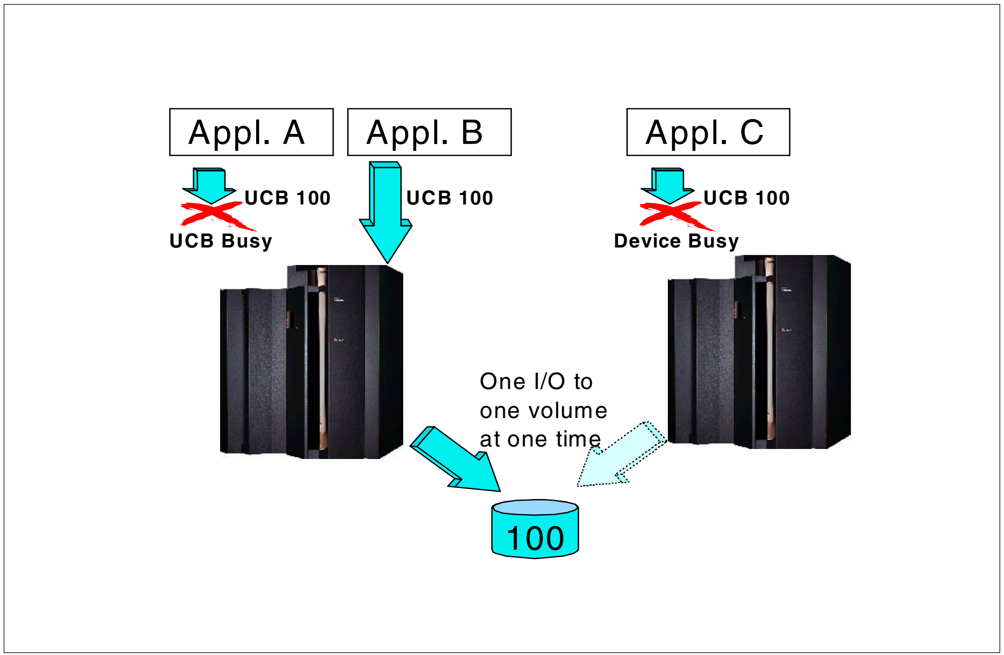


Figure 93. Traditional z/OS behavior

Traditional DASD subsystems (here, we use the term DASD, Direct Access Storage Device, instead of logical volume, since the term DASD is more common between the zSeries users) have allowed for only one channel program to be active to a DASD volume at a time, in order to ensure that data being accessed by one channel program cannot be altered by the activities of some other channel program.

From a performance standpoint, it did not make sense to send more than one I/O at a time to the storage subsystem, because the DASD hardware could process only one I/O at a time.

Knowing this, the z/OS systems did not try to issue another I/O to a DASD volume—represented in z/OS by a Unit Control Block (UCB)—while an I/O was already active for that volume, as indicated by a UCB busy flag (see Figure 93).

Not only were the z/OS systems limited to processing only one I/O at a time, but also, the storage subsystems accepted only one I/O at a time from different system images to a shared DASD volume, for the same reasons mentioned above.

5.2.2 Parallel I/O capability

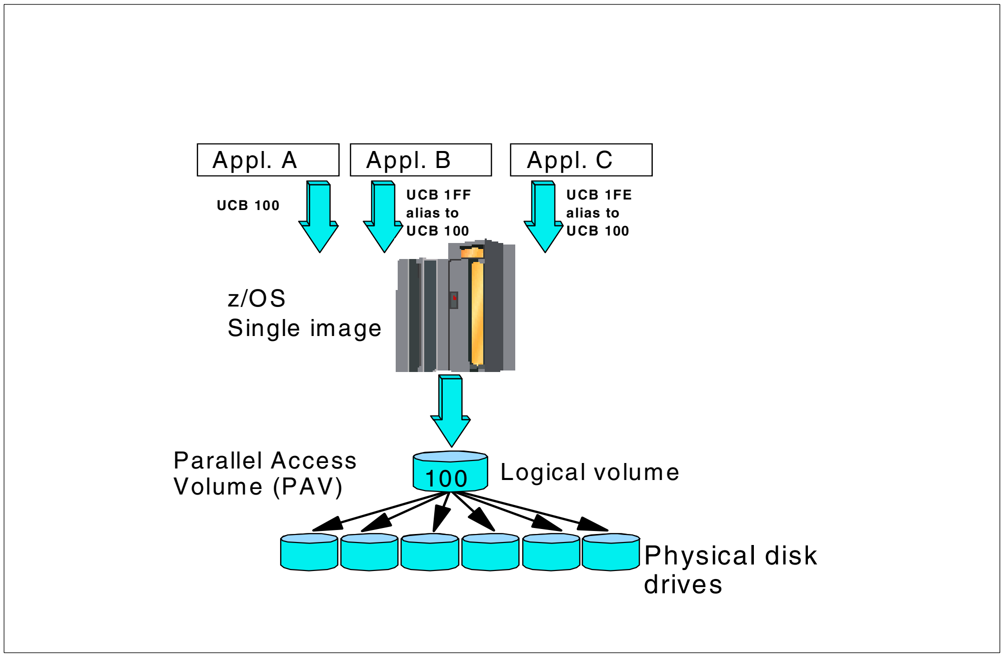


Figure 94. Parallel I/O capability using PAV

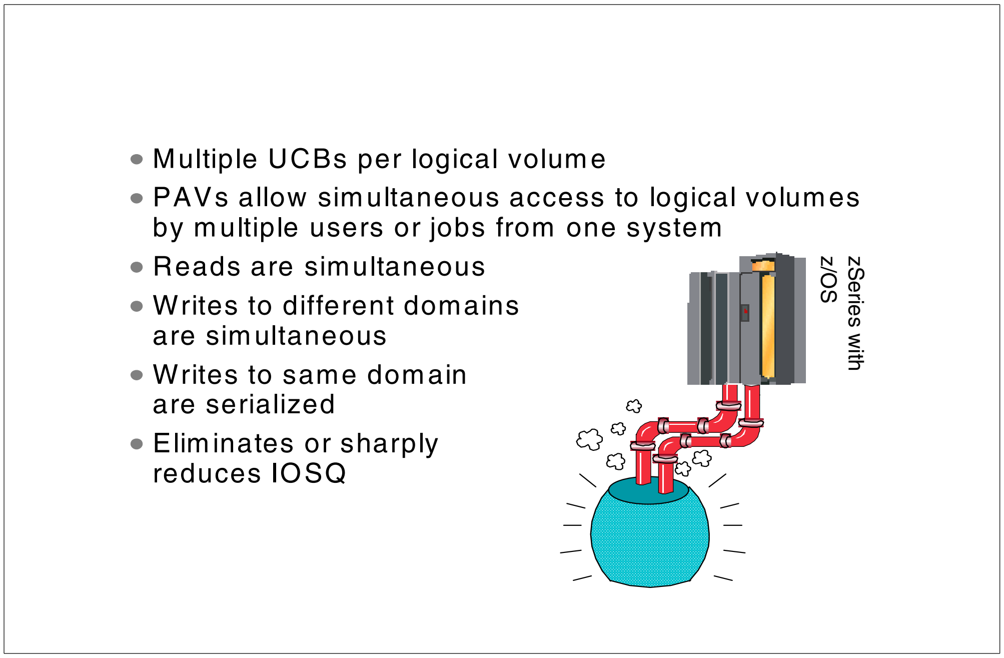
The Enterprise Storage Server is a modern storage subsystem with large cache sizes and disk drives arranged in RAID 5 arrays. Cache I/O is much faster than disk I/O, no mechanical parts (actuator) are involved. And I/Os can take place in parallel, even to the same volume. This is true for reads, and it is also possible for writes, as long as different extents on the volume are accessed.

The Enterprise Storage Server emulates zSeries ECKD volumes over RAID-5 disk arrays. While the zSeries operating systems continue to work with these logical DASD volumes as a unit, its tracks are spread over several physical disk drives. So parallel I/O from different applications to the same logical volume would also be possible for cache misses (when the I/Os have to go to the disk drives, involving mechanical movement of actuators) as long as the logical tracks are on different physical disk drives.

The ESS has the capability to do more than one I/O to an emulated CKD volume. The ESS introduces the new concept of *alias* address*,* in addition to the conventional *base* address. This allows that instead of one UCB per logical volume, a z/OS host can now use several UCBs for the same logical volume. For instance base address 2C00 may have now alias addresses 2CFD, 2CFE and 2CFF. This allows for 4 parallel I/O operations to the same volume.

This feature, that allows parallel I/Os to a volume from one host, is called *Parallel Access Volumes* (PAV).

5.2.3 Benefits of Parallel Access Volume



- Multiple UCBs per logical volume
- PAVs allow simultaneous access to logical volumes by multiple users or jobs from one system
- Reads are simultaneous
- Writes to different domains are simultaneous
- Writes to same domain are serialized
- Eliminates or sharply reduces IOSQ

Figure 95. Parallel Access Volumes (PAV)

z/OS systems queue I/O activity on a Unit Control Block (UCB) that represents the logical device. High I/O activity, particularly to large volumes (3390 model 9), could adversely affect performance, because the volumes are treated as a single resource and serially reused. This could result in large IOSQ times piling up (IOSQ shows these queued I/Os average waiting time). This is because traditionally the operating system does not attempt to start more than one I/O operation at a time to the logical device. The I/O request will be enqueued until the device becomes is no more busy with other I/O.

The Enterprise Storage Server is capable of holding parallel I/Os with PAV. The definition and exploitation of ESS's Parallel Access Volumes requires the operating system software support to define and manage *alias* device addresses and sub channels. When the operating system is at this level of software, it can issue multiple channel programs to a volume, allowing simultaneous access to the logical volume by multiple users or jobs. Reads can be satisfied simultaneously, as well as writes to different domains. The domain of an I/O consists of the specified extents to which the I/O operation applies. Writes to the same domain still have to be serialized to maintain data integrity. Older versions of operating systems that may not have the required software level continue to access the volume one-I/O-at-a-time for all kinds of reads or writes.

ESS's parallel I/O capability can drastically reduce or eliminate IOSQ time in the operating system, allowing for much higher I/O rates to a logical volume, and hence increasing the overall throughput of an ESS (see Figure 96). This cooperation of IBM’s Enterprise Storage Server and IBM’s system software provides additional value to your business.

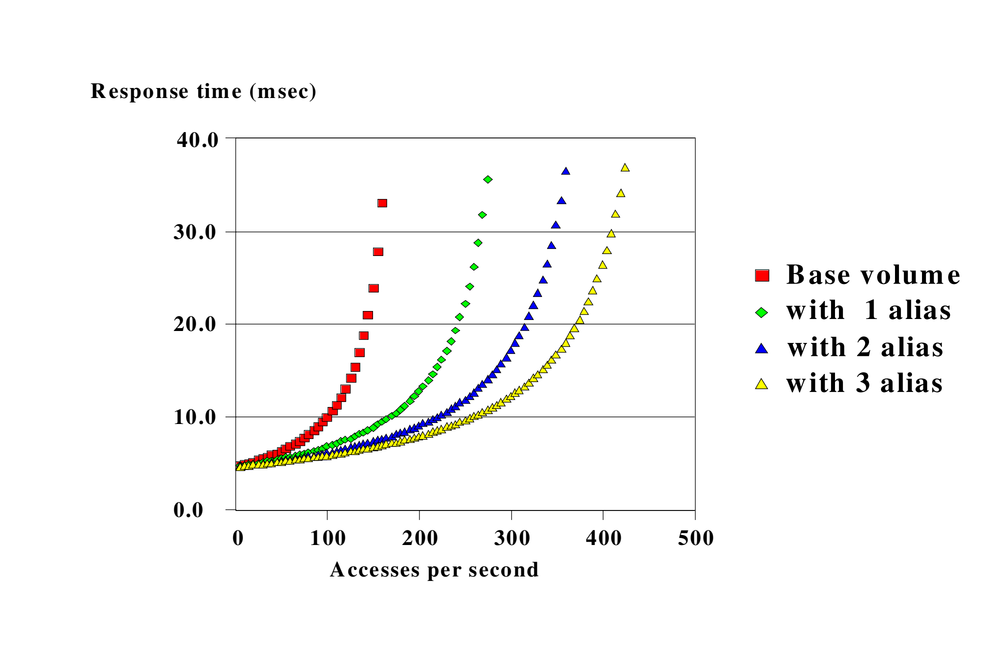
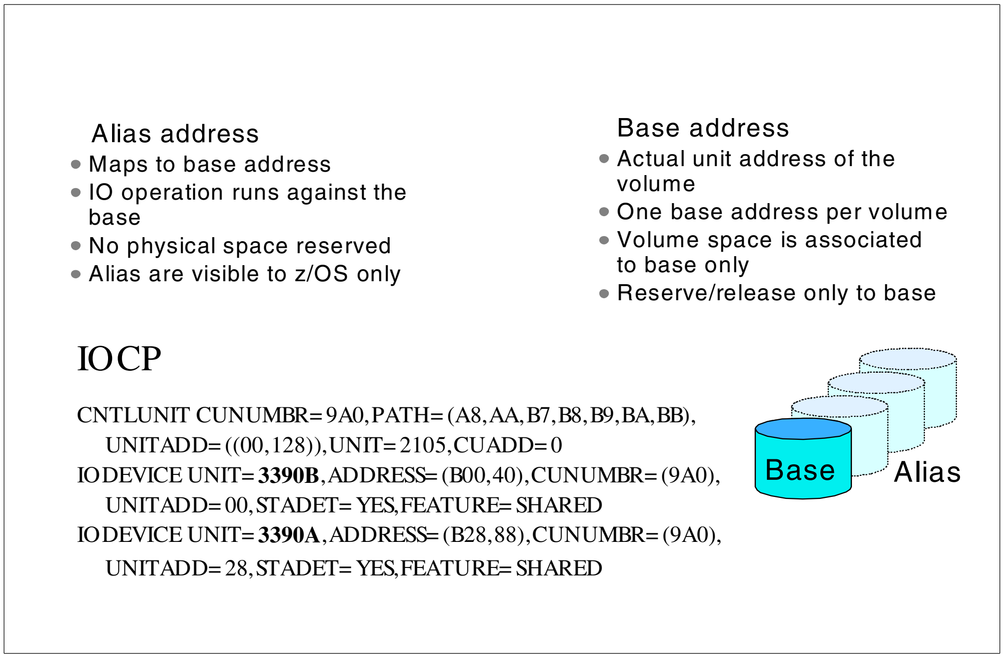


Figure 96. Potential performance impact of PAV

5.2.4 PAV base and alias addresses



Alias address

- Maps to base address
- IO operation runs against the base
- No physical space reserved
- Alias are visible to z/OS only

Base address

- Actual unit address of the volume
- One base address per volume
- Volume space is associated to base only
- Reserve/release only to base

IOCP

```
CNTLUNIT CUNUMBR=9A0,PATH=(A8,AA,B7,B8,B9,BA,BB),
    UNITADD=((00,128)),UNIT=2105,CUADD=0
IODEVICE UNIT=3390B,ADDRESS=(B00,40),CUNUMBR=(9A0),
    UNITADD=00,STADET=YES,FEATURE=SHARED
IODEVICE UNIT=3390A,ADDRESS=(B28,88),CUNUMBR=(9A0),
    UNITADD=28,STADET=YES,FEATURE=SHARED
```

Figure 97. PAV base and alias addresses

The Enterprise Storage Server implementation of PAV introduces two new unit address types: *base* address and *alias* address. Several aliases can be assigned to one base address. Therefore, there can be multiple unit addresses (and hence UCBs) for a volume in z/OS. z/OS can use all these addresses for I/Os to that logical volume.

5.2.4.1 Base address

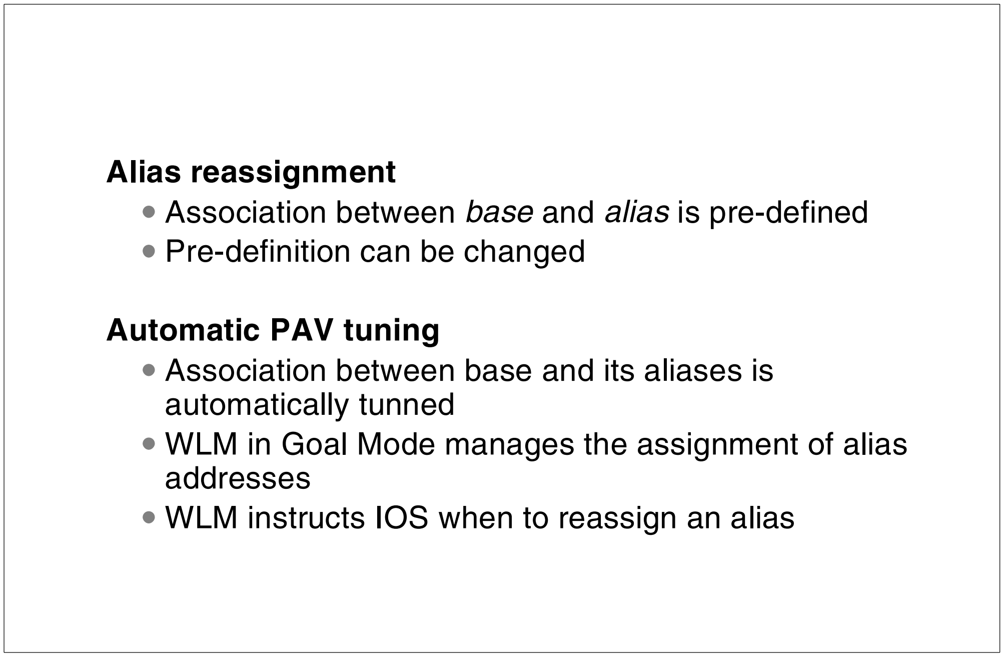
The base address is the conventional unit address of a logical volume. There is only one base address associated with any volume. Disk storage space is associated with the base address. In commands, where you deal with unit addresses—for example, when you set up a PPRC pair—you use the base address.

5.2.4.2 Alias address

An alias address is mapped to a base address. I/O operations to an alias address run against the associated base address storage space. There is no physical space associated with an alias address. You can define more than one alias per base. The ESS Specialist allows you to define up to 255 aliases per base, and the maximum devices (alias plus base) is 256 per LCU (see Section 4.11.16, “Configuring CKD base/aliases” on page 129). Alias addresses are visible only to the I/O Subsystem (IOS). Alias UCBs have the same memory storage requirements as base addresses.

Alias addresses have to be defined to the Enterprise Storage Server and to the zSeries host IODF file. The number of base and alias addresses in both definitions, in the ESS and in the IODF, must match. Base and alias addresses are defined to the host system using the IOCP IODEVICE macro specifying UNIT=3390B and UNIT=3390A respectively (see Figure 97).

5.2.5 PAV tuning



Alias reassignment

- Association between *base* and *alias* is pre-defined
- Pre-definition can be changed

Automatic PAV tuning

- Association between base and its aliases is automatically tunned
- WLM in Goal Mode manages the assignment of alias addresses
- WLM instructs IOS when to reassign an alias

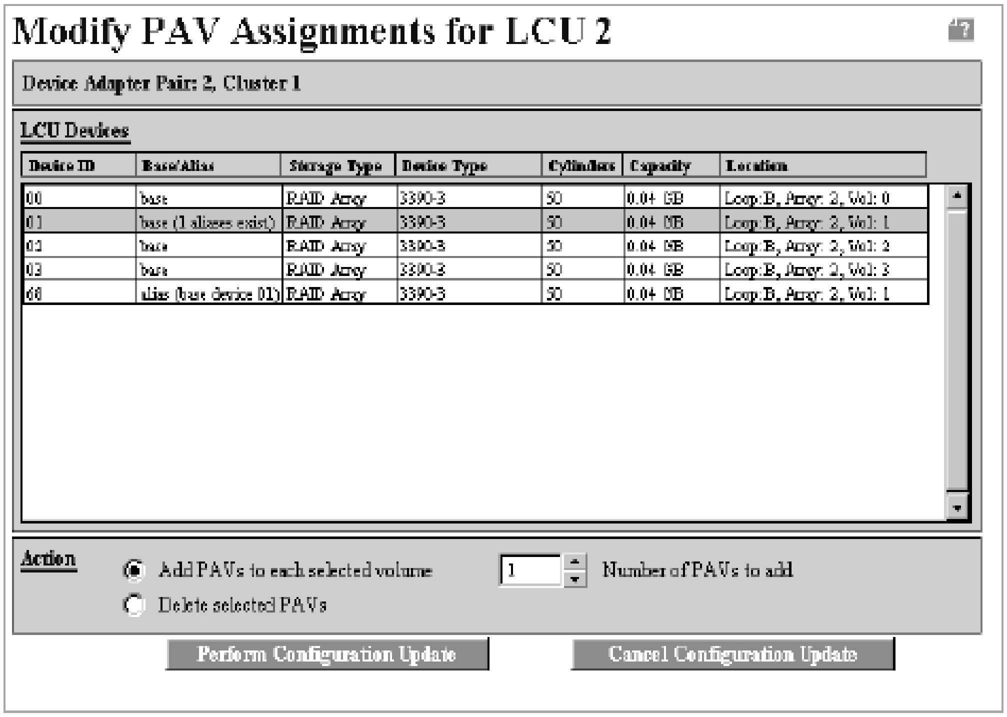
Figure 98. PAV tuning

The association between base addresses and alias addresses is pre-defined in the ESS using of the ESS Specialist. Adding new aliases can be done non-disruptively. The ESS Specialist allows the definition of 0 to 255 aliases per base. So you can have any combination on the number of alias and base for a given LSS, within the 256 device number limit of the LCU.

The association between base and alias addresses is not fixed. Alias addresses can be assigned to different base addresses by the z/OS Workload Manager. If the Workload Manager is not been used then the association becomes a static definition. This means that for each base address you define, the quantity of associated alias addresses will be what you already specified in the IODF and the ESS Specialist when doing the logical configuration procedure.

Automatic PAV tuning
It will not always be easy to predict which volumes should have an alias address assigned, and how many. If your software is at the right level, however, it can automatically manage the aliases according to your goals. z/OS can exploit automatic PAV tuning if you are using the Workload Manager (WLM) in Goal Mode. The WLM can dynamically tune the assignment of alias addresses.The Workload Manager monitors the device performance and is able to dynamically reassign alias addresses from one base to another if predefined goals for a workload are not met. The WLM instructs the IOS to reassign an alias.

5.2.6 Configuring PAVs



Modify PAV Assignments for LCU 2

Device Adapter Pair: 2, Cluster 1

LCU Devices

Device ID	Base/Alias	Storage Type	Device Type	Cylinders	Capacity	Location
00	base	RAID Array	3390-3	50	0.04 GB	Loop:B, Array: 2, Vol: 0
01	base (1 aliases exist)	RAID Array	3390-3	50	0.04 GB	Loop:B, Array: 2, Vol: 1
02	base	RAID Array	3390-3	50	0.04 GB	Loop:B, Array: 2, Vol: 2
03	base	RAID Array	3390-3	50	0.04 GB	Loop:B, Array: 2, Vol: 3
68	alias (base device 01)	RAID Array	3390-3	50	0.04 GB	Loop:B, Array: 2, Vol: 1

Action

● Add PAVs to each selected volume   1   Number of PAVs to add

○ Delete selected PAVs

Figure 99. Modify PAV Assignments panel

Before PAVs can be used, they must first be defined to the ESS using the ESS Specialist (see Figure 99), and to the host IODF file using the Hardware Configuration Definition (HCD) utility. Both, the ESS and the z/OS HCD definitions must match, otherwise you will get an error message.

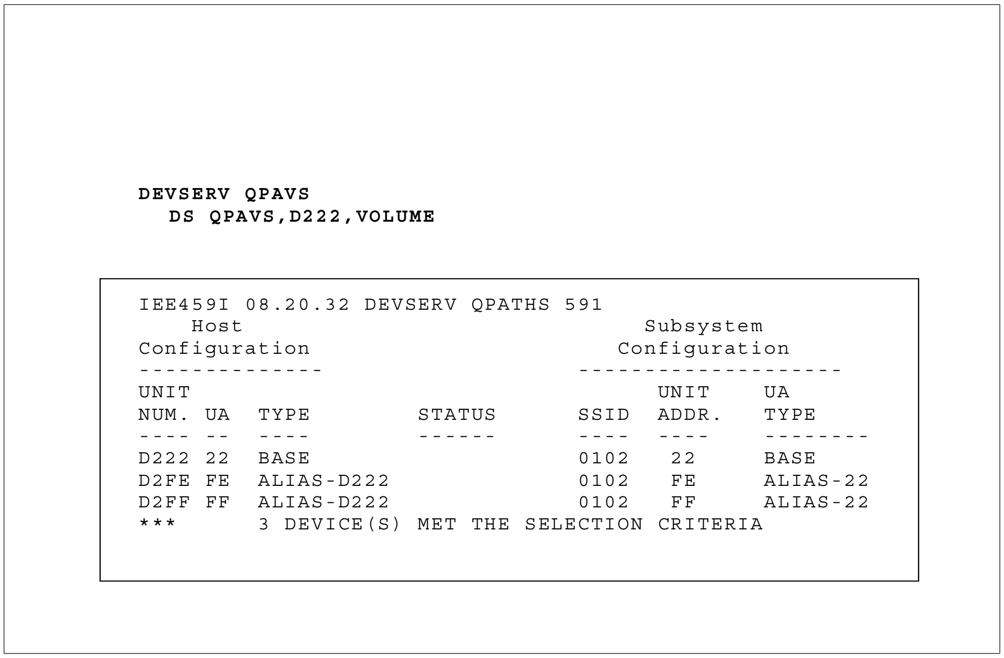
When defining PAV volumes to the ESS, and to HCD, you specify a new device type as already said. The new device type is 3390B (or 3380B) for a PAV base, and 3390A (or 3380A) for a PAV alias. Device support UIMs support these new PAV device types.

You can enable or disable the use of dynamic PAVs. In your HCD definition, you can specify **WLMPAV = YES |NO.** If you stay with the default (WLMPAV=YES), dynamic PAV management by WLM is enabled. In a Parallel Sysplex, if dynamic PAV management is specified for one of the systems, then it is enabled for all the systems in the sysplex, even if they specify **WLMPAV=N**O.

An alias must not be defined for MIH in the IECIOS member. And alias should not be initialized, because they are only known to the I/O Supervisor routines of the z/OS.

The association of alias address (one or more) to base address in the ESS is done using the ESS Specialist *Modify PAV Assignment* panel (see Figure 99). You can modify PAV assignments only after you have created the LCU, defined disk groups, and created base volumes.

5.2.7 Querying PAVs



```
DEVSERV QPAVS
  DS QPAVS,D222,VOLUME
```

```
IEE459I 08.20.32 DEVSERV QPATHS 591
    Host                                 Subsystem
Configuration                          Configuration
--------------                      --------------------
UNIT                                      UNIT    UA
NUM. UA  TYPE        STATUS       SSID  ADDR.   TYPE
---- --  ----        ------       ----  ----    --------
D222 22  BASE                     0102   22     BASE
D2FE FE  ALIAS-D222               0102   FE     ALIAS-22
D2FF FF  ALIAS-D222               0102   FF     ALIAS-22
***      3 DEVICE(S) MET THE SELECTION CRITERIA
```

Figure 100. Querying PAVs

You can use the DEVSERV QPAVS command to verify the PAV definitions (see Figure 100). This command shows you the unit addresses currently assigned to a base. If the hardware and software definitions do not match, the STATUS field will not be empty, but rather will contain a warning such as: INV-ALIAS for an invalid alias or NOT-BASE if the volume is not a PAV volume.

Note that the DEVSERV command shows the *unit number* and *unit address*. The unit number is the address, used by z/OS. This number could be different for different hosts accessing the same logical volume. The unit address is an ESS internal number used to unambiguously identify the logical volume.

In a reverse way, if you are needing to know which base address is associated to a given alias address, then you may find more useful to use the D M=DEV(alias-address) command.

Note: Remember that alias addresses cannot be used in commands like D U, VARY, and so on. For z/VM, where you have PAV guest support, you also have a Q PAV command (see 5.2.12.5, “PAV for VM/ESA guests” on page 161).

5.2.8 PAV assignment

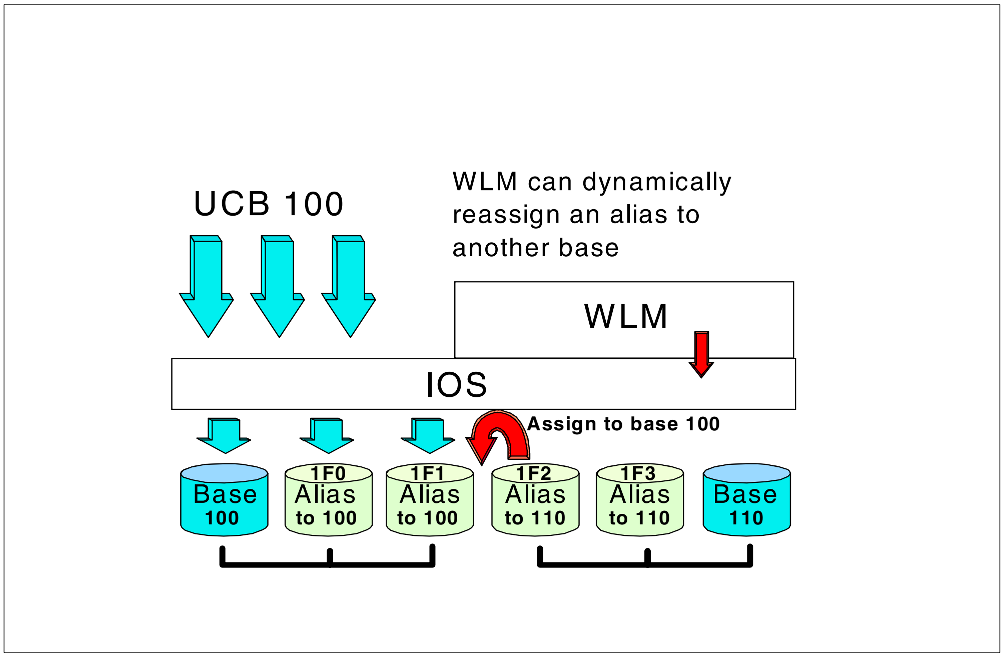


Figure 101. Assignment of alias addresses

z/OS recognizes the aliases that are initially assigned to a base during the NIP (Nucleus Initialization Program) phase. If dynamic PAVs are enabled, the WLM can reassign an alias to another base by instructing the IOS to do so when necessary (see Figure 101).

5.2.9 WLM support for dynamic PAVs

z/OS's Workload Manager in Goal mode tracks the system workload and checks if the workloads are meeting their goals established by the installation.

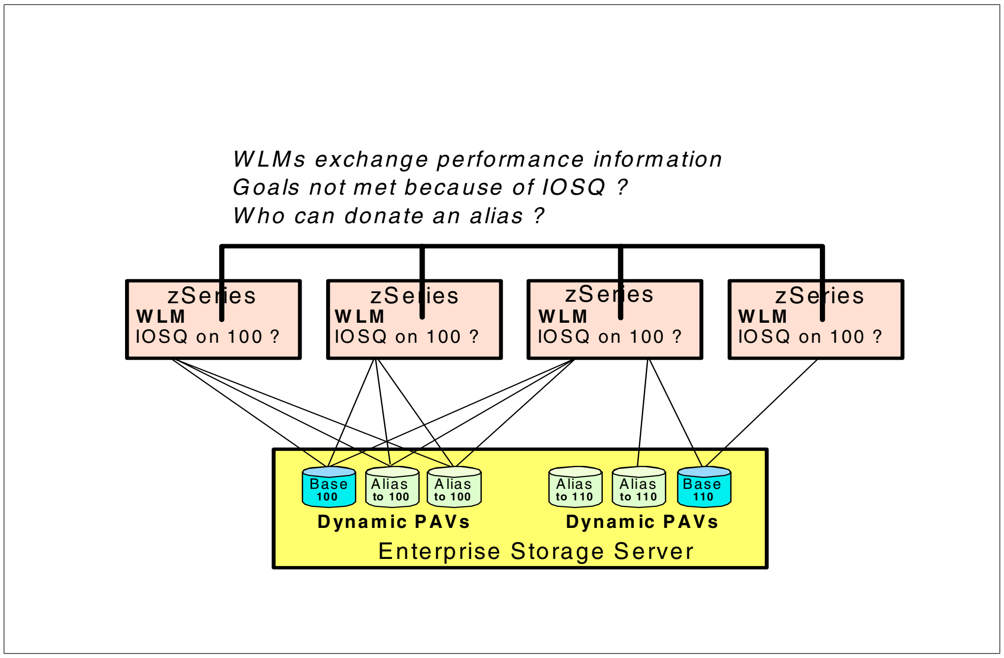


Figure 102. Dynamic PAVs in a sysplex

WLM now also keeps track of the devices utilized by the different workloads, accumulates this information over time, and broadcasts it to the other systems in the same sysplex. If WLM determines that any workload is not meeting its goal due to IOSQ time, WLM will attempt to find an alias device that can be reallocated to help this workload achieve its goal (see Figure 102).

Actually there are two mechanisms to tune the alias assignment:

1. The first mechanism is goal based. This logic attempts to give additional aliases to a PAV enabled device that is experiencing IOS queue delays and is impacting a service class period that is missing its goal. To give additional aliases to the receiver device, a donor device must be found with a less important service class period. A bitmap is maintained with each PAV device that indicates the service classes using the device.
2. The second is to move aliases to high contention PAV enabled devices from low contention PAV devices. High contention devices are identified by having a significant amount of IOS queue time (IOSQ). This tuning is based on efficiency rather than directly helping a workload to meet its goal. Because adjusting the number of aliases for a PAV enabled device affects any system using the device, a sysplex-wide view of performance data is important, and is needed by the adjustment algorithms. Each system in the sysplex broadcasts local performance data to the rest of the sysplex. By combining the data received from other systems with local data, WLM can build the sysplex view.

Note that aliases of an offline device will be considered unbound. WLM uses unbound aliases as the best donor devices. If you run with a device offline to some systems and online to others, you should make the device ineligible for dynamic WLM alias management in HCD.

RMF reports the number of exposures for each device in its Monitor/DASD Activity report and in its Monitor II and Monitor III Device reports. RMF also reports which devices had a change in the number of exposures.

RMF reports all I/O activity against the base address — not by the base and associated aliases. The performance information for the base includes all base and alias activity.

As mentioned before, the workload manager (WLM) must be in goal mode to cause PAVs to be shifted from one logical device to another.

Further information regarding the dynamic WLM and PAVs can be found on the Internet in the WLM/SRM site at `http://www.ibm.com/s390/wlm/`

5.2.10 Reassignment of a PAV alias

To reassign an alias, a token is used by IOS to serialize the alias-change across the Sysplex
Systems informed when move complete

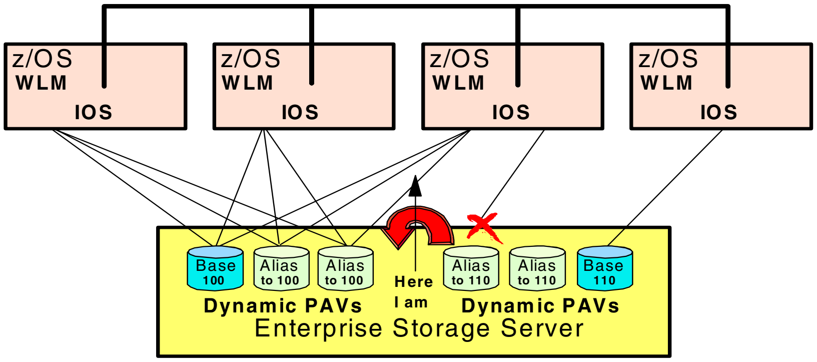
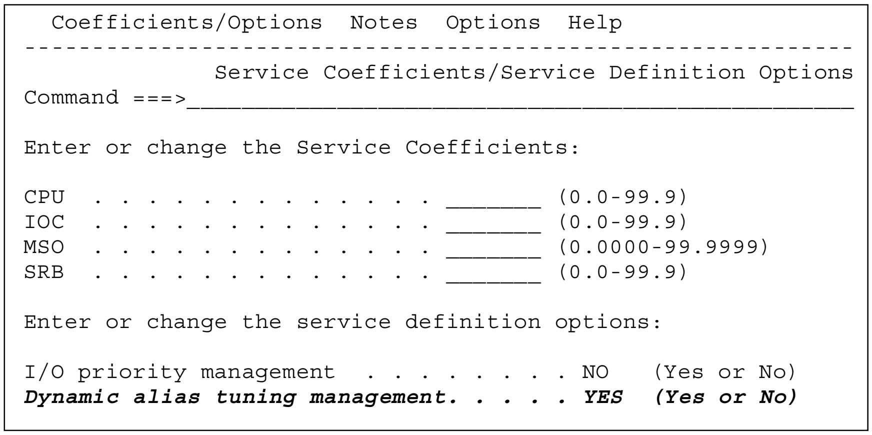


Figure 103. Reassignment of dynamic PAV aliases

The movement of an alias from one base to another is serialized within the sysplex. IOS tracks a token for each PAV enabled device. This token is updated each time an alias change is made for a device. IOS and WLM exchange the token information. When the WLM instructs IOS to move an alias, WLM also presents the token. When IOS has started a move and updated the token, all affected systems are notified of the change through an interrupt.

5.2.11 Mixing PAV types



```
  Coefficients/Options  Notes  Options  Help
-------------------------------------------------------------
              Service Coefficients/Service Definition Options
Command ===>_________________________________________________

Enter or change the Service Coefficients:

CPU  . . . . . . . . . . . . . _______ (0.0-99.9)
IOC  . . . . . . . . . . . . . _______ (0.0-99.9)
MSO  . . . . . . . . . . . . . _______ (0.0000-99.9999)
SRB  . . . . . . . . . . . . . _______ (0.0-99.9)

Enter or change the service definition options:

I/O priority management  . . . . . . . . NO   (Yes or No)
Dynamic alias tuning management. . . . . YES  (Yes or No)
```

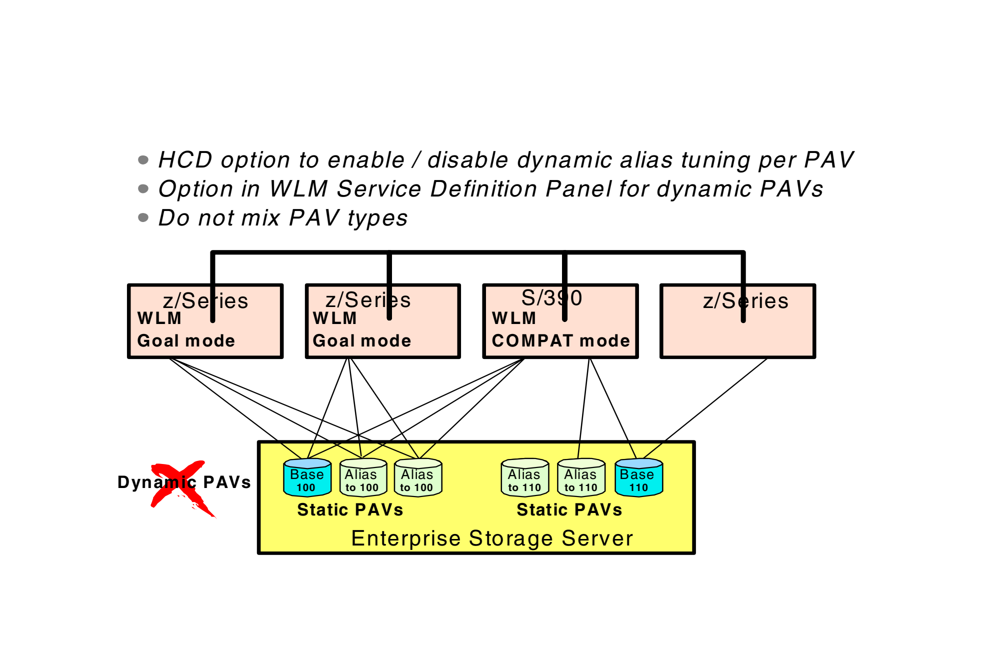
Figure 104. Activation of dynamic alias tuning for the WLM

With HCD, you can enable or disable dynamic alias tuning on a device-by-device basis. On the WLM’s Service Definition ISPF panel, you can globally (sysplex-wide) enable or disable dynamic alias tuning by the WLM as Figure 104 shows. This option can be used to stop WLM from adjusting the number of aliases in general, when devices are shared by systems that do not support dynamic alias tuning.

If you enable alias tuning (this is the default in WLM) for devices shared by zSeries hosts only supporting static PAVs, these systems still recognize the change of an alias and use the new assigned alias for I/Os to the associated base address. However, the WLMs on the systems that do support the dynamic alias tuning will not see the I/O activity to and from these non-supporting systems to the shared devices. Therefore, the WLMs can not take into account this hidden activity when making their judgements.

Without a global view, the WLMs could make a wrong decision. Therefore, you should not use dynamic PAV alias tuning for devices from one system and static PAV for the same devices on another system.

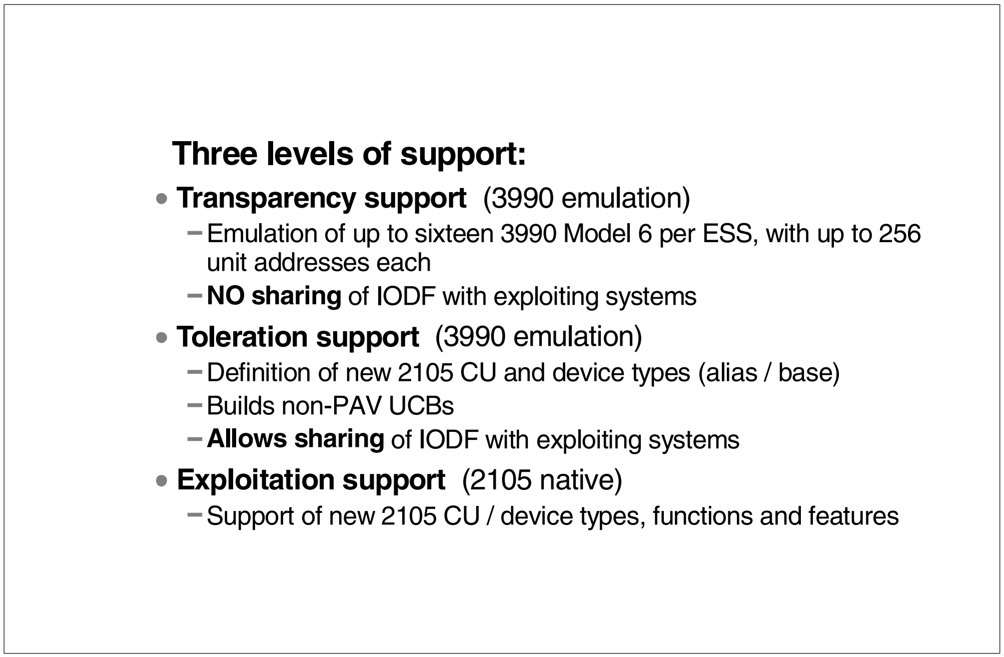
If at least one system in the sysplex specifies dynamic PAV management, then it is enabled for all the systems in the sysplex. There is no consistency checking for this parameter. It is an installation’s responsibility to coordinate definitions consistently across a sysplex. WLM will not attempt to enforce a consistent setting of this option (see Figure 105).



- *HCD option to enable / disable dynamic alias tuning per PAV*
- *Option in WLM Service Definition Panel for dynamic PAVs*
- *Do not mix PAV types*

Figure 105. No mixing of PAV types

5.2.12 PAV support



Three levels of support:

- **Transparency support** (3990 emulation)
  - Emulation of up to sixteen 3990 Model 6 per ESS, with up to 256 unit addresses each
  - **NO sharing** of IODF with exploiting systems
- **Toleration support** (3990 emulation)
  - Definition of new 2105 CU and device types (alias / base)
  - Builds non-PAV UCBs
  - **Allows sharing** of IODF with exploiting systems
- **Exploitation support** (2105 native)
  - Support of new 2105 CU / device types, functions and features

Figure 106. ESS support modes

Figure 106 shows the three modes of operating system support of the Enterprise Storage Server. The mode of support is determined by the host software level and PTFs. In this section we describe the characteristics of these three different support modes, for information on the required software levels please refer to Section 7.2, “z/OS support” on page 233.

5.2.12.1 Transparency support

Transparency provides the base functions of an IBM 3990 model 6 Storage Control. z/OS hosts see the ESS as up to 16 logical IBM 3990 Model 6 subsystems with up to 256 unit addresses per logical subsystem. Note that you cannot share an I/O definition file (IODF) with exploiting systems.

5.2.12.2 Toleration support

With toleration support z/OS hosts recognize the new control unit type 2105 of the ESS, and the new device types 3390 base and 3390 alias.

With toleration support, however, only non-PAV UCBs are built. You can, however, share an IODF with other exploiting systems.

5.2.12.3 Exploitation support

At this support level the ESS is recognized as a 2105 device type and the operating system can exploit the full set of capabilities: definition of PAVs in HCD, and exploitation of PAVs; exploitation of dynamic and automatic tuning of PAVs by the Workload Manager (WLM).

5.2.12.4 Use of shared IODF

In a sysplex, it is quite common to use a shared I/O definition file (IODF). To exploit the new capabilities of the ESS, at least one system must be at the *exploitation* support level. Other systems sharing the IODF must be at least at *toleration* support level. You cannot share an IODF between an *exploitation* system and a *transparency* system. If you plan to share the IODF, you should check if you need to upgrade your system

5.2.12.5 PAV for VM/ESA guests

z/VM has not implemented the exploitation of PAV for itself. However, with VM/ESA 2.4.0, with an enabling APAR, z/OS and OS/390 guests can use PAV volumes and dynamic PAV tuning. Alias and base addresses must be attached to the z/OS guest. You need a separate ATTACH for each alias. You should attach the base and its aliases to the same guest.

A base cannot be attached to SYSTEM if one of its aliases is attached to that guest. This means that you cannot use PAVs for Full Pack mini disks.

There is a new QUERY PAV command available for authorized (class B) users to query base and alias addresses:

```
QUERY PAV rdev
QUERY PAV ALL
```

Response for a PAV base:

```
Device 01D2 is a base Parallel Access Volume device with
the following aliases: 01E6 01EA 01D3
```

Response for a PAV alias:

```
Device 01E7 is an alias Parallel Access Volume device
whose base device is 01A0
```

Refer to Section 7.3, “z/VM support” on page 235 for information on the software levels and PTFs.

5.3 Multiple Allegiance

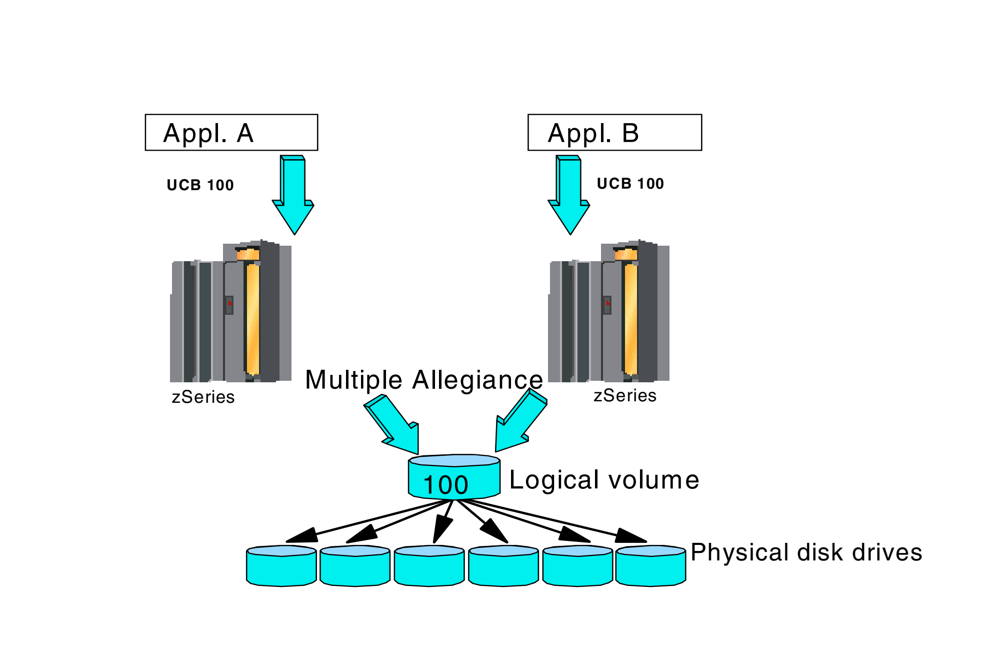


Figure 107. Parallel I/O capability with Multiple Allegiance

Normally, if any zSeries host image (server or LPAR) does an I/O request to a device address for which the storage subsystem is already processing an I/O that came from another zSeries host image, then the storage subsystem will send back a *device busy* indication. This delays the new request and adds to processor and channel overhead (This delay is showed in the RMF *Pend* time column).

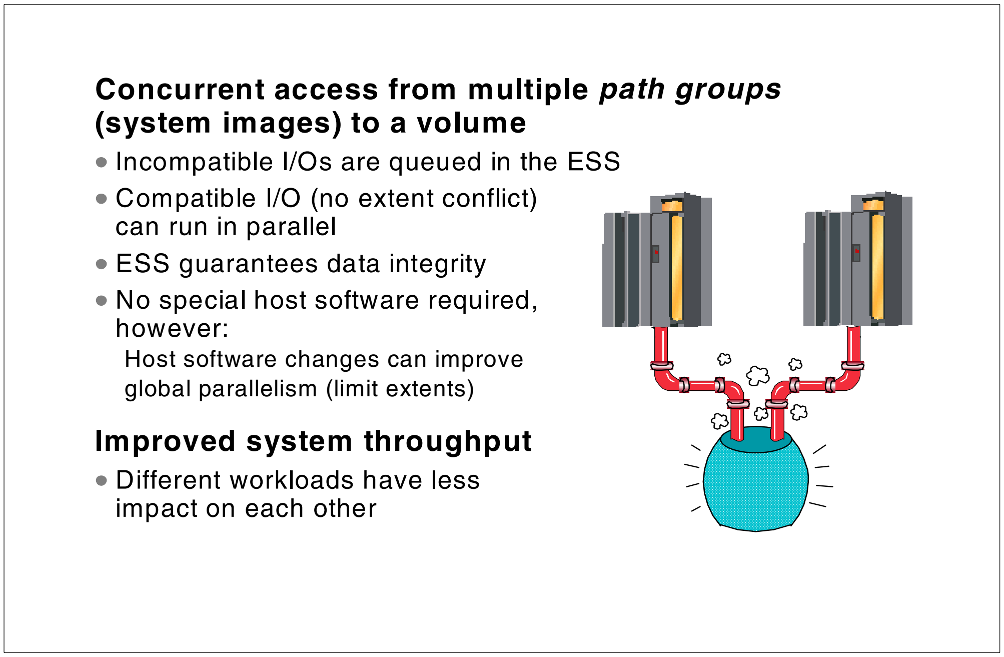
5.3.1 Parallel I/O capability

The Enterprise Storage Server accepts multiple I/O requests from different hosts to the same device address, increasing parallelism and reducing channel overhead.

In previous storage subsystems a device had an implicit allegiance, that is, a relationship created in the control unit between the device and a channel path group when an I/O operation is accepted by the device. The allegiance causes the control unit to guarantee access (no busy status presented) to the device for the remainder of the channel program over the set of paths associated with the allegiance.

Now, with Multiple Allegiance, the requests are accepted by the ESS and all requests will be processed in parallel, unless there is a conflict when writing data to a particular extent of the CDK logical volume. Still, good application software access patterns can improve the global parallelism by avoiding reserves, limiting the extent scope to a minimum, and setting an appropriate file mask, for example, if no write is intended.

In systems without Multiple Allegiance, all except the first I/O request to a shared volume were rejected, and the I/Os were queued in the zSeries channel subsystem, showing up as PEND time in the RMF reports.



Concurrent access from multiple *path groups* (system images) to a volume

- Incompatible I/Os are queued in the ESS
- Compatible I/O (no extent conflict) can run in parallel
- ESS guarantees data integrity
- No special host software required, however:
  - Host software changes can improve global parallelism (limit extents)

Improved system throughput

- Different workloads have less impact on each other

Figure 108. Multiple Allegiance

Multiple Allegiance provides significant benefits for environments running a Sysplex, or zSeries systems sharing access to data volumes (Figure 108).

Multiple Allegiance and PAV can operate together to handle multiple requests from multiple hosts.

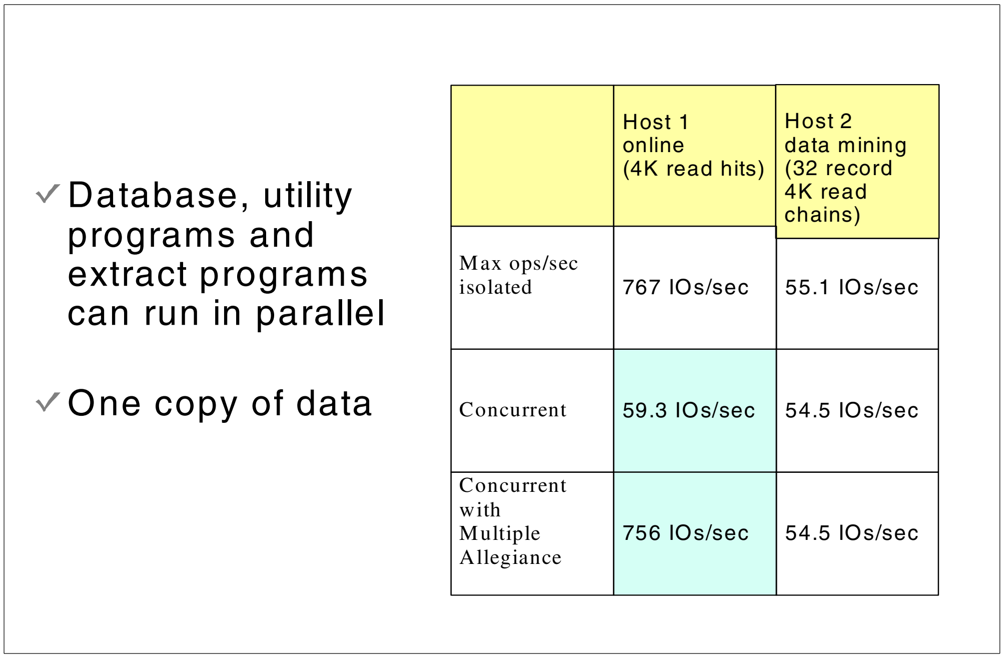
5.3.2 Eligible I/Os for parallel access

ESS distinguishes between compatible channel programs that can operate concurrently and incompatible channel programs that have to be queued to maintain data integrity. In any case the ESS ensures that, despite the concurrent access to a volume, no channel program can alter data that another channel program is using.

5.3.3 Software support

Basically, there is no software support required for the exploitation of ESS's Multiple Allegiance capability. The ESS storage subsystem looks at the extent range of the channel program and whether it intends to read or to write. Whenever possible, ESS will allow the I/Os to run in parallel.

5.3.4 Benefits of Multiple Allegiance



- ✓ Database, utility programs and extract programs can run in parallel
- ✓ One copy of data

	Host 1 online (4K read hits)	Host 2 data mining (32 record 4K read chains)
Max ops/sec isolated	767 IOs/sec	55.1 IOs/sec
Concurrent	59.3 IOs/sec	54.5 IOs/sec
Concurrent with Multiple Allegiance	756 IOs/sec	54.5 IOs/sec

Figure 109. Benefits of MA for mixing workloads

The Enterprise Storage Server ability to run channel programs to the same device in parallel can dramatically reduce IOSQ and pending times in shared environments.

Particularly different workloads—for example, batch and online—running in parallel on different systems, can have an unfavorable impact on each other. In such a case, ESS's Multiple Allegiance can improve the overall throughput. See Figure 109 for an example of a DB2 data mining application running in parallel with normal database access.

The application running long CCW chains (Host 2) drastically slows down the online application in the example when both applications try to access the same volume extents.

ESS's support for parallel I/Os lets both applications run concurrently without impacting each other adversely.

5.4 Priority I/O queuing

The concurrent I/O capability of the Enterprise Storage Servers allows it to execute multiple channel programs concurrently, as long as the data accessed by one channel program is not altered by another channel program.

5.4.1 Queuing of channel programs

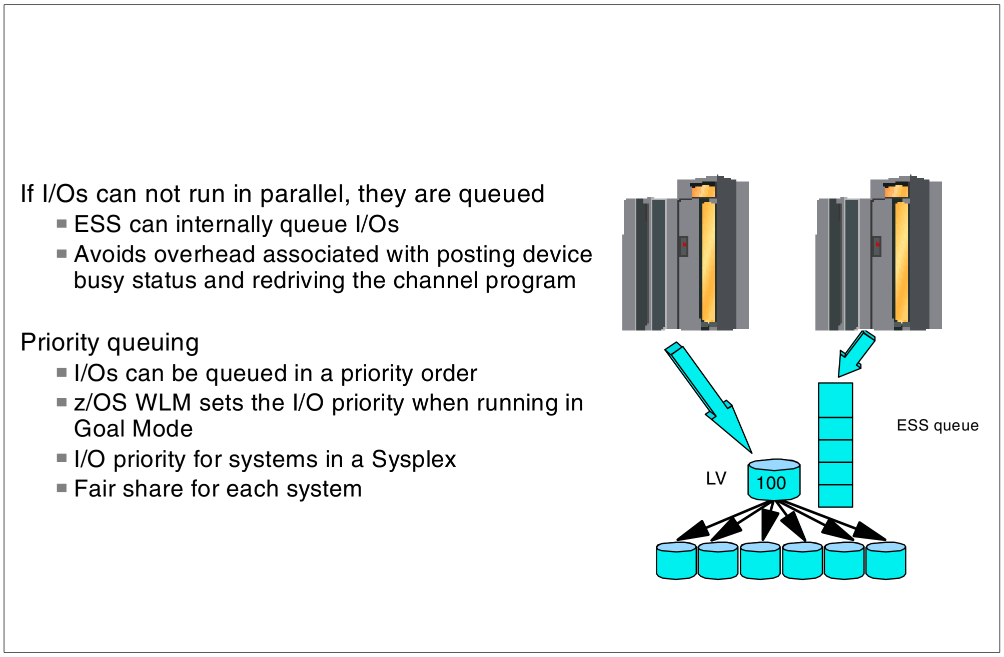
When the channel programs conflict with each other and must be serialized to ensure data consistency, the ESS will internally queue channel programs.

This subsystem I/O queuing capability provides significant benefits:

- Compared to the traditional approach of responding with *device busy* status to an attempt to start a second I/O operation to a device, I/O queuing in the storage subsystem eliminates the overhead associated with posting status indicators and re-driving the queued channel programs.
- Contention in a shared environment is eliminated. Channel programs that cannot execute in parallel are processed in the order they are queued. A fast system cannot monopolize access to a volume also accessed from a slower system. Each system gets a fair share.

5.4.2 Priority queuing

I/Os from different z/OS system images can be queued in a priority order. It is the z/OS Workload Manager that makes use of this priority to privilege I/Os from one system against the others. You can activate I/O Priority Queuing in WLM’s Service Definition settings.



If I/Os can not run in parallel, they are queued

- ESS can internally queue I/Os
- Avoids overhead associated with posting device busy status and redriving the channel program

Priority queuing

- I/Os can be queued in a priority order
- z/OS WLM sets the I/O priority when running in Goal Mode
- I/O priority for systems in a Sysplex
- Fair share for each system

Figure 110. I/O queueing

WLM has to run in Goal mode. z/OS and all releases of OS/390 Version 2, and OS/390 Version 1 Release 3, support I/O Priority Queuing.

When a channel program with a higher priority comes in and is put in front of the queue of channel programs with lower priority, the priority of the low priority programs is also increased. This prevents high priority channel programs from dominating lower priority ones and gives each system a fair share (see Figure 111).

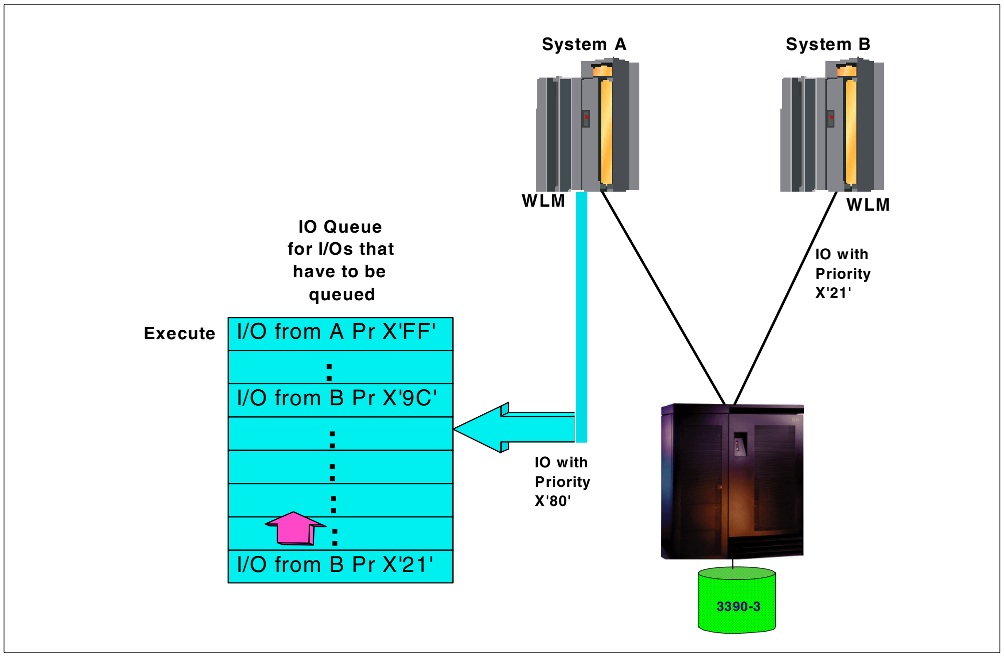


Figure 111. Priority I/O Queuing

5.5 Efficient I/O operations

For the CKD environment, and for the FB environment, I/O operations are solved in the most efficient manner.

5.5.1 z/OS enhanced CCWs

For the z/OS environments, the Enterprise Storage Server supports *channel command words* (CCWs) that reduce the characteristic overhead associated to the previous (3990) CCW chains. Basically, with these CCWs, ESS can read or write more data with fewer CCWs. CCW chains using the old CCWs are converted to the new CCWs whenever possible. Again, the cooperation of IBM z/OS software and the IBM ESS provides the best benefits for the application's performance.

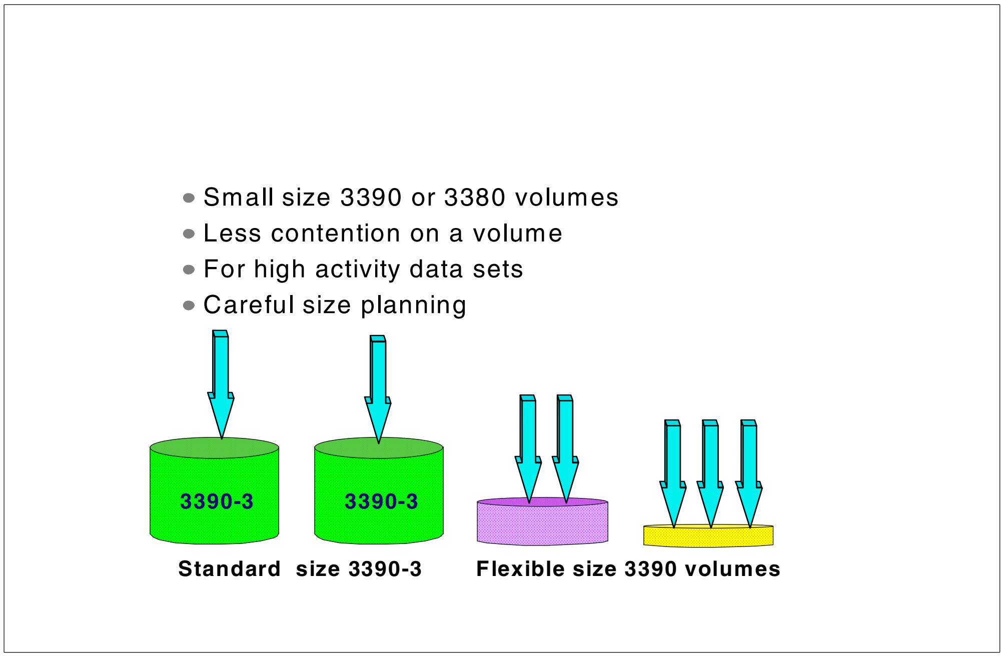
VM/ESA itself does not use the new CCWs. VM/ESA, however, allows a guest to use the new CCWs.

5.5.2 SCSI command tag queuing

Servers connecting to an ESS using the SCSI command set (over Fibre Channel or over parallel SCSI) may use SCS command tag queuing. This function supports multiple outstanding requests to the same LUNs at the same time.

Such requests are processed by the ESS concurrently if possible. Some of these I/O request requests may then be solved by the ESS with a cache hit, and others may be solved on the disk array where the logical volumes are striped.

5.6 Custom volumes



- Small size 3390 or 3380 volumes
- Less contention on a volume
- For high activity data sets
- Careful size planning

Figure 112. CKD custom volumes

As we have seen, the Enterprise Storage Server is able to do several concurrent I/O operations to a logical volume with its PAV and MA functions. This drastically reduces or eliminates IOS queuing and pending times in the z/OS environments.

Even if you can not benefit from these functions, for example, because you did not order the PAV optional feature, or your system does not share the volumes with other systems, you have another option to reduce contention to volumes and hence reduce IOS queue time.

When configuring your CKD volumes in the ESS, you have the option to define *custom volumes*. You can define logical 3390 or 3380 type volumes which do not have the standard number of cylinders of a model 3 or model 9, for example, but instead have a flexible number of cylinders that you can choose.

You probably want to define small volume sizes to reduce contention to the volume. You can put high activity data sets on separate custom volumes. Or you can give each high activity data set its own custom volume.

You should carefully plan the size of the custom volumes, and consider the potential growth of the data sets. You can adjust the size of each custom volume to the data set that you plan to put on this volume. But you might also come to the conclusion that you just need some standard small volume sizes of, perhaps 50, or 100, or 500 cylinders. You have the choice.

5.7 Caching algorithms

With its effective caching algorithms, the Enterprise Storage Server is able to minimize wasted cache space, lessen disk drive utilization, and consequently reduce its backend traffic.

5.7.1 Optimized cache usage

ESS manages its cache in 4 KB segments, so for small data blocks (4 KB and 8 KB are common database block sizes) minimum cache is wasted. In contrast, large cache segments could exhaust cache capacity while filling up with small random reads. Thus the ESS, having smaller cache segments, is able to avoid wasting cache space for situations of small record sizes that are common in the interactive applications.

Also have in mind that the slower a system is able to destage tracks from cache to its backend disk arrays, the more cache it needs to hold the unwritten records. This is not the case for the ESS storage subsystem

5.7.2 Sequential pre-fetch I/O requests

Storage subsystem cache has proven to be of enormous benefits for the CKD servers. It is often of less value for the FB servers because of the way these servers use the server processor memory as cache. In the zSeries environments, storage system cache usually offers significant performance benefits for two main reasons: because zSeries operating systems tend not to keep the most frequently referenced data in processor memory, and because zSeries servers store data in a way that allows disk systems to potentially predict sequential I/O activity and pre-fetch data into system cache.

ESS monitors the channel program patterns to determine if data access pattern is sequential or not. If the access is sequential, then contiguous data is pre-fetched into cache in anticipation of the next read requests. It is common that z/OS set a bit into the channel program notifying the disk subsystem that all subsequent I/O operations will be sequential read requests. ESS supports these bits in the channel program and helps to optimize its pre-fetch process.

ESS uses its sequential predict algorithms and its high backend bandwidth to keep the cache pre-loaded and ready with data for the upcoming I/O operations from the applications.

5.8 Backend high performance design

A performance design based only on cache size and its efficiency may not be properly addressing the workload characteristics of all the FB servers. For these type of servers, the backend efficiency is the key to the unbeatable performance of the Enterprise Storage Server.

5.8.1 Serial Storage Architecture (SSA)

The performance of the disk portion of the disk storage subsystem (disk drives, buses, and disk adapters, generally designated as *backstore* or "*backend*) has a major impact on the performance. In the ESS, these characteristics make the difference for the FB environments where the servers usually do not get so many cache hits in the storage subsystem.

5.8.1.1 Path (loop/bus) bandwidth:

The ESS uses SSA for internal disk communications. SSA is physically configured as a loop. A single SSA loop has a base bandwidth of 160 MB/second. This bandwidth is multiplied by the number of adapters attached to the loop. ESS has a total of 8 SSA adapters and one adapter connects 2 distinct loops. Therefore a total of 8*2*160 MB/second = 2560 MB/second nominal bandwidth is available in the backstore

5.8.1.2 Number of paths:

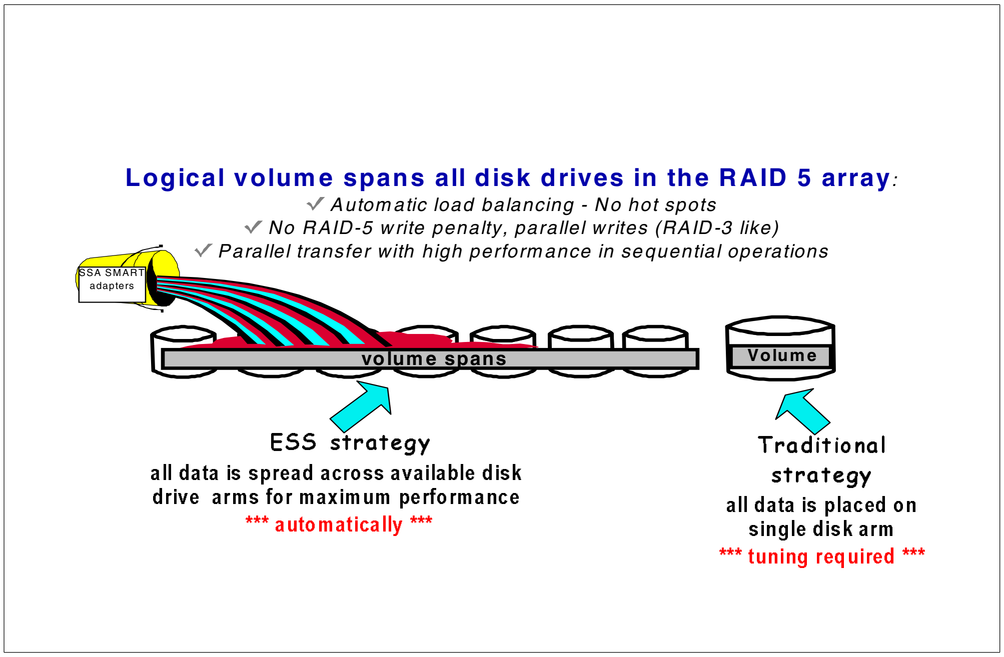
The ESS has up to 64 (SSA) internal data paths to disks, each with a bandwidth of 40 MB/second. (The SSA loop bandwidth of 160 MB/second per adapter consists of two read paths and two write paths at 40 MB/second delivered at each adapter connection). Every path can be transferring data at the same time.

5.8.2 ESS striping

Data striping means storing logically contiguous data across multiple disk drives, which provides significant performance benefits. ESS stripes every RAID-5 protected logical volume across multiple disk drives of the RAID rank. Benefits of this RAID-5 striping implementation include:

- Balanced disk drive utilization
- I/O parallelism for cache misses
- Higher bandwidth for the logical volumes
- Avoidance of disk drive hot spots

All of these lead to higher sustained performance and reduced manual tuning.



Logical volume spans all disk drives in the RAID 5 array:

- ✓ *Automatic load balancing - No hot spots*
- ✓ *No RAID-5 write penalty, parallel writes (RAID-3 like)*
- ✓ *Parallel transfer with high performance in sequential operations*

ESS strategy

all data is spread across available disk drive arms for maximum performance

*** automatically ***

Traditional strategy

all data is placed on single disk arm

*** tuning required ***

Figure 113. RAID-5 logical volume striping

5.9 RISC SMP processor

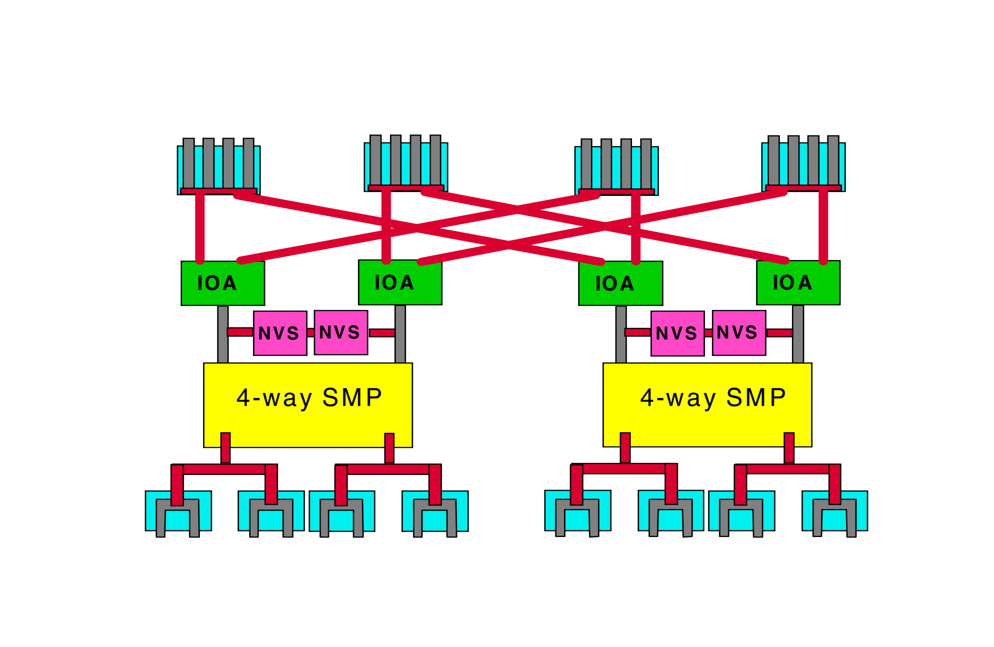


Figure 114. RISC SMP processors

Each cluster contains four high performance RISC processors, configured as 4-way SMP. Each SMP can be configured from 4 GB up to 16 GB of cache and has 4 PCI buses to handle the host adapters, device adapters and NVS

ESCON interfaces, each operates at 200 Mbits/sec. half duplex. Fiber Channel / FICON interfaces, each operates at 1 Gbit/sec. full duplex. ULTRA SCSI interfaces at 40 MB/sec. The SSA device adapters, 160 MB/sec. on each loop.

5.10 FICON host adapters

FICON extends the Enterprise Storage Server ability to deliver bandwidth potential to the volumes needing it, when they need it.

5.10.1 FICON benefits

For the z/Architecture and s/390 servers, the FICON attachment provides many improvements:

- Increased number of concurrent connections. FICON provides channel to ESS multiple concurrent I/O connections. ESCON supports only one I/O connection at any one time
- Increased distance. With FICON, the distance from the channel to the ESS, or channel to switch, or switch to ESS link is increased using dark fiber. The distance for ESCON of 3 Km. is increased to up to 10 Km. (20 Km. with RPQ) for FICON channels using long wavelength laser.
- Increased link bandwidth. FICON has up to 10 times the link bandwidth of ESCON (1 Gigabit/second full duplex, compared to 200 megabits/second half duplex). FICON has up to more than 4 times the effective channel bandwidth for the initial implementation (70 MB/sec. compared to 17 MB/sec. for ESCON).
- No data rate droop effect. For ESCON channels, the droop effect started at 9 Km. For FICON, there is no droop effect even at the maximum supported distance of 103 Km.
- Increased channel device-address support. From 1,024 devices for an ESCON channel to up 16,384 for a FICON channel.
- Greater exploitation of Parallel Access Volume (PAV). FICON allows for greater exploitation of PAV in that more I/O operations can be started for a group of channel paths.
- Greater exploitation of priority I/O queueing. FICON channels use frame and Information Unit (IU) multiplexing control to provide greater exploitation of the priority I/O queueing mechanisms within the FICON capable ESS.
- Better utilization of the links. Frame multiplexing support on FICON channels, switches, and FICON control units like the ESS, provides better utilization of the links.

These improvements can be realized in more powerful and simpler configurations with increased throughput. The z/OS user will notice improvements over ESCON channels, with reduced bottlenecks from the I/O path, allowing the maximum control unit I/O concurrency exploitation:

- IOSQ time (UCB busy) can be reduced configuring more alias device addresses, using Parallel Access Volumes (PAVs). This is possible because FICON channels can address up to 16,384 devices (ESCON channels address up to 1,024).
- Pending time can also be reduced:
  - Channel busy conditions are reduced by FICON channel's multiple starts.
  - Port busy conditions are eliminated by FICON switches frame multiplexing

  - Control unit busy conditions are reduced by FICON adapter’s multiple starts
  - Device busy conditions are also reduced as FICON’s multiple concurrent I/O operations capability can improve the Multiple Allegiance (MA) function exploitation

FICON channels allow a higher I/O throughput, using fewer resources. FICON’s architecture capability to execute multiple concurrent I/O operations, which can be sustained by the FICON link bandwidth.

- A single FICON channel can have I/O operations to multiple logical control units at the same time, by using the FICON protocol frame multiplexing.
- FICON’s CCW and data pre-fetching and pipelining, and protocol frame multiplexing also allows multiple I/O operations to the same logical control unit. As a result, multiple I/O operations can be done concurrently to any logical control unit, even within the same control unit. By using IBM ESS’s Parallel Access Volumes (PAV) function, multiple I/O operations are possible even to the same volume.

5.10.2 FICON I/O operation

An I/O operation executed over a FICON channel has differences with an I/O operation executed over an ESCON channel. Both will be started by the same set of I/O routines of the operating system, but the “mechanics” of the operation differ whether it develops over a FICON channel or over an ESCON channel. In this section we go deeper into the I/O operation components, to understand how FICON improves these components to make the complete I/O more efficient.

Let us first understand some of these basic components that are part of an I/O operation.

Note: The explanations developed in this section are based on the view that a z/OS has of the I/O operation sequence. For other environments the analysis of the I/O sequence is different.

- I/O Supervisor Queue time (IOSQ), measured by the operating system

  The application I/O request may be queued in the operating system, if the I/O device, represented by the UCB, is already being used by another I/O request from the same operating system image (UCB busy). The I/O Supervisor (IOS) does not issue a Start Subchannel (SSCH) command to the *channel subsystem* until the current I/O operation to this device ends, thereby freeing the UCB for use by another I/O operation. The time while in this operating system queue, is the IOSQ time.

- Pending time (PEND), measured by the channel subsystem

  After IOS issues the Start Subchannel command, the channel subsystem may not be able to initiate the I/O operation if any path or device busy condition is encountered:

  - Channel busy, with another I/O operation from another z/OS system image (from the same or from a different CEC)
  - Switch port busy, with another I/O operation from another or the same CEC. This can only occur on an ESCON channel. The use of buffer credits on a FICON native channel eliminates switch port busy

- Control unit adapter busy, with another I/O operation from another or the same CEC
- Device busy, with another I/O operation from another CEC.

- Connect time, measured by the channel subsystem

  This is the time that the channel is connected to the control unit, transferring data for this I/O operation.

- Disconnect time

  The channel is not being used for this I/O operation, as the control unit is disconnected from the channel, waiting for access to the data or to reconnect. Disconnect time often does not exist when a cache hit occurs because data is transferred directly from cache without the need for access to the device.

5.10.2.1 ESCON cache hit I/O operation

Figure 115 shows the components of an ESCON I/O operation when a cache hit occurs. In this case, there is no disconnect time and the I/O operation ends at the end of the connect time, after transferring the requested data.

The ESCON I/O operation may accumulate IOSQ time if the device (UCB) is already in use for another I/O request from this system. The operating system IOS (I/O Supervisor) routines only initiate one I/O operation at a time for a device with the channel subsystem. The new I/O operation cannot be started with the channel subsystem until the I/O interrupt signalling completion of the current outstanding I/O operation has been processed by IOS.

Parallel Access Volume (PAV) reduces IOSQ time and device busy conditions by allowing multiple I/O requests per UCB for access to the same logical volume.

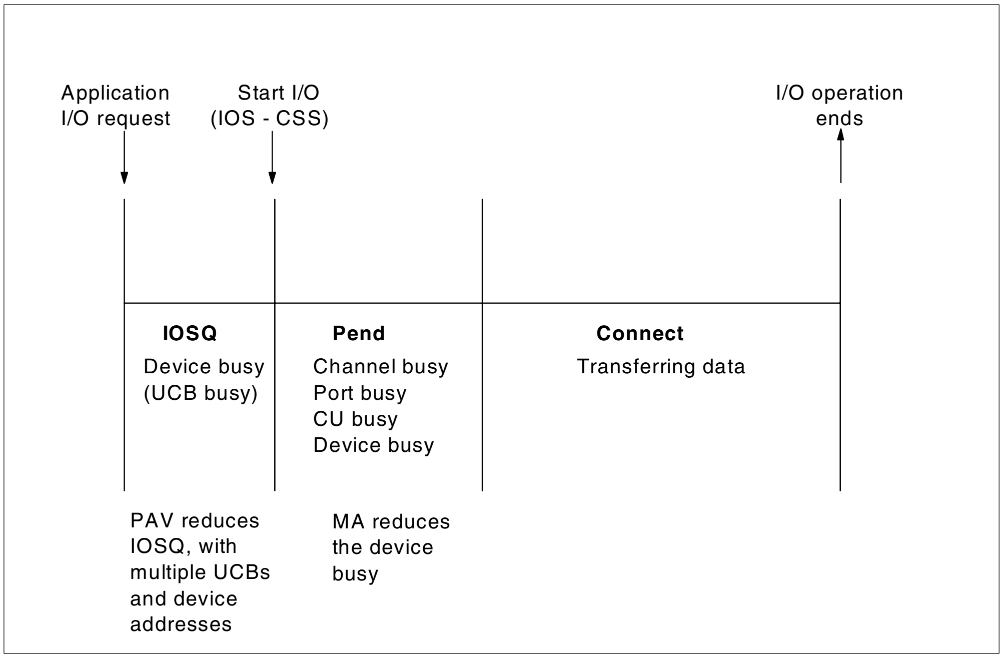


Figure 115. ESCON cache hit I/O operation sequence

Once the request has been accepted by the channel subsystem, it may accumulate PEND time if the channel subsystem is unable to start the request because of either channel busy, port busy, control unit (CU) busy, or device busy.

With the ESS, some control unit busy can be alleviated with I/O queuing by the control unit. In the case of a cache hit, the ESS may queue an I/O request for conditions which in other subsystems would result in CU busy, such as destaging, extent conflict resolution, and so on. This control unit I/O queuing time is accumulated in *disconnect* time, but reported later in *pend* time.

In the ESS, Multiple allegiance alleviates some device busy conditions.The Multiple Allegiance function, enables different operating system images to perform concurrent I/O operations at the same logical volume as long as no extent conflict exists (Note that device busy still occurs if the device is reserved by another operating system image).

When the I/O operation is accepted by the control unit, *connect* time is accumulated as the channel transfers data to/from cache.

The I/O operation completes when the data transfer into cache is complete. No access to the physical volume is required before the end of the I/O operation is signalled in the case of a cache hit.

5.10.2.2 ESCON cache miss I/O operation

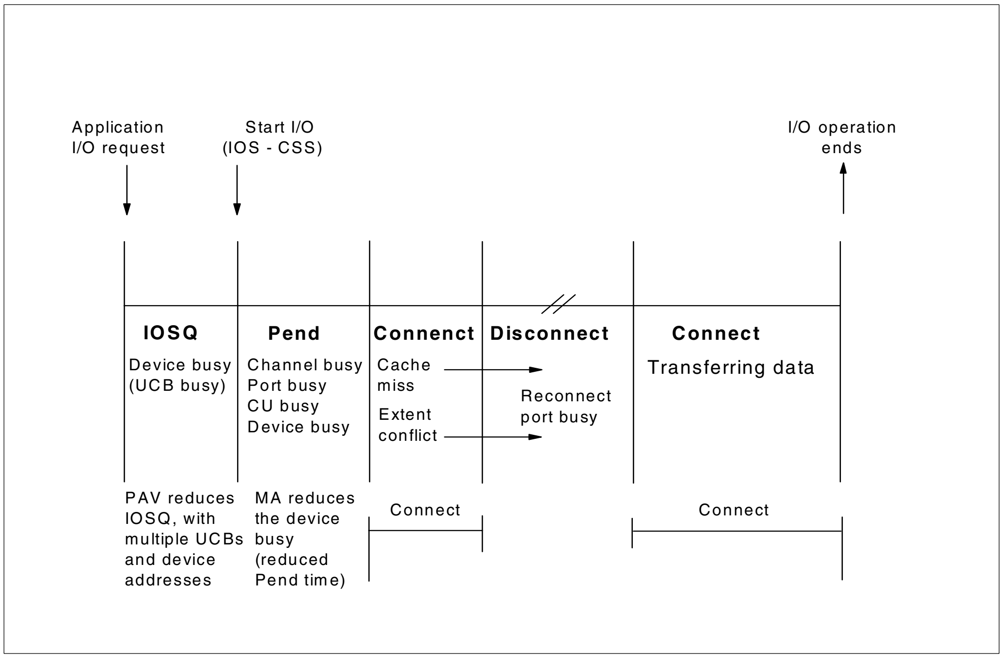


Figure 116. ESCON cache miss I/O operation sequence

Figure 116 shows an ESCON I/O operation sequence when a cache miss occurs. In this case, *connect* time is accumulated as the positioning CCWs are transferred to the ESS. For the ESS, this *connect* time component also includes the extent conflict checking time.

Disconnect time is then accumulated as the physical disk drive positioning operations are performed. Then the control unit must reconnect to the channel for the transfer of data. It is during the attempt to reconnect, that a port busy condition may occur.

Further *connect* time is accumulated as data is transferred over the channel.For ESCON I/O operations, the total connect time component is somehow predictable since the data transfer is directly related to the speed of the channel and the number of bytes transferred.

The I/O operation ends after the requested data has been transferred and the terminating interrupt has been presented by the channel.

5.10.2.3 FICON cache hit I/O operation

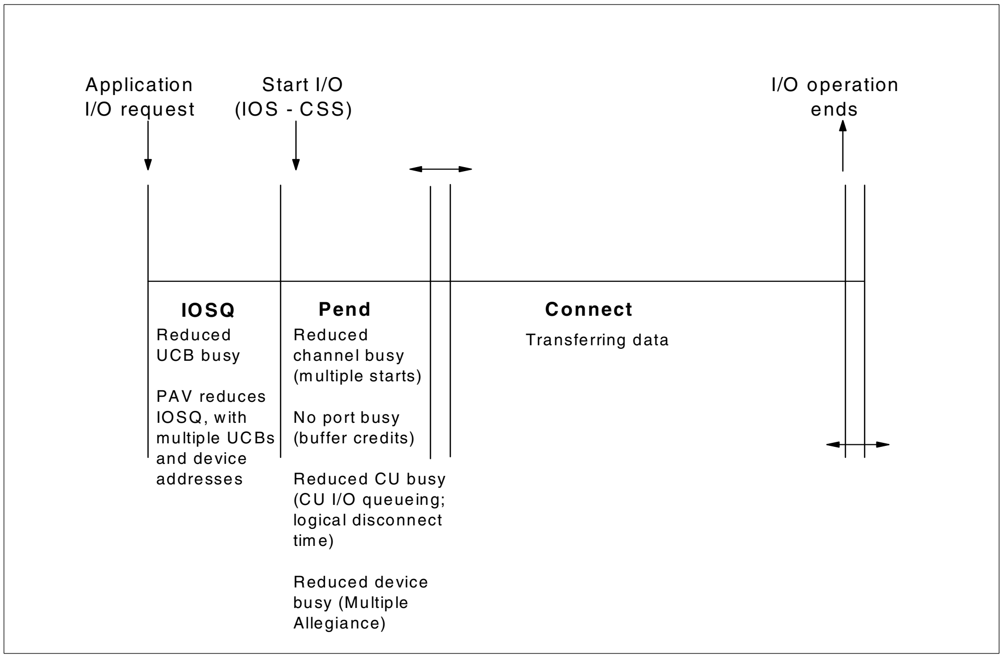

Application I/O request
Start I/O (IOS - CSS)
I/O operation ends
IOSQ
Reduced UCB busy
PAV reduces IOSQ, with multiple UCBs and device addresses
Pend
Reduced channel busy (multiple starts)
No port busy (buffer credits)
Reduced CU busy (CU I/O queueing; logical disconnect time)
Reduced device busy (Multiple Allegiance)
Connect
Transferring data

Figure 117. FICON cache hit I/O operation sequence

Figure 117 shows a FICON I/O operation sequence when a cache hit occurs.. When using FICON, some busy conditions will be reduced or eliminated:

- More devices (and consequently, more UCBs) can be configured (up to 16,384 devices per FICON channel), allowing a higher PAV exploitation. This reduces the number of device busy and UCB busy conditions that are accumulated in the IOSQ time parameter of the z/OS operating systems.
- Channel busy is reduced with FICON’s capability of multiple starts to the same channel path, thereby reducing *pend* time conditions.
- Port busy does not exist on FICON switches. The FICON switch uses switch port buffer credits.
- Control unit busy conditions are reduced with CU I/O queueing in the ESS, also reducing the *pend* time.
- Device busy conditions are reduced by further exploitation of the ESS Multiple Allegiance (MA) function, due to FICON’s multiple concurrent I/O operations capability.

As Fibre Channel frame multiplexing is used by FICON links, some *connect* time is less predictable for individual I/Os. Figure 117 shows this effect when showing the start and end points of the connect component.

5.10.2.4 FICON cache miss I/O operation

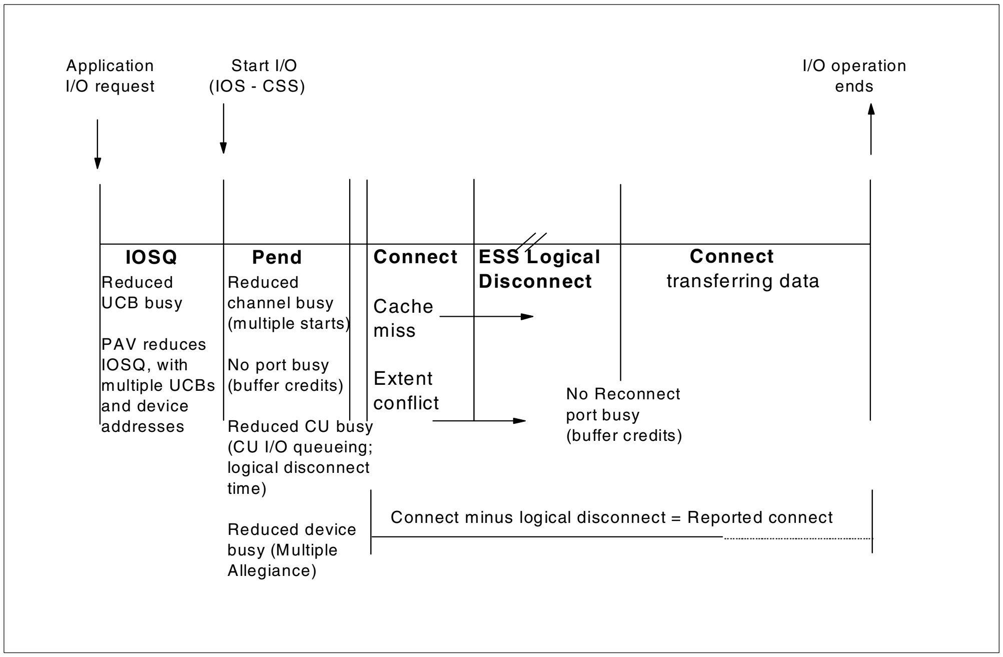

Application I/O request
Start I/O (IOS - CSS)
I/O operation ends
IOSQ
Reduced UCB busy
PAV reduces IOSQ, with multiple UCBs and device addresses
Pend
Reduced channel busy (multiple starts)
No port busy (buffer credits)
Reduced CU busy (CU I/O queueing; logical disconnect time)
Reduced device busy (Multiple Allegiance)
Connect
Cache miss
Extent conflict
ESS Logical Disconnect
No Reconnect port busy (buffer credits)
Connect
transferring data
Connect minus logical disconnect = Reported connect

Figure 118. FICON cache miss I/O operation sequence

Figure 118 shows a FICON I/O operation sequence when a cache miss occurs.

Having all the benefits about reducing busy conditions and times as shown in the previous cache example, a new condition takes place in this kind of I/O operation, removing another busy condition.

In the ESCON cache miss operation, a *disconnect* time component is expected. As the ESS FICON adapter can handle multiple concurrent I/O operations at a time, it will not disconnect from the channel when a cache miss occurs for a single I/O operation. So the ESS adapter remains connected to the channel, being able to transfer data from other I/O operations. This condition is called “logical disconnect”.

With no disconnect time, the port busy condition during ESCON channel reconnect time does not exist also, and this is another improvement over ESCON channels. Note that the channel subsystem reported connect times are not affected by logical disconnected times. The logical disconnect time is accumulated by the ESS as a component of *connect* time, but the *connect* time reported by the channel is calculated by excluding the logical disconnect time. The logical disconnect time is reported as part of the *pend* time.

5.10.3 FICON benefits at your installation

The most obvious benefit of FICON connectivity are the increased per channel bandwidth and greater simplicity of channel fabric. This speed allows to collapse existing ESCON channels into FICON channels approximately 4:1 and more.

Greater simplicity of configuration is also achieved because the FICON architecture allows more devices per channel and an increased bandwidth.

Single stream sequential operations benefit from improvements in throughput, so that elapsed time for key batch, data mining, or dump operations dramatically improve. This provides relieve to those installations where the batch of file maintenance windows are constrained today.

Response time improvements may accrue for some customers, particularly for data stored using larger block sizes. The data transfer portion of the response time is greatly reduced because of the data rate during transfer being six times faster than ESCON. This improvement leads to significant connect time reduction. The larger the transfer, the greater the reduction as a percentage of the total I/O service time.

Pend time caused by director port busy is totally eliminated because there are no more collisions in the director with FICON architecture.

PAV and FICON work together to allow multiple data transfers to the same volume at the same time over the same channel, providing greater parallelism and greater bandwidth while simplifying the configurations and the storage management activities.

5.11 High performance disk drives

Figure 119. IBM Ultrastar disk drive

The disk drives used in the Enterprise Storage Server are IBM’s latest high performance SSA disk drives.

Three disk drive capacities are used in the ESS 9.1 GB, 18.2 GB and 36.4 GB, and they have the following characteristics:

- 10 krpm: delivers an outstanding sustained data rate up to 29.5 MB/second, and a maximum of 44.3 MB/second
- Maximum average seek time of 5.4 msec., for the 36.4GB DDMs, and 4.9 msec. for the 9.1GB and 18.2GB DDMs.
- The track-to-track minimum seek is 0.3 msec.
- The latency is only 3ms. because of the 10 krpm.
- 40 MB/second bandwidth SSA interface
- Buffer of 2 MB for read and writes

Note: As the technology characteristics of the disk drives improve constantly, you should check the latest specifications for the disks that come with the ESS. You can refer to: `http://www.storage.ibm.com/hardsoft/diskdrdl.htm` for updated information and additional details of the IBM Ultrastar disk drives.

5.12 Configuring for performance

Here are recommendations for configuring for performance.

5.12.1 Host adapters

Always spread the host connections across all the host adapter bays. This recommendation is for two reasons:

1. The bays are connected to different PCI buses in each cluster, and by spreading the adapter connections to the host adapters across the bays, you also spread the load and improve overall performance and throughput.
2. Distribute the connections to the host adapters across the bays in the following sequence:

   Bay 1 - Bay 3 - Bay 2 - Bay 4
3. If you need to replace or upgrade a host adapter in a bay, then you have to quiesce all the adapters in that bay. If you spread them evenly, then you will only have to quiesce a quarter of your adapters. For example, for an ESCON configuration with 8 ESCON links spread across the 4 bays, then the loss of 2 ESCON links out of 8 may have only a small impact, compared with all 8 if they were all installed in one bay.
4. Take into consideration that HAs in bays 1 and 3 share the same internal bus in each cluster, and HAs in bays 2 and 4 share the other internal bus in the cluster. This is especially important for open systems to avoid the situation where all the activity for a cluster comes from bays 1 and 3, or from bays 2 and 4.

For the zSeries environments, there are certain specific recommendations in order to fully take advantage of the ESS performance capacity. Else you will be using just part of the total ESS performance capability.

When configuring for ESCON, consider these recommendations:

- Configure 16 ESCON channels to the ESS
- Use 8-path groups
- Plug channels for an 8-path group into four HAs (i.e. use one HA per bay)
- Each 8-path group should access its LCU on one cluster
- One 8-path group is better than two 4-path groups

This way, both the channels connected to the same HA will serve only even or only odd LCUs, which is the best, and access will be distributed over the 4 HA bays.

When configuring for FICON consider the following recommendations:

- Define a minimum of four channel paths per CU. Fewer channel paths will not allow exploitation of full ESS bandwidth. A more typical configuration would have eight FICON channels.
- Spread FICON host adapters across all adapter bays. This should result in minimally one host adapter per bay, or in a typically configured ESS, two host adapters per bay.
- Define a minimum of four FICON channels per path group.

See Section 5.10, “FICON host adapters” on page 173 and Section 4.16, “FICON host connectivity” on page 139 for more details on ESS FICON attachment characteristics and configuration.

5.12.2 RAID ranks

For almost all environments, you will get the best performance and availability from using RAID ranks. This applies to both open systems servers and zSeries servers environments. The common perception that RAID-5 has poor performance characteristics is not true for the ESS. For both the zSeries and the open systems servers, the fast write capability of the ESS masks all RAID-5 operations, so you get high performance for all writes.

For the best performance, spread your I/O activity across all the RAID ranks.

If you must optimize your configuration of RAID ranks for maximum performance, you can create a configuration where you connect only one RAID rank to each SSA adapter within the ESS (although loop A and loop B are both attached to the same SSA adapter each loop has its own processor). Each RAID rank will be serviced by its own logical subsystem (or LCU in zSeries terms). This way, you get the maximum benefit from the 16 high performance loops by each SSA adapter only having to manage 8 disks.

5.12.3 JBOD

A group of non-RAID disk drives can be used in a situation where you need high random write performance and you do not need RAID protection. A good example of this could be the IMS Write Ahead Data Set (WADS), where the data set can be duplexed by software and you have a high amount of update writes. There may be other examples in UNIX, NT, and OS/400. Although the write performance is very good because we do not have any RAID-5 write penalty, read performance is likely to be worse than a RAID array, which can spread the data across multiple disk drives. In terms of sequential read and write throughput, the RAID array will be better— because of the spread of data over the array and the use of stripe writes to eliminate the RAID-5 write penalty (RAID-3 like).

Note: Always remember that with JBOD data is not protected. Should a disk drive fail, then you loose all the data in the disk group.

5.12.4 Cache

Cache size can be 8, 16, 24 or 32 GB. With different cache sizes, the Enterprise Storage Server can be configured for the appropriate balance between disk capacity and cache size. Alternatively, if you don’t increase in disk capacity and double the cache size, ESS improves the cache hit ratio, resulting in better I/O response times to the application.

The large ESS cache size, together with the ESS advanced cache management algorithms, provide high performance for a variety of workloads.

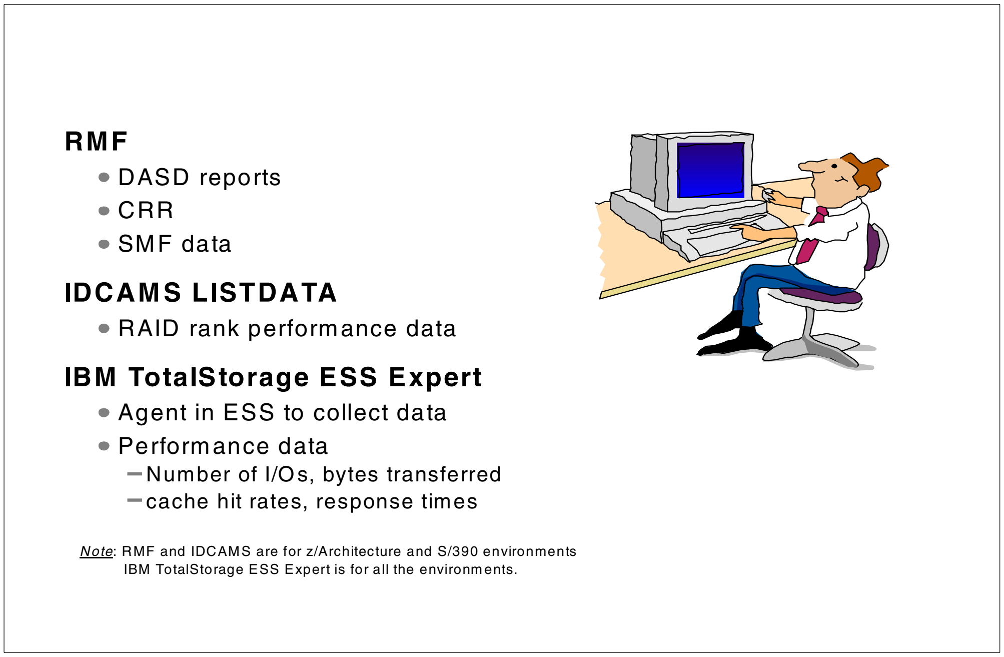
5.12.5 Logical volumes

Use the standard facility of interleaving for zSeries volumes when defining the RAID ranks. This spreads the zSeries volume cylinders across the array and can maximize the benefits of having multiple disk drives.

Use *custom volumes* where you need to place high performance data sets that should be on their own logical volume. This is particularly useful if you are on a release of OS/390 that does not support PAVs, or if you are on VM/ESA or VSE/ESA. You can define a custom volume just large enough to hold your critical data set, thus reducing wasted space.

5.13 Measurement tools

You may wish to perform performance measurements and decide what tool you might be using for this purpose. There are a couple of tools available either for the zSeries environment and for the open systems environments.



RMF

- DASD reports
- CRR
- SMF data

IDCAMS LISTDATA

- RAID rank performance data

IBM TotalStorage ESS Expert

- Agent in ESS to collect data
- Performance data
  - Number of I/Os, bytes transferred
  - cache hit rates, response times

Note: RMF and IDCAMS are for z/Architecture and S/390 environments
IBM TotalStorage ESS Expert is for all the environments.

Figure 120. Performance measurement tools

Today's storage subsystems have very complex features such as multiple host adapters (fiber channel, ESCON, FICON, SCSI), HDD device adapters, large cache memory sizes, multiple clusters, RAID protected disk arrays, all tied together to provide large bandwidth to satisfy performance requirements. This multiplicity of features is especially true with ESS. Working with more than one tool may result useful when managing today storage servers.

For more information on ESS performance refer to *IBM Enterprise Storage Server Performance Monitoring and Tuning Guide,* SG24-5656.

5.13.1 RMF

If you are in a z/OS environment, RMF will help you look at ESS and discover how well logical volumes and control units are performing. You will get information such as I/O rate at the device and logical control unit level, and response time for specific volumes. Also information on the number of I/O requests that are completed within cache memory (cache hits).

Even RMF will report on ESS data in a similar way to a 3990, some specific considerations apply for ESS. Device addresses are reported with response time, connect, disconnect, PEND and, IOSQ times as usual. Alias addresses for PAVs are not reported, but RMF will report the number of alias (or in RMF terms, exposures) that have been used by a device and whether the number of exposures has changed during the reporting interval (The MX field shows the number of exposures, which is the number of aliases + 1, the base). RMF cache statistics are collected and reported by logical subsystem (LCU in zSeries terminology). So a fully configured ESS will produce 16 sets of cache data.

Note: For FICON attachment there are changes in the way the components of the I/O operations add up to the values reported by RMF. You may refer to Section 5.10.2, “FICON I/O operation” on page 174 for detailed explanation of the FICON I/O operation components.

If you are using z/OS, RMF can work together with ESS to retrieve performance statistics data, examining its cache report. However, these statistics will be related to the activity of the z/OS images, and will not show the activity of any other open systems servers that may be attached to the same ESS. In this situations the IBM TotalStorage ESS Expert may be a helpful complement, because it gathers information from the various diverse host operating systems running on processors attached to the ESS.

ESS Cache and NVS reporting

RMF cache reporting and the results of a LISTDATA STATUS command, report a cache size of 4, 8, 12, 16 GB and an NVS size of 192 MB — half the actual size. This is because the information returned represents only the cluster to which the logical control unit is attached. Each LCU on the cluster reflects the cache and NVS size of that cluster. z/OS users will find that only the SETCACHE CFW ON | OFF command is supported, while other SETCACHE command options (for example, DEVICE, SUBSYSTEM, DFW, NVS) are not accepted.

FICON

RMF has been changed to support FICON channels. With APAR OW35586, RMF extends the information in the *Channel Path Activity* report of all monitors by reporting about data transfer rates and bus utilization values for FICON channels. There are five new measurements reported by RMF in a FICON channel environment:

- Bus utilization
- Read bandwidth for a partition in MB/second
- Read bandwidth total (all logical partitions on the processor) in MB/second
- Write bandwidth for a partition in MB/second
- Write bandwidth total in MB/second

Also there has been some SMF record changes for FICON. For more information about the new fields reported by RMF for FICON channels and SMF record changes you may refer to *FICON Native Implementation and Reference Guide,* SG24-6266.

5.13.2 IBM TotalStorage ESS Expert

IBM TotalStorage Enterprise Storage Server Expert (ESS Expert), a product of the IBM TotalStorage family, enables you to gather various information about your Enterprise Storage Server. ESS Expert helps you in the following activities:

1. Performance management
2. Asset management
3. Capacity management

This section gives an introduction and brief description of the Performance Management tasks the ESS Experts offers. For more extensive and detailed information on the IBM TotalStorage Expert and uses, see *IBM StorWatch Expert Hands-On Usage Guide,* SG24-6102, and *IBM Enterprise Storage Server Performance Monitoring and Tuning Guide,* SG24-5656.

5.13.2.1 How does ESS Expert work

Figure 121 shows how ESS Expert works with your ESSs. ESS Expert communicates with your ESSs through a TCP/IP network. Therefore, you can gather information on ESSs across your local or wide area network as long as you have a communication path between the ESS Expert and the ESSs. ESS Expert itself runs under a Windows NT or AIX V4.3.2 operating system. However, you do not have to operate ESS Expert right where it runs, since the user interface to the ESS Expert is a Web browser interface. In other words, you can operate ESS Expert through Netscape Navigator or Microsoft Internet Explorer.

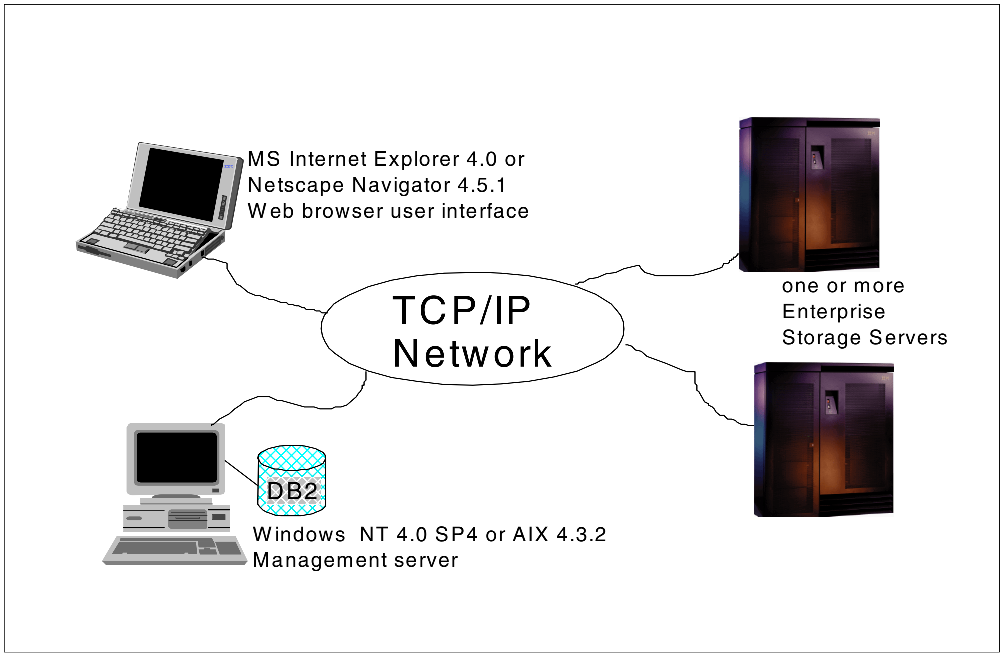

MS Internet Explorer 4.0 or
Netscape Navigator 4.5.1
Web browser user interface
TCP/IP Network
one or more Enterprise Storage Servers
DB2
Windows NT 4.0 SP4 or AIX 4.3.2
Management server

Figure 121. How Expert communicates with ESS

ESS Expert collects from your ESSs information about their capacity or performance. ESS Expert stores the collected information into a DB2 database and produces the reports that you request.

5.13.2.2 Expert reports

ESS Expert provides basically three types of reports:

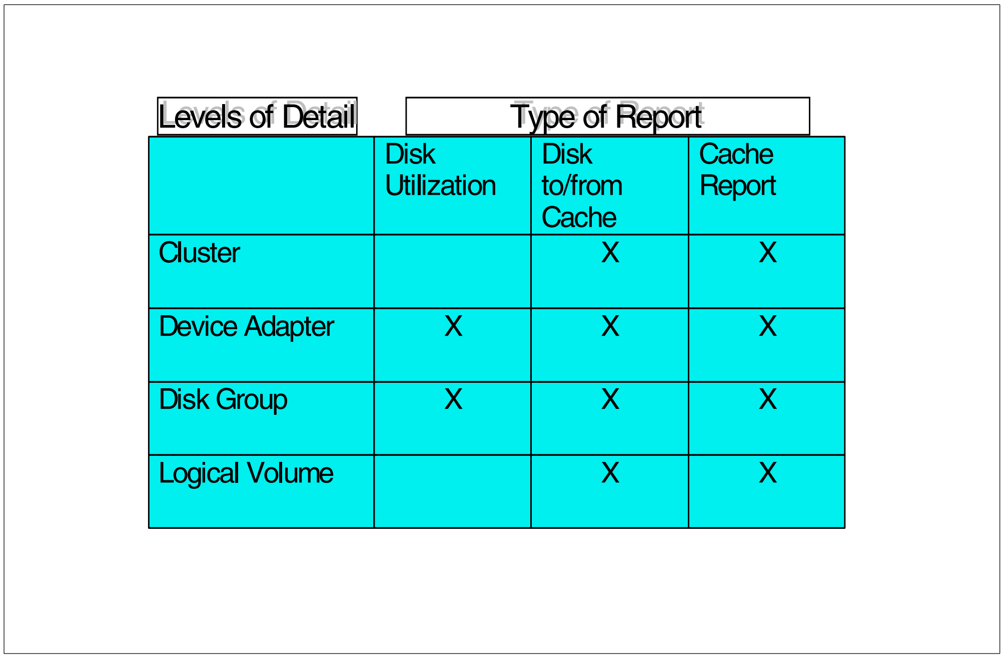
- Disk Utilization
- Disk-Cache
- Cache (bandwidth to/from the cache)

Additionally ESS Expert provides four levels of reports under each type:

- Cluster
- SSA Device Adapter
- Disk Group (RAID5 or non-RAID)
- Logical Volume.

For example, you may want to know what is the disk drive utilization for a given RAID-5 disk group to determine whether or not some of its files should be moved to other arrays, in order to alleviate the I/O activity.

The following matrix (Figure 122) shows the possible combination of performance reports you may want to get.



Levels of Detail	Type of Report		
	Disk Utilization	Disk to/from Cache	Cache Report
Cluster		X	X
Device Adapter	X	X	X
Disk Group	X	X	X
Logical Volume		X	X

Figure 122. IBM TotalStorage ESS Expert report matrix

Figure 123 shows the ESS components reported by the IBM TotalStorage ESS Expert.

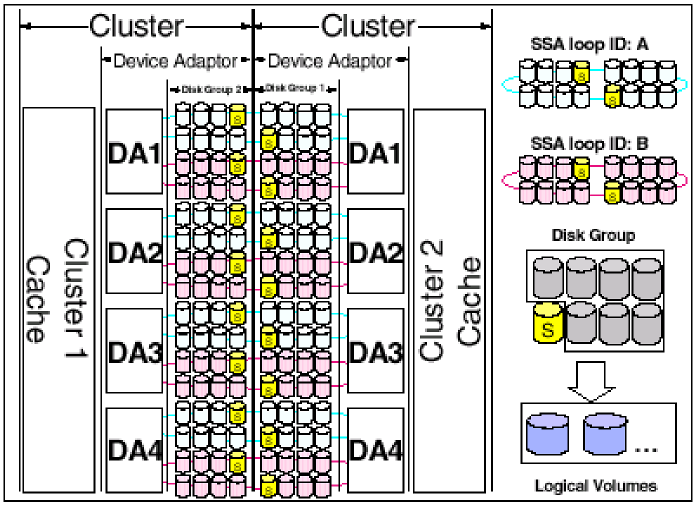


Figure 123. ESS Expert reported components

5.13.2.3 Cluster reports

Cluster reports show entire subsystem performance.

An ESS has two clusters, and each cluster independently provides major functions for the storage subsystem. Some examples include directing host adapters for data transferring to/from host processors, managing cache resources, and directing lower device interfaces for data transferring to/from backend disk drives.

This way, the cluster level of report gives you a performance overview from the viewpoint of the subsystem. Please note that *Disk Utilization* reports do not have *Cluster* level reports, as *Disk Utilization* reports deal with physical device performance.

5.13.2.4 SSA Device Adapter reports

Device Adapter reports show I/O distribution across clusters.

Each cluster on an ESS has four device adapters. As you can see Figure 123, a device adapter in a cluster is paired with the other cluster's device adapter, and each device adapter pair has up to two SSA disk loops. Though they make a pair, they usually work independently, and each manages separate groups of DDMs under normal operation. *Device Adapter* level reports help you understand the I/O workload distribution among device adapters/loops on either cluster of the ESS.

5.13.2.5 Disk Group reports

Disk Group reports show I/O distribution in a device adapter.

A disk group is an array of seven or eight DDMs which are on the same SSA loop. Ranks are configured from the disk groups. A rank can be either RAID-5 format, or non-RAID which is referred as “Just a Bunch Of Disks” (JBODs). The rank belongs to a device adapter, and the device adapter controls it under normal operation. The *Disk Group* level of reports helps you understand I/O workload distribution among disk groups in a loop in a certain device adapter.

Note: ESS Expert reports show a disk group and a disk number. If a disk group is configured as JBODs, ESS Expert will show a disk number for each disk drive. If a rank is in RAID-5, ESS Expert will show the disk number as “N/A”.

5.13.2.6 Logical Volumes reports

Logical Volume reports show I/O distribution in a disk group.

A logical volume belongs to a disk group. Host systems view logical volumes as their logical devices. A disk group has multiple logical volumes, depending on your definitions. So this level of reports helps you understand I/O workload distribution among logical volumes in a disk group. Please note that *Disk Utilization* reports do not have this level of report, as *Disk Utilization* reports deal with physical device performance.

5.13.2.7 Examples

Please refer to Appendix A, “IBM TotalStorage ESS Expert” on page 281 for further information and examples of ESS Expert reports.

Chapter 6. Copy functions

The Enterprise Storage Server supports several hardware-assisted copy functions for two purposes: mirroring for disaster recovery solutions, and copy functions that provide an instant copy of the data. The ability to make an instant copy is also called time zero (T0) copy.

In this chapter we see the advanced copy functions that the Enterprise Storage Server has available for the open systems environments and for the zSeries environments

6.1 Copy Services functions

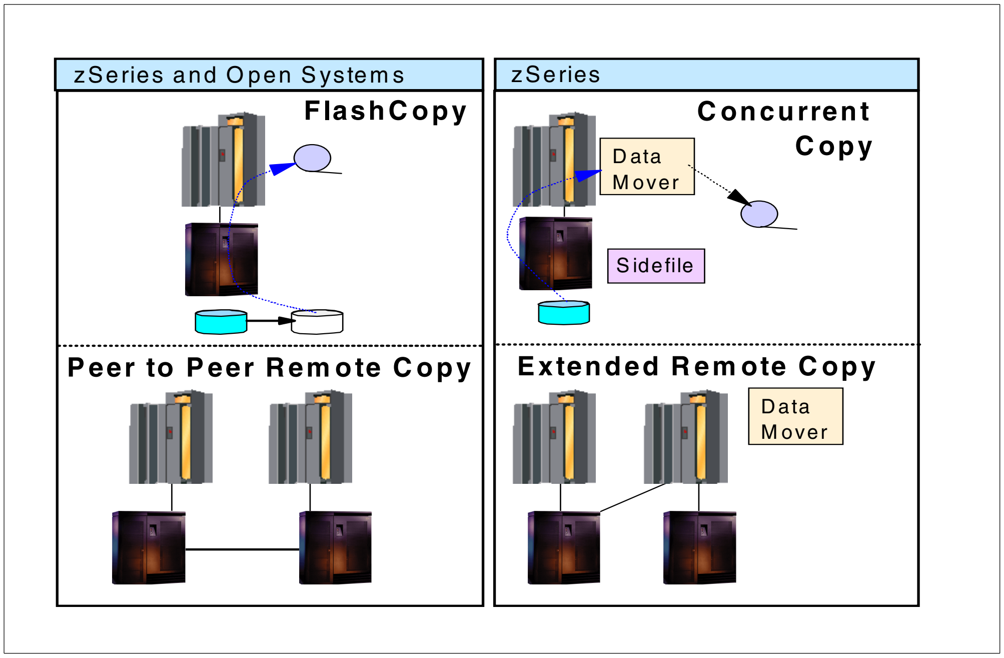


Figure 124. ESS advanced copy functions

The hardware assisted copy functions of the Enterprise Storage Server (ESS) are (see Figure 124):

- *FlashCopy* allows you to make a point-in-time copy of a volume — open systems and zSeries environments. Point-in-time copy gives you an instantaneous copy, or view, of what the original data looked like at a specific point-in-time. This is known as a T0 (time-zero) copy.
- *Concurrent Copy* also allows for a point-in-time T0 copy — zSeries environments only. It works on a volume or data set basis and uses cache side files in the ESS.
- *Peer-to-Peer Remote Copy* (PPRC) maintains synchronous mirror copies of volumes on remote ESSs — open systems and zSeries environments. In contrast, PPRC is a T1 copy because the secondary is continuously updated.
- *Extended Remote Copy* (XRC) maintains asynchronous mirror copies of volumes on remote storage systems over large distances — zSeries environments only. XRC is also a T1 copy because the secondary is continuously updated.

ESS advanced copy functions are enterprise-level functions that give you leading edge solutions to meet your needs for disaster recovery, data migration, and data duplication.

For further information on the characteristics and on the implementation of the ESS copy functions you may refer to *Implementing ESS Copy Services on UNIX and Windows NT/2000,* SG24-5757 and *Implementing ESS Copy Services on S/390,* SG24-5680.

6.2 ESS Copy Services overview

The ESS Specialist Copy Services runs inside the Enterprise Storage Server. It has a Web browser interface that provides a means to set up the Peer-to-Peer Remote Copy and FlashCopy functions, and manage them for both the zSeries and the open systems environments using a standard Web browser.

For some open system servers the ESS Specialist Copy Services functions can also be invoked using a Command Line Interface (CLI).

For the zSeries environments, the ESS Copy Services functions can be invoked by means of specific commands (TSO/E commands), system utilities (ICKDSF, DFSMSdss, IXFP SNAP command), or macros (ANTRQST macro). We will refer to them later, when describing the way of invoking each copy function.

6.2.1 Web interface

The ESS Specialist Copy Services has a Web browser interface that provides a means to set up and manage the FlashCopy and the Peer-to-Peer Remote Copy advanced copy functions of the Enterprise Storage Server for all the open systems and the zSeries environments.

Using a Web browser gives the possibility to easily control the ESS copy functionality over the network from any platform for which the browse is supported.

The ESS Copy Services require one of the following Internet browsers:

- Netscape Communicator 4.6 or above
- Microsoft Internet Explorer (MSIE) 5.0 or above

6.2.2 Command Line Interface (CLI)

The ESS Copy Services Command Line Interface (CLI) is available for the supported open system host platforms. Using the Command Line Interface, you are able to communicate with the ESS Copy Services server from the host's command line. An example would be to automate tasks like doing a FlashCopy by invoking the Copy Services commands within customized scripts.

The Command Line Interface is available for the following operating systems:

- AIX
- Sun Solaris
- HP-UX
- Windows NT 4.0 and Windows 200

6.3 The Web interface

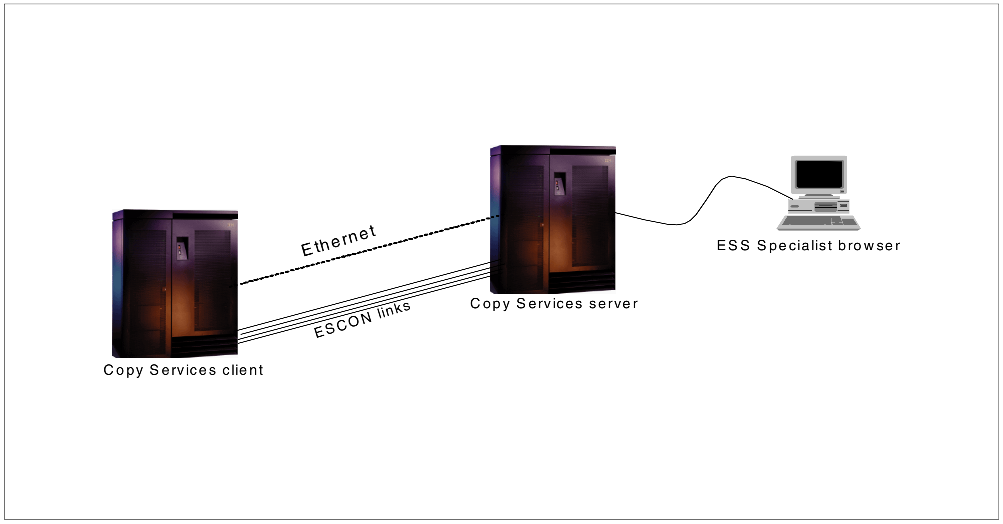


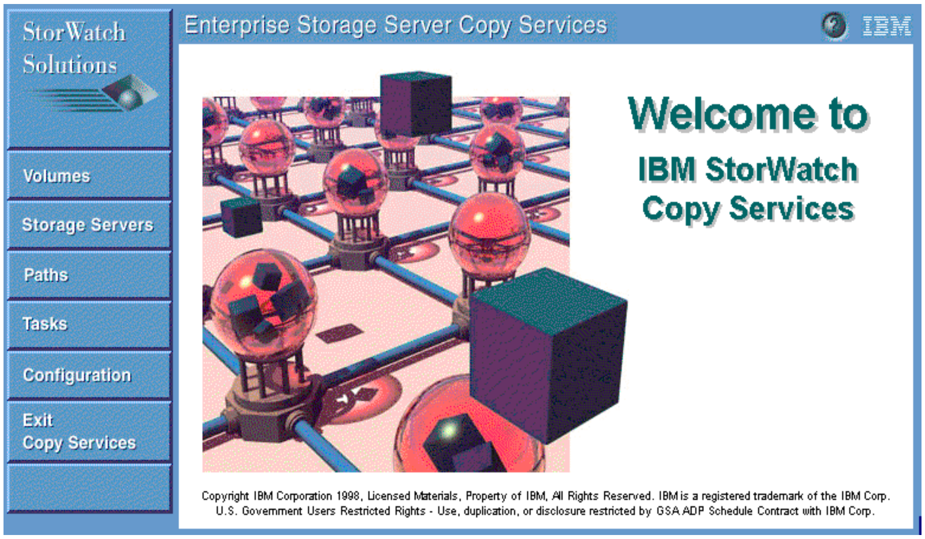
Figure 125. ESS Copy Services server and clients

6.3.1 Client/server setup

With the ESS Copy services running inside the ESSs, one ESS has to be defined as the Copy Services server. Optionally, there could be a second ESS defined to be the backup Copy Sevices server. On each ESS that is intended to use Copy Services, there is a Copy Services client running who communicates to the server.

The ESS Copy Services server holds the Copy Services related information. The server can be any ESS, a primary ESS, a secondary ESS, or an ESS without any PPRC or FlashCopy volumes. The server ESS, however, needs Ethernet connections to each client ESS (see Figure 125). To define the server ESS, you must specify its TCP/IP address in one of the configuration screens of the ESS Specialist for each client ESS.

6.3.2 ESS Copy Services panels



Welcome to IBM StorWatch Copy Services

Copyright IBM Corporation 1998, Licensed Materials, Property of IBM, All Rights Reserved. IBM is a registered trademark of the IBM Corp.
U.S. Government Users Restricted Rights - Use, duplication, or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Figure 126. ESS Copy Services Welcome menu

Whether you will be creating PPRC or FlashCopy volumes, ESS Specialist Copy Services provides several screens to use for both functions.

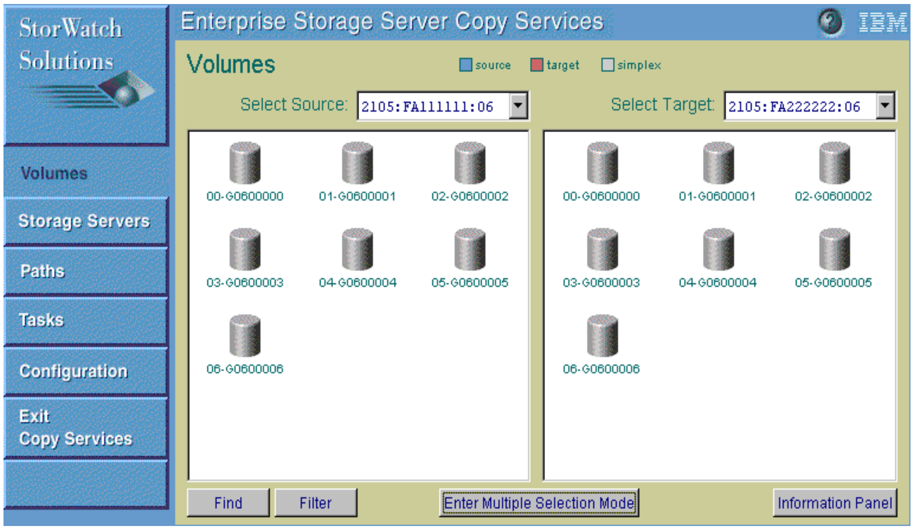
To start the Copy Services function, you must select the Copy Services option from the ESS Specialist panel. When you select this, then the *ESS Copy Services Welcome* panel will be presented (see Figure 126).

As you can see on the left side of Figure 126, you can choose any of the following options from the welcome panel:

Volumes:	Operate on volumes
Storage Servers:	Operate on groups of volumes
Paths:	Operate on paths
Tasks:	Manage tasks
Configuration:	View log, identify server
Exit Copy Services:	Return to ESS Specialist

In this section we will describe some of the initial common screens. For a complete description of the specific screens you will be seeing when invoking the individual copy functions of PPRC or FlashCopy, we recommend you to refer to the publications *Implementing ESS Copy Services on UNIX and Windows NT/ 2000,* SG24-5757 and *Implementing ESS Copy Services on S/390,* SG24-5680.

6.3.2.1 The Volumes panel



Select Source: 2105:FA111111:06

Select Target: 2105:FA222222:06

Figure 127. ESS Copy Services - Volumes panel

From the *Volumes* panel you will be able to:

- Get information and status about volumes defined in a logical storage subsystem (LSS) of the ESS
- Select source and target volume for a PPRC or FlashCopy task
- Filter the output of the volume display to a selected range
- Search for a specific volume based on its unique volume ID.
- Establish, terminate and suspend PPRC Copy pairs and optionally save the operation task
- Establish and withdraw FlashCopy pairs and optionally save this operation as a task
- Enter the multiple selection mode for PPRC and FlashCopy

Figure 127 shows the volumes displayed in the source and target areas of the *Volumes* panel once the corresponding LSS selections have been made. From this menu you will be able to select the source and target volumes for PPRC and FlashCopy tasks.

6.3.2.2 The Storage Server panel

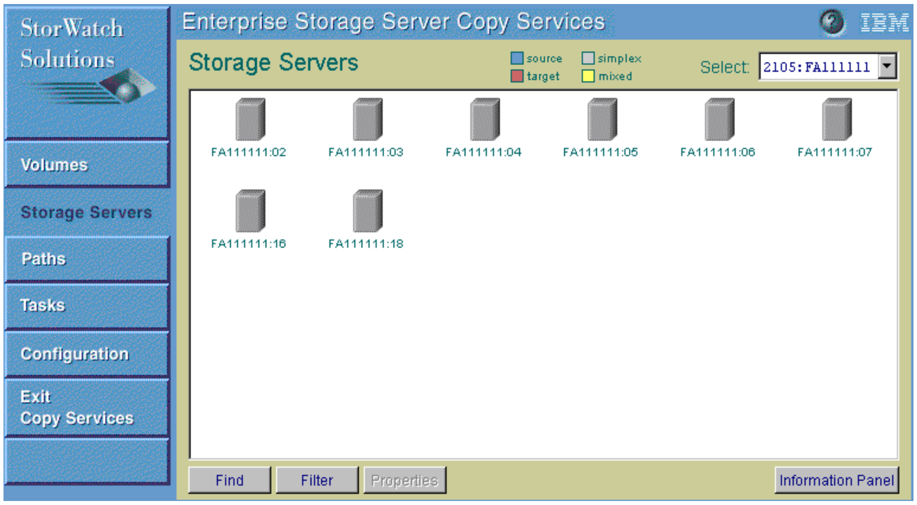


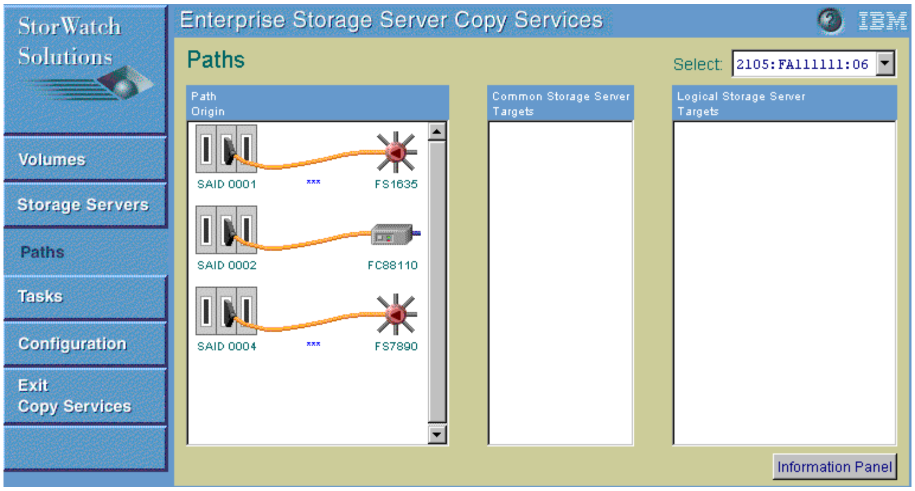
Figure 128. ESS Copy Services - Storage Servers panel

The *Storage Server*s panel displays the Enterprise Storage Servers and the logical subsystems within the storage network. The storage network includes all the ESSs that are configured to use the same ESS Copy Sevices server. With the *Storage Server*s panel you will be able to:

- View all Enterprise Storage Servers within the storage network
- View all logical subsystems (LSSs) within the storage network
- Get information about a logical subsystem and its status
- View and modify the copy properties of a logical subsystem
- Filter the output of a selected range
- Search for a specific logical subsystem based on its unique address
- Establish, terminate, and suspend PPRC Copy pairs, and optionally save them as a task

Figure 128 shows the *Storage Servers* panel output for a selected ESS. The color indicates the state of the LSS, whether it contains volumes that are currently in any copy relationship (source, target, mixed) or not part of a copy pair at all.

6.3.2.3 The Paths panel



Select: 2105:FA111111:06

Figure 129. ESS Copy Services - Paths panel

A path is used to send data between the source and target PPRC pairs. The physical path consists of the ESCON connection between two ESSs while the logical path describes the connection of the PPRC source and targets. There could be multiple logical paths established over a single physical path.

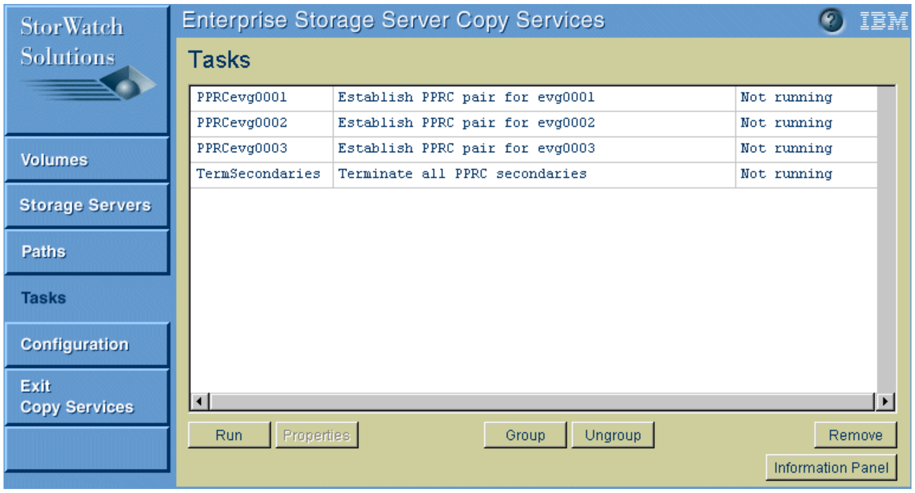
From the *Paths* panel you will be able to:

- Establish PPRC paths
- Remove PPRC paths
- View information about PPRC paths

A path is always specified between two logical subsystems of the ESSs. Volumes on these logical subsystems can use the paths defined to transfer PPRC data. For availability and performance reasons it is recommended to define multiple physical paths between PPRC source and targets.

Figure 129 shows all configured ESCON adapters once an LSS was selected.

6.3.2.4 The Tasks panel



PPRCevg0001	Establish PPRC pair for evg0001	Not running
PPRCevg0002	Establish PPRC pair for evg0002	Not running
PPRCevg0003	Establish PPRC pair for evg0003	Not running
TermSecondaries	Terminate all PPRC secondaries	Not running

Figure 130. ESS Copy Services - Task introduction panel

By using the Volumes, Storage Servers, and Paths panels of the ESS Copy Services, you can create, save and run one or more tasks, which perform particular functions when you choose to run them. When you run a task, the ESS Copy Services server tests whether there are existing conflicts. If conflicts exist, then the server ends this task.

Figure 130 shows a sample *Tasks* introduction panel. As can be seen in the figure, there are four tasks that have been previously defined. From this panel you can select and run a previously defined task.

6.4 Command Line Interface (CLI)

The ESS Copy Services Command Line Interface (CLI) provides you a mean to communicate with the ESS Copy Services server from the host's command line. It is available for the AIX, Sun Solaris, HP-UX, Windows NT and Windows 2000 operating environments (see 7.12, "Command Line Interface (CLI)" on page 252 for software requirements).

The Command Line Interface is Java based, and therefore the Java runtime environment needs to be installed on each host system from which you want to issue the commands.

The host system does not necessarily need to be connected to storage assigned to one of the host adapter ports of the ESS. The only requirement is that the host from where you want to invoke the commands is connected to the ESS that is defined as the primary Copy Services server via a local area network (LAN). However some options can only be used when the host is physically connected to the ESS storage.

Copy Services commands

For both the UNIX and Windows, the same Copy Services commands are available. The command set for UNIX operating systems consists of shell scripts with the *.sh ending; the commands for the Windows operating systems are batch files with the *.bat ending. Functionally they are identical, but there may be some differences in the parameters you can specify when invoking the commands.

- **rsExecuteTask** (.sh .bat)

  Accepts and executes one or more pre-defined Copy Services tasks. Waits for these tasks to complete execution.

- **rsList2105s** (.sh .bat)

  Displays the mapping of host physical volume name to 2105 (ESS) volume serial number

- **rsPrimeServer** (.sh .bat)

  Notifies the Copy Services server of the mapping of host disk name to 2105 (ESS) volume serial number. This command is useful when the Copy Services Web screens are used to perform FlashCopy and/or PPRC functions. Collects the mapping of the host physical volume names to the ESS Specialist Copy Services server. This permits a host volume view from the ESS Specialist Copy Services Web screen.

- **rsQuery** (.sh .bat)

  Queries the FlashCopy and PPRC status of one or more volumes

- **rsQueryComplete** (.sh .bat)

  Accepts a pre-defined Copy Services server task name and determines whether all volumes defined in that task have completed their PPRC copy initialization. If not, this command waits for that initialization to complete.

- **rsTestConnection** (.sh .bat)

  Determines whether the Copy Services server can successfully be contacted.

Scripting the Command Line Interface

You can enhance the functionality of the Copy Services Command Line Interface by incorporating its use in your own customized scripts. Common applications of the CLI might include batch, automation, and custom utilities.

6.5 FlashCopy

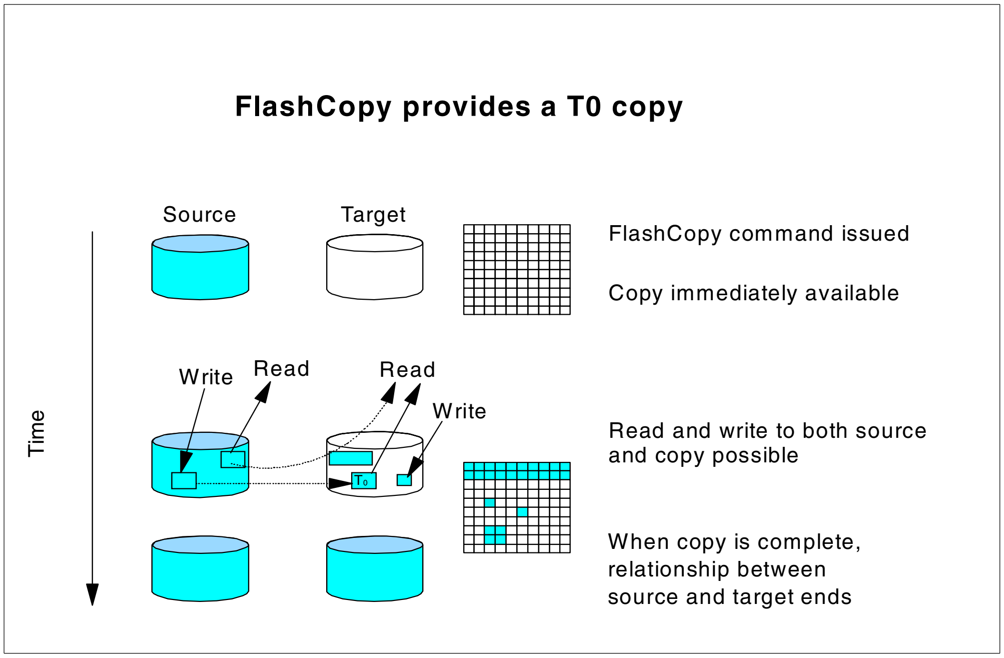

FlashCopy provides a T0 copy
Source
Target
FlashCopy command issued
Copy immediately available
Time
Write
Read
Read
Write
T_0
Read and write to both source and copy possible
When copy is complete, relationship between source and target ends

Figure 131. FlashCopy

6.5.1 Operation

FlashCopy provides a point-in-time copy of an ESS logical volume. The point-in-time copy functions give you an instantaneous copy, or “view”, of what the original data looked like at a specific point-in-time. This is known as the T0 (time-zero) copy.

When a FlashCopy is invoked, the command returns to the operating system as soon as the FlashCopy pair has been established and the necessary control bitmaps have been created. This process takes only a few seconds to complete. Thereafter, you have access to a T0 copy of the source volume. As soon as the pair has been established, you can read and write to both the source and the target logical volumes.

The point-in-time copy created by FlashCopy is typically used where you need a copy of production data to be produced with minimal application downtime. It can be used for online backup, testing of new applications, or for creating a database for data-mining purposes. The copy looks exactly like the original source volume and is an instantly available, binary copy. See Figure 131 for an illustration of FlashCopy concepts.

FlashCopy is possible only between logical volumes in the same logical subsystem (LSS). The source and target volumes can be on the same or on different arrays, but only if they are part of the same LSS. A source volume and the target can be involved in only one FlashCopy relationship at a time. When you set up the copy, a relationship is established between the source and the target volume and a bitmap of the source volume is created.

Once this relationship is established and the bitmap created, the target volume copy can be accessed as though all the data had been physically copied. While a relationship between source and target volumes exists, a background task copies the tracks from the source to the target. The relationship ends when the physical background copy task has completed.

You can suppress the background copy task using the *Do not perform background copy* (NOCOPY) option. This may be useful if you need the copy only for a short time, such as making a backup to tape. If you start a FlashCopy with the *Do not perform background copy* option, you must withdraw the pair (a function you can select) to end the relationship between source and target. Note that you still need a target volume of at least the same size as the source volume.

You cannot create a FlashCopy on one type of operating system and make it available to a different type of operating system. You can make the target available to another host running the same type of operating system.

6.5.2 Consistency

At the time when FlashCopy is started, the target volume is basically empty. The background copy task copies data from the source to the target. The FlashCopy bitmap keeps track of which data has been copied from source to target. If an application wants to read some data from the target that has not yet been copied to the target, the data is read from the source; otherwise, the read is satisfied from the target volume (See Figure 131). When the bitmap is updated for a particular piece of data, it signifies that source data has been copied to the target and updated on the source. Further updates to the same area are ignored by FlashCopy. This is the essence of the T0 point-in-time copy mechanism.

When an application updates a track on the source that has not yet been copied, the track is copied to the target volume (See Figure 131). Reads that are subsequently directed to this track on the target volume are now satisfied from the target volume instead of the source volume. After some time, all tracks will have been copied to the target volume, and the FlashCopy relationship will end.

Please note FlashCopy can operate at extent level in zOS environments.

Applications do not have to be stopped for FlashCopy. However, you have to manage the data consistency between different volumes. Either applications have to be frozen consistently, or you have to invoke built functions which maintain that consistency.

Today, more than ever, organizations require their applications to be available 24 hours per day, seven days per week (24x7). They require high availability, minimal application downtime for maintenance, and the ability to perform data backups with the shortest possible application outage.

6.6 Invocation of FlashCopy

FlashCopy can be invoked by different methods in different environments.

6.6.1 zSeries environments

In the z/OS environments, FlashCopy can be invoked by four different methods.

- DFSMSdss can be used to invoke FlashCopy in batch via the ADRDSSU program, with the *Copy Full* utility function
- TSO/E commands that are unique to FlashCopy. The three new commands are:
  - FCESTABL — used to establish a FlashCopy relationship
  - FCQUERY — used to query the status of a device
  - FCWITHDR — used to withdraw a FlashCopy relationship
- DFSMSdfp Advanced Services ANTRQST macro which calls the System Data Mover API. The new ANTRQST macro supports the establish, query and withdraw FlashCopy requests.
- ESS Specialist Copy Services Web interface. The ESS Copy Services that runs in the ESS provides a Web browser interface that can be used to control the FlashCopy functions in a zSeries environment. See 6.3, “The Web interface” on page 194 for further information.

In the z/VM environments, FlashCopy can be managed by use of the ESS Specialist Copy Services Web browser interface. FlashCopy is also supported for guest use (z/OS’s DFSMSdss COPY) for dedicated (attached) volumes or for full pack mini disks.

VSE/ESA Version 2 Release 5 provides the support of the FlashCopy function. Once the FlashCopy function is purchased for the ESS, support is available for VSE Version 2 Release 5 at no additional VSE cost. VSE/ESA support is provided by the IXFP SNAP command.

You may refer to *Implementing ESS Copy Services on S/390,* SG24-5680 for detailed information on invoking FlashCopy services in the zSeries environments

6.6.2 Open systems environments

There are two different methods of using the ESS Copy Services in the open systems environments:

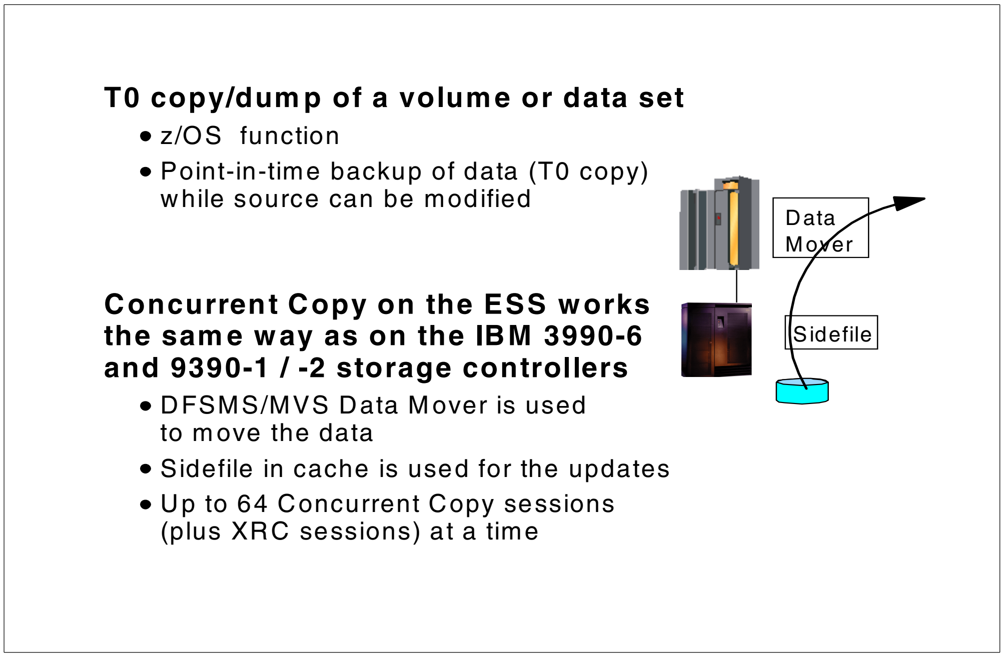
- ESS Specialist Copy Services Web interface. The ESS Copy Services that runs in the ESS provides a Web browser interface that can be used to control the FlashCopy functions in the open systems environments. For this situation you will be establishing a FlashCopy pair using the *Volumes* panel, or using the *Task* panel that the Web browser can present you. See 6.3, “The Web interface” on page 194.
- A Java-based Command Line Interface (CLI). The CLI interface allows administrators to execute Java-based Copy Services commands from a command line. This command line interface is currently available for the following operating systems: AIX, Sun Solaris, HP-UX and Windows NT and 2000. For this situation you will be starting a pre-define task from the command line or from one of your own customized scripts. See 6.4, “Command Line Interface (CLI)” on page 200.

You may refer to *Implementing ESS Copy Services on UNIX and Windows NT/ 2000,* SG24-5757 for detailed information on invoking FlashCopy services in the opens systems environments.

6.7 Concurrent Copy

Concurrent Copy is a copy function for the z/OS operating environments. Similar to FlashCopy, it creates an instantaneous point-in-time (T0) copy of the source. Concurrent Copy, however, works not only on a full volume basis, but also at a data set level. Also the target is not restricted only to DASD volumes in the same storage controller. For Concurrent Copy the target can also be a tape cartridge or a DASD volume on another physical storage controller.

The System Data Mover (SDM), a DFSMS/MVS component, reads the data from the source (volume or data set) and copies it to the target.



T0 copy/dump of a volume or data set

- z/OS function
- Point-in-time backup of data (T0 copy) while source can be modified

Concurrent Copy on the ESS works the same way as on the IBM 3990-6 and 9390-1 / -2 storage controllers

- DFSMS/MVS Data Mover is used to move the data
- Sidefile in cache is used for the updates
- Up to 64 Concurrent Copy sessions (plus XRC sessions) at a time

Figure 132. Concurrent Copy

6.7.1 Concurrent Copy process

For the copy process, we must distinguish between the *logical* completion of the copy and the *physical* completion. The copy process is logically complete when the System Data Mover has figured out what to copy. This is a very short process. After the logical completion, updates to the source are allowed while the System Data Mover, in cooperation with the Enterprise Storage Server, ensures that the copy reflects the state of the data when the Concurrent Copy command was issued. When an update to the source is to be performed and this data has not yet been copied to the target, the original data is first copied to a sidefile in cache before the source is updated, as shown in Figure 132.

6.7.2 Concurrent Copy on the ESS

Concurrent Copy on the Enterprise Storage Server works the same way as on the IBM 3990 Model 6 and the IBM 9390 models 1 and 2. Concurrent Copy is initiated using the CONCURRENT keyword in DFSMSdss or in applications that internally call DFSMSdss as the copy program, for example, DB2's COPY utility.

As on the IBM 3990s and 9390s, the System Data Mover establishes a *session* with the storage control unit. There can be up to 64 sessions active at a time (including sessions for Extended Remote Copy XRC copy function).

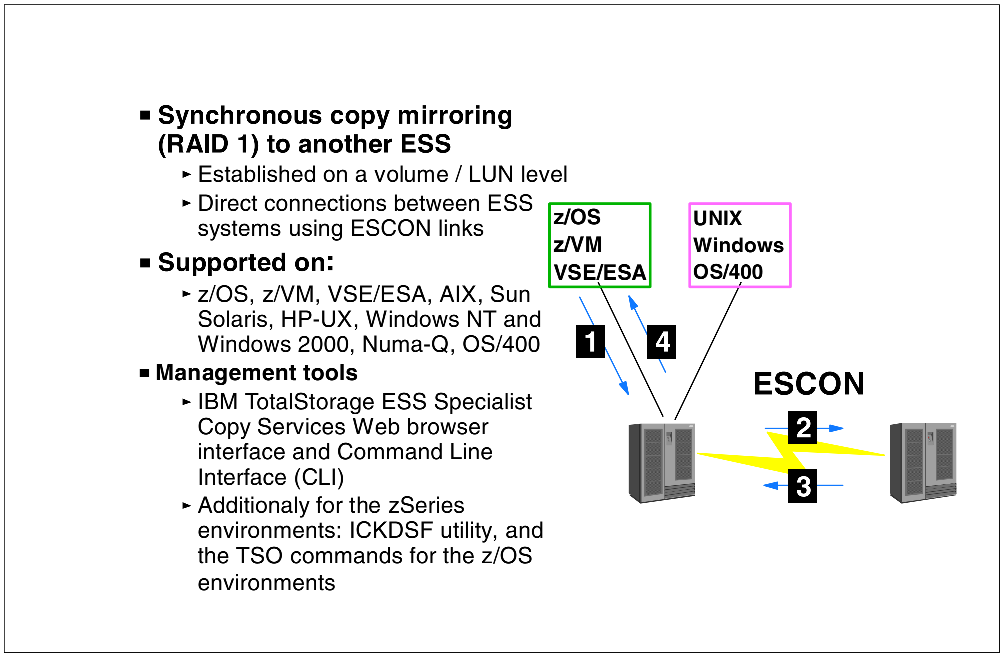
If you used Concurrent Copy on an IBM 3990 or 9390, or if you used Virtual Concurrent Copy on an IBM RVA, no changes are required when migrating to an ESS.

6.7.3 Concurrent Copy and FlashCopy

If DFSMSdss is instructed to do a Concurrent Copy by specifying the CONCURRENT keyword, and the copy is for a full volume with the target within the same logical storage subsystem of the ESS, then DFSMSdss will choose the fastest copy process and start a FlashCopy copy process instead of Concurrent Copy.

6.8 Peer-to-Peer Remote Copy

Peer-to-Peer Remote Copy (PPRC) is an established data mirroring technology that has been used for many years in the zSeries environments. Now it is also available for the open systems environments. It is used primarily to protect an organization’s data against disk subsystem loss or, in the worst case, complete site failure.



- **Synchronous copy mirroring (RAID 1) to another ESS**
  - Established on a volume / LUN level
  - Direct connections between ESS systems using ESCON links
- **Supported on:**
  - z/OS, z/VM, VSE/ESA, AIX, Sun Solaris, HP-UX, Windows NT and Windows 2000, Numa-Q, OS/400
- **Management tools**
  - IBM TotalStorage ESS Specialist Copy Services Web browser interface and Command Line Interface (CLI)
  - Additionaly for the zSeries environments: ICKDSF utility, and the TSO commands for the z/OS environments

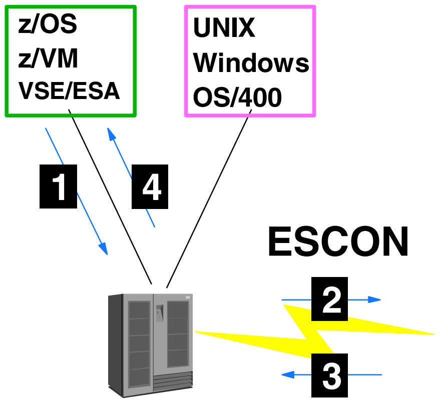
Figure 133. ESS synchronous volume copy - PPRC

6.8.1 PPRC overview

PPRC is a synchronous real time mirroring of data from one logical volume (or LUN) to another logical volume. The logical volumes can be in the same ESS or in another ESS located at another site, some distance away. Mirroring is done at a logical volume level.

PPRC is application independent. Because the copy function occurs at the storage subsystem level, the application does not need to know its existence.

The PPRC protocol guarantees that the secondary copy is up-to-date by ensuring that the primary copy will be written only if the primary storage subsystem receives acknowledgement that the secondary copy has been written.



- **zSeries and open systems** support
- Standard distance increased to **103km** (more by RPQ)
- PPRC workload spread across **1** to **8** paths per volume/LUN to secondary
- **Much faster** initial copy and resynch times
- **FlashCopy** from PPRC primary or secondary
- PPRC primary supports full **PAV, Multiple Allegiance, Priority I/O queueing** for z/OS environments
- Full support by z/OS **Geographically Dispersed Parallel Sysplex** (GDPS)
  - GDPS and RCMF

Figure 134. PPRC improvements with ESS

The sequence when updating records is (see Figure 134):

1. *Write to primary volume* (to ESS cache and NVS). The application system writes data to a primary volume on an ESS at the application site, and cache hit occurs.
2. *Write to secondary* (to ESS cache and NVS). PPRC dispatches a write hit over an ESCON channel to the secondary ESS at the recovery site.
3. *Signal write complete on the secondary.* The recovery site ESS signals *write complete* to the application site ESS when the updated data is in its cache and NVS.
4. *Post I/O complete*. When the application site ESS receives the *write complete* from the recovery site ESS, it returns I/O complete status to the application system.

Destage from cache to the backend disk drive modules on both the application and recovery site ESS is performed asynchronously.

When any problem is met in the process, PPRC suspends automatically the mirroring function and applies the critical attribute behavior described into 6.8.3, “The PPRC critical attribute” on page 211.

6.8.2 PPRC volume states

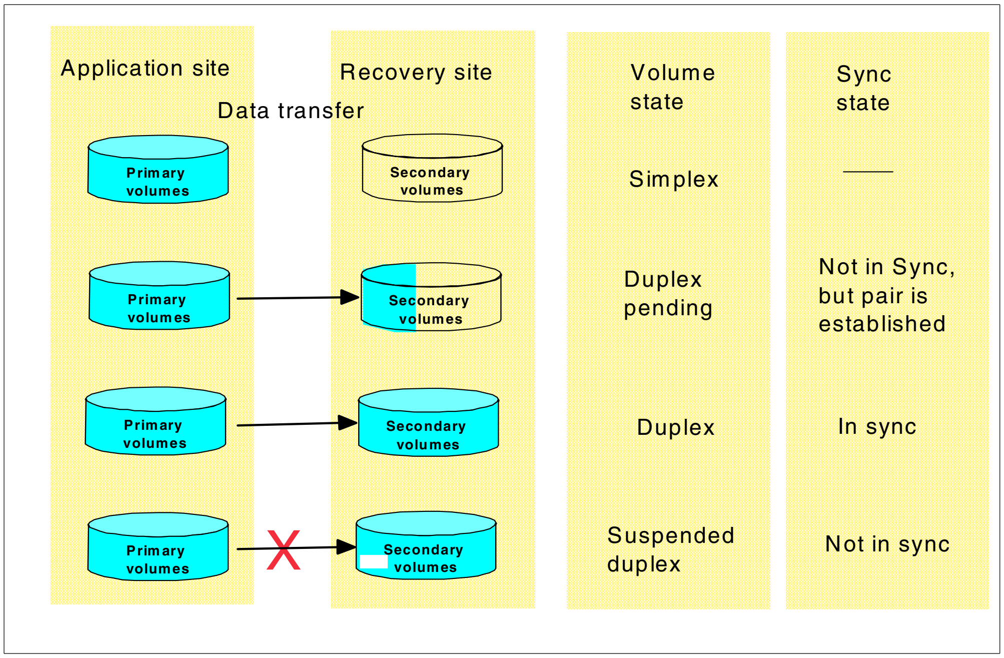


Figure 135. PPRC volume states

Volumes within the Enterprise Storage Server used for PPRC can be in one of the following states (Figure 135):

Simplex
The *simplex* state is the initial state of the volumes before they are used in any PPRC relationship, or after the PPRC relationship has been withdrawn. Both volumes are accessible only when in *simplex* state, else just the primary is accessible.

Duplex pending
Volumes are in *duplex pending* state after the PPRC copy relationship is established, but the source and target volume are still out of synchronization (*sync*). In that case, data still needs to be copied from the source to the target volume of a PPRC pair. This situation occurs either after the PPRC relationship was just established (or reestablished for *suspended* volumes), or when the PPRC volume pair reestablishes after a storage subsystem failure. The PPRC secondary volume is not accessible when the pair is in *duplex pending* state.

Duplex
This is the state of a volume pair that is in *sync*; that is, both source and target volumes containing exactly the same data. Sometimes this state is also referred as the *full copy* mode. The PPRC secondary volume is not accessible when the pair is in *duplex* state.

Suspended
Volumes are in *suspended* state either when the source and target storage subsystems cannot communicate anymore, and therefore the PPRC pair could not be kept in *sync*, or when the PPRC pair was suspended manually. During the

suspended state, the primary volume's storage server keeps a bitmap reflecting all the tracks that were updated in the source volume. This bit map will be used for reestablishment of the PPRC pair later on.

The PPRC secondary volume is not accessible when the pair is in *suspended* state. This does not mean that access is inhibited by hardware. In fact the secondary volumes are offline to anyone and cannot be varied online to anyone. However, if a software program or utility is able to use the device while it is offline, then this program will be able to read the volumes data. This is what happens with TSO FlashCopy when having a PPRC secondary as a source. This is also what happens with FDR (Fast Dump Restore) an OEM software from Innovation DP, that is able to do a full -volume copy of an offline device.

FlashCopy of PPRC secondary

You can FlashCopy the PPRC secondary when it is in *suspended* mode to another volume, and use it the way a FlashCopy target can be used. It is necessary to comply with the FlashCopy source consistency requirements.

6.8.3 The PPRC critical attribute

The *critical attribute* of a pair is setup when the pair is established or re-synchronized. This parameter defines the behavior of PPRC in case of an error. There are two alternatives for PPRC when an error occurs:

- Either suspend the pair and do not accept any further writes on the primary address (CRIT=YES),
- Or suspend the pair and accept further writes to the primary (CRIT=NO), even when the updates cannot be delivered to the secondary.

There are two global strategies for the process triggered by CRIT=YES, which can be parametrized on an ESS:

- Either CRIT=YES - Paths (Light version), which suspends the pair, and does not accept any further writes if the logical control units cannot longer communicate. Otherwise it has a behavior similar to CRIT= NO. The assumption is that the problem is at the device level and not a disaster that has affected the whole storage subsystem.
- Or CRIT=YES - All (Heavy version), which suspends the pair and does not accept any further writes to the primary volume if data cannot be sent to the secondary volume.

The CRIT = YES - All alternative assures that primary and the secondary will always have the identical data, but at the potential cost of a production outage. However, this does not prevent a Data Base Management System like DB2, for instance, to flag into its log files (which are mirrored) the need to do a recovery on a table space which has met a primary physical I/O error when the secondary is still valid. After a rolling disaster this mechanism can make the recovery task very difficult when you have many table spaces with many indexes. This is what is called as the *dying scenario*.

GDPS / PPRC

A much more flexible solution, based on policies to be applied for certain outage conditions, exists with GDPS/PPRC for the z/OS environments.

GDPS/PPRC has the capability of freezing the whole mirroring in a current consistent state, *just before the I/O error*. There are three possible behaviors:

1. Freeze-and-Stop: no data loss at all, at the price of a global outage.
2. Freeze-and-Go: Suspends all the secondary volumes in a consistent way, and continues the production on the primary volumes, at the price of potential data loss when a disaster occurs.
3. Freeze-and-Stop-Conditional, which launches some fast analysis mechanisms before making decision.

These behaviors are the result of automated application of pre-determined policies that are setup by GDPS and the production analysts.

6.9 Invocation of PPRC

PPRC can be managed in different environments.

6.9.1 zSeries environments

In the z/OS environments PPRC can be managed using TSO commands or the ICKDSF utility. Also in this environment the ANTRQST macro supports PPRC requests.

For the z/VM and VSE/ESA environments the support to manage PPRC is by means of the ICKDSF utility. ICKDSF is also necessary for the stand-alone situations.

For all the zSeries environments (z/OS, z/VM and VSE/ESA) the ESS Specialist Copy Services Web interface can be used to control the PPRC operation.

TSO commands

Specific TSO/E commands can be used in the z/OS environments to control PPRC. These commands are extremely powerful, so it is important that they are used correctly and directed to the correct devices. We recommend that you place the TSO commands in a RACF-protected library to restrict PPRC TSO commands to authorized storage administrators only.

The TSO commands specific for PPRC are:

- CESTPATH — Used to establish ESCON paths between an application site (source) logical storage subsystem (LSS) and a recovery site (target) LSS
- CESTPAIR — Used to specify the primary and the secondary volumes of a pair and initiate or re-establish the copy process
- CSUSPEND — Used to suspend PPRC operations between a volume pair. PPRC stops transferring data to the secondary volume
- CQUERY — Used to query the status of one volume of a PPRC pair, or the status of all paths associated with an LSS.
- CGROUP — Used to control operations for all PPRC volume pairs on a single primary and secondary LSS pairing
- CRECOVER — Used to allow the recovery system to gain control of a logical volume on its ESS. This command is issued from the recovery system
- CDELPAIR — Used to specify the primary and secondary volumes to remove from PPRC pairing
- CDELPATH — Used to delete all established ESCON paths between the application site ESS and the recovery site ESS

ANTRQST macro

For the z/OS environments, the DFSMSdfp Advanced Services ANTRQST macro which calls the System Data Mover API can be used. The new ANTRQST macro supports the PPRC requests CESTPATH RESETHP(YES|NO) parameter, and the CESTPAIR ONLINSEC (YES|NO) parameter.

ICKDSF

ICKDSF Release 16 plus APAR PQ26800 supports PPRC operations in the z/OS, z/VM, VSE/ESA and stand-alone environments.

The PPRC functions are supported through the PPRCOPY command. This command uses the DELPAIR, ESTPAIR, DELPATH, ESTPATH, SUSPEND, RECOVER, and QUERY parameters.

ICKDSF does not support consistency grouping. Therefore you cannot request any long busy states on failing devices or freeze control unit pairings, and thus it cannot have the CGROUP command or parameter.

ICKDSF provides a subset of the z/OS PPRC commands for stand-alone use when no operating system is available. This subset could be used to issue PPRCOPY RECOVER commands at a secondary site.

However, because ICKDSF contains a subset of the PPRC commands and functions, then for the z/VM and the VSE/ESA environments you must establish measures to control recovery in the event of a disaster, for example, procedures to respond to failure scenarios that might otherwise be addressed by the TSO commands.

ESS Copy Services Web interface

ESS Specialist Copy Services Web interface. The ESS Copy Services that runs in the ESS provides a Web browser interface that can be used to control the PPRC functions in a zSeries environment. See 6.3, “The Web interface” on page 194 for further information.

You may refer to *Implementing ESS Copy Services on S/390,* SG24-5680 for detailed information on invoking PPRC copy services in the zSeries environments

6.9.2 Open systems environments

There are two different methods of using the ESS Copy Services in the open systems environments:

- ESS Specialist Copy Services Web interface. The ESS Copy Services that runs in the ESS provides a Web browser interface that can be used to control the PPRC functions in the open systems environments. For this situation you will be able to establish PPRC pairs in three different ways:
  - From the *Volumes* panel (based on volumes)
  - From the *Storage Servers* panel (based on entire logical subsystems)
  - From the *Task* panel (once a task for PPRC is created)

Remember that previously to establishing the pairs, you must define the paths and make them available. This is done using the *Paths* panel provides by the Web browser interface. See 6.3, “The Web interface” on page 194 for more information.

- A Java-based Command Line Interface (CLI). The CLI interface allows administrators to execute Java-based Copy Services commands from a command line. This command line interface is currently available for the following operating systems: AIX, Sun Solaris, HP-UX and Windows NT and 2000. For this situation you will be starting a pre-define PPRC tasks from the command line or from one of your own customized scripts. See 6.4, “Command Line Interface (CLI)” on page 200.

You may refer to *Implementing ESS Copy Services on UNIX and Windows NT/2000,* SG24-5757 for detailed information on invoking PPRC copy services in the opens systems environments.

6.10 PPRC Implementation on the ESS

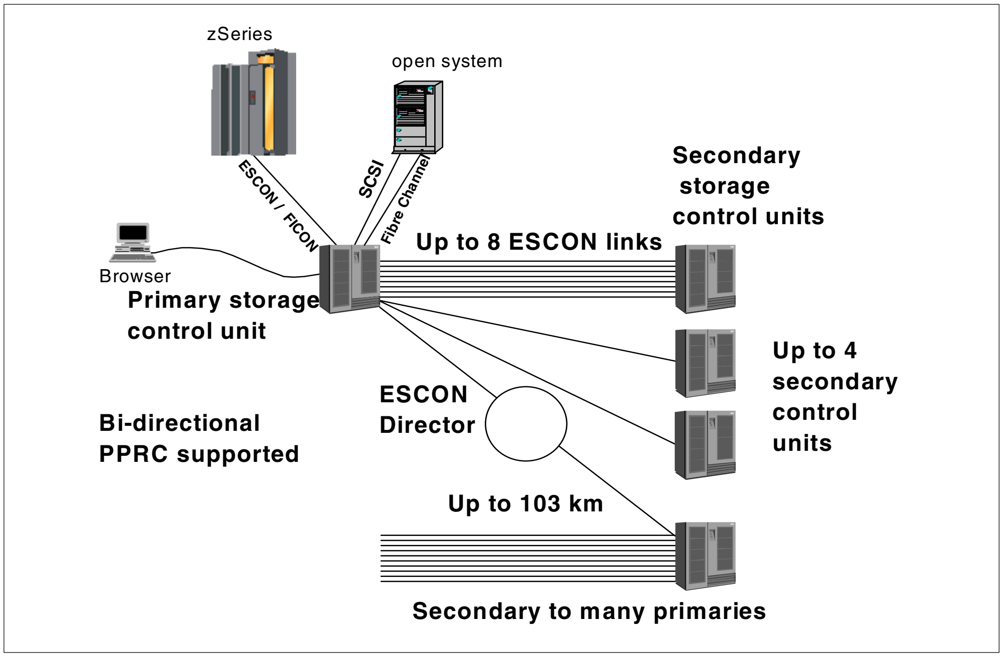

zSeries
open system
ESCON / FICON
SCSI
Fibre Channel
Secondary storage control units
Up to 8 ESCON links
Browser
Primary storage control unit
Up to 4 secondary control units
ESCON Director
Bi-directional PPRC supported
Up to 103 km
Secondary to many primaries

Figure 136. PPRC configuration options with ESS

As with other PPRC implementations, you can establish PPRC pairs only between storage control units of the same type, which means that you can connect an ESS with another ESS only.

6.10.1 ESCON links

ESCON links between ESS subsystems are required. Up to eight ESCON links are supported between two ESS storage subsystems. The local ESS is usually called *primary* if it contains at least one PPRC source volume, while the remote ESS is called *secondary* if it contains at least one PPRC target volume. An ESS can act as primary and secondary at the same time if it has PPRC source and target volumes. This mode of operation is called bi-directional PPRC.

A primary ESS can be connected to up to four secondary ESSs (see Figure 136). A secondary ESS can be connected to as many primary ESSs as ESCON links are available.

PPRC links are unidirectional. This means, a physical ESCON link can be used to transmit data from the primary ESS to the secondary ESS. If you want to set up a bi-directional PPRC configuration with source and target volumes on each ESS, you need ESCON PPRC links in each direction (see Figure 137). The number of links depends on the write activity to the primary volumes.

Primary PPRC ESCON ports are dedicated for PPRC use. A PPRC port operates in control unit mode when it is talking to a host. In this mode, an ESCON port can also receive data from the primary control unit when the ESS port is connected to an ESCON director. So, the ESCON port on the secondary ESS does not need to be dedicated for PPRC use.

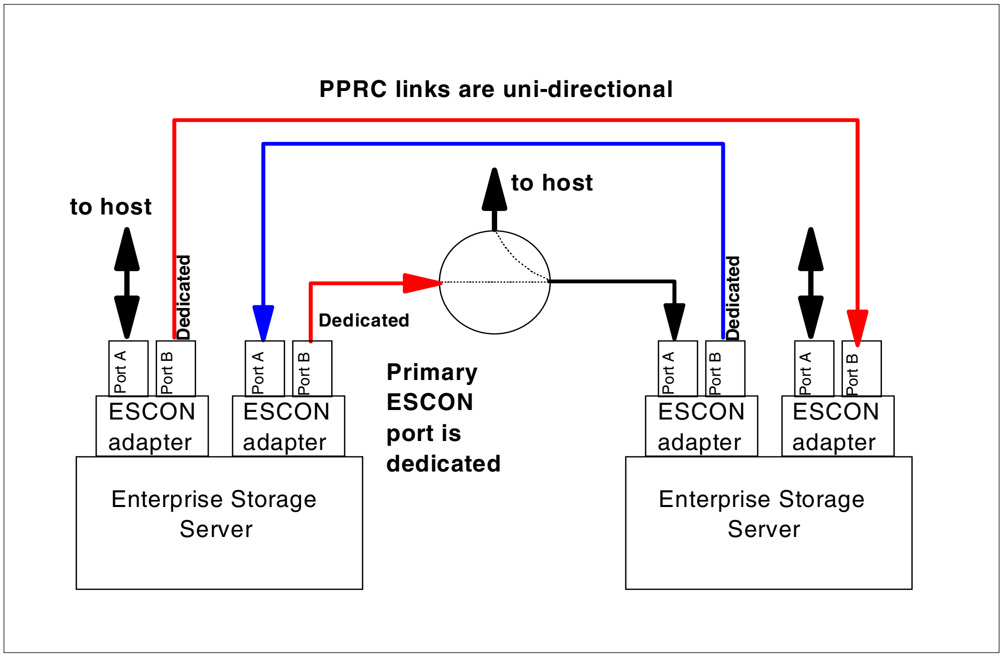


Figure 137. PPRC links

An ESCON port is operating in channel mode when it is used on the primary ESS for PPRC I/O to the secondary ESS. The switching between control unit mode and channel mode is dynamic.

If there is any logical path defined for an ESCON port to a zSeries host, you cannot switch this port to channel mode for PPRC to use as primary port. You must first configure all logical paths from the zSeries host to that port offline. Now you can define a PPRC logical path over this ESCON port from the primary to the secondary ESS. When you establish the logical path, the port will automatically switch to channel mode.

6.10.2 PPRC logical paths

Before PPRC pairs can be established, logical paths must be defined between the logical control unit images. The Enterprise Storage Server supports up to 16 logical CKD control unit images and up to 16 FB controller images. An ESCON adapter supports up to 64 logical paths. You establish logical paths between control unit images of the same type over physical ESCON links (see Figure 138).

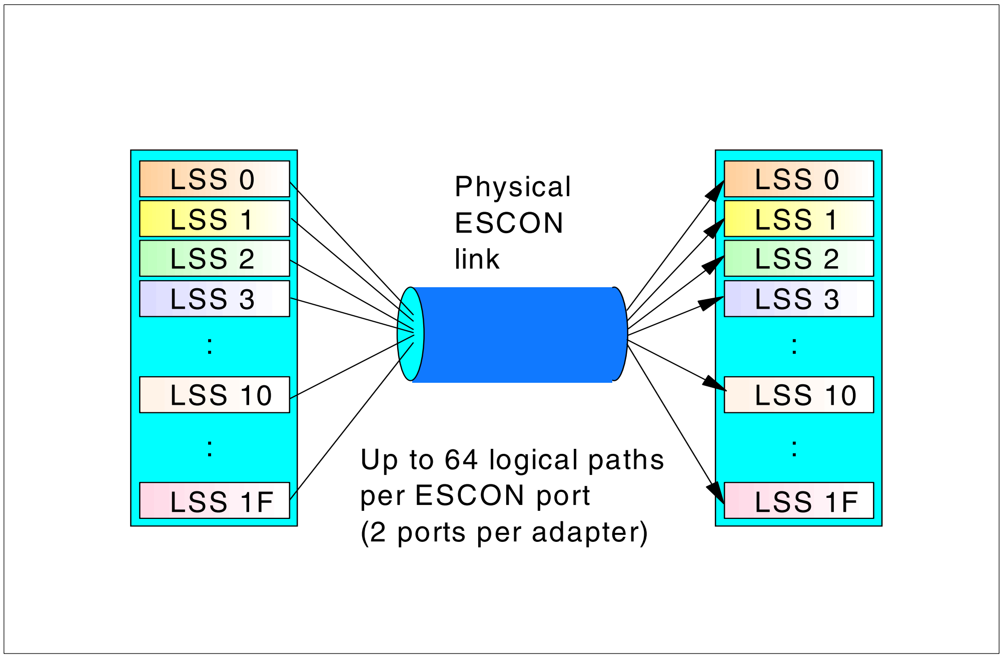
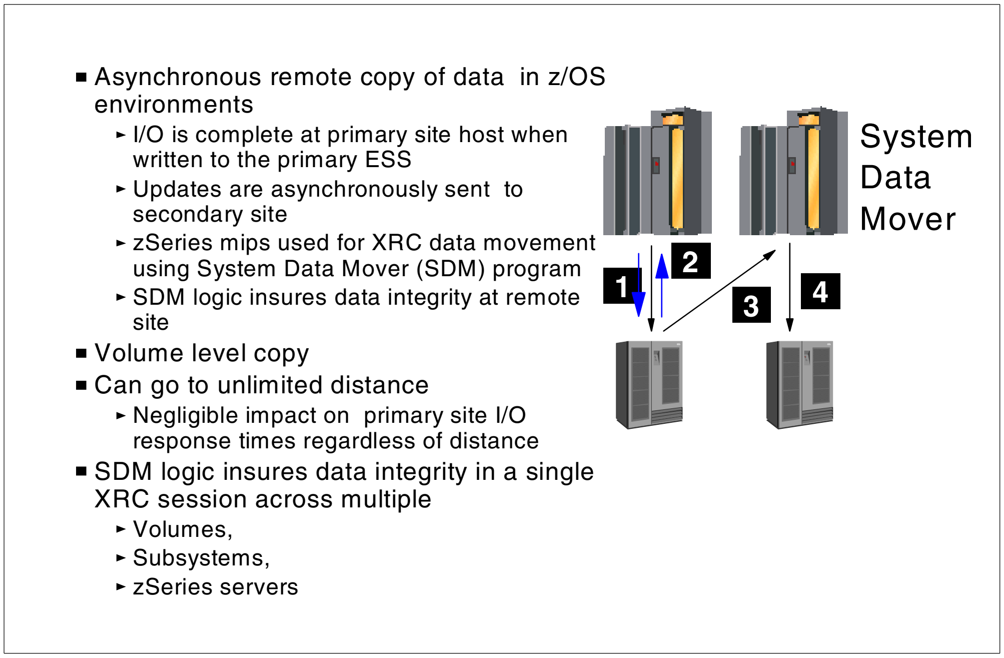


Figure 138. PPRC logical paths

6.11 Extended Remote Copy (XRC)



- Asynchronous remote copy of data in z/OS environments
  - I/O is complete at primary site host when written to the primary ESS
  - Updates are asynchronously sent to secondary site
  - zSeries mips used for XRC data movement using System Data Mover (SDM) program
  - SDM logic insures data integrity at remote site
- Volume level copy
- Can go to unlimited distance
  - Negligible impact on primary site I/O response times regardless of distance
- SDM logic insures data integrity in a single XRC session across multiple
  - Volumes,
  - Subsystems,
  - zSeries servers

Figure 139. Extended Remote Copy (XRC)

6.11.1 Overview

Extended Remote Copy (XRC) is an asynchronous remote mirroring solution. It uses the System Data Mover (SDM) of the DFSMS/MVS program, and hence, works only in the z/OS environments.

Application systems accessing the same source volumes need to have the same internal clock time, which is provided by a Sysplex-timer, within the sysplex. Each write I/O on an XRC mirrored volume gets a time stamp.

Applications doing write I/Os to primary (source) volumes — (1) in Figure 139 — get a Device End status (write I/O complete) as soon as the data has been secured in cache and NVS of the primary ESS (2). The System Data Mover, that may be running at the recovery site host, reads out the updates to the XRC source volumes from the cache (3) and sends it to the secondary volume on a remote storage control (4).

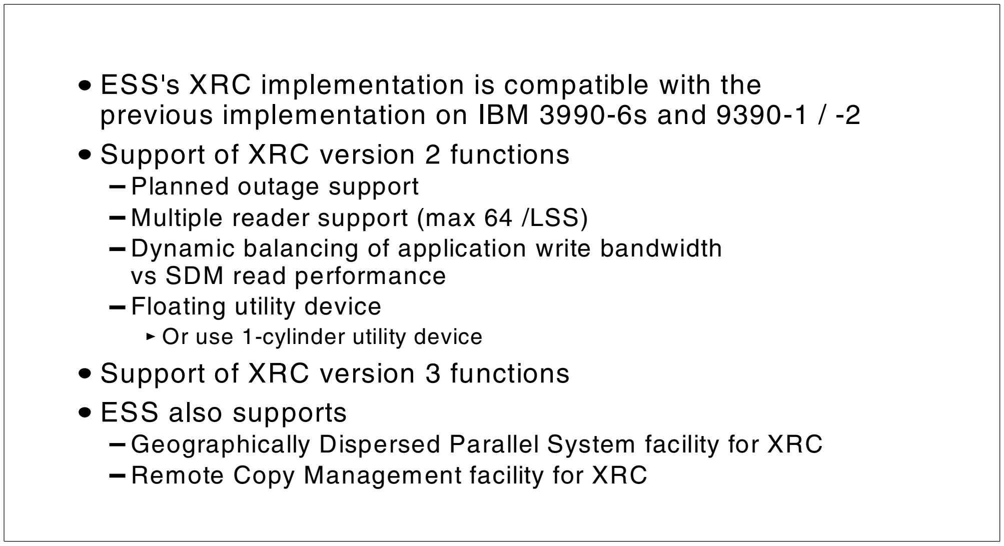
The System Data Mover needs to have access to all primary storage control units with XRC volumes the Data Mover has to handle, as well as to the target storage control units. In this way, the Data Mover has the higher authority to all control units involved in the remote mirroring process and can assure data integrity across several primary storage control units. The I/O replication in the right sequence to the target volumes is guaranteed by the System Data Mover.

XRC was already available on the IBM 3990 Model 6 and the 9390 Models 1 and 2 storage control units.

XRC is designed as a solution that offers the highest levels of data integrity and data availability in a disaster recovery, workload movement, and/or device migration environment. It provides real time data shadowing over extended distances. With its single command recovery mechanism furnishing both a fast and effective recovery, XRC provides a complete disaster recovery solution.

6.11.2 XRC Implementation on the ESS

The implementation of XRC on the ESS is compatible with XRC's implementation on previous IBM 3990 Model 6 and 9390 Models 1 and 2.



- ESS's XRC implementation is compatible with the previous implementation on IBM 3990-6s and 9390-1 / -2
- Support of XRC version 2 functions
  - Planned outage support
  - Multiple reader support (max 64 /LSS)
  - Dynamic balancing of application write bandwidth vs SDM read performance
  - Floating utility device
    - Or use 1-cylinder utility device
- Support of XRC version 3 functions
- ESS also supports
  - Geographically Dispersed Parallel System facility for XRC
  - Remote Copy Management facility for XRC

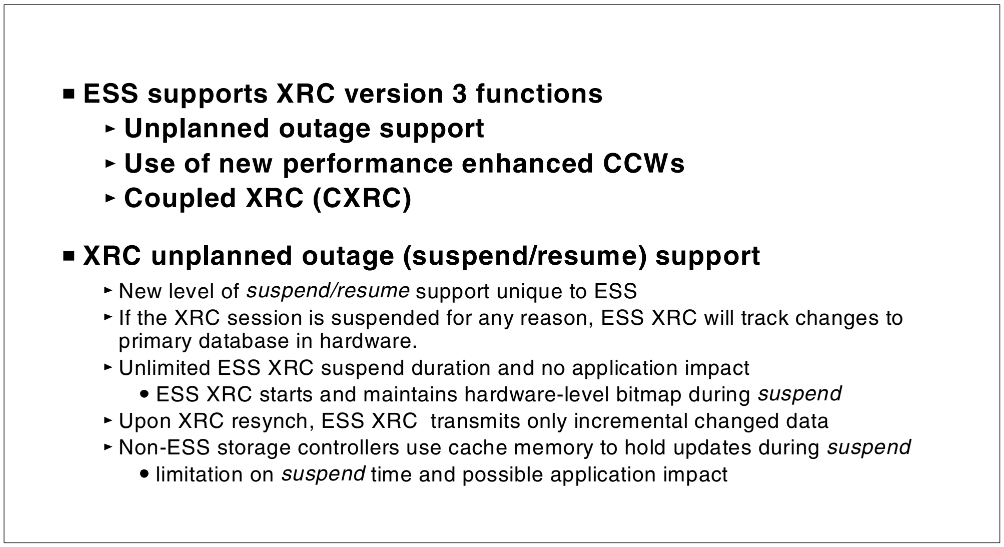
Figure 140. XRC implementation on the ESS

6.11.3 Support of XRC version 2 functions

ESS supports all the XRC version 2 enhancements:

- Planned outage support is the capability to suspend for a while the SDM function (to do maintenance host, for example) and to re-synchronize later the suspended volumes with the updates without a full copy of the mirrored volumes.
- The System Data Mover reader supports up to 64 readers per logical subsystem
- Dynamic balancing of application write bandwidth with SDM read performance
- Floating utility device: this facility is per default on ESS
- Use of 1-cylinder utility device (ESS's custom volume capability)

6.11.4 Support of XRC version 3 functions



- **ESS supports XRC version 3 functions**
  - **Unplanned outage support**
  - **Use of new performance enhanced CCWs**
  - **Coupled XRC (CXRC)**
- **XRC unplanned outage (suspend/resume) support**
  - New level of *suspend/resume* support unique to ESS
  - If the XRC session is suspended for any reason, ESS XRC will track changes to primary database in hardware.
  - Unlimited ESS XRC suspend duration and no application impact
    - ESS XRC starts and maintains hardware-level bitmap during *suspend*
  - Upon XRC resynch, ESS XRC transmits only incremental changed data
  - Non-ESS storage controllers use cache memory to hold updates during *suspend*
    - limitation on *suspend* time and possible application impact

Figure 141. ESS XRC unplanned outage support

The Enterprise Storage Server provides support for XRC version 3 functions.

6.11.4.1 Unplanned outage support

On an IBM 3990 Model 6, XRC pairs could be suspended only for a short time or when the System Data Mover was still active. This was because the bitmap of changed cylinders was maintained by the System Data Mover in the software. This software implementation allowed a re synchronization of pairs only during a planned outage of the System Data Mover or the secondary subsystem.

The Enterprise Storage Server starts and maintains a bitmap of changed tracks in the hardware, in non-volatile storage (NVS), when the connection to the System Data Mover is lost or an XRC pair is suspended by command.

The bitmap is used in the re-synchronization process when you issue the `XADD SUSPENDED` command to re-synchronize all suspended XRC pairs. Copying only the changed tracks is much faster compared to a full copy of all data. With ESS's XRC support, a re-synchronization is now possible for a planned as well as an unplanned outage of one of the components needed for XRC to operate.

6.11.4.2 New performance enhanced CCWs

The Enterprise Storage Server supports new performance enhanced channel command words (CCWs) that allow reading or writing more data with fewer CCWs and thus reducing the overhead of previous CCW chains.

The System Data Mover will take advantage of these performance enhanced CCWs for XRC operations on an ESS.

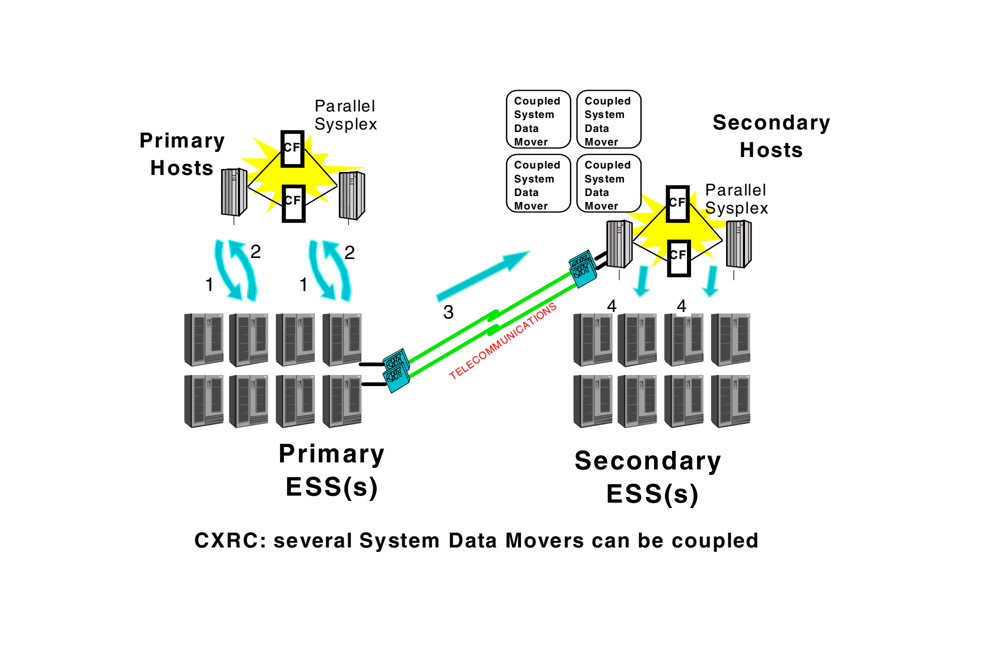

Parallel Sysplex
Primary Hosts
CF
CF
Coupled System Data Mover
Coupled System Data Mover
Coupled System Data Mover
Coupled System Data Mover
Secondary Hosts
Parallel Sysplex
CF
CF
1
2
1
2
3
4
4
TELECOMMUNICATIONS
Primary ESS(s)
Secondary ESS(s)
CXRC: several System Data Movers can be coupled

Figure 142. CXRC performance and scalability

6.11.4.3 Coupled Extended Remote Copy (CXRC)

Coupled Extended Remote Copy (CXRC) expands the capability of Extended Remote Copy (XRC) so that very large customers who have configurations consisting of thousands of primary volumes can be assured that all their volumes can be recovered to a consistent point in time (see Figure 142).

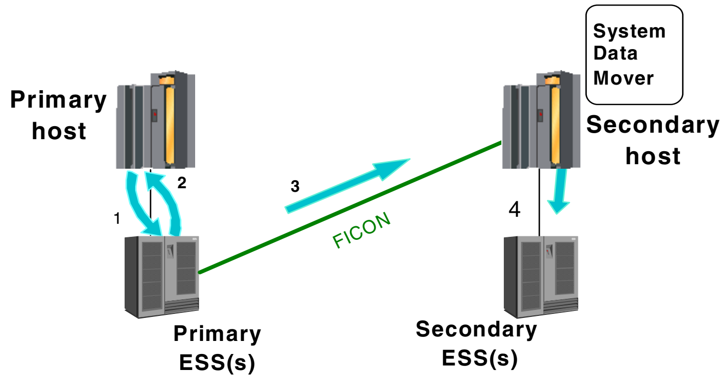
Where XRC was an effective solution mirroring hundreds of volumes, CXRC becomes an equally effective solution for mirroring thousands of volumes. All of which can be recovered to a consistent point in time.

In a disaster recovery situation, for the recovered data of a particular application to be usable, the data must be recovered to a consistent point in time. Since many customer applications are interrelated, the most efficient situation is to recover all data to the same point in time. CXRC provides this capability by allowing multiple XRC sessions to be *coupled* with a coordination of consistency between them. This way all the volumes in all the coupled sessions can be recovered to a consistent point in time.

Migration for Customers Currently Using XRC

Existing XRC customers will be able to continue to use XRC in the same manner. The new CXRC support will not directly affect existing XRC sessions. The new support is provided by new commands and keywords and enhancements to existing commands. When the new support is not in use, all existing commands and keywords continue to work as currently defined. Once the support has been installed, customers can choose to start one or more additional XRC sessions and couple them into a master session.

6.11.5 ESS XRC FICON support



- **ESS XRC FICON support**
  - Native ESS FICON
  - ESS XRC runs on FICON channel
  - FICON speed, distance, throughput benefits
  - ESCON and FICON XRC connections coexist on the same ESS
  - XRC ESCON and FICON are both DWDM-capable

Figure 143. ESS/XRC FICON support

The SDM takes a great benefit of the higher bandwidth FICON channels bring to read the updates from the primary ESS. The transfer bandwidth for a single thread read is for XRC at least five times better than with ESCON due to the large blocking factor of transfers.

The improved bandwidth comes along with the longer distance support (up to 100 Km.) than the previous ESCON channels, when using direct channel attachment. These capabilities position XRC as a disaster / recovery solution in the range of metropolitan areas. With ESCON the recommendation was channel extenders through telecom lines beyond 25 Km. Now, with FICON we the same recommendation applies beyond 100 Km.

The CNT channel extenders support ESCON channels, in a compatible way with the currently available bandwidths on the telecom networks.

6.11.6 Invocation and management of XRC

XRC can be controlled by using TSO/E commands, or through the DFSMSdfp Advanced Services ANTRQST Macro, which calls the System Data Mover API.

Managing a large set of mirrored volumes over long distance requires automatization of monitoring and decision making. The GDPS/XRC automation package, developed from customer requirements, offers a standard solution to that demand.

For more details on the implementation of XRC, please refer to *Implementing ESS Copy Services on S/390,* SG24-5680.

6.12 iSeries 400 use of ESS copy functions

The iSeries 400 servers can take advantage of the Enterprise Storage Server advanced copy functions FlashCopy and PPRC. Because of the particular way in which the storage is managed by the iSeries servers, some considerations apply.

6.12.1 The Load Source Unit (LSU)

The Load Source Unit (LSU) is a special DASD in the iSeries. This is the device that is used to IPL the system (among other things). It is similar to a boot drive. All other user data can be located on external DASD units, but the LSU must be an internal drive. This is because the system cannot be IPLed from an I/O adapter (IOA) supporting external drives.

Due to the nature of iSeries Single Level Storage, it is necessary to consider the LSU as a special case. On other open system platforms, such as UNIX and NT, each volume can be identified with its contents. The iSeries is different, as all storage is considered as a single large address space. The LSU is within this address space.

Therefore, if you want to use facilities such as Peer-to-Peer Remote Copy (PPRC) or FlashCopy to do a hardware copy of the volumes attached to the iSeries, you then must mirror the LSU from the internal drive into the ESS, to ensure the whole single level storage is copied.

6.12.2 Mirrored internal DASD support

Support has been added to the iSeries 400 to allow internal DASD, like the LSU, to be mirrored to external DASD. This requires that the external DASD reports as unprotected, even though in practice, it may actually be protected in a RAID-5 rank within the ESS.

Note: Currently the PPRC mirroring support is only available for configurations where the iSeries and the ESS connect with SCSI adapters via an RPQ. For configurations with fibre channel attachment, it is not currently supported.

Before implementing remote load source mirroring, you should check for the latest maintenance required to support this function. The maintenance APARs, PTFs and PSP buckets can be found at `http://www.as400service.ibm.com`

6.12.3 LSU mirroring

The primary reason for using remote load source mirroring is to get a copy of the LSU into the ESS, so that the entire DASD space in single level storage can be duplicated by the hardware facilities such as PPRC and FlashCopy.

When using remote load source mirroring, normal OS/400 rules for mirroring apply. Both the source and target disk drives must be the same size, although they can be of different drive types and speeds It is simply capacity which must match.

To allow the LSU to be mirrored, the target disk must be unprotected, as OS/400 does not allow to mirror any disk to another which is already protected. This must be done when you first define the LUN in the ESS. Once the LUN has been specified, you can not change the designated protection. Normally you will define only the LUN for the LSU as being unprotected. All other LUNs will be defined as protected, reflecting their true status to OS/400. To do this, you select *Unprotected* in the *Add Volumes* panel of the ESS Specialist when doing the logical configuration.

For more details on the setup of mirrored configurations you can see *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220.

6.12.4 FlashCopy

There is no native iSeries command interface to initiate FlashCopy service on the iSeries. This copy function is managed using the IBM TotalStorage ESS Specialist.

Note: You may consider that iSeries backup and recovery offers a very similar function to FlashCopy. This function is called *Save-while-active* and also provides a T0 point-in-time copy of the data.

6.12.5 PPRC

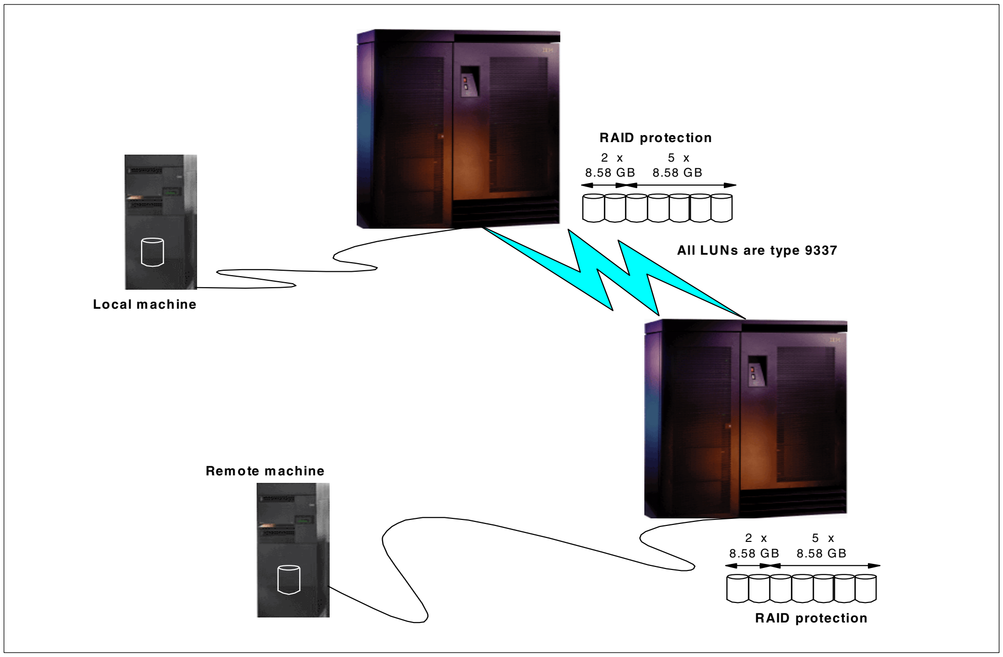


Figure 144. iSeries and ESS PPRC implementation

Peer-to-Peer Remote Copy (PPRC) is currently available for SCSI attached ESS via RPQ. Figure 144 shows an example of iSeries and ESS in a PPRC implementation. Because the attachment between the iSeries and the ESS is a SCSI attachment, then all the DASDs are 9337s.

In this implementation, the local iSeries has only one internal volume, and this is the load source unit (LSU). Within the local ESS there is the remote load source mirrored pair and the rest of the local iSeries LUNs.

On the remote site there is another ESS. This remote ESS contains the PPRC copy of the remote load source unit and also the PPRC copies of all the other LUNs from the local iSeries.

In the event of a disaster at the production site (the local site in Figure 144), then the backup iSeries at the remote site recovers to ESS LUNs at that same remote site. This recover process includes an *Abnormal* IPL on the iSeries. This IPL could be many hours. This time can be reduced by implementing OS/400 availability functions to protect applications and database. For example, journaling, commitment control, system managed access paths (SMAP).

Note: in a configuration with remote load source mirroring, the system can only IPL from, or perform a main storage dump to, an internal LSU.

For more information on the implementation of the ESS advanced copy functions in an iSeries 400 environment, see *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220.

6.13 Geographically Dispersed Parallel Sysplex (GDPS)

How would a shutdown of your z/OS system affect your business? Do you put off system maintenance and upgrades to avoid system downtime? What about a site disaster? Is your business-critical processing and data protected from a site disaster? In today's highly competitive e-business world, outages will have a devastating impact on a business - they could mean its demise. Many companies have inadequate business continuance plans developed on the premise that back office and manual processes will keep the business running till computer systems are available. Characteristics of these recovery models allow critical applications to recover within 24-48 hours, with data loss potentially exceeding 24 hours, and full business recovery taking days or weeks. As companies transform their business to compete in the e-marketplace, business continuity strategies and availability requirements must be reevaluated to ensure that they are based on today's business requirements.

In e-business, two of the most stringent demands for survival are continuous availability and near transparent disaster recovery (D/R). Systems that deliver continuous availability combine the characteristics of high availability and continuous operations to always deliver high levels of service (24x7x365). High availability is the attribute of a system to provide service at agreed upon levels and mask unplanned outages from end users. Continuous operations, on the other hand is the attribute of a system to continuously operate and mask planned outages from end users. To attain the highest levels of continuous availability and near-transparent D/R, the solution must be based on geographical clusters and data mirroring. These technologies are the backbone of the Geographically Dispersed Parallel Sysplex (GDPS) solution.

GDPS complements a multi-site Parallel Sysplex by providing a single, automated solution to dynamically manage storage subsystem mirroring, processors, and network resources to allow a business to attain *continuous availability* and *near transparent business continuity* (disaster recovery) without data loss. GDPS is designed to minimize and potentially eliminate the impact of any failure including disasters, or a planned site outage. It provides the ability to perform a controlled site switch for both planned and unplanned site outages, with no data loss, maintaining full data integrity across multiple volumes and storage subsystems and the ability to perform a normal Data Base Management System (DBMS) restart - not DBMS recovery - at the opposite site. GDPS is application independent and, therefore, covers the customer's complete application environment.

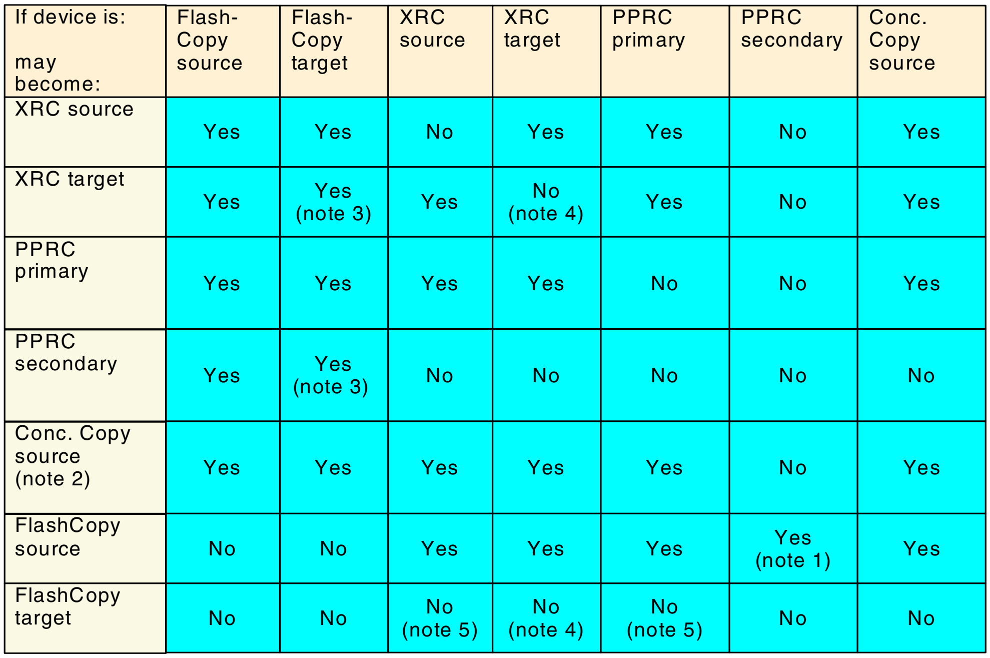
GDPS is enabled by means of key IBM technologies:

- Parallel Sysplex
- Systems Automation for OS/390
- Enterprise Storage Server
- PPRC (Peer-to-Peer Remote Copy)
- XRC (Extended Remote Copy)
- Optical Dense Wavelength Division Multiplexer (DWDM)

6.14 Combination of copy services

Some of the Enterprise Storage Server copy services can be combined with others. You can, for example make a FlashCopy copy of a PPRC or XRC primary or secondary volume. This provides for an easy and fast way, for example, to create test data on the remote site, or copies of the production data on the secondary site for data mining.

If the remote site is used purely as a disaster backup with no primary volumes, the storage subsystem resources are not fully utilized. The cache, for example, is hardly used. Therefore, it might be a good idea to run data mining applications on the secondary site to utilize the cache for read operations.



If device is: may become:	Flash-Copy source	Flash-Copy target	XRC source	XRC target	PPRC primary	PPRC secondary	Conc. Copy source
XRC source	Yes	Yes	No	Yes	Yes	No	Yes
XRC target	Yes	Yes (note 3)	Yes	No (note 4)	Yes	No	Yes
PPRC primary	Yes	Yes	Yes	Yes	No	No	Yes
PPRC secondary	Yes	Yes (note 3)	No	No	No	No	No
Conc. Copy source (note 2)	Yes	Yes	Yes	Yes	Yes	No	Yes
FlashCopy source	No	No	Yes	Yes	Yes	Yes (note 1)	Yes
FlashCopy target	No	No	No (note 5)	No (note 4)	No (note 5)	No	No

Figure 145. Advanced Copy Services combinations

Figure 145 shows a list of valid combinations of advanced copy functions. The top row indicates the copy function in which the device is. The left most column indicates the copy function in which the device can/cannot participate.

Notes

1. The primary control unit could see long busy time while the PPRC secondary is destaging all cache modified data in order to become a FlashCopy pair.
2. These columns and rows should be interpreted on an extent basis in the zSeries environments.
3. The operation is allowed. Updates to the affected extents will result in implicit removal of the FlashCopy relationship.
4. Host software disallows this combination.
5. Since PPRC and XRC intercept host updates, the target of a FlashCopy cannot be the primary volume of a PPRC or XRC pair. Using *suspend* will not help. If the target of a FlashCopy is intended to be the primary volume, delete the pairing (remove the PPRC or XRC relationship), perform the FlashCopy, then re-establish the pairing. This will force all tracks to be re-read and sent to the secondary volume.

To better understand Figure 145 consider that it reflects a notion of sequence. In other words, you must look at the current state of the volumes and then decide what is valid. An example would be a volume that is already a FlashCopy target and it is made a PPRC primary. This is allowed and works. However, if the volume is already a PPRC primary and you perform a FlashCopy to the volume, this is not allowed and does not transfer the data to the remote location. For his particular situation, the reason is the fact that FlashCopy operates at the stage/destage level and PPRC operates at write intercept time at the top end of cache.

6.15 Priced Copy Services features

The following optional ESS Copy Services require the ordering of priced features to enable the function on an Enterprise Storage Server.

- FlashCopy
- Peer-to-Peer Remote Copy (PPRC)
- Extended Remote Copy (XRC)

The PPRC enabling feature is required for both primary and secondary ESS control units.

The XRC enabling feature is required on each ESS with XRC primary volumes. If the ESS acts as a secondary control unit only, you do not need this feature. In this case, you can use XRC to migrate your data from an IBM 3990 Model 6 control unit to an ESS. If you use XRC for a disaster recovery protection, however, keep in mind that without the XRC feature on the secondary control unit, you cannot use XRC to copy your data back from the secondary to the primary, if you intend to do so.

Chapter 7. Software support

The Enterprise Storage Server (ESS) introduces advanced new functions for both open systems and zSeries. In this chapter we list the software levels required to support the ESS and exploit its new functions.

As the information presented in this chapter changes over time, we strongly recommend that you get the most updated information on the latest required software level, by referring the following sources of information:

For open systems environments visit: `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

For zSeries environments we recommend you to review the latest PSP (Preventive Service Planning) bucket, that you may ask from your IBM Software Representative. This documentation will provide you the software levels required and also the latest PTFs required.

7.1 zSeries environment support

zSeries environments must be at the correct support level to exploit all the ESS functions. Note that, for these environments, it is possible that some PTFs (Program Temporary Fix) may be needed in order to connect an ESS at all, so you must ensure that you contact your local IBM Support Center prior to installation.

The Enterprise Storage Server supports the following mainframe operating systems:

- OS/390
- Multiple Virtual Storage/Enterprise Storage Architecture (MVS/ESA)
- Virtual Machine/Enterprise Storage Architecture (VM/ESA)
- Virtual Storage Extended/Enterprise Storage Architecture (VSE/ESA)
- Transaction Processing Facility (TPF)

The ESS also supports the following operating systems for the new IBM e server zSeries 900 (z900) server:

- z/OS Version 1 Release 1
- OS/390 Version 2 Release 6 or higher
- VM
  - z/VM version 3 release 1
  - VM/ESA version 2 release 2
  - VM/ESA version 2 release 3
  - VM/ESA version 2 release 4
- VSE/ESA version 2 release 3 and higher
- Transaction Processing Facility (TPF) version 4 release 1

See the preventive service planning (PSP bucket) for operating system support, and planning information that includes APARs and programming temporary fixes (PTFs).

In the following sections you will find specific more detailed information on the particular software requirements to fully support the ESS and its advanced functions and features.

7.2 z/OS support

All releases of z/OS and OS/390 are capable of supporting the IBM Enterprise Storage Server. In order to provide standard PAV support, OS/390 V1R3 is the minimum software level. In order to provide dynamic PAV support, OS/390 V2R7 with DFSMS/MVS 1.5 is the minimum software level; and the workload manager (WLM) must run in goal mode. For additional information on the Workload Manager functions and use visit the Workload Manager Web site on the Internet at `http://www.ibm.com/s390/wlm`.

MVS/ESA 5.1.0 is the minimum level capable of supporting the ESS

There are three modes of software support of the IBM Enterprise Storage Server: transparency, toleration and exploitation (see to 5.2.12, "PAV support" on page 160).

7.2.1 Transparency support

Transparency provides the base functions of the IBM 3990 Model 6 storage control. z/OS sees up to 16 logical 3990-6s with up to 256 unit addresses per logical subsystem.

DFSMS/MVS 1.2, 1.3 and1.4 provides this capability with the appropriate PTFs applied. ICKDSF (refresh) release 16 and DFSORT for MVS/ESA release 13 are also prerequisites for this level.

An I/O Definition File (IODF) cannot be shared with an exploiting system.

7.2.2 Toleration support

Toleration permits a level of OS/390 which does not support PAV to share the IBM Enterprise Storage Server with releases of z/OS and OS/390 which do. The OS/390 host that does not support PAV recognizes the ESS as a 2105 control unit and recognizes the base and alias addresses needed to support PAV. With toleration support, only non-PAV UCBs are built. However, an IODF can be shared with exploitation-capable systems.

Only DFSMS/MVS 1.2 provides this capability with the appropriate PTFs applied.

7.2.3 Exploitation support

Exploitation support is when a release of OS/390 supports PAV.

z/OS version 1 release 1 and OS/390 1.3 — 2.7 with DFSMS/MVS 1.3 — 1.5 recognize the IBM Enterprise Storage Server as a 2105 device type and can exploit the new PAV capability with the appropriate PTFs applied.

ICKDSF (refresh) release 16, EREP version 3 release 5, and DFSORT for MVS/ESA release 13 are prerequisites too.

7.2.4 Other related support products

Several components of OS/390 software have been changed to support the ESS.

7.2.4.1 ICKDSF

The formatting of CKD volumes is performed when you set up the ESS and define CKD volumes through the IBM TotalStorage ESS Specialist.

To use the volumes in OS/390 only a *minimal init* by ICKDSF is required to write a label and VTOC index.

The following ICKDSF functions are not supported and not required on an ESS:

- ANALYZE with DRIVETEST
- INSTALL
- REVAL
- RESETICD

7.2.4.2 Access Method Services

The AMS LISTDATA command provides new *Rank Counter* reports. This is how you can get some information on the activities of a RAID rank. While z/OS performance monitoring software only provides a view of the logical volumes, this rank information shows the activity of the physical drives.

7.2.4.3 EREP

EREP provides problem incidents reports and uses the new device type *2105* for the ESS.

7.2.4.4 Media Manager and AOM

Both components take advantage of the new performance enhanced CCW on ESS, and they limit extent access to a minimum to increase I/O parallelism.

7.2.4.5 DFSMSdss and DASD ERP

Both of these components also make use of the performance enhanced CCWs for their operations.

7.3 z/VM support

z/VM 3.1 supports both transparency and exploitation modes for the IBM Enterprise Storage Server.

VM/ESA 2.2 and 2.3 contain transparency support for the ESS.

VM/ESA 2.4 is the minimum level which allows an OS/390 guest to use PAVs. VM/ESA 2.4 is also the minimum level which allows a TPF guest to use its exploitation support. A PTF is needed for each release of VM/ESA.

Refer to figure Figure 146 for an overview of supported functions and features for the VM environment.

	2105 support	**New ccw support for guest**	**PAV support for guest**	**FCopy support for guest**	**FCopy support native**	**ICKDSF**
VM/ESA 2.2.0	*requires APAR*	*NO*	*NO*	*NO*	*NO*	*requires APAR*
VM/ESA 2.3.0	*requires APAR*	*YES*	*NO*	*NO*	*NO*	*requires APAR*
VM/ESA 2.4.0	*no APAR*	*YES*	*requires APAR*	*requires APAR*	*NO*	*requires APAR*
z/VM 3.1.0	*no APAR*	*YES*	*YES*	*YES*	*YES*	*requires APAR*

Figure 146. VM support overview

7.3.1 Exploitation of ESS functions

7.3.1.1 PPRC

Synchronous remote copy (PPRC) is supported in VM by the use of ICKDSF or the IBM TotalStorage ESS Specialist.

7.3.1.2 Multiple Allegiance and Priority I/O Queueing

Multiple Allegiance and *Priority I/O Queueing* are ESS hardware functions independent of software support. So VM operating systems can take advantage of this in a shared environment. The priority, however, is not set by VM. So, there is no I/O Priority Queueing control by VM.

7.3.1.3 Guest support

z/VM and VM/ESA do not recognize the ESS by its new device type 2105, but see it as an IBM 3990 Model 6. z/VM and VM/ESA 2.2.0-2.4.0 with PTFs applied supports the ESS as a 3990-6. However, when both operating systems sense the control unit, the returned function bits can be interpreted by guest systems to see what functions are supported on this control unit.

z/VM and VM/ESA 2.3.0 — 2.4.0 with enabling PTFs applied will allow guest systems to use the new performance enhanced CCWs.

7.3.1.4 Parallel Access Volumes

Guest use of *Parallel Access Volumes* is supported in z/VM and VM/ESA 2.4.0 with enabling PTFs applied.

Support for *Parallel Access Volumes* includes:

1. The new CP QUERY PAV command, which displays information about the Parallel Access Volume devices on the system.
2. Enhancements to the CP QUERY DASD DETAILS command to display additional information if the queried device is a Parallel Access Volume.
3. A new CP Monitor Record, which has been added to Domain 6 (I/O) to record state change interrupts that indicate a change in the Parallel Access Volumes information: – Record 20 – MRIODSTC – State change

7.3.1.5 FlashCopy

VM/ESA

FlashCopy is supported for guest use (z/OS and OS/390) for dedicated (attached) volumes or for *full pack mini disks*. FlashCopy is supported for guests in VM/ESA 2.4.0 with an enabling PTF applied. The Web browser ESS Specialist interface can also be used to copy any volume within the same logical subsystem.

z/VM

z/VM allows a native CP user to initiate a FlashCopy function of a source device to a target device on an IBM Enterprise Storage Server. FlashCopy support includes the new CP FLASHCOPY command.

7.4 VSE/ESA support

VSE/ESA sees the Enterprise Storage Server as an IBM 3990 Model 6 storage control. VSE/ESA 2.1.0 is the minimum level that supports the ESS with Transparency mode

7.4.1 VSE/ESA support level

To use DASD volumes on an ESS, your VSE/ESA system must be at least at level VSE/ESA 2.1.0.

7.4.2 Exploitation of ESS functions

7.4.2.1 FlashCopy

VSE/ESA provides FlashCopy support with VSE/ESA 2.5.

Flashcopy can be set using the IBM TotalStorage ESS Specialist.

VSE/ESA uses the IXFP SNAP command to invoke FlashCopy. In order to prevent its casual use, you may use the VSE/ESA STACK command to hide the IXFP command in the following fashion:

STACK IXFP | * " IXFP " ...command reserved by Systems Programming

If a user were to issue the command IXFP SNAP,120:123 after using the above STACK command, the result would be that the command would be treated as a comment command and logged on SYSLOG. Then you may have batch jobs for invoking FlashCopy and use the UNSTACK command at the beginning of the job step to allow the IXFP command to be used and then reissue the STACK command at the end of the step.

More information on the STACK command is available at `http://www-1.ibm.com/servers/eserver/zseries/os/vse/pdf/vseesaht.pdf`. You arrive here from the System Hints and Tips from VSE entry on the VSE/ESA Web page at: `http://www-1.ibm.com/servers/eserver/zseries/os/vse/library/library.htm`

7.4.2.2 PPRC

Before version 2 release 5, VSE/ESA does not exploit the ESS unique functions, but it supports Peer-to-Peer Remote Copy (PPRC). Both PPRC and FlashCopy can be set up by using the IBM TotalStorage ESS Specialist. You can still use ICKDSF to manage PPRC.

7.4.2.3 Multiple Allegiance and Priority I/O Queueing

Multiple Allegiance and *Priority I/O Queueing* are ESS hardware functions independent of software support. So, VSE/ESA benefits from these functions in a shared environment. The priority, however, is not set by VSE/ESA. So, there is no I/O Priority Queuing for VSE initiated I/Os.

7.5 TPF support

The inherent performance of the Enterprise Storage Server makes it an ideal storage subsystem for a TPF environment.

7.5.1 Control unit emulation mode

To use an ESS in a TPF system, at least one logical subsystem in the ESS has to be defined to operate in IBM 3990 Model 3 TPF control unit emulation mode. The volumes defined in this logical subsystem can be used by TPF.

7.5.2 Multi Path Locking Facility

The ESS supports the *Multi Path Locking Facility* as previously available on IBM 3990 control units for TPF environments.

7.5.3 TPF support levels

The ESS is supported by TPF 4.1. Without additional PTFs applied, TPF 4.1 provides *transparency* support only, so due to the fact that FlashCopy requires *exploitation* support, FlashCopy is not supported by native TPF.

There are PTFs available for *exploitation* support of TPF 4.1. With the PTFs applied, the ESS function bits are interpreted and TFP will use the new performance enhanced CCWs.

Multiple Allegiance was a function already available for TPF environments on IBM 3990 systems as an RPQ. TPF benefits from ESS's *Multiple Allegiance* and *I/O Queuing functions*.

7.6 Open systems support

The Enterprise Storage Server supports the majority of the open systems environments. Constantly new levels of operating systems, servers of different manufacturers, file systems, host adapters and cluster software are announced in the market. Consequently, storage solutions must be tested with these new environments in order to have the appropriated technical support.

IBM keeps the following Web page updated with technical information that must be reviewed when you are interested to know the ESS support for specific open systems environments:

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

For this reason, it is not the intention of this section to describe all the supported server models, operating system levels, and host adapters. We strongly recommend that you go to the Web page, to get the latest support information. Following we review the current minimum open systems software support.

7.6.1 Compaq AlphaServer platforms

Using SCSI adapters, Tru64 UNIX 4.0D and Open VMS 6.2 are the minimum software level required for using the ESS with a Compaq AlphaServer system. Since the ESS uses the standard device drivers available with OpenVMS and Tru64 UNIX, there is no need for installing additional driver(s). A driver for the host SCSI adapter may be needed id a new, different adapter must be installed for the ESS at the same time.

Using fibre channel adapters, Tru64 UNIX 4.0F and Open VMS 7.2 are the minimum software level required for using a ESS with a Compaq AlphaServer system. Since the ESS uses the standard device drivers available with OpenVMS and Tru64 UNIX, there is no need for installing additional driver(s).

Note: A complete list of supported releases of Tru64 UNIX and OpenVMS can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.2 Data General AViiON platforms

Using SCSI adapters, DG/UX 4.2 is the minimum software level required for using the ESS with a Data General AViiON system. Since the ESS uses the standard device drivers available with DG/UX, there is no need for installing additional driver(s). A driver for the host SCSI adapter may be needed id a new, different adapter must be installed for the ESS at the same time.

Note: A complete list of supported releases of DG/UX can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.3 Hewlett Packard platforms

Using SCSI adapters, HP/UX10.20 is the minimum software level required for using the ESS with a Hewlet Packard system. Since the ESS uses the standard device drivers available with HP/UX, there is no need for installing additional driver(s). A driver for the host SCSI adapter may be needed id a new, different adapter must be installed for the ESS at the same time.

Remember to review the installation scripts to set up the ODM file so that performance is optimized. See the *IBM Enterprise Storage Server Host System Attachment Guide,* SC26-7296 for instructions on running this script.

Using fibre channel adapters HP/UX 11.0 is the minimum software level required for using the ESS with a Hewlett Packard system.

Remember that the IBM Subsystem Device Driver (SDD) gives HP/UX severs the ability to balance I/O across a secondary path and, in the process, improve subsystem performance. Not only does this provide better performance, it also improves availability. Should the primary path device have a problem, the workload is automatically switched to the secondary path to the ESS from the HP/UX sever.

Note: A complete list of supported releases of HP/UX can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.4 Intel based PC servers

Using SCSI adapters, Windows NT Server 4.0, Windows NT 4.0, Server Edition, or Windows 2000 Server is the minimum software level required for using the ESS with an Intel based (or equivalent) PC server system. Since the ESS uses the standard device drivers available with Windows NT Server 4.0, there is no need for installing additional driver.

Using fibre channel adapters Windows NT Server 4.0, Windows NT Server 4.0, Enterprise Edition, or Windows 2000 Server is the minimum software level required for using the ESS with an Intel based PC server.

Remember that the IBM Subsystem Device Driver (SDD) gives Windows NT Server 4.0 severs the ability to balance I/O across a secondary path and, in the process, improve subsystem performance. Not only does this provide better performance, it also improves availability. Should the primary path device have a problem, the workload is automatically switched to the secondary path to the ESS from the Windows NT Server 4.0 server.

Using SCSI adapters, Novell NetWare 4.2 is the minimum software level required for using the ESS with an Intel based (or equivalent) PC server system. Since the ESS uses the standard device drivers available with NetWare there is no need for installing additional driver.

Using fibre channel adapters Novell NetWare 4.2 is the minimum software level required for using the ESS with an Intel based PC server.

Note: A complete list of supported releases of Windows NT, Windows 2000 and Novell NetWare can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.5 Sun Microsystems platforms

Using SCSI adapters, Solaris 2.5.1 is the minimum software level required for using the ESS with a Sun Microsystems SPARCserver, SPARCenter, Ultra, or Enterprise Series server. Since the ESS uses the standard device drivers available with Solaris, there is no need for installing additional driver(s). A driver for the host SCSI adapter may be needed if a new, different adapter must be installed for the ESS at the same time. The Sun host file **sd.conf** may need to be edited to see all of the ESS LUNs. Without editing this file, LUN 0 can be seen, but not LUN1 and above.

Using fibre channel adapters, Solaris 2.6 is the minimum software level required for using the ESS with a Sun Microsystems SPARCsever, SPARCenter, Ultra, or Enterprise Series server.

Remember that the IBM Subsystem Device Driver (SDD) gives Sun Solaris severs the ability to balance I/O across a secondary path and, in the process, improve subsystem performance. Not only does this provide better performance, it also improves availability. Should the primary path device have a problem, the workload is automatically switched to the secondary path to the ESS from the Sun Microsystems server.

Note: A complete list of supported releases of Sun Microsystems can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.6 IBM RS/6000, RS/6000 SP and pSeries

Using SCSI adapters, AIX 4.2.1, 4.3 and later releases are the minimum software level required for using the ESS with a RS/6000, RS/6000 SP, and pSeries servers. HACMP 4.2.2 is the minimum software level used. PSSP 3.1 is the minimum software level for the IBM RS/6000 and pSeries SP if used. (If support for PSSP 2.4 is required, an RPQ must be submitted indicating this need).

Using fibre channel adapters, AIX 4.3.3 is the minimum software level required for using the ESS with a RS/6000, RS/6000 SP, or pSeries server. HACMP 4.3.1 is the minimum software level where used. PSSP 3.1.1 is the minimum software level for the IBM RS/6000 and pSeries SP if used.

Remember to review the installation scripts to set up the ODM file so that performance is optimized. See the *IBM Enterprise Storage Server Host System Attachment Guide,* SC26-7296 for instructions.

Remember that the IBM Subsystem Device Driver (SDD) gives AIX severs the ability to balance I/O across a secondary path and, in the process, improve subsystem performance. Not only does this provide better performance, it also improves availability. Should the primary path device have a problem, the workload is automatically switched to the secondary path to the ESS from the AIX server.

Note: A complete list of supported releases of AIX, HACMP and PSSP can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.6.7 IBM iSeries 400 and AS/400

Using SCSI adapters (# 6501), OS/400 Version 3.2, 4.1 or higher is required for using the ESS with an AS/400 or iSeries 400 server. For *remote load source* support OS/400 Version 4.3, or higher, with PTFs is required.

Using Fibre Channel adapter (# 2766), OS/400 5.1 is required.

Note: A complete list of supported releases of OS/400 can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

Also, the latest maintenance information (APARs, PTFs and PSP buckets) for the iSeries servers can be found at `http://www.as400service.ibm.com`

7.6.8 IBM NUMA-Q platforms

DYNIX/ptx 4.5.1 is the minimum software level supporting an ESS attached to a NUMA-Q server.

NUMA-Q servers only have fibre channel paths to natively attach to an ESS. They can also use the NUMA-Q Fibre Channel bridge to attach to a SCSI adapter on the ESS.

Note: A complete list of supported releases of DYNIX/ptx can be found on the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.7 FICON support

FICON channels provide enhanced performance for the execution of channel programs that allow the use of CCW and data pre-fetching and pipelining. The new FICON channel protocols are fully compatible with existing channel programs and Access Methods. IBM software has been changed to exploit this function but an installation should review the readiness of its non-IBM software.

When planning for FICON attachment, you may find helpful to refer to *FICON Native Implementation and Reference Guide,* SG24-6266.

7.7.1 z/OS and OS/390

FICON native channels are supported by z/OS version 1 release 1 and OS/390 version 2 release 8 and later releases.

Important: When planning to install the ESS with FICON attachment, the latest Preventive Service Planning (PSP) bucket for the processor or control unit should always be reviewed prior to installation. The PSP you can get it from your IBM Software Representative.

There is no unique PSP bucket for FICON. For both z/OS and OS/390, the FICON support information is contained in the PSP buckets for the processor and control units, listed in Table 5.

Table 5. Processor and ESS PSP buckets for z/OS and S/390 FICON support

Processor / ESS	PSP upgrade	PSP subset
zSeries 900	2064device	2064/OS390
9672 G5/G6	9672device	9672OS390G5+
ESS	2105device	

DFSMS

FICON device support for the ESS 2105 is provided with DFSMS APARs. The changes to DFSMS components, including IEBCOPY, IEWFETCH, ICKDSF, and so on, in support of FICON are to take advantage of the performance enhancements inherent in the operation of the FICON channel.

Note: refer to the latest PSP bucket for updated information on the required maintenance for the DFSMS component.

7.7.2 z/VM and VM/ESA

z/VM version 3 release 1 and VM/ESA 2.4.0. contain support for FICON. VM/ESA 2.3.0 requires an APAR for FICON support. The most complete and updated list of required maintenance (APARs and PTFs) will be available in the PSP bucket.

Important: The latest copy of the PSP bucket for the processor or control unit should always be reviewed prior to installation. The PSP buckets are listed in Table 6.

Table 6. Processor PSP buckets for z/VM and VM/ESA FICON support

Processor	PSP upgrade	PSP subset
zSeries 900	2064device	2064z/vm 2064vm/esa
9672 G5/G6	9672device	9672vm/esa

7.7.3 VSE/ESA

VSE/ESA 2.6 contains support for FICON. The most complete and updated list of required maintenance (APARs and PTFs) will be available in the PSP bucket.

Important: The latest copy of the PSP bucket for the processor or control unit should always be reviewed prior to installation. The PSP buckets are listed in Table 7.

Table 7. Processor PSP buckets for z/VM and VSE/ESA FICON support

Processor	PSP upgrade	PSP subset
zSeries 900	2064device	2064vse/esa
9672 G5/G6	9672device	9672vse/esa

7.7.4 TPF

TPF 4.1 contains support for FICON.

7.8 IBM TotalStorage ESS Specialist

The IBM TotalStorage ESS Specialist (ESS Specialist) is the Web user interface that is included with the Enterprise Storage Server. You use the ESS Specialist to view machine resources, problem status and machine configuration — and modifying the ESS configuration.

The IBM TotalStorage ESS Specialist also provides the means for invoking Copy Services to establish Peer-to-Peer Remote Copy and FlashCopy without having to involve the operating system running in the host server.

Using an Internet browser, such as Netscape Navigator or Microsoft Internet Explorer, the storage system administrator can access the ESS Specialist from a desktop or mobile computer as supported by the network.

The ESS Web interfaces support both the Netscape Navigator and the MSIE versions listed in Table 8. The key is that the browser must support Java 1.1. (Netscape Navigator 4.04 or later provides this support. Microsoft Internet Explorer 4.00 does too.)

Table 8. Web browsers supported by ESS Specialist

Netscape level	MSIE level
Netscape 4.04 with JDK 1.1 fixpack	MSIE 4.x with Microsoft Java Virtual Machine (JVM) 4.0 or 5.0
Netscape 4.05 with JDK 1.1 fixpack	MSIE 5.x with Microsoft JVM 4.0 or 5.0
Netscape 4.06 (no fixpack required)	
Netscape 4.5x (nofixpack required)	
Netscape 4.7x (no fixpack required)	

Note: ESS Specialist does not support Netscape 6.0.

When the Internet Explorer is the browser accessing the ESS from an Internet domain different from that of the ESS, the ESS will return an error 1114 to the browser when the user attempts an update — e.g. accessing domain **ESS1.illinois.abccorp.com** from **user1** at domain **indiana.abccorp.co**m. This is a browser restriction. Netscape Navigator or some other browser should be used to circumvent this restriction.

7.9 IBM TotalStorage ESS Expert

The IBM TotalStorage Enterprise Storage Server Expert (ESS Expert) gathers and presents information that can significantly help storage administrators manage, one, or more, Enterprise Storage Server. Capabilities are provided for performance management, asset management, and capacity management (see Appendix A, “IBM TotalStorage ESS Expert” on page 281).

For installation and use of the ESS Expert, you may refer to *IBM StorWatch Expert Hands-On Usage Guide,* SG24-6102.

The ESS Expert can be installed on the following platforms:

- Windows NT 4.0 with Service Pack 6
- AIX 4.3.3

For additional information, and the more frequently updated, we recommend you to visit the ESS Expert home page on the Internet at

`http://www.ibm.com/storage/software/storwatch/ess/`

When the ESS Expert installs, in addition to the Expert code, the installation process installs DB2 UDB which provides the database storage capability for the asset, capacity, and performance information, the IBM WebSphere Application Server which provides the capability for remote viewing and management, and the Sage product from Viador Corp, which provides capability for graphing and for customizing queries. The combined memory requirement for these applications is such that we recommend, in Windows NT environments, that the Expert be run in a server of 256 MB memory in production environments.

7.9.1 AIX

The software requirements for the ESS Expert in the AIX environment are:

PROGRAM	NOTE
AIX operating system 4.3.3	Installed separately
TCP/IP with Domain Name Server (DNS) host name	Included with AIX 4.3.3

The ESS Expert component is designed to operate with the following programs, that are all included in the IBM TotalStorage Enterprise Storage Server Expert distribution CD:

- IBM HTTP Server 1.3.6.2 *
- JDK 1.1.8 PTF 7 *
- IBM WebSphere Application Server 2.031 Standard Edition *
- DB2 Universal Database (UDB) 7.1 Workgroup Edition
- Viador Sage 4.1 *

Note: These programs are installed by the ESS Expert installation and cannot pre-exist on the target server because explicit releases are included and different levels cannot be present on the server. If any of these products are installed, then you must un-install them first.

It is okay to have another Web server (which is not the IBM HTTP Server) installed on your machine, but this Web server and the IBM HTTP Server should not run on the same port. The default port used by most Web servers is 80. You must make sure that the port entered for the IBM HTTP Server during installation time is not used by another Web server.

For the most updated information we recommend you to visit the ESS Expert home page on the Internet at `http://www.ibm.com/storage/software/storwatch/ess/`

7.9.2 Windows

The following table shows the programs that are required by ESS Expert

Program	Level	Note
Windows NT operating system	Version 4.0 with Service Pack 6	To be installed before installing ESS Expert
TCP/IP with Domain Name Server (DNS) host name		To be installed before installing ESS Expert

The ESS Expert component is designed to operate with the following programs, that are all included in the IBM TotalStorage Enterprise Storage Server Expert distribution CD:

- IBM HTTP Server 1.3.6.2 *
- JDK 1.1.8 *
- WebSphere Application Server 2.031 Standard Edition *
- DB2 Universal Database (UDB) 7.1 Workgroup Edition
- Viador Sage 4.1 *

Note: These programs are installed by the ESS Expert installation and cannot pre-exist on the target server because explicit releases are included and different levels cannot be present on the server. If any of these products are installed, then you must un-install them first.

For the most updated information we recommend you to visit the ESS Expert home page on the Internet at `http://www.ibm.com/storage/software/storwatch/ess/`

7.9.3 Client access to the ESS Expert

The ESS Expert provides a Web based user interface which is designed to operate with the following programs:

- Netscape 4.7 for Windows NT 4.0 with Service Pack 6
- Microsoft Internet Explorer 5.0 for Windows NT 4.0 with Service Pack 6

7.10 Peer-to-Peer Remote Copy (PPRC)

Peer-to-Peer Remote Copy (PPRC) is a hardware solution that enables the asynchronous shadowing of application system data from one site to a second site (see 6.8, “Peer-to-Peer Remote Copy” on page 208).

This solution is available for open systems environments and for zSeries environments. It is an optional feature that can be ordered with the Enterprise Storage Server (see Appendix E.5.2, “PPRC and FlashCopy” on page 310).

7.10.1 Open systems environments

Listed in this section are the operating systems requirements for the open system environments. If you are planning to implement ESS Copy Services on any of the supported open environments, we recommend you to see *Implementing ESS Copy Services on UNIX and Windows NT/2000,* SG24-5757 and *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220.

For the most current information on supported servers and maintenance requirements you must visit the ESS site at: `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

7.10.1.1 AIX support

For SCSI attachment of the ESS to pSeries, RS/6000 and RS/6000 SP servers:

- AIX Version 4.2.1
- AIX Version 4.3.1 or higher

When the attachment is Fibre Channel:

- AIX Version 4.3.3 or higher

7.10.1.2 Windows NT

For servers running Windows NT:

- Microsoft Windows NT Server 4.0 requires Service pack 4 or 5
- Microsoft Windows NT Server 4.0 Enterprise Edition requires Service pack 4 or 5

7.10.1.3 SUN

For servers running Sun Solaris:

- Solaris 2.6
- Solaris 7
- Solaris 8

7.10.1.4 HP/UX

For servers running HP/UX:

- HP/UX 10.20
- HP/UX 11.00

7.10.1.5 OS/400

For SCSI attached iSeries 400 servers:

- OS/400 4.5 with PTFs

Currently the iSeries PPRC support is available for SCSI attached disks via RPQ.

7.10.1.6 NUMA-Q

For servers running NUMA-Q:

- DYNIX/ptx 4.4.7

7.10.2 zSeries environments

Listed in this section are the operating systems requirements for the zSeries environments. If you are planning to implement ESS Copy Services on a zSeries environment, we recommend you to refer to *Implementing ESS Copy Services on S/390,* SG24-5680, and visit the DFSMS/MVS Copy Services home page on the Internet at `http://www.storage.ibm.com/software/sms/sdm/sdmtech.htm`

Also, for the required software releases and latest PTFs, you should review a most recent PSP (Preventive Service Planning) bucket, that you can get from your IBM Software Representative.

7.10.2.1 z/OS

z/OS supports PPRC at levels 1.1 and 1.3.

7.10.2.2 OS/390

The support for PPRC in the OS/390 is provided by DFSMS 1.3.0 and above, plus PTFs.

You must also have ICKDSF release 16 plus APARs.

For the most updated information you must visit the DFSMS/MVS Copy Services home page on the Internet at:

`http://www.storage.ibm.com/software/sms/sdm/sdmtech.htm`

7.10.2.3 VM/ESA

PPRC is supported in native VM/ESA 2.2.0 and above by the use of ICKDSF or the ESS Specialist. ICKDSF must be at release 16 plus APARs.

7.10.2.4 z/VM

PPRC is supported in native z/VM and above by the use of ICKDSF or the ESS Specialist. ICKDSF must be at release 16 plus APARs.

7.10.2.5 VSE/ESA

Much like VM/ESA PPRC support is provided by ICKDSF release 16 plus APARs. VSE/ESA 2.3.1 is the minimum release level for PPRC support plus APARs. See also 7.4.2.2, “PPRC” on page 237.

7.11 FlashCopy support

FlashCopy is a function that makes a point-in-time (T0) copy of data (see 6.5, “FlashCopy” on page 202).

This solution is available for open systems environments and for zSeries environments. It is an optional feature that can be ordered with the Enterprise Storage Server (see Appendix E.5.2, “PPRC and FlashCopy” on page 310).

7.11.1 zSeries environments

7.11.1.1 z/OS and OS/390

Listed in this section are the operating systems requirements for the zSeries environments. If you are planning to set up Copy Services on an ESS, we recommend you to refer to *Implementing ESS Copy Services on S/390,* SG24-5680.

For the most updated requirements information we recommend you to visit the DFSMS/MVS Copy Services home page on the Internet at:

```
http://www.storage.ibm.com/software/sms/sdm/sdmtech.htm
```

Also, for the required software releases and latest PTFs, you should review a most recent PSP (Preventive Service Planning) bucket, that you can get from your IBM Software Representative.

FlashCopy requires ESS exploitation support.

- z/OS fully exploits this function
- For OS/390, DFSMS/MVS version 1 release 3, and subsequent releases, provide FlashCopy support

As part of the set up planning for FlashCopy implementation, you should get the latest list of required maintenance (minimum software level, PTFs, and APARs). This information you find it in the PSP bucket, that you can get from your IBM Software Representative. Also you may refer to the DFSMS Copy Services home page.

Under z/OS and OS/390, FlashCopy can be controlled using DFSMSdss or TSO commands besides the ESS Specialist

7.11.1.2 z/VM

VM supports the ESS as an IBM 3990 Model 6 Storage Control, so due that FlashCopy requires exploitation support, FlashCopy is supported for guest use only (DFSMSdss COPY). See also 7.3.1.5, “FlashCopy” on page 236.

7.11.1.3 VSE/ESA

VSE/ESA version 2 release 5 provides the support of the FlashCopy function of the Enterprise Storage Server. FlashCopy supports fast duplication of disk volumes. VSE/ESA support is provided by the IXFP SNAP command. See also 7.4.2.1, “FlashCopy” on page 237.

7.11.2 Open systems environments

Listed in this section are the operating systems requirements for the open systems. If you are planning to set up Copy Services on an ESS, we recommend you to see *Implementing ESS Copy Services on UNIX and Windows NT/2000,* SG24-5757 and *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220.You should also refer to the most complete and current information on all supported servers at:

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

7.11.2.1 AIX Support.

For SCSI attachment of the ESS to pSeries, RS/6000 and RS/6000 SP servers:

- AIX Version 4.2.1
- AIX Version 4.3.1 or higher

When the attachment is Fibre Channel:

- AIX Version 4.3.3 or higher

7.11.2.2 Windows NT

For servers running Windows NT:

- Microsoft Windows NT Server 4.0 requires Service pack 4 or 5
- Microsoft Windows NT Server 4.0 Enterprise Edition requires Service pack 4 or 5.

7.11.2.3 SUN

For servers running Sun Solaris:

- Solaris 2.6
- Solaris 7
- Solaris 8

7.11.2.4 HP/UX

For servers running HP/UX:

- HP/UX 10.20
- HP/UX 11.00

7.11.2.5 OS/400

For SCSI attached iSeries 400 servers:

- OS/400 4.5 with PTFs

7.11.2.6 NUMA-Q

For servers running NUMA-Q:

- DYNIX/ptx 4.4.7

7.12 Command Line Interface (CLI)

The ESS Specialist Web Copy Services Interface is available for all supported platforms. The Web Copy Services Interface function is a Java/CORBA based application that runs on the host server and requires a TCP/IP connection to each ESS under it. Copy Services also includes a Command Line Interface feature (CLI.) Using the CLI, users are able to communicate with the ESS Copy Services from the server command line. See 6.6, “Invocation of FlashCopy” on page 204 and 6.9, “Invocation of PPRC” on page 213.

The following operating system environments are supported with the Command Line Interface for invocation of the Copy Services:

- AIX 4.2.1, 4.3.1, 4.3.2, 4.3.3 with Java 1.1.8
- Windows NT 4.0 or Windows 2000 with Java 1.1.8
- HP-UX 10.20 or 11.0 (32 & 64bit) with Java 1.1.8
- Sun Solaris 2.6, 7, 8 both 32 & 64 bit with Java 1.1.8

If you are planning to use Copy Services Command Line Interface please refer to:

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

7.13 Subsystem Device Driver (SDD)

The IBM Subsystem Device Driver (SDD) is a program that manages redundant connections between the host server and the Enterprise Storage Server, providing performance enhancement and availability. This program comes standard with the ESS.

Currently there are two versions of the IBM SDD. SDD version 1.2.0 and SDD version 1.2.1.

If you plan to use SDD, then you should refer to the IBM Subsystem Device Driver (SDD)/IBM Data Path Optimizer section of the Internet Web page at: `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm` for supported functions and servers, and updated software requirements information.

Currently, support is available for using the IBM SDD on either AIX, Windows NT, Windows 2000, HP-UX and Sun Solaris platforms. You can find the procedure for this support and additional software maintenance level information at: `http://www.ibm.com/storage/support/techsup/swtechsup.nsf/support/sddupdates`.

AIX 4.3.3., 4.3.2., 4.2.1. For AIX 4.3.3, 4.3.2 and 4.2.1, you must install SDD version 1.2.0. However if HACMP is installed and activated, you must have 4.3.3 and SDD version 1.2.1.0. For AIX 4.3.3, you must have RML 04 or higher. For the allowed lower AIX levels, please refer to *IBM Subsystem Device Driver User's Guide* (this publication is available only in the Internet at `http://www.ibm.com/storage/support/techsup/swtechsup.nsf/support/sddupdates)`.

Microsoft Windows NT. SDD version 1.2.1 is required to support Windows NT clustering. Windows NT clustering requires Windows NT 4.0 Enterprise Edition. It is important to note that SDD 1.2.1 does not support load balancing in a Windows NT clustering environment. You must have Windows NT 4.0 Service pack 3.0 or higher installed on your system.

Windows 2000. Windows 2000 requires SDD version 1.2.0. For Windows 2000, you cannot run the Subsystem Device Driver in a non-concurrent environment in which more than one host is attached to a LUN on an ESS. However, concurrent multi-host environment is supported.

HP/UX 11.0. Version 1.2.0 of SDD is required for HP/UX. The SDD version 1.2.0 supports only 64-bit mode application on HP/UX 11.0 servers. The Subsystem Device Driver resides above the HP SCSI disk driver in the protocol stack.

SUN Solaris 2.6, 7 and 8. The IBM SDD supports only 32-bit application on a Solaris 2.6, but supports both 32-bit and 64-bit on Solaris 7 and 8.

HACMP support for SDD 1.2. The information of the HACMP hardware and software support for the ESS can be obtained from the HACMP SupportLine Web page on the IBM Intranet. Customers who have SupportLine service can also obtain this information by calling the IBM Software Support Center.

7.14 GDPS

GDPS is designed to minimize and potentially eliminate the impact of any failure, including disasters, or a planned site outage (for a detailed description of GDPS see 6.13, “Geographically Dispersed Parallel Sysplex (GDPS)” on page 226).

7.14.1 Software

The following are the software requirements:

- MVS/JES2 SP5.2.2 / DFSMS 1.3, OS/390 V1 R1, or later — this contains the PPRC command set support and freeze support.
- System Automation for OS/390 1.2 or later (formerly AOC) — System Automation is used to automate workload startup and shutdown. This also applies for Cgroup Freeze/Run.
- NetView 2.4, NetView 3.1, or TME 10 NetView for OS/390 or later — used to automate GDPS processing. Although GDPS prerequisites NetView and System Automation, if an enterprise does not use NetView and System Automation, GDPS will provide interfaces to coordinate processing between GDPS and an enterprise's existing automation for their coexistence.
- Site Monitor 3.18 (including HMC Bridge) or later, Site Manager 2.14 (including HMC Bridge) or later, or Availability and Operations Manager (AOM) — used to IPL, query status, system reset, and system restart a member of the sysplex.
- TSO/E any release — used to issue commands to manage storage subsystems

7.14.2 Enterprise

The following enterprise setup is required:

- The multi-site Parallel Sysplex cluster must be configured with redundant hardware in both sites.
- The cross-site links should be distributed across a minimum of two diverse routings for availability.
- The critical data for restart must be mirrored via PPRC and reside on disk that supports the PPRC
- Mono-directional PPRC
- CGROUP FREEZE/RUN function.
- The enterprise must establish procedures for mirroring any tape resident data sets.
- The workload startup and shutdown must be automated using System Automation for OS/390 or equivalent. This applies also for Cgroup *freeze/run*.

7.15 DB2 split mirror

This function, implemented as a fix in DB2 UDB for OS/390 V6.1, allows the command SET LOG SUSPEND. The command forces out log buffers, updates the high written RBA in the BSDS, and sets a latch that prevents writes to the log and to application data. This creates a point of consistency which allows a completely consistent backup copy.

This function allows controlled suspension of application WRITEs and is available for DB2 V5.1 as a user modification, and for DB2 V6.1 with APARs, and is fully implemented in DB2 V7.

Chapter 8. Installation planning

The installation and implementation of the Enterprise Storage Server requires careful planning. In this chapter you will find a quick reference to the major steps and considerations required to get the ESS installed and operational.

When starting your ESS installation planning you will find very helpful to refer to the *IBM Enterprise Storage Server Introduction and Planning Guide,* GC26-7294

8.1 Major planning steps

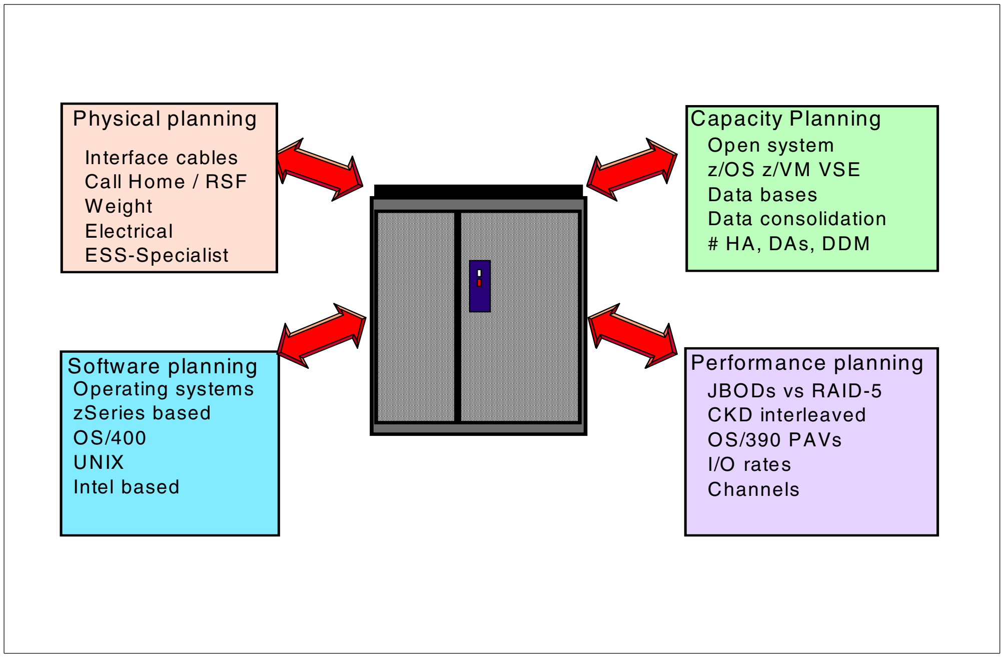


Figure 147. Installation/implementation planning

The major planning steps, as shown in Figure 147, on page 258 are:

- *Capacity planning*. The Enterprise Storage Server is a storage subsystem designed for mainframes and open systems. This characteristic makes the ESS your first choice when planning for storage consolidation. In this way, capacity planning becomes an important step for all the servers that share the ESS can fully benefit from its features and capabilities. Since the ESS is able to be used by different host environments, you must bring together everyone in your shop that will store data in the Enterprise Storage Server. This encompasses, for example, zSeries and the UNIX staff, as well as Windows NT administrators or database managers. Every person involved in the capacity planning of the ESS must plan for the requirements and needs of their particular host system.
- *Physical planning*. This step includes the planning for all interface cables to the different hosts, the adapters, and the prerequisites for the *Call Home* and *Remote Support Facilities*. You also need to plan the connection to the ESS Ethernet ports to be able to configure the ESS. It is also important to plan what power feature will be ordered, all the power outlets, and the size and weight of the ESS and prepare for it. Your IBM Installation Planning Representative will assist you with the physical planning.
- *Software planning*. In this step you will have to check if your host software is ready to support an ESS. You have to make sure that you are at the appropriate software release level and maintenance to exploit the functions of the ESS.
- *Performance planning*. Here you have to consider things like JBODs or RAID ranks, CKD interleaved ranks, Parallel Access Volumes, number and type of channels (zSeries ESCON and FICON), and number and type of interfaces

(open systems, SCSI and Fibre Channel) that will be attached to the Enterprise Storage Server.

- *Disaster recovery planning*. If you are setting up an environment for disaster recovery, then you must plan additional things. You will have to plan for the remote site, the links, the readiness of the procedures for a disaster recovery situation. You will also plan for the data that will be protected.

8.2 ESS installation planning

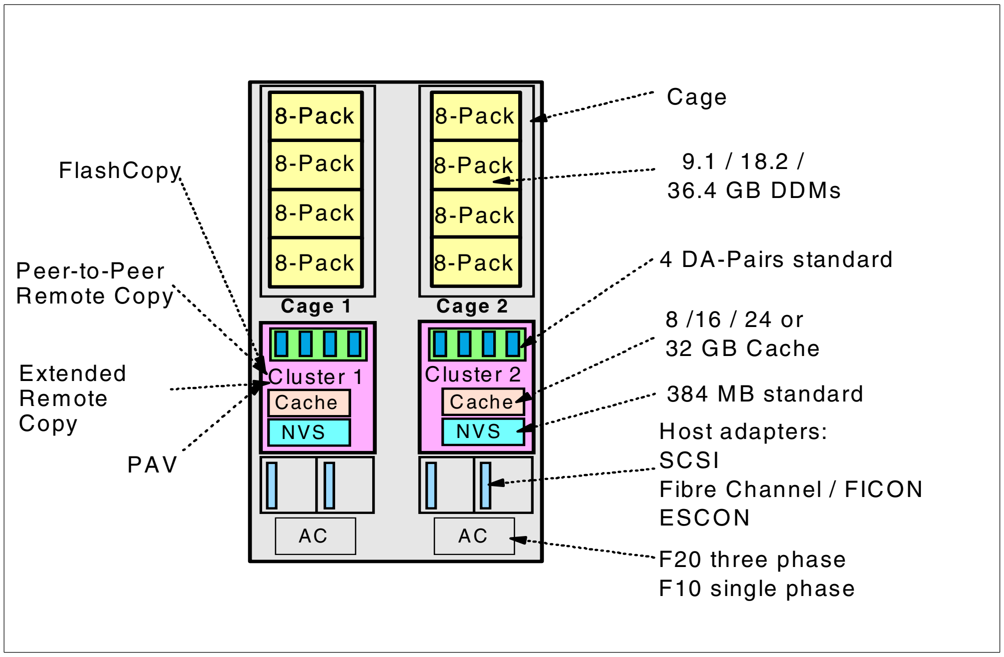


Figure 148. Planning considerations

The following is an outline of the planning steps that will aid you in planning for an ESS installation. You may find helpful to refer to the *IBM Enterprise Storage Server Configuration Planner,* SC26-7353 when planning for the ESS installation.

8.2.1 ESS planning steps

1. Order the ESS manuals that deal with installation planning. You should also order the manuals for the IBM TotalStorage ESS Specialist.
2. Next, you have to get together with all the people involved in storage planning and administration from the various platforms that plan to place data on an ESS. All of them speak a different language when dealing with storage. zSeries people talk about ESCON and FICON channels, and CKD volumes, while open systems people talk about SCSI and Fibre Channel, and LUNs.
3. For your capacity planning, you must find out the storage requirements for all these various users (this should also include storage capacity growth planning). This information determines the capacity feature codes and 8-packs you need for the ESS. For standard configurations, it will also determine how the ranks will be pre-formatted: for CKD use, or for open systems (FB) use.
4. If you have any IBM 7133s, you need to decide whether you want to attach them to the ESS. If you decide to attach existing 7133s then you will need to configure the ESS with reserved loops and also order the 2105-100 racks to hold the 7133 drawers. You may need to upgrade the existing drawers for attaching to the ESS.
5. The next step deals with some basic performance planning aspects. Which standard configuration is most suited to your requirements?

6. The number of systems you want to attach to the ESS, their performance and availability requirements, will determine the number and type of host adapters. If you want high availability and high performance for your UNIX and Windows NT data, consider the use of the IBM SubSystem Device Driver.
7. Next you need to plan for all the cables required to attach the ESS. This includes the cables for the host attachment, types of cables, and cable length. Naturally, the number of each specific type of cable is dependent on the number of host attachments you plan to have for using each cable type. To know the types of adapters and cables required for the attachment of the various opens system servers, you must refer to `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`
8. You must plan for the Ethernet LAN required to configure the ESS. Note: Some of the following steps will not be required, if you plan to take advantage of the IBM service to provide a private network for the ESS connections:
   - You may need to plan for the workstation, the cable, hub and Ethernet adapter to connect the workstation running a Web browser to the ESS.
   - You may need a Web browser that supports Java 1.1 to be able to access the IBM TotalStorage ESS Specialist Web server in the ESS to configure the ESS.
   - You may need to supply a TCP/IP address for each of the ESS clusters.
9. Check whether your environment is ready for the installation of an ESS. (Particularly, open systems environments may not be set up to handle the necessary power and other requirements.)
   - The ESS is available as a single-phase power box, and as a three-phase power box. Depending on the model, it might be necessary to use a three-phase power supply. You might also need a three-phase Un-interrupted Power Supply (UPS).
   - The ESS requires a raised floor, and since it is quite heavy, the floor where you plan to install the ESS must be able to sustain the weight of the ESS.
10. Plan for the modems, cables, and telephone connection required for your ESS's *Call Home* facility and power outlets.
11. Check with your IBM service organization to see if your host system software is ready for an ESS attachment to the host. The information in the Web will be useful for these activities
    - For z/Series operating systems, you need to determine, for example, what software support level (transparent support, toleration support, or exploitation support) is available, or what level is required to exploit the new performance functions of Parallel Access Volumes. Use the PSP buckets to know the software level and maintenance requirements you must meet.
    - For open systems, check whether any software updates must be applied to your system. You will find the information you need at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`
12. For z/OS, you need to decide if you are going to exploit ESS's Parallel Access Volumes (PAV). Look at your RMF data, and if you find *Avg IOSQ* times there, you should consider the use of PAV. (Note: PAV is an optional feature for the ESS, so you will have to order it.)

13. Next, you should discuss with all the storage administrators and database people what copy or backup techniques they plan to use. The ESS can copy a volume within a few seconds with its FlashCopy copy function. This may change the way you are making your backups on an ESS. Since FlashCopy is an optional feature, you need to order it if you plan to use it.

14. You should discuss with the higher management levels of your enterprise what protection level is required for your enterprise data. This can lead to the decision to implement a disaster recovery solution. The ESS provides a synchronous remote copy function for disaster backup of your data. For z/OS environments, there is also an asynchronous remote copy function available. In any case, the remote copy function is a separate feature that must be ordered to use it. If you have decided to take advantage of such a disaster recovery solution, further planning is required. This includes planning for the secondary ESS, ESCON connections between the Enterprise Storage Servers, and Ethernet connections between the ESSs.

15. An important planning topic is how you are going to migrate your data from the previous storage systems onto the ESS. You can either choose migration techniques for each environment, or decide to take advantage of IBM's migration services. Check with your IBM representative regarding what migration services are available in your country.

8.2.2 ESS pre-installation steps

The installation of the IBM Enterprise Storage Server is performed by the IBM System Support Representative (SSR). However, there are checks that need to be performed before the ESS is actually delivered, unpacked and installed.

1. The ESS is a big machine that weighs more than 1 metric ton. The customer site must be ready and prepared to accommodate such a machine. The final position of the ESS must be prepared and marked. The ESS is also serviced from the front and rear, so enough service clearance must be allocated.

2. It is also very important to know and plan how the ESS will enter the customer's site to its final position in the computer room. The ESS cannot be dismantled so that it would fit in a door. The path the ESS would take on its way to the computer room must be traced and all corridors, turns and doors must be carefully measured to ensure the ESS will fit through.

3. A site inspection of the final position of the ESS must be conducted to ensure everything is ready. This inspection includes power test and environmental test.

8.2.3 Physical installation

Physical installation begins once all the pre-installation steps are done, the site was inspected and meets the requirements, and the ESS is delivered. The IBM System Support Representative will prepare the ESS physically and run the tests needed to verify that the ESS is working properly.

8.2.4 ESS configuration steps

Once the Enterprise Storage Server has arrived, and set up by the SSR, it is now ready to be configured, and attached to the hosts. To configure the ESS, the workstation running the Web browser must be connected to the ESS Ethernet LAN and must access the IBM TotalStorage ESS Specialist Web server by entering the ESS TCP/IP host name of one of the ESS clusters.

1. If you choose one of the standard logical configurations, most of the logical configuration steps have already been performed for you by the IBM SSR (by means of the Batch Configuration Tool). Ranks are predefined either to be CKD or FB. What is left to do for open systems connections is the configuration of the SCSI and/or Fibre Channel ports.
2. Consider the use of *custom volumes*, particularly for zSeries systems. If you have applications with very high I/O activity to certain data sets, consider the placement of these data sets on separate small custom volumes.
3. If you did not choose one of the standard configurations, you will have to do the full logical configuration of the Enterprise Storage Server. This includes:
   - Formatting for CKD (zSeries) or FB (open systems)
   - Volume emulation type (3390 or 3380 for zSeries, 9337 or 2105 for iSeries). LUN size for the other open systems
   - Rank type: RAID 5 or JBOD
   - Use of the interleaved partition
   - Use of custom volumes
4. If you are going to use Parallel Access Volumes in zSeries, you must define *alias* addresses for the *base* volumes. You have to define the *aliases* in the ESS as well as in the HCD.
5. For zSeries systems, you must enter the ESS logical configuration in HCD to create an IODF. You must define the logical control units there, and the alias addresses for PAV volumes.
6. After you have completed the logical configuration of the ESS and prepared your host software for the ESS attachment, you can attach the ESS to the hosts. Depending on the host operating system, this task can be performed concurrent to your normal production, or you may have to plan for the shutdown of the host systems to connect the ESS.
7. The next step includes the things required to make the logical volumes of the ESS usable by your host systems:
   - For the CKD type servers, you must format the volumes with ICKDSF. You only need to do a minimal INIT.
   - For UNIX systems, you can add the ESS logical volumes (logical disks) to logical volume groups and logical volumes in UNIX; you can create file systems on them, and so on.
   - In Windows NT, you can assign drive letters to the logical volumes.

Chapter 9. Migration

Migrating data to the Enterprise Storage Server can be done using standard host utilities. Each operating system sees the logical volumes in an ESS as normal logical devices. For the CKD type of servers this is an IBM 3390 or 3380. For the iSeries servers, these are 9337 or 2105. For UNIX systems the logical disks are recognized as SCSI or Fibre Channel attached drives (hdisk in AIX, for example), and Windows NT also sees them as SCSI or Fibre Channel attached disks.

Data migration requires careful planning. The major steps are:

1. Education (including documentation)
2. Software readiness
3. Configuration planning
4. Data migration planning
5. Hardware and software preparation
6. Data migration

Refer to the *IBM Enterprise Storage Server Host System Attachment Guide,* SC26-7296 publication for platform specific details on migrating data to the ESS.

9.1 Defining and connecting the system

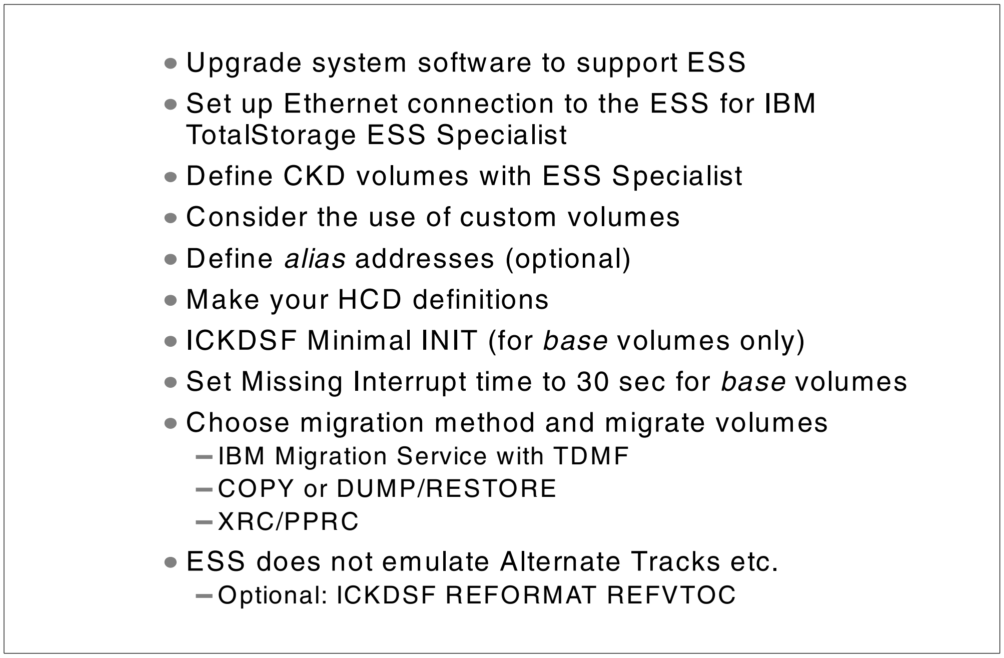
Before you can migrate your data to an ESS, you first have to configure the ESS and connect it to a host.

Before you can do the configuration, you need either an Ethernet connection from your intranet to the ESS, or a private network for your ESS systems. You will have to specify TCP/IP addresses for each ESS cluster.

During the configuration process, you specify—for each disk rank in the ESS you want to use—the logical disks or logical volumes you will be using. The ESS then formats the logical volumes according to your definitions.

You can use either the standard logical configurations done by the IBM SSR tool or use the ESS Specialist to configure the volumes.. For detailed information on using the ESS Specialist when configuring the ESS, please refer to the *IBM Enterprise Storage Server Web Interface Users Guide for the ESS Specialist and ESS Copy Services,* SC26-7346*.*

9.2 Data migration in z/OS environments



- Upgrade system software to support ESS
- Set up Ethernet connection to the ESS for IBM TotalStorage ESS Specialist
- Define CKD volumes with ESS Specialist
- Consider the use of custom volumes
- Define *alias* addresses (optional)
- Make your HCD definitions
- ICKDSF Minimal INIT (for *base* volumes only)
- Set Missing Interrupt time to 30 sec for *base* volumes
- Choose migration method and migrate volumes
  - IBM Migration Service with TDMF
  - COPY or DUMP/RESTORE
  - XRC/PPRC
- ESS does not emulate Alternate Tracks etc.
  - Optional: ICKDSF REFORMAT REFVTOC

Figure 149. Preparation for zSeries data migration

To access the logical volumes on an ESS from a zSeries system, you need an IODF that includes the logical CKD subsystems of the ESS. This is done with the HCD dialog.

9.2.1 Defining and initializing the volumes

Before you start to define DASD devices in the HCD and the ESS, consider if you are going to use Parallel Access Volumes. Using your performance reporter (RMF, for example),check whether your volumes have high *Avg IOSQ* times. If this is the case, you should use Parallel Access Volumes (PAVs) for better performance. If you intend to use PAVs, you need to plan for the *alias* addresses that must be defined.

After you have finished the volume configuration in the ESS and in the HCD, the logical volumes must be initialized with ICKDSF. Only a Minimal INIT is required.

9.2.2 Migration considerations

Some of the ESS advanced functions require a certain level of software support. This should be considered before using the ESS.

9.2.2.1 Software support

Depending on the ESS functions you want to exploit, you might have to upgrade your software to the right level. Always check the latest PSP (Preventive Service Planning) Buckets for software support levels. See Chapter 7, “Software support” on page 231.

For the *base* volumes on an ESS, set the missing interrupt time to 30 seconds.

9.2.2.2 Custom volumes

The ESS allows you to define small 3390 or 3380 volumes. You define how many cylinders you need for a logical volume. Having small logical volumes can drastically reduce contention for a volume, particularly when several data sets with high activity reside on the same volume. On an ESS, you can place each highly active data set on a separate *custom volume* without wasting a lot of space.

Before migrating your volumes 1:1 onto ESS, you should consider if there are candidate data sets for custom volumes. Having identified such data sets, you can plan for the size of the custom volumes.

9.2.2.3 Considerations after data migration

The ESS does not emulate Alternate Tracks, Device Support Tracks, and Service Tracks. This is similar to the implementation on IBM 3990 Model 6 with RAMAC devices. The IBM RVA did emulate these Alternate, Device Support, and RAS tracks. This sometimes caused some problems when migrating volumes 1:1 from one storage subsystem to another, when back level system software was used which did not update the Volume Table of Content (VTOC) to reflect the correct number of tracks. It is always a good idea to refresh the VTOC after a volume migration with the ICKDSF REFORMAT REFVTOC command. This refresh sets the number of tracks to the correct value.

TDMF is one of the programs that does not manage the alternate track difference; for this reason, a REFORMAT REFVTOC will be needed after migrating to an ESS if the source volume was an IBM RVA, or a non-IBM subsystem.

9.2.3 Migration methods

There are several ways to migrate data. Depending on your requirements and your environment, one of these methods may be adequate for you.

9.2.3.1 IBM migration service

The easiest way to do data migration is to let IBM do it for you. In several countries IBM offers a migration service. Data can be migrated from any previous S/390 storage subsystem to ESS. IBM uses the Transparent Data Migration Facility (TDMF) tool to do the data migration. Data is migrated while your normal production work continues. When all data is copied onto the ESS, you can restart your systems using the new volumes on ESS.

9.2.3.2 Copy, and Dump/Restore

The classic approach is to dump all source volumes to cartridge and restore them to ESS volumes. This is the slowest approach and requires the application systems to be down during the dump/restore process. The advantage of this approach is that you do not need to attach both storage subsystems (the old one and the ESS) at the same time.

A much faster migration method is to do a volume copy from the old volumes onto ESS volumes using, for example, the DFSMSdss COPY program. This migration method also requires that both storage subsystems are online to the system that does the migration. Application systems must be down during the migration process

While DFDSS is by far the simplest way to move most data, the following alternatives are also available:

- IDCAMS EXPORT/IMPORT (VSAM)
- IDCAMS REPRO (VSAM, SAM, BDAM)
- IEBCOPY (PDSs - including load module libraries - and PDSEs)
- ICEGENER (SAM) - part of DFSORT
- IEBGENER (SAM)
- Specialized data base utilities (e.g. for CICS, DB2, or IMS)

9.2.3.3 Migrating data with XRC

If your system is z/OS and your data currently resides on DASDs behind an IBM 3990 Model 6 or IBM 9390 storage controller, you can use XRC to migrate your volumes to ESS. This is the most convenient way to do data migration. Your application systems can continue to run while you are migrating your data. When old and new volumes are synchronized, you can shut down your systems and restart them using the new volumes on ESS.

While you need a special enabling feature for the ESS if you want to use the ESS as a primary control unit, this feature is not required when the ESS is a secondary control unit as in a migration scenario.

Migrating volumes with XRC is quite easy. You just need to allocate a State Data Set, for example hlq.XCOPY.session_id.STATE., and solve RACF demands. The use of XRC commands like XSTART, XEND, XADDPAIR, and XRECOVER must be allowed. The Data Mover task ANTAS001 must be allowed by RACF to read from the source volumes, and it needs update authority to the State Data Set.

The system where the System Data Mover task runs needs access to both source and target storage control units. The target volumes must be online to the System Data Mover system. The volumes can have any VOLSER.

You can start an XRC session with the command:

```
XSTART session_ID ERRORLEVEL VOLUME SESSIONTYPE(MIGRATE) HLQ(hlq)
```

Any name you choose can be used for session_ID, but it must match the session_ID in the State Data Set.

Now you can add pairs to be synchronized with the command:

```
XADDPAIR session_ID VOLUME(source target)
```

After all pairs are synchronized, you can check this with the XQUERY command. You need to choose a time when you can shut down your application systems to do the switch to the new volumes.

After you have stopped your application systems, issue the command sequence:

```
XEND session_ID
XRECOVER session_ID
```

The XRECOVER command will re-label the target volumes with the source volume's *volser*. If the source volumes are still online to the System Data Mover system, the target volumes will go offline.

Now you can restart your systems using the new volumes.

For more information about XRC see *DFSMS/MVS Remote Copy Guide and Reference,* SC35-0169.

9.3 Moving data in z/VM and VSE/ESA

When considering the technique to move the data, z/VM and VSE/ESA offer alternatives.

9.3.0.1 z/VM commands and utility programs

Data migration can be greatly simplified by using DFSMS/VM which is a no-charge feature of VM/ESA. This lets you manage and control the use of VM disk space. It will also move mini disks from one media to another.

Using DIRMAINT also provides tools that will manage the movement of CMS mini disks from one media to another.

DFSMS/VM contains services which include a data mover, an auto-mated move process, and an interactive user interface.

DASD Dump Restore (DDR) is a service utility shipped with VM that you can use to dump data from disk to tape, restore data from tape to disk, and copy data between like disk drive volumes. You cannot use DDR to copy data between disk devices with different track formats.

CMDISK is a DIRMAINT command you can use to move mini disks from any device type supported by VM to any other type.

COPYFILE is a CMS command you can use to copy files or mini disks between devices with the same or different track formats.

SPTAPE is a CP command you can use to dump spool files to tape and to load them from tape to disk.

9.3.0.2 VSE/ESA commands and utilities

You can use several dialogs in the VSE Interactive Interface to set up the jobs to move data. You can reorganize your data and eliminate space fragmentation by using the backup/restore technique. Plan to use the following Backup/Restore dialogs:

- Export and import VSAM files
- Back up and restore VSAM files
- Back up and restore ICCF libraries
- Back up and restore the system history file
- Back up and restore the system residence library
- Create a loadable tape with the system residence library and system history file ready to restore

VSE/FASTCOPY can be used to move volumes and files between devices with identical track formats. Also, VSE/DITTO can be used to copy files.

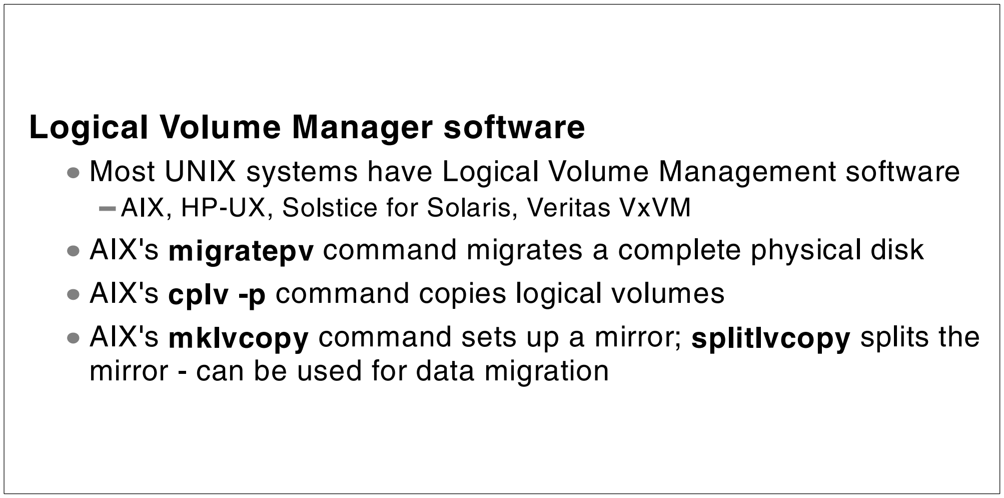
VSE/POWER commands can be used to transfer the SPOOL queue from one device to another. VSE/VSAM can move any VSAM dataset using either REPRO or EXPORT/IMPORT functions.

9.4 Data migration in UNIX environments

No special tools or methods are required for moving data to Enterprise Storage Server disks. The migration of data is done using standard host operating system commands. The UNIX hosts see the ESS logical devices (or logical volumes) just like normal physical SCSI disks.

Before you can put any data on an ESS, you first have to define the logical FB volumes in ESS using the ESS Specialist. To be able to do so, a Web browser is required, as well as an Ethernet connection to the ESS.

9.4.1 Migration methods



Logical Volume Manager software

- Most UNIX systems have Logical Volume Management software
  - AIX, HP-UX, Solstice for Solaris, Veritas VxVM
- AIX's **migratepv** command migrates a complete physical disk
- AIX's **cplv -p** command copies logical volumes
- AIX's **mklvcopy** command sets up a mirror; **splitlvcopy** splits the mirror - can be used for data migration

Figure 150. UNIX data migration methods

For UNIX hosts, there are a number of methods of copying or moving data from one disk to another (see *IBM Enterprise Storage Server Introduction and Planning Guide,* GC26-7294 for detailed discussion on the migration options). Some common migration methods are:

9.4.1.1 Volume management software

Most UNIX systems provide specific tools for the movement of large amounts of data. These tools can directly control the disks attached to the system. AIX’s Logical Volume Manager (LVM) is an example for such a tool. Logical Volume management software is available for most of the UNIX systems, like HP-UX, Solstice from Sun Microsystems for Solaris, and Veritas Volume Manager (VxVM) from Solaris. The LVM provides another layer of storage. It provides logical volumes that consist of physical partitions spread over several physical disks.

The AIX LVM provides a `migratepv` command to migrate complete physical volume data from one disk to another.

The AIX LVM also provides a command (`cplv`) to migrate logical volumes to new logical volumes, created on an ESS, for example. Do not be confused by the term logical volume as it is used in UNIX and the term logical volume used in the ESS documentation for a logical disk, which is actually seen by the UNIX operating system as a physical disk.

One of the facilities of the AIX LVM is RAID 1 data mirroring in software. This facilitates data movement to new disks. You can use the `mklvcopy` command to set up a mirror of the whole logical volume onto another logical volume, defined on logical disks (we prefer this term here instead of logical volume) on an ESS. Once the synchronization of the copy is complete, the mirror can be split up by the `splitlvcopy` command.

9.4.1.2 Standard UNIX commands for data migration

If you do not have a Logical Volume Manager, you can use standard UNIX commands to copy or migrate your data onto an ESS.

You can do a direct copy with the cpio -p command. The cpio command is used for archiving and copying data. The -p option allows data to be copied between file systems without the creation of an intermediate archive. For a copy operation, your host must have access to the old disks and the new disks on an ESS. This procedure does not require application downtime.

The backup (in AIX) or dump (on other UNIX systems) and restore commands are commonly used to archive and restore data. They do not support a direct disk-to-disk copy operation, but require an intermediate device such as a tape drive or a spare disk to hold the archive created by the backup command.

There are other UNIX commands such as the `tar` command that also provide archival facilities that can be used for data migration. These commands require an intermediate device to hold the archive before you can restore it onto an ESS.

For more information about the use of these commands see, for example, *AIX Storage Management,* GG24-4484*.*

9.4.2 7133 attachment

Only 7133 Models 020 and D40 disks can be incorporated to the ESS. The 7133 disks have to be installed in a new 2105-100 frame. The 2105-100 can only be attached to loops which have been reserved using the Reserved Loop Feature (FC9904). Remember that this is mutually exclusive with the Expansion Enclosure.

There are unique considerations for customers with existing 7133 disk subsystems who want to use their 7133 disks with an ESS. If data in the 7133 disks are still important, then the data must be saved to tape or cartridge. This is because the 7133 disks need to be reformatted before it can be used on the ESS (refer to Section 2.6.3, “7133-D40 drawers” on page 25 and Section 4.5, “Mixing with 2105-100 racks” on page 96 for detailed considerations for attaching existing 7133 Serial Storage Disks).

9.5 Migrating from SCSI to Fibre Channel

The early Enterprise Storage Servers were shipped with either SCSI or ESCON host adapters only. This was before native Fibre Channel attachment became available. If you still have data on your ESS that is been access thru SCSI attachment, and you are currently planning to swap to Fibre Channel attachment, this section will help you start planning.

There are many possible ESS environments, and each environment requires a different detailed procedure for migration. This section presents the general procedure for migrating from SCSI to Fibre Channel. Note that migration from SCSI to Fibre Channel is generally a disruptive procedure. You can do a non-disruptive upgrades also.

If you are planning to migrate from SCSI to Fibre Channel, then we recommend you to refer the following material:

- The white paper, *ESS Fibre Channel Scenarios*, that you can get from the Internet at `http://www.storage.ibm.com/hardsoft/products/ess/support/essfcmig.pdf`
- *Implementing Fibre Channel Attachment on the ESS,* SG24-6113

9.5.1 Migration procedures

Enumerated below are the general procedures for migrating from SCSI to Fibre:

- First of all, you need to make both host and ESS ready for Fibre Channel attachment. Apply all software requirements to the host system and upgrade the microcode of the ESS as required.
- Install the Fibre Channel adapters on the ESS. If not yet installed, you must also install the Fibre Channel adapters on the host(s) and have the adapters recognized by the host(s).
- Perform all the operating system tasks to free the disks. A sample of this procedure is to unmount the file, vary off the disks and export the volume group.
- Using the IBM TotalStorage ESS Specialist, unassign the disks. This makes the disks inaccessible to the server. However, the disks are still configured inside the ESS and the data in the disks is still intact.
- Using the ESS Specialist, configure the Fibre Channel adapters using the actual *World Wide Port Name* (WWPN) from the hosts Fibre Channel adapters.
- If multiple Fibre Channel hosts adapters are being connected to the ESS, the second and subsequent connections must be added as *new* Fibre Channel hosts using the actual World Wide Port Name from each of the hosts Fibre Channel adapters.
- Using the ESS Specialist, manually assign the volumes to the *new* hosts.
- Restart the server and make the volumes available to the server. This can be done by mounting the files.

9.6 Migrating from SDG to Fibre Channel

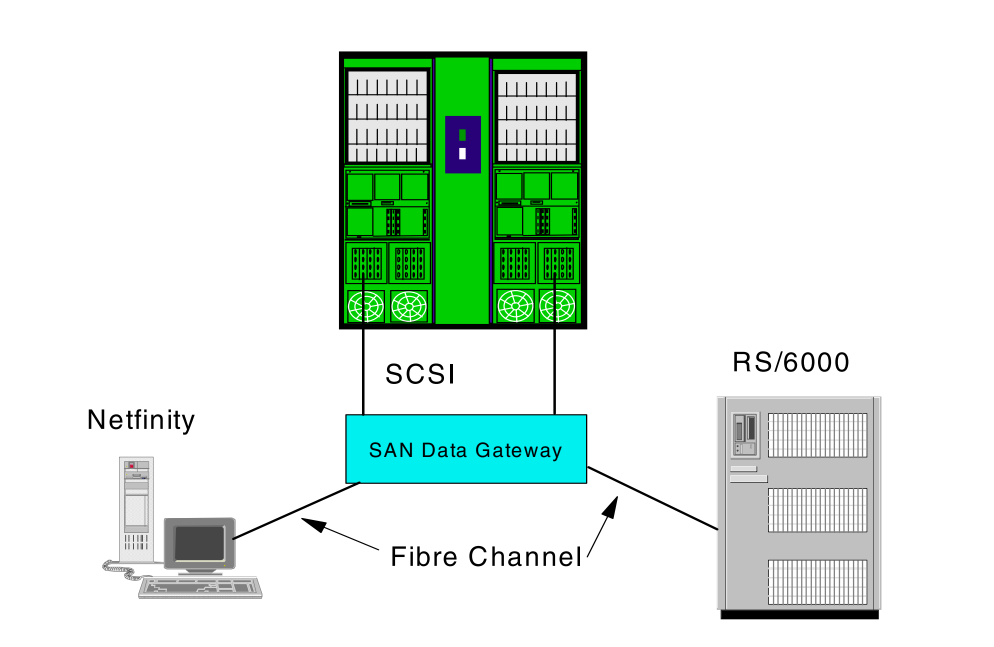


Figure 151. ESS and SAN Data Gateway setup

Figure 151 shows a SAN Data Gateway (SDG) setup that is to be migrated to native Fibre Channel attachment. With the availability of the native Fibre Channel attachment on the ESS, the need to migrate to end-to-end Fibre Connection becomes inevitable. Listed below are the general steps involved in migrating volumes from SDG to native Fibre Channel.

1. Define the new Fibre Channel host adapter to the ESS using the ESS Specialist.
2. Unassign the volumes from the SCSI host (different operating systems have different procedures in ‘unassigning’ the volumes).
3. Assign the volumes to the newly defined host. In practice, this means assigning the volumes to the WWPN of the host adapter.

9.7 Migrating from ESCON to FICON

FICON is supported on the Enterprise Storage Server models F10 and F20. Migration to FICON is by one of the following procedures:

- Upgrade from an ESS model E to model F with FICON adapters (the upgrade from model E to model F is disruptive).
- Upgrade an existing ESS model F to FICON.

If the current ESS is an F model with ESCON adapters, the migration to a FICON configuration is straightforward. The migration steps are described in this section.

Note: Keep in mind that although ESCON and FICON paths can co-exist in the same path group (that is, the set of paths to one operating system image) on the ESS, the intermixed path group configuration is only supported for the duration of the migration. The ESS F model fully supports an intermixed configuration of ESCON and FICON host adapters; the limitation applies only to the intermix of ESCON and FICON paths in a path group, that is, intermixed paths to the same operating system image.

A sample ESCON configuration is shown in Figure 152. Each image on each of the four CECs has eight ESCON channel paths to each ESS *logical control unit* (LCU). Four paths are cabled through ESCON Director #2 and four paths are cabled through ESCON Director #3.

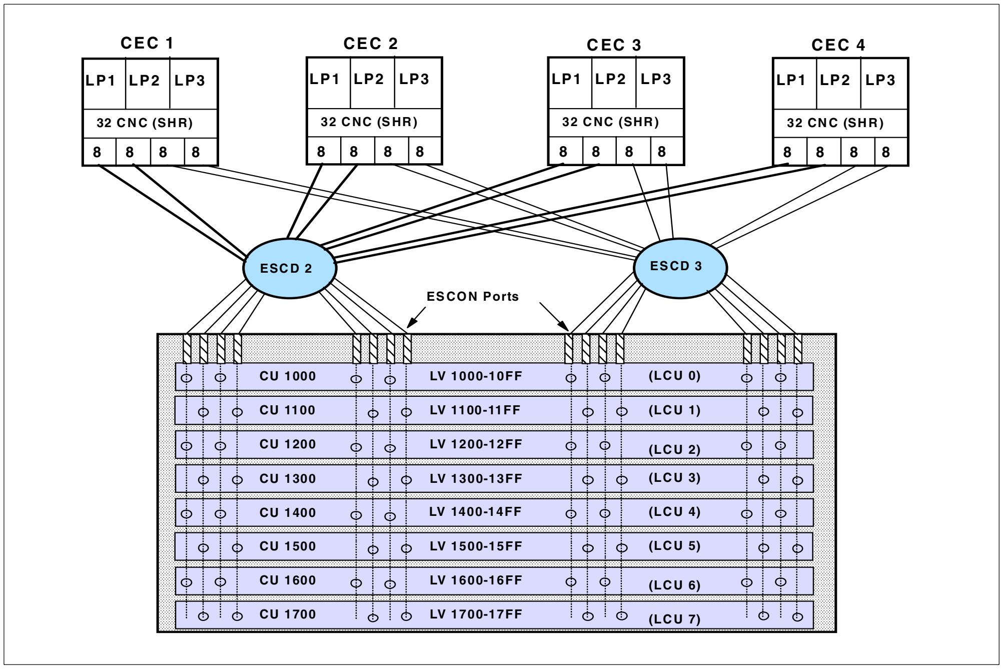


Figure 152. ESS configuration with ESCON adapters

If you plan to install four FICON channels per CEC and four FICON host adapters on the ESS, then the interim configuration is shown in Figure 153. In the interim configuration, the same images can access the ESS devices over a combination of both ESCON and FICON paths. This configuration is only supported for the duration of the migration to FICON.

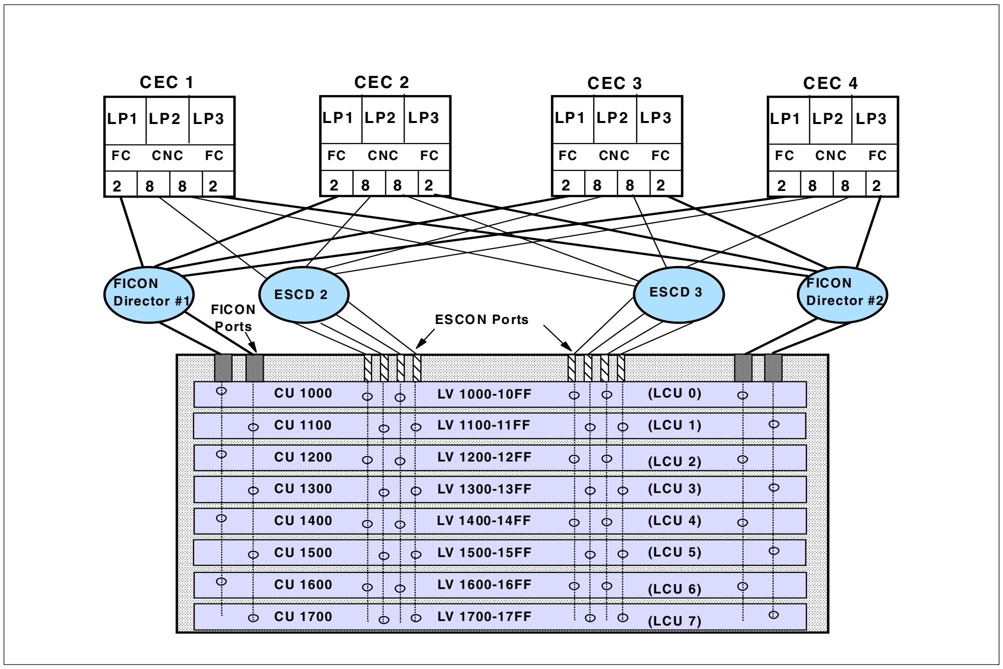


Figure 153. Interim configuration with ESCON and FICON intermix

The following steps are needed:

- Upgrade the software to levels that support FICON (see Section 7.7, "FICON support" on page 243).
- Install the FICON channel on the CECs. This is a non-disruptive procedure.
- Install FICON host adapters on the ESS, and upgrade the microcode to the required level (your IBM SSR will do this).
- Install FICON Directors. Note that the FICON channels could be connected point-to-point to the ESS, but this would require more ESS adapters.
- Perform a dynamic I/O re-configuration change to add the FICON channel paths to the ESS control unit definition, while keeping the devices and ESCON paths online.

The final configuration is shown in Figure 154.

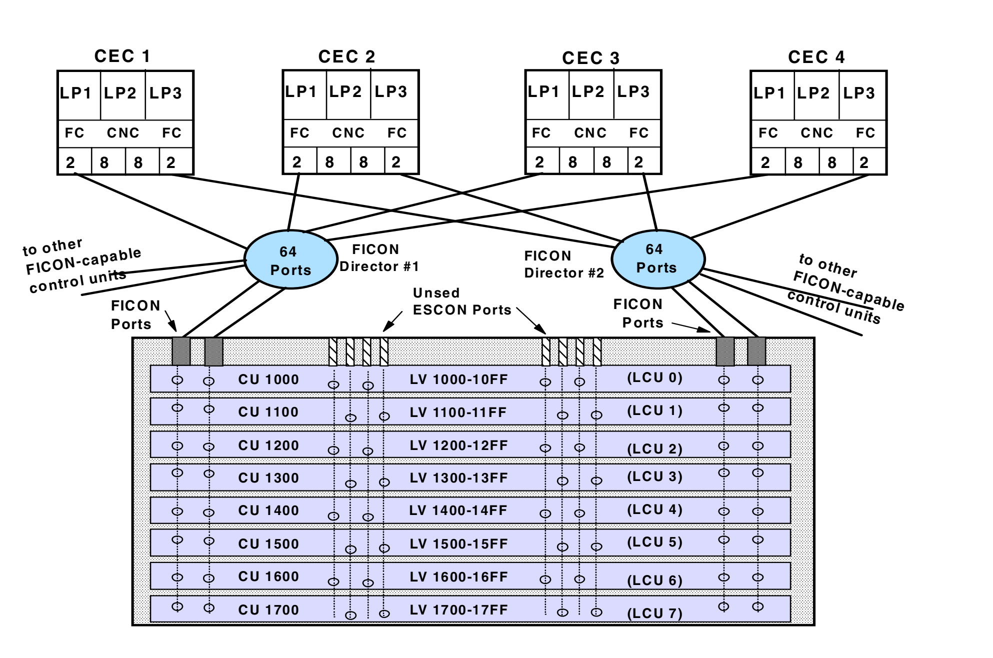


Figure 154. Target FICON ESS migration configuration

Note that the intermixed ESCON and FICON channel paths in the same path group are only supported to facilitate a non-disruptive migration, and should not be used for any extended length of time. Re-connections for devices in an intermixed configuration are not optimal from a performance perspective and it is strongly recommended that the customer move from an intermixed configuration as soon as possible.

9.8 IBM migration services

In several countries IBM offers migration services for different environments. Check with your IBM Sales Representative, or contact your local IBM Global Services (IGS) representatives for additional information. You may also get information from the Internet at `http://www.ibm.com/ibmlink` (select your geographic area and the look for **ServOffer** under **InfoLink**).

Appendix A. IBM TotalStorage ESS Expert

The IBM TotalStorage Enterprise Storage Server Expert (ESS Expert) gathers and presents information that can significantly help you manage the performance, asset, and capacity of your Enterprise Storage Server.

For more information on the ESS Expert you can refer to *IBM StorWatch Expert Hands-On Usage Guide,* SG24-6102 and visit:

```
http://www.ibm.com/storage/software/storwatch/ess/
```

A.1 Asset management

When storage administrators have to manage multiple storage systems, it can be time consuming to track all of the storage system names, microcode levels, model numbers, and trends in growth. The asset management capabilities of ESS Expert provide the capability to:

- Discover (in the enterprise network) all of the ESS storage systems, and identify them by serial number, name, and model number
- Identify the number of clusters and expansion features on each
- Track the microcode level of each cluster

This information can save administrators time in keeping track of their storage system hardware environment.

A.2 Capacity management

ESS Expert provides information that storage administrators can use to manage capacity. Some of the information that ESS Expert provides is:

- Storage capacity including:
  - Storage that is assigned to application server hosts
  - Storage that is free space
- Capacity assigned to each SCSI and Fiber Channel attached application server and capacity shared with other SCSI and Fiber Channel attached application servers. ESS Expert also provides the names and types of each application server host that can access the Enterprise Storage Server
- A view of volumes per host. This view lists the volumes accessible to a particular SCSI or Fiber Channel attached host for each Enterprise Storage Server.
- A volume capability that provides information about a particular volume. This function also identifies all the application SCSI and Fiber Channel attached server hosts that can access it
- Trends and projections of total, assigned, and free space over time for each Enterprise Storage Server.
- Trends in growth.

The capacity tracking and reporting capabilities of ESS Expert help increase the storage administrators productivity, identify trends over time, and help identify when additional space must be added to the Enterprise Storage Server, or that data must be moved off to free up space for higher priority files. This can help administrators be proactive and avoid application outages due to out of space conditions.

A.3 Performance management

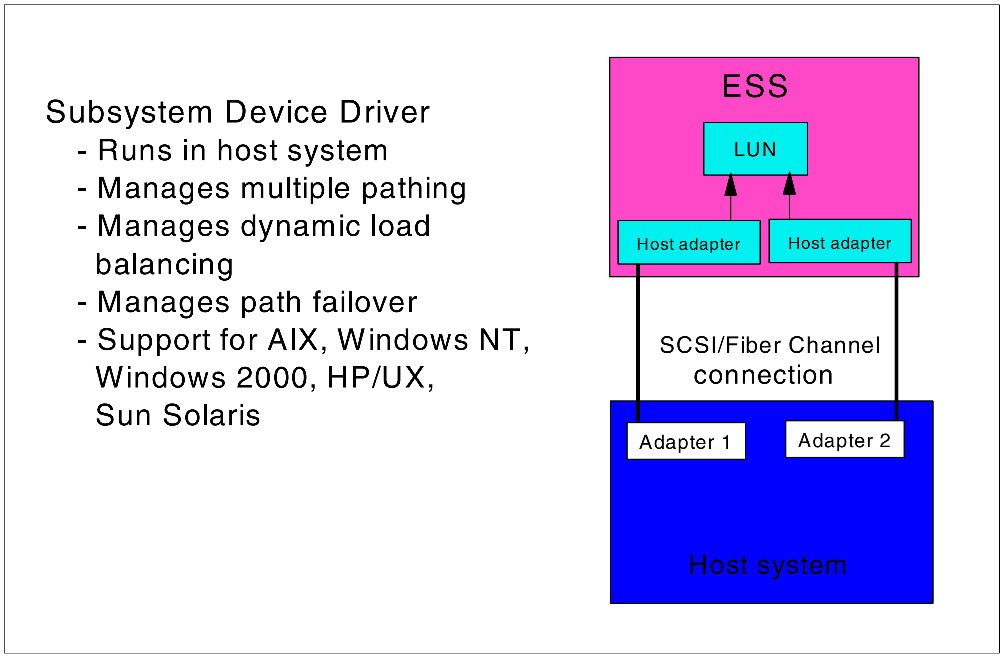
The ESS Expert gathers performance information from the Enterprise Storage Server and stores it in a relational database. Administrators can generate and view reports of performance information to help them make informed decisions about volume placement and capacity planning, as well as identifying ways to improve Enterprise Storage Server performance. The kind of performance information that is gathered and presented is:

- Disk utilization
- Number of disk lower interface I/O requests for each array, and for each adapter (the internal adapter that provides I/O to the physical disks)
- Read and write cache hit ratio
- Cache to disk operations (stage/destage)
- Disk lower interface read and write response time

Storage administrators can use this information to determine ways to improve ESS performance, make decisions about where to allocate new space, and identify time periods of heavy usage.

For description of the ESS Expert reports, how to get them and interpret the information they provide, you can refer to *IBM StorWatch Expert Hands-On Usage Guide,* SG24-6102.

Appendix B. Subsystem Device Driver (SSD)



Subsystem Device Driver

- Runs in host system
- Manages multiple pathing
- Manages dynamic load balancing
- Manages path failover
- Support for AIX, Windows NT, Windows 2000, HP/UX, Sun Solaris

Figure 155. Subsystem Device Driver (SDD)

The IBM Subsystem Device Driver (SDD) software, is a host resident software that manages redundant connections between the host server and the Enterprise Storage Server, providing enhanced performance and data availability.

SDD provides ESS users in Windows NT and 2000, AIX, HP/UX, and Sun Solaris environments with:

- An enhanced data availability capability for customers that have more than one path from a system server to the ESS. It eliminates a potential single point of failure by automatically rerouting I/O operations when a path failure occurs.
- Load balancing between the paths when there is more than one path from a system server to the Enterprise Storage Server. This may eliminate I/O bottlenecks that occur when many I/O operations are directed to common devices via the same I/O path.

The Subsystem Device Driver, (formerly the Data Path Optimizer) may operate under different modes/configurations:

Concurrent data access mode - A system configuration where simultaneous access to data on common LUNs by more than one host, is controlled by system application software such as Oracle Parallel Server, or file access software that has the ability to deal with address conflicts. The LUN is not involved in access resolution.

Non concurrent data access mode - A system configuration where there is no inherent system software control of simultaneous access to the data on a common LUN by more than one host. Therefore, access conflicts must be controlled at the LUN level by a hardware locking facility such as SCSI Reserve/ Release.

IBM Subsystem Device Driver is required on the supported system servers that are attached to the ESS. Some operating systems and file systems natively provide similar benefits, for example, z/OS, OS/400, NUMA-Q Dynix, and HP/UX.

It is important to note that the IBM Subsystem Device Driver does not support boot from a Subsystem Device Driver pseudo device. Also, the SDD does not support placing a system paging file in a SDD pseudo device.

For more information you can refer to the *IBM Subsystem Device Driver User's Guide* (this publication is available only in the internet at `http://www.ibm.com/storage/support/techsup/swtechsup.nsf/support/sddupdates).`

Load balancing

SDD automatically adjust data routing for optimum performance. Multi-path load balancing of data flow prevents a single path from becoming overloaded, causing input/output congestion that occurs when many I/O operations are directed to common devices along the same input/output path. Normally, selection is performed on a global rotating basis; however, the same path is used when two sequential write operations are detected.

Path fail-over and online recovery

SDD automatically and non-disruptively can redirect data to an alternate data path. In most cases, host servers are configured with multiple host adapters with either SCSI or Fibre Channel connection to an ESS that in turn would provide internal component redundancy. With dual clusters and multiple host adapters, the ESS provides more flexibility in the number of input/output paths that are available.

When a path failure occurs, the IBM SDD automatically reroutes the I/O operations from the failed path to the other remaining paths. This eliminates the possibility of a data path being a single point of failure.

Appendix C. Split mirror backup

This section refers to z/OS, DB2, SAP R/3 and the Enterprise Storage Server addressing database availability.

ESS advanced functions such as FlashCopy and Peer-to-Peer Remote Copy are synergistically tied with the software layer to provide solutions addressing data availability and disaster recoverability.

This appendix gives an overview of how you can protect your application and keep your data available in a disaster situation. This overview is based on a white paper from SAP AG. that is available at `http://www.storage.ibm.com/hardsoft/products/sap/smsm200.pdf`.

C.1 Log write suspend

Before reviewing the implementation of the solution let us understand the DB2 `suspend` command.

This is a new function that allows controlled suspension of the application writes. It is implemented in DB2 UDB for OS/390 version 6.1 and version 7 allowing a new command function — `set log suspend`.

The command forces out log buffers, updates the high written RBA in the BSDS, and sets a latch that prevents writes to the log and to application data. This creates a point of consistency which allows a completely consistent backup copy.

The actions taken when the operator executes the `set log suspend` and `resume` commands are:

SUSPEND log writes command

```
DB2 5.1 usermod: SET ARCIVE TIME(0)
DB2 6.1 & 7: SET LOG SUSPEND
   actions:
       Set a log-write latch
       Force out LOG buffers
       Update high-written RBA in the BSDS
       Echo back high-written RBA in a response message
   after successful execution:
       Application READ I/O is possible
       Application WRITE I/O is not  possible
       Resume write must be enforced by operator command
```

RESUME log writes command

```
DB2 5.1 usermod: SET ARCIVE TIME(1)
DB2 6.1 & 7: SET LOG RESUME
    actions:
       Release log-write latch
       Echo back a response message
   after sucessful execution:
       Application READ I/O is possible
       Application WRITE I/O is possible
```

C.2 Split-mirror backup

This is the backup and recovery high availability solution which uses the ESS advanced copy functions FlashCopy and PPRC, along with the backup copies of the logs, the BSDS and ICF catalog.

Please refer to Figure 156 while reading the implementation process.

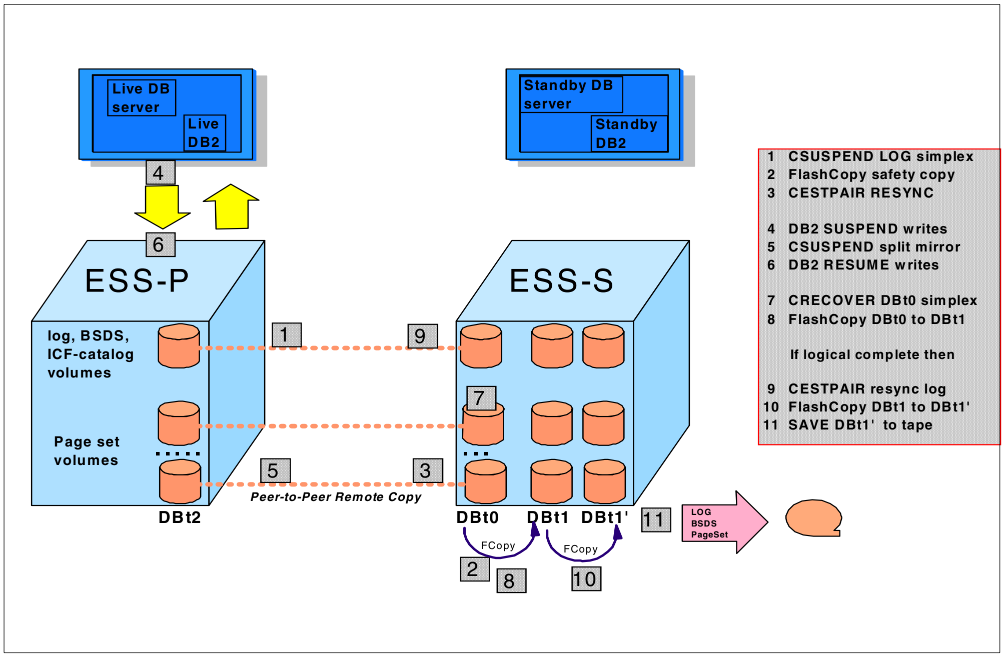


Figure 156. ESS split mirror backup

Initial situation

The live system is in normal read/write operation. Only the LOG volumes (DB2 logs, BSDS, ICFCAT and user catalog) are synchronously mirrored with PPRC. After the last re-synchronization (backup) all other PPRC volume pairs were suspended and recovered, and are now accessible as usual simplex volumes.

Split mirror backup process

On the primary site we `Csuspend` the LOG volume pairs. This allows us to operate on the primary and secondary side independently without affecting each other.

On the secondary site we make the LOG volumes simplex (`Crecover`) and vary them together with all other DBt0 and DBt1 volumes online (step 1).

The next step is to FlashCopy the DBt0 volumes to the DBt1 volumes (step 2). This safety copy will help us to recover from a disaster that may happen during the re-synchronization of our mirror. A user or application logical error during the re-synchronization would immediately make our mirror also inconsistent. The safety copy DBt1 will enable us to recover the database with a DB2 conditional restart on the secondary side, while the live system is available for error analysis.

At this point in time we are ready for the re-synchronization of the mirror.

We again establish all PPRC pairs in *resync* mode, which copies only those tracks updated during the period of suspension (step 3).

As soon as this process is 99% finished, we cause DB2 to suspend all writes (step 4) coming from the R/3 application. This DB2 function throttles down application writes, but no R/3 application process will recognize this — they will only slow down as long as we cause DB2 to resume write operations.

Once the remaining 1% of updated tracks is also synchronized, we suspend all PPRC pairs (step 5). Immediately after the mirror is split, we cause DB2 to resume all write operations (step 6). Note that only in the short period between step 4 and step 6 application writes were slowed down.

The live system is now back to normal read/write operation.

On the secondary side we now make all DBt0 volumes *simplex* and vary them together with all other DBt0 and DBt1 volumes online (step 7).

Because we need the DBt0 volumes as fast as possible back to establish the PPRC pair between the LOG volumes, we copy with FlashCopy all DBt0 volumes to DBt1 (step 8).

As soon as the FlashCopy relationship is established — *logical complete* — FlashCopy allows read/write access although the tracks are not yet physically copied.

Before a task can update a track on the source that has not yet been copied, FlashCopy copies the track to the target volume. The following reads to this *old* track on the target volume will be satisfied from the target volume. After some time, all tracks will have been copied to the target volume, and the FlashCopy relationship will end.

Immediately after the *logical complete* message we again establish the PPRC pairs in *resync* mode (step 9) for the LOG volumes.

Our environment is now back to normal processing.

In our test scenario, we assumed that the standby system also has to be available as soon as possible. Therefore after the first FlashCopy is physically completed we start a second FlashCopy from DBt1 to DBt1' (step 10). We have to wait for the physical completion of the FlashCopy started in step 8, because a source and target volume can be involved in only one FlashCopy relationship at a time.

As soon as this second FlashCopy is logically completed, we restart our standby system and, after the copy is physically completed, we move the contents of the DBt1' volumes to tape (step 11).

Our Split-Mirror backup process is finished.

Appendix D. Implementing disaster recovery

In this appendix we are going to overview several topics based on Copy Services implementation. All of them are related to implementing remote copy technologies for a disaster / recovery solution.

- Remote copy technologies positioning
- Considerations on implementing a disaster/recovery solution
- Distance factor analysis for remote copy
- Practicing combination of copy services.

Please refer to the following redbooks for further details:

- *Implementing ESS Copy Services on UNIX and Windows NT/2000, SG24-5757*
- *Implementing ESS Copy Services on S/390,* SG24-5680
- *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220

D.1 Preliminary questions

Suppose you want to implement remote copy technologies to address a disaster / recovery project. Here are some good preliminary questions to ask yourself?

1. Suppose you have chosen a mirroring technology, what are the requirements to have permanently the shortest possible recovery insurance in case of disaster?
2. How could you get an immediate return of invest on your solution?
3. Suppose you have to progress to this technology, what trade-offs would you choose in a stepped implementation approach?
4. Suppose this technology is so fast moving, that you could have to continuously invest on, how would you agree to support this challenge?
5. Suppose you finally discover that remote copy is just a new kind of application, but with very unusual new requirements, like continuous management, how would you deal with this new way of thinking?

These questions, and many others, should have been asked and the answers studied carefully before any choice of technology, and of course any implementation. IBM Global Services and IBM Storage Server Group Advanced Technical Support people can help you in that operation.

D.2 Remote copy technologies positioning

Remote copy is a storage based disaster recovery and workload migration solution that provides the capability to copy data in real time from a primary location to a remote secondary location. The Remote Copy architecture is an open architecture that is available to all vendors and the solution is supported for all vendors who implement the architecture. Remote copy offers two options for disaster recovery and workload migration needs: Extended Remote Copy (XRC) and Peer-to-Peer Remote Copy (PPRC).

D.2.1 XRC solution profile

XRC operates with two systems: a primary application system at one location and a recovery system at another location. These locations may physically reside in the same site or be completely remote from each other. The remote copy function provides a real time copy of all primary application data at a record level to the secondary site. The primary storage control must be an XRC-capable controller. The secondary storage control can be any supported storage subsystem. XRC has a host component that may be located at the primary site, the secondary site, or an intermediate site. In a channel extender environment most customers will place the data mover at the secondary site. The rationale is that it is more efficient for the data mover to read remotely and write locally. In addition, since a host processor is required at the secondary site to handle recovery from a disaster, the data mover can take advantage of the host capability without impacting the host load at the primary site.

In general, XRC is used in configurations where disaster/recovery support is to have little or no impact on production application program response time and where disaster/recovery is required over an extended distance (20 Km. or more). For large configurations, it is recommended that the GDPS/XRC or RCMF/XRC functions be used to assist in the management and control of such environments.The rationale is that the SDM does not see any production site or primary disk problems, other than the complete failure of a subsystem. If devices start to fail, then the SDM will not see that. Consequently a rolling disaster could roll for some time before the SDM will be aware of any problem. Therefore the GDPS/XRC “watchdog” is a very valuable function.

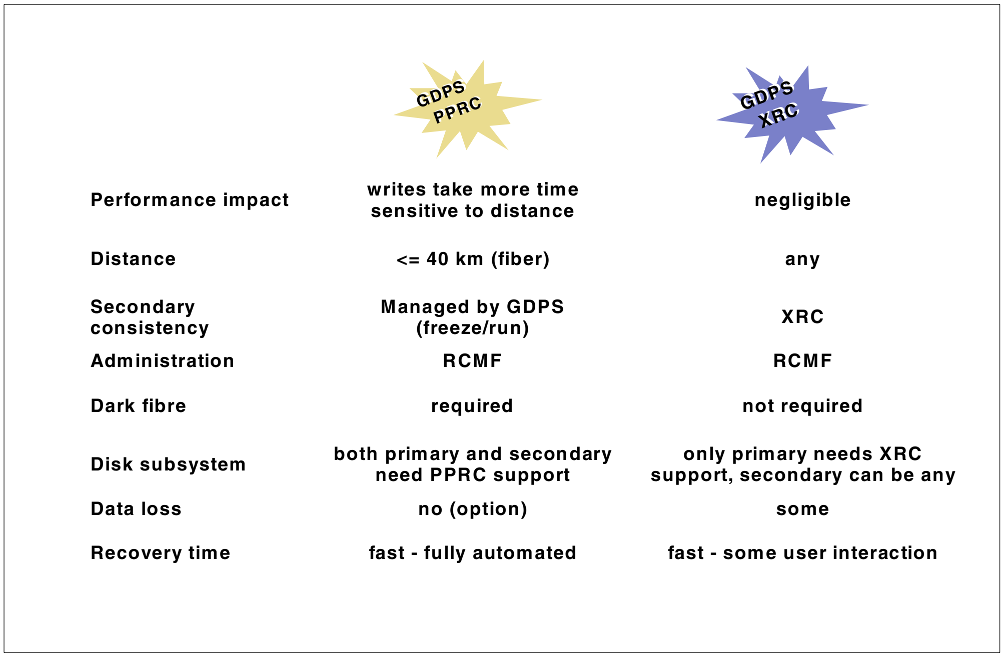
The software component of XRC, called the system data mover, resides in a z/OS host LPAR. The system data mover component periodically obtains application updates from the primary storage controller, gathers them into a time consistent collection of records referred to as a consistency group, writes the consistency group to a journal data set, and applies the updates to a set of target volumes. The consistency group provides a common recovery point for all volumes in the XRC session since recovery always occurs on a consistency group boundary. The consistency group allows recovery of multiple data sets across volume and storage control unit boundaries.

XRC is an asynchronous disaster/recovery solution that means that the secondary volumes may be a few seconds out-of-synch with updates to the primary volumes. In the event of a disaster, updates that are in the journal data sets will be applied to the secondary volume that minimizes the data loss by this solution. Recovery of large configurations is generally completed in five minutes or less and the data across all secondary volumes and storage controllers is consistent to the same consistency time. After the recovery, standard data base restart procedures allow the secondary system to resume operation with minimal data loss.

D.2.2 PPRC solution profile

The PPRC function is provided primarily by microcode support in both the primary and secondary storage controllers. Both the primary and secondary storage controller must be peers of each other (i.e.the same manufacturer). A primary storage controller volume is kept consistent with a secondary storage controller volume by ensuring that an application update is not considered complete until it has been written to both the primary and the secondary volumes. The updates are performed in a synchronous fashion to both which ensures that the data is consistent on both volumes. PPRC requires a host automation function such as GDPS/PPRC or RCMF/PPRC to manage consistency across multiple control units. In the event of a disaster, the automation function will freeze activity across all control units to force consistency and allow recovery to take place. PPRC is used in configurations for which moderate application response time impact is acceptable and the distance between the primary and secondary sites is reasonably short (less than 103 Km.). As the distance increases, the impact to application program response time increases.

D.2.3 Recovery environment profiles



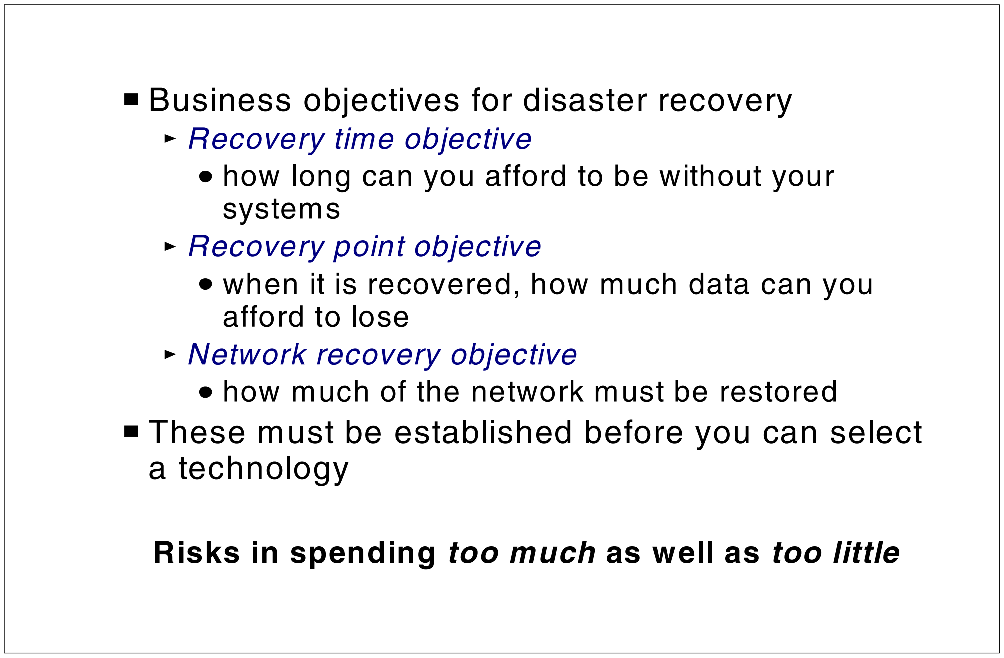
	GDPS PPRC	GDPS XRC
Performance impact	writes take more time sensitive to distance	negligible
Distance	<= 40 km (fiber)	any
Secondary consistency	Managed by GDPS (freeze/run)	XRC
Administration	RCMF	RCMF
Dark fibre	required	not required
Disk subsystem	both primary and secondary need PPRC support	only primary needs XRC support, secondary can be any
Data loss	no (option)	some
Recovery time	fast - fully automated	fast - some user interaction

Figure 157. Positioning GDPS/PPRC and GDPS/XRC

There are many reasons for having a consistent set of volumes at the secondary at a given point in time. If a disaster were to occur, it is critical that all secondary data is consistent at the time of the disaster. Both XRC and PPRC provide dynamically data consistency. But in case of disaster the secondary consistency is not automatic for PPRC, as it is for XRC: You need to do something. Critical YES keeps primary and secondary identical, but because it is a volume level attribute, you cannot prevent rolling disaster effects (logical contamination that makes life difficult when trying to get applications going again). Only the GDPS *Freeze* function gives you secondary consistency without any logical data contamination.

Many installations have the requirement to conduct a periodic disaster/recovery test. Some installations will create periodic backups of secondary volumes to be able to recover the data if a subsequent physical or logical failure were to occur.

D.3 Implementing a disaster/recovery solution



- Business objectives for disaster recovery
  - *Recovery time objective*
    - how long can you afford to be without your systems
  - *Recovery point objective*
    - when it is recovered, how much data can you afford to lose
  - *Network recovery objective*
    - how much of the network must be restored
- These must be established before you can select a technology

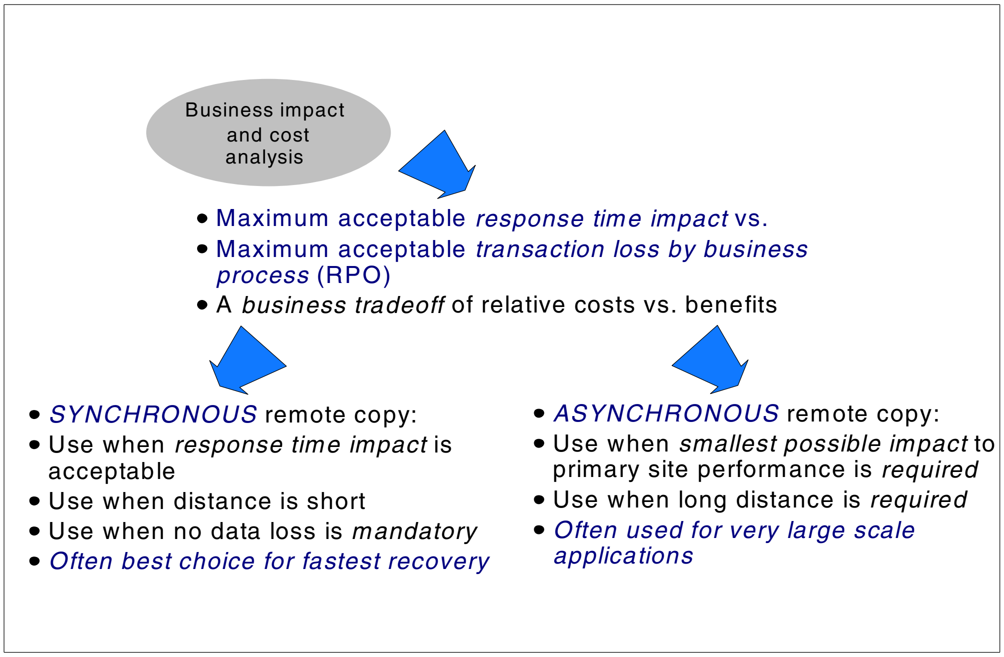
Risks in spending *too much* as well as *too little*

Figure 158. Business continuance objectives

In a disaster/recovery project you have first to determine the corporate strategic objectives as described in Figure 158. These three parameters condition your further strategies for both resource and technology determination: *recovery time objective* (RTO), *recovery point objective* (RPO), *network recovery objective* (NRO).

The next step is to choose a synchronous or asynchronous technology (see Figure 159). Both technologies offer consistency. But you have in any case to manage it.

The most commonly addressed factors are performance and price, But having consistent mirrored data, should any disaster occurs, has the greatest importance because it guarantees a simple restart with a RTO short lead time instead of a very long and hazardous recovery. The management of this consistency is accomplished by an effectively tested automation, which applies predefined policies in case of problem. This investment determines the effectiveness of your disaster/recovery solution and must be in relationship with the corporate objectives.



Business impact and cost analysis

- Maximum acceptable *response time impact* vs.
- Maximum acceptable *transaction loss by business process* (RPO)
- A *business tradeoff* of relative costs vs. benefits

- *SYNCHRONOUS* remote copy:
- Use when *response time impact* is acceptable
- Use when distance is short
- Use when no data loss is *mandatory*
- *Often best choice for fastest recovery*

- *ASYNCHRONOUS* remote copy:
- Use when *smallest possible impact* to primary site performance is *required*
- Use when long distance is *required*
- *Often used for very large scale applications*

Figure 159. Choosing a technology: synchronous versus asynchronous

Example

As an example, look at the following sections, which are extracted from an IBM Advanced Technical Support document about benchmarking a very large configuration at various distances, with specific exploitation requirement to established the feasibility of the solution, and its flexibility.The project has been successfully used to validate the hardware, software, network and system operational requirements to mirror 8000 volumes over a 1500 to 3000 mile distance utilizing the XRC solution.

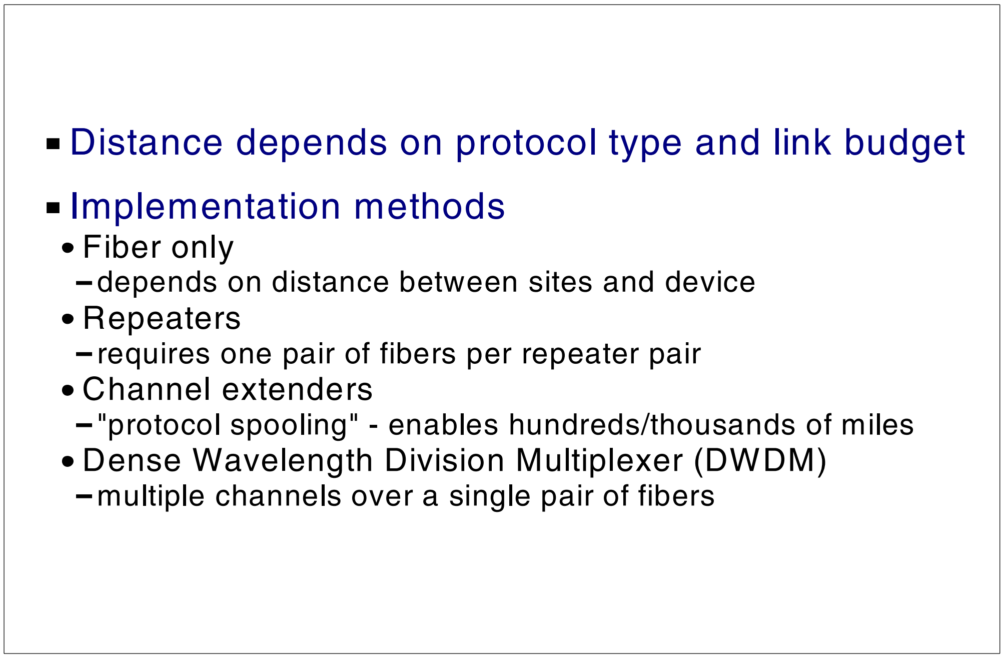
“The environment that was configured for this project used an approach that minimized the disaster/recovery exposure time while providing a backup of the data. This environment was one where XRC was used to copy data from the primary volumes to the secondary volumes, which was done at various distances. PPRC was used to copy data from the secondary volumes to a tertiary set of volumes that was done at local distance. The rationale for this configuration was based on the requirement that the disaster /recovery solution would have little or no impact on primary application I/O. This could only be achieved by using XRC. In addition, the environment would be one that would support sites located either near distance or at a distance up to 3000 miles apart. Again, this can only be achieved by using XRC. Use of PPRC for the local copy took advantage of the synchronous nature of PPRC that ensured no data loss while not causing any application performance impact.”

“At the secondary site, the requirement was that disaster/recovery readiness could be verified at any point in time with minimal impact on the primary disaster/recovery activity. To achieve this objective, PPRC was used in making a local copy of the secondary DASD. When a disaster/recovery test would be done, the XRC session would be *paused* to allow GDPS/PPRC to freeze the PPRC session. The XRC session would then be *resumed* with little or no impact seen by the application system, thus maintaining disaster/recovery readiness at all times.

Meanwhile, the tertiary copy of the data would be in a consistent state that would allow a disaster/recovery readiness test to be conducted.”

This project had very specific requirements, which have involved the chosen solution: no impact on application performance, and immediate access to a copy of the mirrored data for recovery tests. This access to a copy of mirrored data, which can be obtained by various technics with a minimal impact on mirroring, is one of the main return on invest many companies practice.

D.4 The distance factor



- Distance depends on protocol type and link budget
- Implementation methods
  - Fiber only
    - depends on distance between sites and device
  - Repeaters
    - requires one pair of fibers per repeater pair
  - Channel extenders
    - "protocol spooling" - enables hundreds/thousands of miles
  - Dense Wavelength Division Multiplexer (DWDM)
    - multiple channels over a single pair of fibers

Figure 160. Methods of distance extension

This section analyzes the distance parameter between the recovery and the production sites. Figure 160 shows various technologies. All of them have a very fast evolution and consequently, we recommend you refer to last level of information: a couple of months can bring dramatic changes in this domain.

D.4.1 Distance extension technics

Extended distance is an integral part of disaster recovery applications.

The maximum distance supported depends on the device, its protocol and the link budget which is expressed in decibel (dB) budget terms. We have described four categories of implementing distance solutions. Fiber only applies when the distance between two locations is less than 3 Km. for ESCON devices, 10 Km. for FICON. Repeaters such as IBM 9036 and the XDF feature for the 9032 ESCON Director can extend the distance but requires one pair of fibers for each channel. Channel extenders (example is boxes offered by CNT) enable distances at hundreds or thousands of miles, but each link is either T3 (45Mbps) or OC3 (155 Mbps). DWDM can provide extended distances of 50–70 Km. and provide the capability to multiplex multiple channels over a single pair of fibers.

The channel extenders can go at an unlimited distance. Figure 161 displays graphically the different distances based on protocol for the ESCON and FICON implementations (standard and RPQ). The GDPS implementations are limited by the Sysplex Timer (ETR) protocol.

Consider that signals need time to travel, and the longer the distance the more time needed for the signal to travel. The data droop effect makes ESCON practical only up to distances no more than 9 Km. For FICON there is no data rate droop effect even at the maximum supported distance of 103 Km.

.

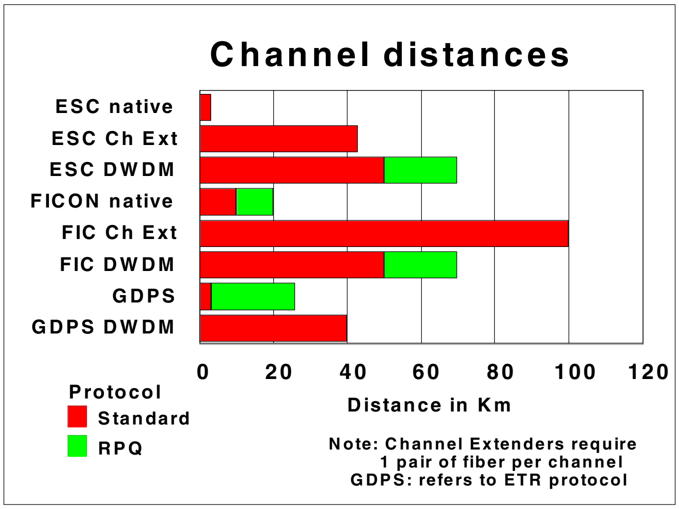
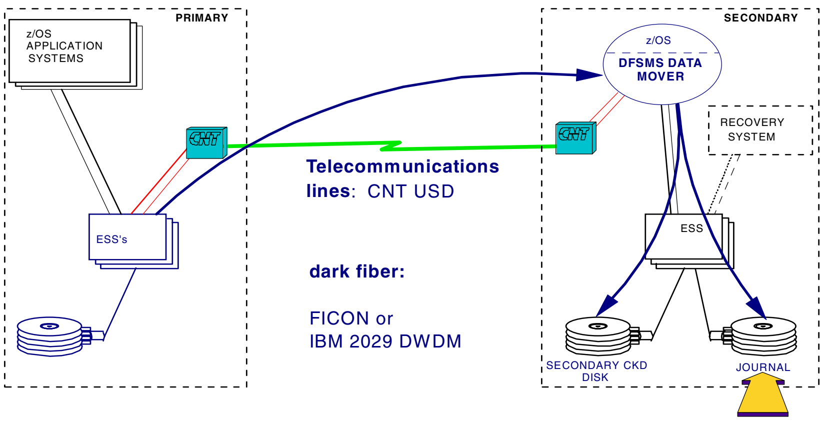


Figure 161. Distance options with fiber

Figure 162 shows the potentially unlimited distances for XRC solutions, when using CNT Channel Extenders with telecommunication lines.



XRC architecture provides cross-subsystem sequence integrity

Figure 162. IBM XRC telecommunication distance

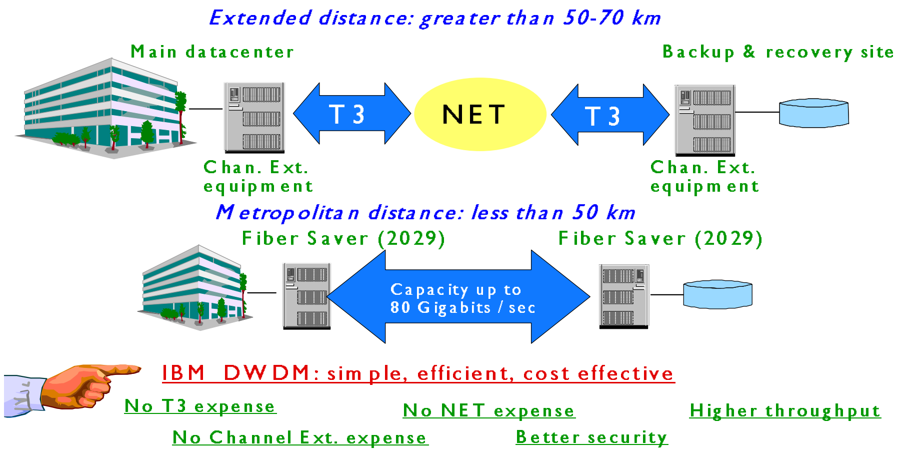


Figure 163. Disk mirroring (ESCON) — two solutions

Figure 163 compares two potential solutions, one for distances greater than 50 Km. (31 miles) and one for distances shorter than 50 Km.

One solution uses channel extension equipment and telecommunication lines with speeds up to 620 Mbps. In this case, T3 lines are shown which are only rated at 45 Mbps throughput. The channel extension equipment simulates ("fakes" out) the ESCON protocol to enable distances greater than the 43 Km. which is the supported distance for most ESCON devices. The key is one pair of extenders and one OC3 (155 Mbps) link for every channel being extended. This solution can be expensive plus the maximum channel speed is less than the ESCON channel speed of 200 Mbps.

The other solution depicts a Fiber Saver implementation that has a capacity of up to 80 Gigabits per second. This solution provides the capability to support up to 64 channels, all at rated speeds over two pair of fiber.

DWDM benefits include savings of channel extension equipment, T3 or OC3 line charges, and higher throughput since all channels are supported at rated speeds. Additionally, better security is available since the link between sites is dedicated and not shared.

D.4.2 Considerations on the distance factor

Before determining the distance between application and recovery centers it is useful to have a preview of some of the involved requirements.

New domains to manage

The distance brings new domains to manage. From that viewpoint, there are two classes of technologies:

- One deals with local management of recovery at native channel distance,
- The other requires channel extenders which can involve a network management.

Many switch directors, like ESCON Channel Directors (ESCD), have additive channel extension functions, and a specific device can have several functions. Consequently the comparison is sometimes uneasy. To differentiate technologies in the disaster/recovery environment, you have to consider that it involves a continuous process, that you have to manage all associated resources (links, devices, people), to address their outages, and eventually to deal with the security over public networks.

Speed of transfers on the link

The distance can determine the way you are going to use a chosen technology.

It is commonly admitted a propagation delay of 30 micro-seconds per kilometer. Over long distances, this overhead becomes noticeable. Moreover a pair of channel extenders brings an added overhead, which appears currently as a distance elongation of around 50 Km. So 600 Km require 19.5 milliseconds, which prohibits using synchronous technologies in update write mode.

Bandwidth

The bandwidth factor is tightly associated to distance. It has a cost. Its influence has a direct retro effect on the corporate objectives to be retained, and leads to making trade offs. But these trade-offs have to be periodically reviewed.

As soon distance exists between both sites, you should determine the required bandwidth to restart a full mirroring of the data at the most intensive production activity in order to evaluate properly your RPO. As the cost of this bandwidth increases with the distance, any capacity limitation trade off has an impact on this RPO. In case of any re synchronization, you increase the duration of the process, and consequently more data not yet synchronized can be lost in case of disaster when your recovery center is at longer distance.

Examples

The design of solutions has the same fast evolution than the technology. So do the behavior of people in making decisions about disaster/recovery.

- Some companies determine the distance between sites from their strategic requirements after a detailed analysis of the market at a specific time to fix a budget. This budget is invested in a bandwidth. Then they choose an *adaptive solution*, which maximizes their insurance coverage for a given investment. For this financement they adjust a reasonably short RTO (delay before restarting) with a medium RPO (amount of data they can lose without being out of business) provided with they trust the consistency of their data at the recovery center. This is the *point-in-time* remote copy solution we are going to illustrate for the open systems environment. They transmit periodical fixed checkpoints of their data captured by FlashCopy over medium to long distance by PPRC. The delay between successive transmission, which determines the RPO, depends directly of the quantity of data to be transmitted for a given bandwidth.
- Some other companies develop a solution based on combined Copy Services, the purpose of which is having no or minimal data loss in case of disaster. In the z/OS environments, the combination of local PPRC at short distance and of XRC in medium to very long distance is the other example we are going to show.

D.5 Combination of copy services

The following examples reflect the always moving balance between the current state of the available products, and the disaster / recovery demand: new functions can change important parameters considered when implementing the solutions.

Please note as a given solution is built up from several pieces (ESS Copy Services, hardware, links, software, automation, procedures, etc.), any change on one topic can interact with the others, and consequently requires a global evaluation.

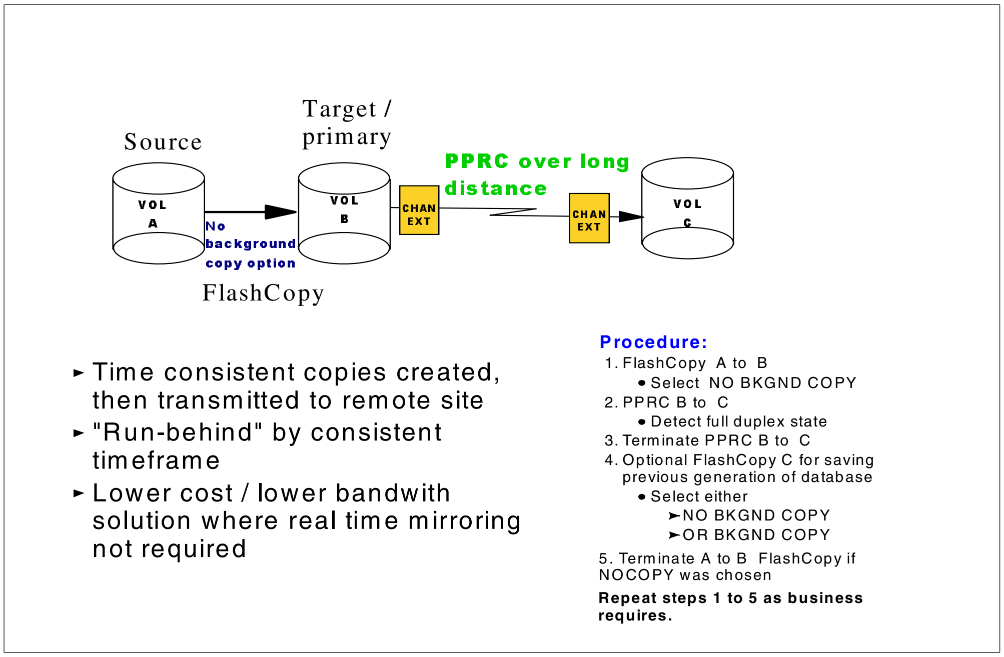
D.5.1 Point-in-time remote copy: FlashCopy + PPRC

Customers are realizing that, in some cases, the FlashCopy + PPRC approach may be a better solution than mirroring the entire database. The perception of benefits from continuously copying the entire database (thinking that there will be a slightly more in-synch database) does not always result in improved elapsed recovery, as continuous copying also creates circumstances in which database recovery at the secondary site might be required. FlashCopy + PPRC, because it always starts with a known consistent point in time copy, can assure a quick, predictable, consistent elapsed time database restart at the remote site.

Enterprise Storage Server open systems users can use ESS Copy Services — Peer-to-Peer Remote Copy and FlashCopy — to meet their remote copy requirements for long distance. In most cases, by combining the functions of ESS FlashCopy with ESS PPRC, customers can achieve nearly the same effect as continuous asynchronous long distance open systems copy. The FlashCopy +

PPRC method also minimizes telecom costs by transmitting database log volumes only (as opposed to mirroring the entire database). As soon as the log volume arrives at the remote site, the log is applied to a preexisting copy of the database.

Another rationale has been the effective motivation to develop this implementation: with a given bandwidth, transmitting files is more efficient. Consequently, as there are no instantaneous peak of write updates to deal with, you can mirror more applications with the same bandwidth.



- Time consistent copies created, then transmitted to remote site
- "Run-behind" by consistent timeframe
- Lower cost / lower bandwith solution where real time mirroring not required

Procedure:

1. FlashCopy A to B
   - Select NO BKGND COPY
2. PPRC B to C
   - Detect full duplex state
3. Terminate PPRC B to C
4. Optional FlashCopy C for saving previous generation of database
   - Select either
     - NO BKGND COPY
     - OR BKGND COPY

5. Terminate A to B FlashCopy if NOCOPY was chosen

Repeat steps 1 to 5 as business requires.

Figure 164. ESS PPRC + FlashCopy

FlashCopy + PPRC Method (Figure 164)

- Data is written to the primary ESS databases as normal.
- Using the standard log file switch capability on application database software, application log files/volumes are dynamically switched or temporarily quiesced.
- If the log volumes are dynamically switched, then any open system supported method (TCP/IP File Transfer Program, PPRC, or other) may be used to transport the log files to a remote site. The primary site continues running on the new log volumes without impact. Once those log volumes arrive at the secondary site, a shadow process there applies the logs to the secondary site database.
- If the production log volumes or databases are temporarily quiesced (application QUIESCE, or exploitation of the DB2/390 or DB2 UDB SET WRITE SUSPEND functions), you may FlashCopy these volumes to assure database integrity and coherency.
- If PPRC is used, Inrange 9801 Storage Network Subsystem can be used as the channel extender used to go to the secondary site.

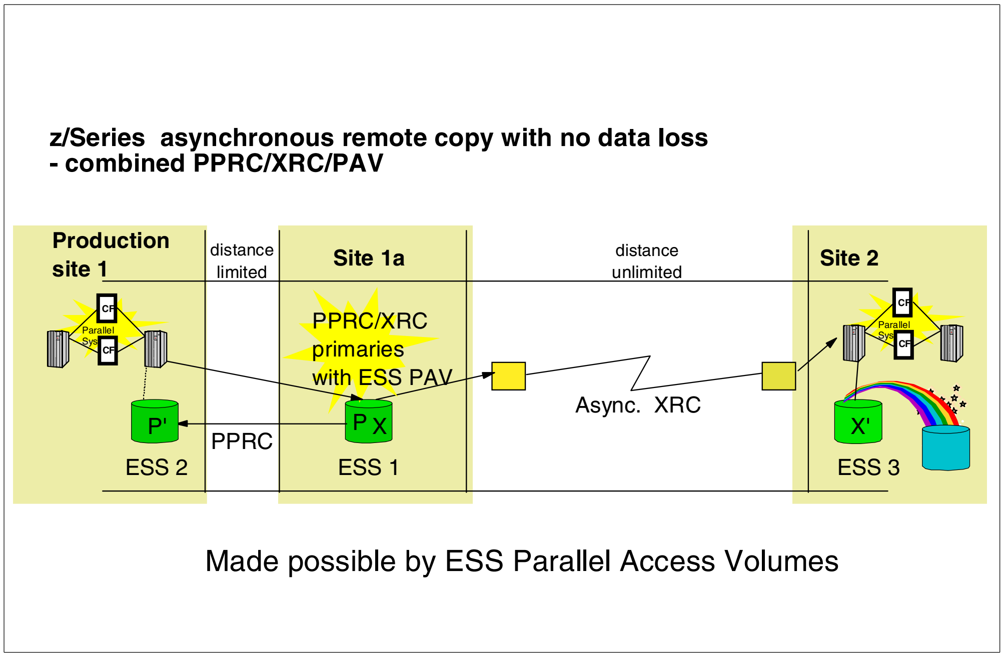
- If there is high-speed TCP/IP connection to the remote site (especially useful for very long distances), then standard TCP/IP File Transfer Program or equivalent may also be used.
- Once complete, this process is repeated at customer-desired intervals.

Advantages of FlashCopy + PPRC Method

- It can be cheaper and effective enough to send only the log volumes, as in many cases there may be no real benefit to duplicate sending the actual writes to the database. The log files contain everything that is needed for database update.
- Applying log file updates to a database always leaves the database in a known condition, as there are no partial updates or broken database writes due to remote disk mirroring that can cause time consuming database recovery.
- A switch to remote site is very expensive, given the fixed costs involved. Such costs are motivating customers to investigate other methods, such as FlashCopy + PPRC, for providing a better cost justified solution to business continuance.
- Databases all have the ability to dynamically and non-disruptively switch database log files.
- FlashCopy is very fast on a small number of (log) volumes. In any case faster than on a large number of volumes.
- Log volumes are much smaller than the entire database, hence much lower telecom costs are incurred. Moreover, by sending only the log volumes, write data is sent only once across a PPRC link, as opposed to a scenario in which the entire database is mirrored. Thus, customer cost is minimized.
- In most cases, customers find that they can achieve acceptable recovery levels for open systems data at lower telecom costs by using FlashCopy + PPRC. Alternatively, this method may be used to include much more applications for a given amount of telecom bandwidth.
- Finally, IBM DB2/390 and IBM DB2 UDB for open systems have a special modification that can support DB2 applications to initialize a FlashCopy without taking an application quiesce to assure data integrity. We recommend you to see SG24-6180, *DB2/390 V6 Technical Update* for a description of the DB2/390 *set write suspend* function in more detail. DB2 UDB V7 for UNIX/NT has also shipped the same functionality.

D.5.2 Intermediate site remote copy

The objective of this zSeries environment (see Figure 165) is fail over capability beyond ESCON distance without XRC data loss. The solution is based on an ESCON short distance PPRC between two close but independent sites combined to a medium or long distance XRC. Moreover, it requires automation investments. But it brings a significant added value to the long distance XRC technology, which maintains a consistent state of mirrored data in case of disaster at the price of a variable lag time from the primary, which generates potential data loss.



z/Series asynchronous remote copy with no data loss
- combined PPRC/XRC/PAV

Made possible by ESS Parallel Access Volumes

Figure 165. Remote copy with intermediate site in z/Series environment

Description of production and recovery centers

- Site 1 — is the primary production location and contains hosts and secondary PPRC volumes.
- Site 1a — is an intermediate site which contains the application volumes, which are primary for both PPRC and XRC. As primary disks are outside of site 1, XRC can retrieve latest updates in case site 1 is down
- Sites 1 and 1a are independent. Their distance is few kilometers: distance factor on performance negligible.
- The global Sysplex (1-1a) is managed by GDPS/PPRC, which allows a near continuous availability and offers freezing capability. This is recommended as soon as there are more than one primary ESS.
- Site 2 — is the failover site and contains XRC-SDM into another Sysplex, which is also used for other purposes.
- No connection between site 1 and site 2.
- RCMF/PPRC and RCMF/XRC are recommended to have a central point of control, to initialize and maintain the remote copy configuration: they manage configurations instead of pairs.
- PAV addresses potential performance concerns: performance with PAV and combined PPRC + XRC is better than without PAV and without any remote copy function active.
- According to the available bandwidth, XRC operates either in continuous mode or in continuous suspend/resume loop (which allows an efficient telecom lines utilization).

Trunk topology between sites 1 and 1a

Two channel sets of eight paths come from hosts in site 1 to ESS1 in site 1a, primary for both PPRC and XRC (see Figure 165). Then two sets of four paths link ESS1 in site 1a to ESS2 in site 1 for PPRC: the primary is ESS 1 and the secondary ESS 2.

Figure 166 shows the PPRC and host to primary connectivity.

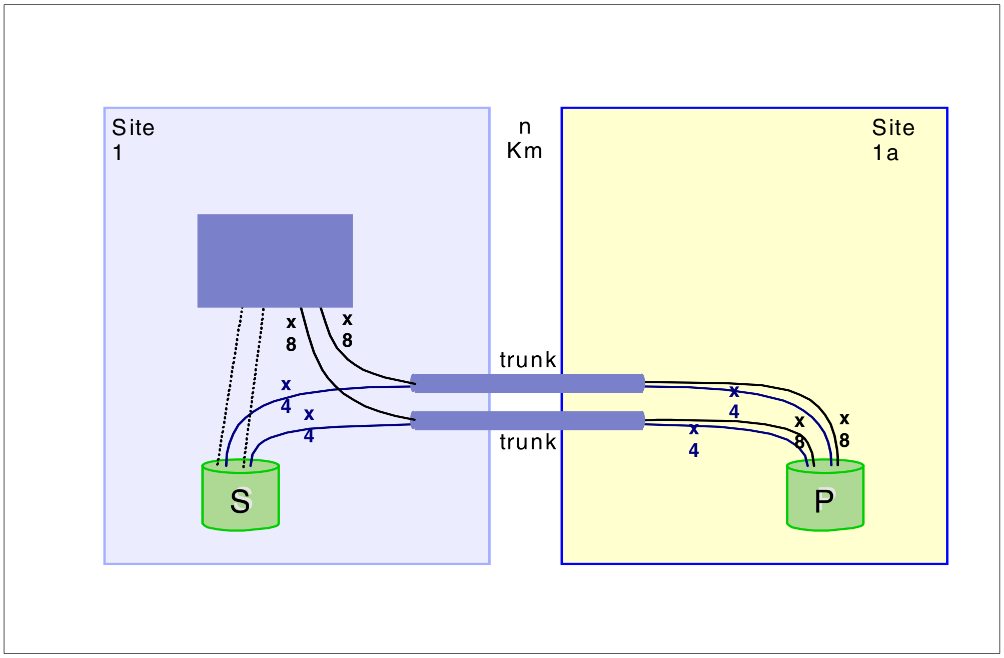
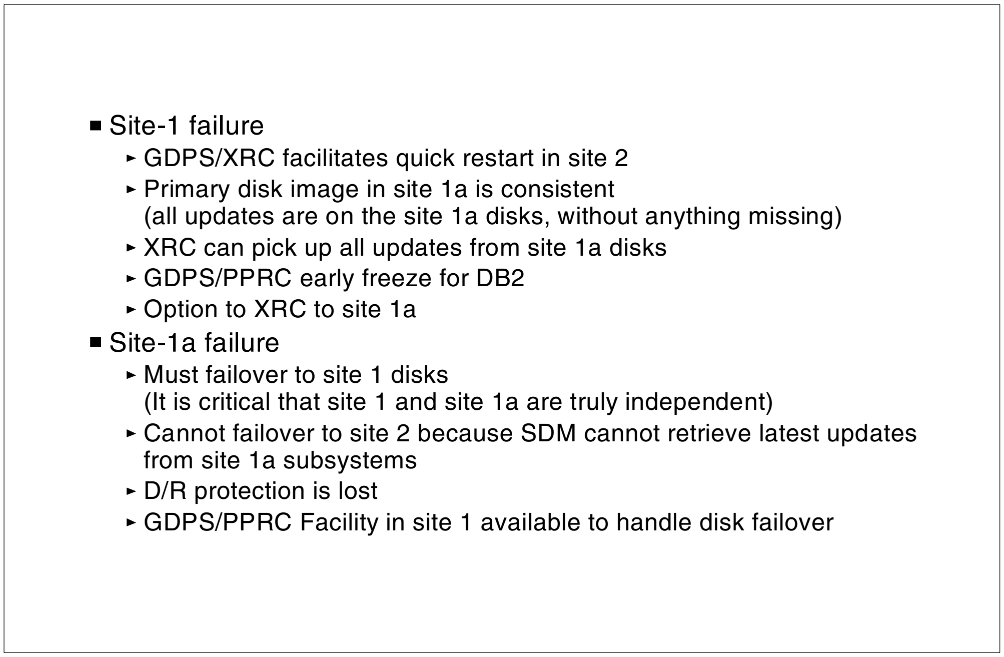


Figure 166. Site 1 to site 1a connectivity

Component failure impact analysis

Figure 167 shows the resulting decision making for failure processing based on the PPRC component failure impact analysis discussion presented below. The failure of both sites 1 and 1a has not been considered as relevant in that installation.



- Site-1 failure
  - GDPS/XRC facilitates quick restart in site 2
  - Primary disk image in site 1a is consistent (all updates are on the site 1a disks, without anything missing)
  - XRC can pick up all updates from site 1a disks
  - GDPS/PPRC early freeze for DB2
  - Option to XRC to site 1a
- Site-1a failure
  - Must failover to site 1 disks (It is critical that site 1 and site 1a are truly independent)
  - Cannot failover to site 2 because SDM cannot retrieve latest updates from site 1a subsystems
  - D/R protection is lost
  - GDPS/PPRC Facility in site 1 available to handle disk failover

Figure 167. PPRC/XRC failure processing

1. **Site 1 failure**
   - Situation assessment. The worst case is a serious problem in site 1. That requires a fail over to site 2. The secondary ESS2 in site 1 is not needed.
   - PPRC Critical attribute considerations. Using critical YES-All should impact the production. In fact this facility does not seem to be justified.
   - GDPS options. The freeze-and-go or freeze-and-stop-conditional put the ESS1 production disks in a consistent state without affecting production. This function can be accomplished across any number of primary LSSs.
2. **Links failure between site 1 and site 1a**
   - Situation assessment. There must be serious problems in site 1 or site 1a. As the same trunks are used for host access to primary ESS1, and for PPRC replication from ESS1 to ESS2, the writes cannot be reported complete unless data is on both subsystems. Most likely, the production is down due to a loss of host access to primary ESS1.
   - PPRC Critical attributes considerations. Primary ESS1 and secondary ESS2 will not get out of synchronization, regardless of critical attribute setting.
   - GDPS options. Probably they are not effective, due to the loss of host access to primary ESS1.

3. **Site 1a failure**

   - Situation assessment. This can be an hardware problem with ESS1 in site 1a. That precludes a future fail over to site 2. In that case the site 1 production readiness is essential for that issue. The question is: can ESS1 suspend the pairs and continue to accept updates? In that case potential data loss should occur.
   - PPRC Critical attribute considerations. CRIT=NO and CRIT=YES-Paths may not be adequate. CRIT=YES-All keeps primary/secondary identical. But *dying scenario* (for instance a DB2 flagging all his table spaces in recover mode into the log, still mirrored for a while) may be partially replicated, complicating the subsequent recovery.
   - GDPS options. Freeze-and-go may lose data in fail over to site 1 ESS2. Freeze-and-stop-conditional offers no data loss and no logical data contamination. Note GDPS has the capacity to initiate the Freeze process also on serious I/O errors.

Appendix E. Feature codes

The Enterprise Storage Server provides a wide range of configurations scaling from 420 GB to 11.2 TB. Currently there are 32 standard capacity configurations you can order. For these configurations there are also optional features you can order, like the desired cache size, disk drive capacity, and host adapter types. This allows you to configure the ESS according to your specific workload and attachment requirements.

This part of the book summarizes the features used to order these standard configurations and options.

E.1 ESS models

There are two models of the Enterprise Storage Server:

- **IBM 2105-F10** Enterprise Storage Server (Single-phase power)
- **IBM 2105-F20** Enterprise Storage Server (Three-phase power)

When ordering the ESS you must specify the machine model, with the standard capacity of your choice, and the host adapters you will require. You will also specify the cache you need, and optional features. Note that the F10 model, cannot be configured with the 2105 Expansion Frame (feature code 2100).

E.2 Standard configurations

You have a set of standard capacity configurations, currently 32, that meet most customer requirements. These configurations are ordered by feature code, that determines the total capacity and the type of disk drives used.

E.2.1 Standard capacity configurations

The capacities shown in Table 9, are the capacities that each configuration has available for user data, already RAID-5 protected.

Table 9. Standard capacity configurations (gigabytes))

Capacity	Exp. Rack	Models	Type/No. 8pack	Feature Code
420	No	E10/E20 F10/F20	9.1/8	9601
840	No	E20/F20	9.1/16	9602
420	No	E10/E20 F10/F20	18.2/4	9621
630	No	E10/E20 F10/F20	18.2/6	9622
840	No	E10/E20 F10/F20	18.2/8	9623
1,050	No	E20/F20	18.2/10	9630
1,260	No	E20/F20	18.2/12	9624
1,470	No	E20/F20	18.2/14	9631
1,680	No	E20/F20	18.2/16	9625
2,170	Yes	E20/F20	18.2/20	9632

Capacity	Exp. Rack	Models	Type/No. 8pack	Feature Code
2,660	Yes	E20/F20	18.2/24	9626
3,150	Yes	E20/F20	18.2/28	9633
3,640	Yes	E20/F20	18.2/32	9627
4,130	Yes	E20/F20	18.2/36	9634
4,620	Yes	E20/F20	18.2/40	9628
5,110	Yes	E20/F20	18.2/44	9635
5,600	Yes	E20/F20	18.2/48	9629
840	No	E10/F10 E20/F20	36.4/4	9646
1,260	No	E10/F10 E20/F20	36.4/6	9647
1,680	No	E10/F10 E20/F20	36.4/8	9641
2,100	No	E20/F20	36.4/10	9650
2,520	No	E20/F20	36.4/12	9642
2,940	No	E20/F20	36.4/14	9651
3,360	No	E20/F20	36.4/16	9643
4,340	Yes	E20/F20	36.4/20	9652
5,320	Yes	E20/F20	36.4/24	9648
6,300	Yes	E20/F20	36.4/28	9653
7,280	Yes	E20/F20	36.4/32	9644
8,260	Yes	E20/F20	36.4/36	9654
9,240	Yes	E20/F20	36.4/40	9649
10,220	Yes	E20/F20	36.4/44	9655
11,200	Yes	E20/F20	36.4/48	9645

Table 10 shows the capacities of the disk 8-packs when configured as RAID-5 ranks. The capacities shown, are fully available for user data. Note that the RAID-5 ranks in the base frame of the ESS are configured as 6+P+S, while the RAID-5 ranks in the expansion frame of the ESS are configured as 7+P.

Table 10. Raid-5 array capacities (gigabytes)

Disk Capacity	6+P (Base Frame)	7+P (Expansion Frame)
9.1	52.5	61
18.2	105	122.5
36.4	210	245

Capacity upgrades

Field installed capacity upgrades are enabled for configurations of the same disk drive size (9.1 GB, 18.2 GB, and 36.4 GB). This capacity upgrade between any configuration of like disk drives can be performed concurrently with normal I/O operations on the ESS.

Capacity upgrades are available within a drive family to any larger standard configuration, for example, specify number #9601 to #9602, #9621 to #9622 through #9629, and so on.

E.3 Step Ahead configurations

The ESS provides a wide range of step ahead configurations scaling from 420 GB (w/210 GB Step Ahead) to 10,220 GB (w/980 GB Step Ahead). When you order an ESS step ahead standard configuration you are shipped the next larger standard configuration ESS.

Table 11. Step Ahead capacity increments (gigabytes)

Base	Step Ahead	Exp. Frame	8-packs	Specify
420	210	No	18.2/4+2	9521
630	210	No	18.2/6+2	9522
840	210	No	18.2/8+2	9523
1,050	210	No	18.2/10+2	9530
1,260	210	No	18.2/12+2	9524
1,470	210	No	18.2/14+2	9531
1,680	490	Yes	18.2/16+4	9525
2,170	490	Yes	18.2/20+4	9532
2,660	490	Yes	18.2/24+4	9526
3,150	490	Yes	18.2/28+4	9533
3,640	490	Yes	18.2/32+4	9527
4,130	490	Yes	18.2/36+4	9534
4,620	490	Yes	18.2/40+4	9528
5,110	490	Yes	18.2/44+4	9535
840	420	No	36.6/4+2	9546
1,260	420	No	36.6/6+2	9547
1,680	420	No	36.6/8+2	9541
2,100	420	No	36.6/10+2	9550
2,520	420	No	36.6/12+2	9542
2,940	420	No	36.6/14+2	9551
3,360	980	Yes	36.4/16+4	9543
4,340	980	Yes	36.4/20+4	9552
5,320	980	Yes	36.4/24+4	9548
6,300	980	Yes	36.4/28+4	9553
7,280	980	Yes	36.4/32+4	9544
8,260	980	Yes	36.4/36+4	9554

Base	Step Ahead	Exp. Frame	8-packs	Specify
9,240	980	Yes	36.4/40+4	9549
10,220	980	Yes	36.4/44+4	9555

E.4 Major feature codes

For each of the standard capacity configuration you select, you will also select a cache size, and the type and quantity of host adapters.

E.4.1 Cache sizes

The ESS can be configured with four different cache capacities, according to Table 12

Table 12. ESS cache capacities (gigabytes)

Capacity	Feature
8 GB	4002
16 GB	4004
24 GB	4005
32 GB	4006

E.4.2 Host adapters

The ESS supports the intermix of FICON, ESCON, SCSI, and Fibre Channel attachments, making the ESS the natural fit for server consolidation requirements. Table 13 shows the feature codes for the different host adapters that can be ordered with the ESS. The ESS can be configured with up to 16 of these host adapters.

Table 13. ESS host adapters

Host Adapter	Feature
Fibre Channel / FICON (long wave)	3021
Fibre Channel / FICON (short wave)	3023
Enhanced ESCON	3012
SCSI	3002

E.4.2.1 Fibre Channel / FICON (Long Wave) host adapter (fc 3021)

Each Fibre Channel / FICON (long wave) host adapter provides one port with an SC-type connector. The interface supports 100 MB/second, full duplex data transfer over long-wave fibre links. The adapter supports the SCSI-FCP ULP on point-to-point, fabric, and arbitrated loop (private loop) topologies, and the FICON ULP on point-to-point and fabric topologies. SCSI-FCP and FICON are not supported simultaneously on an adapter. Minimum number of this feature per ESS, none. Maximum number of this feature per ESS, sixteen. No cable order is required, as one 31 meter (100 foot) single mode 9 micron fibre cable with SC Duplex connectors is included with each feature.

E.4.2.2 Fibre Channel / FICON (Short Wave) host adapter (fc 3023)

Each Fibre Channel / FICON (short wave) host adapter provides one port with an SC-type connector. The interface supports 100 MB/second, full duplex data transfer over short-wave fibre lines. The adapter supports the SCSI-FCP ULP on point-to-point, fabric, and arbitrated loop (private loop) topologies, and the FICON ULP on point-to-point and fabric topologies. SCSI-FCP and FICON are not supported simultaneously on an adapter. Minimum number of this feature per ESS, none. Maximum number of this feature per ESS, sixteen. No cable order is required, as one 31 meter (100 foot) multimode 50 micron fibre cable with SC Duplex connectors is included with each feature.

E.4.2.3 Enhanced ESCON host adapter (fc 3012)

Each Enhanced ESCON host adapter supports two ESCON links with each link supporting 64 logical paths. The ESCON attachment uses an LED-type interface. ESCON cables must be ordered separately. Minimum number of this feature per ESS, none. Maximum number of this feature per ESS, sixteen.

E.4.2.4 SCSI host adapter (fc 3002)

The dual-port Ultra SCSI host adapter supports the Ultra SCSI protocol (SCSI-2 Fast/Wide differential is a subset of Ultra SCSI and is therefore supported as well). If the server has SCSI-2 (fast/wide differential) adapters and/or Ultra SCSI adapters, they can attach to the Enterprise Storage Server. Minimum number of this feature per ESS, none. Maximum number of this feature per ESS, sixteen.

SCSI host attachment cables are required. When ordering the ESS you use the feature codes 9701 to 9710 to request the necessary cables. You can order up to 32 of the 97xx cables in any combination. For additional cables you use the feature codes 2801 to 2810. The Host Adapters and Cables table, at `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm` shows the cable feature codes that correspond to the different server SCSI adapters. Table 14 lists the feature codes for the different SCSI cables that can be ordered for the ESS.

Table 14. SCSI cables feature codes

Description	Feature code	Additional cables
10 Meter Cable - Ultra SCSI	9701	2801
20 Meter Cable - Ultra SCSI	9702	2802
10 Meter Cable - SCSI-2 Fast/Wide	9703	2803
20 Meter Cable - SCSI-2 Fast/Wide	9704	2804
10 Meter Cable - SCSI-2 Fast/Wide (AS/400)	9705	2805
20 Meter Cable - SCSI-2 Fast/Wide (AS/400)	9706	2806
10 Meter Cable - SCSI-2 Fast/Wide (Sun/HP (dual) PCI)	9707	2807
20 Meter Cable - SCSI-2 Fast/Wide (Sun/HP (dual) PCI)	9708	2808
10 Meter Cable - SCSI-2 Fast/Wide (HP PCI)	9709	2809
20 Meter Cable - SCSI-2 Fast/Wide (HP PCI)	9710	2810

E.5 Advanced functions

The advanced functions of the ESS are ordered as optional features, either initially when ordering the ESS, or later as a field upgrade to the ESS.

E.5.1 PAV and XRC

Parallel Access Volumes (PAV) and Extended Remote Copy (XRC) are applicable to the z/OS environments only. The license for PAV (feature codes 1800-1805) and XRC (feature codes1810-1815) must be equal to or greater than the capacity of the CKD configured portion of the ESS. Table 15 lists the feature codes for the different CKD capacities that will be configured on the ESS.

Table 15. PAV and XRC feature codes

ESS CKD Capacity	PAV feature code	XRC feature code
Up to 0.5 TB	1800	1810
Up to 1 TB	1801	1811
Up to 2 TB	1802	1812
Up to 4 TB	1803	1813
Up to 8 TB	1804	1814
Larger than 8 TB	1805	1815

E.5.2 PPRC and FlashCopy

Peer-to-Peer Copy (PPRC) and FlashCopy are applicable to all environments. The license for PPRC (feature codes 1820-1825) and FlashCopy (feature codes 1830-1835) must be equal to or greater than the total capacity of the ESS. Table 16 lists the feature codes for the different ESS capacities.

Note: Implementation of PPRC requires the PPRC feature to be installed on both the primary and secondary ESS.

Table 16. PPRC and FlashCopy feature codes

ESS Total Capacity	PPRC feature code	FlashCopy feature code
up to 0.5 TB	1820	1830
up to 1 TB	1821	1831
up to 2 TB	1822	1832
up to 4 TB	1823	1833
up to 8 TB	1824	1834
larger than 8 TB	1825	1835

E.6 Additional feature codes

There are additional feature codes used to configure the ESS. Some of these features are so called specifies, and they allow the plant to setup the ESS with the appropriate cables and parts characteristics for the location where the ESS will be operating. This section lists very briefly these specify features, plus some additional features used to configure the ESS.

E.6.1 Specify features

E.6.1.1 Operator Panel Language groups (2924)

A specify code for the Operator Panel Language group is not required. The standard default Operator Panel provided is English. The specify codes 2928 to 2980 can be used when an alternative to the configurator default is required.

E.6.1.2 Modem specify features for Remote Support (9301-9318)

Feature number 2715 supplies a modem for the ESS. To designate the country specific modem required, one of the 9301 to 9318 features is used. Minimum, one. Maximum, one.

E.6.1.3 Convenience outlet line cords for service (9401-9412)

One of the country specific convenience cord features 9401 to 9412 for the service provider is specified as a function of the Service Alert capabilities of the ESS. This feature will provide two convenience cords for the Remote Support Switch modem and the service provider. Minimum, one. Maximum, one.

E.6.1.4 Power cords

The ESS power control subsystem is fully redundant. Power is supplied through two 30, 50, or 60 amp line cords, with either line cord capable of providing 100% of the needed power. One of the following power cord specify codes must be selected:

- 9851 for three phase, 50/60 Hz, 50 amp
- 9853 for three phase, 50/60 Hz, 60 amp (US and Japan)
- 9854 for three phase, 50/60 Hz, 60 amp
- 9855 for three phase, 50/60 Hz, 30 amp

E.6.1.5 Input voltage

The input voltage feature determines the type of AC/DC power supply used within the ESS enclosures and must be selected using one of the 9870 to 9871 specify codes.

- 9870 for nominal AC Voltage: 200V-240V
- 9871 for nominal AC Voltage: 380V-480V

E.6.2 Special feature codes

E.6.2.1 Remote Power Control (1001)

Provides a logic card to the enclosure to support remote power control within the storage facility.

E.6.2.2 IBM 7133 Drawer attachment

Attachment of IBM 7133 Model 020 and D40 drawers is provided through the attachment of the 2105 Model 100 rack to the ESS.

- Feature number 1121 provides for the attachment of one 2105-100 to the ESS. The 2105-100 can support up to six customer-supplied 7133-020 or 7133-D40 drawers in this configuration.
- Feature number 1122 provides for the attachment of a second 2105-100 to the ESS. The second 2105-100 can also support up to six customer supplied 7133-020 or 7133-D40 drawers, for a total of twelve drawers in this configuration.

Feature numbers 1121 and 1122 must be ordered against the 2105-100 and apply to new orders only.

Attach Model 100 to Model E10, E20, F10, or F20 (1121)

Up to two 2105-100 racks can be attached to the ESS Model E10, E20, F10, or F20. Each 2105-100 must include battery backup power, feature code 1000. The 2105-100 that attaches directly to the ESS (first in string) must include this feature code 1121. This feature provides the installation, formatting, and checkout of up to six customer provided 7133-020 or 7133-D40 drawers.

The ESS cannot already have the Expansion Enclosure (feature code 2100) attached.

E.6.2.3 Reserve loop (9904)

This feature reserves two SSA storage loops for future attachment of 2105 Model 100 with 7133 Model 020 and/or Model D40 drawers. Maximum, two.

Attach second Model 100 to Model E10, E20, F10, or F20 (1122)

A Second 2105-100 rack can be attached to the first 2105-100 and must include feature code 1122. A maximum of twelve 7133 drawers can be installed, six in each rack.

E.6.2.4 Expansion Enclosure (2100)

This feature provides an expansion enclosure for the primary Enterprise Storage Server Model E20 or F20.

E.6.2.5 Remote Support Facility (2715)

The Remote Support Facility (feature code 2715) provides call home and remote support capability for the ESS. This feature provides a switch, modem, cables, and connectors along with ESSNet items (PC, hub, and Ethernet cables) to attach the first ESS in the installation. The Remote Support Facility will support 1-7 Model E10s, E20s, F10s, or F20s. It must be ordered on the first storage server in the installation.

Each ESS Model E10, E20, F10, or F20 must specify either feature code 2715 or 2716.

E.6.2.6 Remote Support Facility attachment (2716)

The Remote Support Facility Attachment (feature code 2716) provides the modem cables along with ESSNet items (Ethernet cables) to attach additional ESSs to the Remote Support Facility for remote serviceability. This feature provides the modem cables and Ethernet cables to attach the second thru seventh ESS to the Remote Support Facility. It must be ordered on the second thru seventh ESS in the installation.

Appendix F. Related publications

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

F.1 IBM Redbooks

For information on ordering these publications see F.4, “How to get IBM Redbooks” on page 314.

- *Implementing ESS Copy Services on UNIX and Windows NT/2000,* SG24-5757
- *Implementing ESS Copy Services on S/390,* SG24-5680
- *Implementing the Enterprise Storage Server in Your Environment,* SG24-5420
- *Implementing Fibre Channel Attachment on the ESS,* SG24-6113
- *Introduction to IBM S/390 FICON,* SG24-5176
- *IBM S/390 FICON Implementation Guide,* SG24-5169
- *FICON (FCV Mode) Planning Guide,* SG24-5445
- *IBM Enterprise Storage Server Performance Monitoring and Tuning Guide,* SG24-5656
- *IBM StorWatch Expert Hands-On Usage Guide,* SG24-6102
- *ESS Solutions for Open Systems Storage: Compaq AlphaServer, HP, and Sun,* SG24-6119
- *FICON Native Implementation and Reference Guide,* SG24-6266
- *Enterprise Storage Solutions Handbook,* SG24-5250
- *IBM e(logo)server iSeries in Storage Area Networks: A Guide to Implementing FC Disk and Tape with iSeries,* SG24-6220

  Note: The above publication (SG24-6220) is a Redpiece and is expected to be published as an IBM Redbook before the end of the year 2001.

F.2 Other resources

The following publications are available with CD-ROM Kit SK2T-8770 and are shipped with the ESS.

- *IBM Enterprise Storage Server Introduction and Planning Guide,* GC26-7294
- *IBM Enterprise Storage Server User’s Guide,* SC26-7295
- *IBM Enterprise Storage Server Host System Attachment Guide,* SC26-7296
- *IBM Enterprise Storage Server SCSI Command Reference,* SC26-7297
- *IBM Enterprise Storage Server Quick Configuration Guide,* SC26-7354
- *IBM Enterprise Storage Server System/390 Command Reference,* SC26-7298
- *IBM Storage Solution Safety Notices,* GC26-7229

The above publications are also available on the ESS Web site: `http://www.storage.ibm.com/storage/hardsoft/products/ess/refinfo.htm`

You can also order the following publications, that are also relevant as further information sources. They are not included in the CD-ROM:

- *DFSMS/MVS Software Support for the IBM Enterprise Storage Server,* SC26-7318
- *DFSMS/MVS Version 1 Advanced Copy Services,* SC35-0355
- *IBM Enterprise Storage Server Web Interface Users Guide for the ESS Specialist and ESS Copy Services,* SC26-7346

The publications above can be ordered from your IBM representative, or via the Publication Notification System (PNS) Web site at: `http://www.ibm.com/shop/publications/pns`

The *IBM Enterprise Storage Server Configuration Planner,* SC26-7353, is available only at the ESS Web site at: `http://www.storage.ibm.com/storage/hardsoft/products/ess/refinfo.htm`

F.3 Referenced Web sites

These Web sites are also relevant as further information sources:

- `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`
  ESS supported servers and adapters, cables and software requirements
- `http://www.ibm.com/storage/support/techsup/swtechsup.nsf/support/sddupdates`
  Subsystem Device Driver documentation and downloads
- `http://www.ibm.com/storage`
  IBM Storage
- `http://www.storage.ibm.com/hdd/index.htm`
  Hard disk drives
- `http://www.ibm.com/storage/software/storwatch/ess`
  StorWatch Expert
- `http://www.storage.ibm.com/software/sms/sdm/sdmtech.htm`
  DFSMS/MVS Copy Services

F.4 How to get IBM Redbooks

Search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also 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.

F.4.1 IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Appendix G. Special notices

This publication is intended to help customers, IBM personnel, and business partners to understand the IBM 2105 Enterprise Storage Server and its powerful functions. The information in this publication is not intended as the specification of any programming interfaces that are provided by the Enterprise Storage Server product name(s) 2105 Model F10 and F20. See the PUBLICATIONS section of the IBM Programming Announcement for the Enterprise Storage Server product name(s) 2105 Model F10 and F20 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:

e (logo)®
IBM ®
AIX
AS/400
400
AT
CICS
CT
CUA
Current
DB2
DB2 Universal Database
DFSMS/MVS
DFSMS/VM
DFSMSdfp
DFSMSdss
DFSORT
ECKD
Enterprise Storage Server
Enterprise Systems Connection Architecture
ESCON
FICON
Lotus
Magstar
MVS/DFP
MVS/ESA
Redbooks
Redbooks Logo
Netfinity
NetView
Notes
OS/390
OS/400
Parallel Sysplex
RACF
RAMAC
RMF
RS/6000
S/390
Seascape
SP
StorWatch
Sysplex Timer
System/370
System/390
TME
Ultrastar
VM/ESA
VSE/ESA
Wave
WebSphere

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.

Glossary

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

A

allegiance. The zSeries 900 and S/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.

APAR. Authorized program analysis report. Used to report bugs and fixes in the z/OS and OS/390 environments,.

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.

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, Fibre Channel or FICON 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. A sequence of CCWs make up a Channel Program perform I/O operations on zSeries 900 environments.

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 on zSeries 900 and S/390 environments.

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 zSeries 900 and S/390 term for the interconnection between a channel and its associated controllers.

channel subsystem. The zSeries 900 and S/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 zSeries 900 and S/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 for the z/OS and OS/390.

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.

Combination E/F Ports, G_Ports. These ports are sometimes found in Fibre Channel Switched Fabrics and are used either as E_Ports, when the link is connected to another switch, or as F_Ports, when the link is connected to an N_Port for a host or device. This port automatically determines what mode to run in after determining what it is connected to.

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.

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

D

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

DA. The SSA loops of the ESS are physically and logically connected to the Device Adapters (also see 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.

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. Acronym for Direct Access Storage Device. This term is common in the zSeries 900 and S/390 environments to designate a logical volume.

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.

DDM. see disk drive module.

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 zSeries 900 and S/390 term for a disk drive.

device address. The zSeries 900 and S/390 term for the field of an ESCON or FICON device-level frame that selects a specific device on a control-unit image. The one or two left most digits are the address of the channel to which the device is attached. The two right most 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. The zSeries 900 and S/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 in zSeries 900 and S/390 environments.

DFDSS. Data Facility Data Set Services (see DFSMSdss)

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 (DDM). The primary nonvolatile storage medium that you use for any host data that is stored within a subsystem. Number and type of DDMs within a storage facility may vary. DDMs are the whole replaceable units (FRUs) that hold the HDDs

disk-group. In the ESS, a group of 7 or 8 DDMs that are not yet formatted as ranks.

disk 8-pack. In the ESS, a group of 7 or 8 DDMs that are not yet formatted as ranks.

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.

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

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. See Environmental Record Editing and Printing

Program.

ERP. Error recovery procedure.

ESCD. ESCON Director. A high speed switch for ESCON links.

ESCM. See ESCON Manager.

ESCON. Enterprise Systems Connection Architecture. An zSeries 900 and S/390 computer peripheral interface. The I/O interface utilizes S/390 logical protocols over a serial interface that configures attached units to a communication fabric.

ESCON Channel. A channel that has an ESCON channel-to-controller I/O interface that uses optical cables as a transmission medium.

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

Expansion Ports, E_ports. These ports are found in Fibre Channel Switched Fabrics and are used to interconnect the individual switch or routing elements. They are not the source or destination of IUs, but instead function like the F_ports and FL_ports to relay the IUs from one switch or routing element to another. E_ports can only attach to other E_ports. The ESS Fibre Channel adapters do not support the E_port functionality, which is found only in fabrics or hubs.

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

F_Node Fabric Node - a fabric attached node.

F_Port Fabric Port - a port used to attach a NodePort (N_Port) to a switch fabric.

Fabric Fibre Channel employs a fabric to connect devices. A fabric can be as simple as a single cable connecting two devices. The term is most often used to describe a more complex network utilizing hubs, switches and gateways.

Fabric Login Fabric Login (FLOGI) is used by an N_Port to determine if a fabric is present and, if so, to initiate a session with the fabric by exchanging service parameters with the fabric. Fabric Login is performed by an N_Port following link initialization and before communication with other N_Ports is attempted.

FC see Fibre Channel

FC-0 Lowest level of the Fibre Channel Physical standard, covering the physical characteristics of the interface and media

FC-1 Middle level of the Fibre Channel Physical standard, defining the 8B/10B encoding/decoding and transmission protocol.

FC-2 Highest level of the Fibre Channel Physical

standard, defining the rules for signaling protocol and describing transfer of frame, sequence and exchanges.

FC-3 The hierarchical level in the Fibre Channel standard that provides common services such as striping definition.

FC-4 The hierarchical level in the Fibre Channel standard that specifies the mapping of upper-layer protocols to levels below.

FCA Fiber Channel Association.

FC-AL. Fibre Channel - Arbitrated Loop. An implementation of the fibre channel standard that uses a ring topology for the communication fabric. A shared gigabit media for up to 127 nodes one of which may be attached to a switch fabric.

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.

FCP. Fibre Channel Protocol. A SCSI protocol mapped on to the FCP-4 Upper Layer of the Fibre Channel transport.

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).

fibre channel ports. There are five basic kinds of ports defined in the Fibre Channel architecture, as well as some vendor-specific variations. The five basic ports are as follows:

Fabric-Loop Ports, FL_ports. These ports are just like the F_ports, except that they connect to an FC-AL topology. FL_ports can only attach to NL_ports. The ESS Fibre Channel adapters do not support the FL_port functionality, which is found only in fabrics or hubs.

Fabric Ports, F_ports. These ports are found in Fibre Channel Switched Fabrics. They are not the source or destination of IUs, but instead function only as a "middleman" to relay the IUs from the sender to the receiver. F_ports can only attach to N_ports. The ESS Fibre Channel adapters do not support the F_port functionality, which is found only in fabrics.

fiber optic cable. A fiber, or bundle of fibers, in a structure built to meet optic, mechanical, and environmental specifications.

FICON. An IO interface based on the Fibre Channel architecture. In this new interface, the ESCON protocols have been mapped to the FC-4 layer, i.e. the Upper Level Protocol layer, of the Fibre Channel Architecture. It is used in the S/390 and z/series environments.

FICON Channel. A channel that has a FICON channel-to-controller I/O interface that uses optical cables as a transmission medium.

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

G

GB. See gigabyte.

gigabyte. 1 073 741 824 bytes.

group. A group consist of eight DDMs.

GTF. Generalized trace facility.

H

HA. (1) Home address. (2) Host adapter.

hard disk drive (HDD). A non-volatile storage medium within a storage server used to keep the data records. In the ESS these are 3.5” disks, coated with thin layers of special substances, where the information is magnetically recorded and read. HDDs are packed in replaceable units called DDMs.

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

HDD. See hard disk drive. Also sometimes referred as head and disk drive.

home address (HA). 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. The server where the application programs run.

host adapter (HA). 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. See 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. A 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.

iSeries 400. The new integrated application servers family of processors of the IBM ⓔserver brand. These servers are successors to the IBM AS/400 family of servers.

ITSO. International Technical Support Organization.

J

JBOD. A disk group configured without the disk redundancy of the RAID-5 arrangement. When configured as JBOD, each disk in the disk group is a rank in itself.

JCL. See 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. See 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. One thousand meters.

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.

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

logical address. On an ESCON or FICON 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 partition (LPAR). The term for a set of functions that create the programming environment that is defined by the S/390 and zSeries 900 architecture. zSeries 900 and S/390 architectures use 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.

LSS. See logical subsystem.

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 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.

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. See 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.

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. See Multiple Virtual Storage.

N

Node-Loop Ports, NL_ports. These ports are just like the N_ports, except that they connect to a Fibre Channel Arbitrated Loop (FC-AL) topology. NL_ports can only attach to other NL_ports or to FL_ports. The ESS Fibre Channel adapters support the NL_port functionality when connected directly to a loop.

Node Ports, N_ports. These ports are found in Fibre Channel Nodes, which are defined to be the source or destination of Information Units (IUs). I/O devices and host systems interconnected in point-to-point or switched topologies use N_ports for their connections. N_ports can only attach to other N_ports or to F_ports. The ESS Fibre Channel adapters support the N_port functionality when connected directly to a host or to a fabric.

non-disruptive. 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. See 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.

P

path group. The s/390 and z/Architecture 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 S/390 and z/Architecture 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 for the MVS and z/OS environments.

P/DAS. PPRC dynamic address switching.

PDSE. See 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. See 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. A fix to a bug in a program or routine.

R

RACF. Resource access control facility.

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

rank. A disk group upon which a RAID-5 array is configured. For JBOD, each DDM becomes a rank.

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. A S/390 and z/Architecture 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. A S/390 disk storage unit.

S

SAM. Sequential access method. These routines access the data records in storage in a sequential manner.

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.

SIM. See 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. MVS and z/OS routines that collect records on the activities of the system.

SMS. Storage Management Subsystem.

SRM. System resources manager. Routines in the MVS and z/OS operating systems, that manage the use of the complex resources (CPU, IO, storage) based on user parameters.

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.

SSR. System Support Representative. The IBM person who does the hardware installation and maintenance.

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

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 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, multi-byte, 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. An interactive tool for the MVS and z/OS environment.

U

Ultra-SCSI. An enhanced small computer system interface.

unit address. The s/390 and z/Architecture 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.

V

vital product data (VPD). Nonvolatile data that is stored in various locations in the DASD subsystem. It includes information that uniquely defines the system, like configuration data, machine serial number, machine features, microcode information.

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.

VSAM. Virtual storage access method. A S/390 organization of the data in storage.

VTOC. Volume table of contents. In a DASD, the place where the space and allocation information of the volume is maintained.

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.

Z

z/Architecture. The IBM 64-bit real architecture implemented in the new IBM eserver zSeries 900 enterprise e-business servers.

z/OS. The IBM operating systems for the z/Architecture family of processors.

zSeries 900. The new enterprise e-business server family of processors of the IBM eserver brand. These servers are the successors to the IBM 9672 G5 and G6 family of processors.

List of abbreviations

CKD	Count Key Data
CPC	Cluster Processing Complex
CPI	Common Parts Interconnect
CU	Control Unit
DA	Device Adapter
DASD	Direct Access Storage Device
DDM	Disk Drive Module
DPO	Data Path Optimizer
EREP	Environmental Record Editing and Printing
ERP	Error and Recovery Procedure
ESCM	ESCON Manager
ESCD	ESCON Director
ESCON	Enterprise Connection
ESS	Enterprise Storage Server
FB	Fixed Block (Architecture)
FBA	Fixed Block Architecture
FC	Fibre Channel
FCAL	Fibre Channel Arbitrated Loop
FCP	Fixed Channel Protocol
FICON	Fiber Connection
GB	Gigabyte
GDPS	Geographically Dispersed Parallel Sysplex
HA	Host Adapter
HDD	Head and Disk Drive
IBM	International Business Machines Corporation
IML	Initial Microcode Load
IPL	Initial Program Load
ESCM	ESCON Manager
ITSO	International Technical Support Organization
IODF	I/O Definition File
IOS	Input Output Supervisor
JCL	Job Control Language
JBOD	Just a Bunch Of Disks
KB	Kilobyte
Km	Kilometer
LAN	Local Area Network
LD	Logical Disk
LED	Light Emitting Diode
LV	Logical Volume
LCU	Logical Control Unit
LIC	Licensed Internal Code
LP	Logical Path
LPAR	Logical Partition
LSS	Logical SubSystem
LUN	Logical Unit Number
LRU	Least Recently Used
LV	Logical Volume
MB	Megabyte
MIH	Missing Interrupt Handler
NAS	Network Attached Storage
NVS	Non-volatile Storage
PAV	Parallel Access Volume
PPRC	Peer to Peer Remote Copy
PTF	Program Temporary Fix
RAID	Redundant Array of Independent Disks
RVA	RAMAC Virtual Array
SAM	Sequential Access Method
SAN	Storage Area Network
SCSI	Small Systems Computer Interface
SS	SubSystem
SSA	Serial Storage Architecture
SDD	SubSystem Device Driver
SSID	SubSystem IDentification
SSR	IBM System Support Representative
TPF	Transaction Processing Facility
UCB	Unit Control Block
VPD	Vital Product Data
VSAM	Virtual Storage Access Method
VTOC	Volume Table of Contents
WLM	Workload Manager
XRC	Extended Remote Copy

Index

IBM Enterprise Storage Server

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

IBM Enterprise Storage Server

Understand the ESS characteristics and design

Discover its powerful performance and copy features

Learn the basics before planning and configuring

This IBM Redbook describes, in detail, the architecture, hardware, and advanced functions of the Enterprise Storage Server (ESS). It covers Fibre Channel and FICON attachment. The information will help the IT Specialist in the field to understand the ESS, its components, its overall design, and how all of this works together. This understanding will be useful for proper planning and logical configuration at the time of installation.

The IT Specialist will find useful information when availability and disaster recovery solutions are to be implemented, with the explanation of the advanced copy functions that ESS has available: FlashCopy, Peer-to-Peer Remote Copy, Extended Remote Copy, and Concurrent Copy. Also the performance features of the ESS are explained, making all of these topics very helpful for the IT Specialist for the optimization of resources in the computing center.

This book has been updated to support the announcement and general availability of the new Fibre Channel/FICON host adapters. Also, it has been updated to the ESS models F10 and F20, and to generally all that has been available since the initial announcement of Enterprise Storage Server.

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

SG24-5465-01 ISBN 0738424277