Global.Name
registry

APPENDIX.D.2.1.3

IBM ESS Storage

A high-performance disk storage solution for systems across the enterprise

IBM Enterprise Storage Server, Models F10 and F20

Highlights

Provides superior storage sharing for UNIX**, Windows NT**, Windows** 2000, Novell NetWare**, AS/400*, and S/390* servers

Provides high performance with two powerful four-way RISC SMP processors, large cache, and serial disk attachment

Features industry-standard, state-of-the-art copy services—including FlashCopy*, Peer-to-Peer Remote Copy, and Extended Remote Copy—for rapid backup and disaster recovery

Uses redundant hardware and RAID 5 disk arrays to provide high availability for mission-critical business applications

Provides fast data transfer rates with attached hosts via Fibre Channel, UltraSCSI, ESCON*, and FICON interfaces

Increases administrative productivity by centralizing operations management and providing users with a single interface via a Web browser

Enables enterprises with multiple heterogeneous hosts to scale up to 11 TB while maintaining excellent performance

Shared storage for all major types of servers

The IBM Enterprise Storage Server* is a second-generation Seascape* disk storage system that provides industry-leading availability, performance, manageability, and scalability. Virtually all types of servers can concurrently attach to the Enterprise Storage Server—including S/390, Windows NT, Windows 2000, Novell NetWare, AS/400, and many types of UNIX servers. As a result, the Enterprise Storage Server is ideal for organizations with growing e-business operations that are being handled by multiple heterogeneous servers.

Enterprise-strength storage for distributed systems

With more business-critical information processing being performed on distributed systems (running several different operating systems), the IBM Enterprise Storage Server addresses the need to protect and manage distributed data with the same level of performance previously reserved for the mainframe environment. The IBM Enterprise Storage Server does more than simply enable shared storage across enterprise platforms—it can improve the performance, availability, scalability, and manageability of enterprise-wide storage resources through a variety of powerful functions:

- *FlashCopy* provides fast data duplication capability. This option helps eliminate the need to stop applications for extended periods of time in order to perform backups and restores.

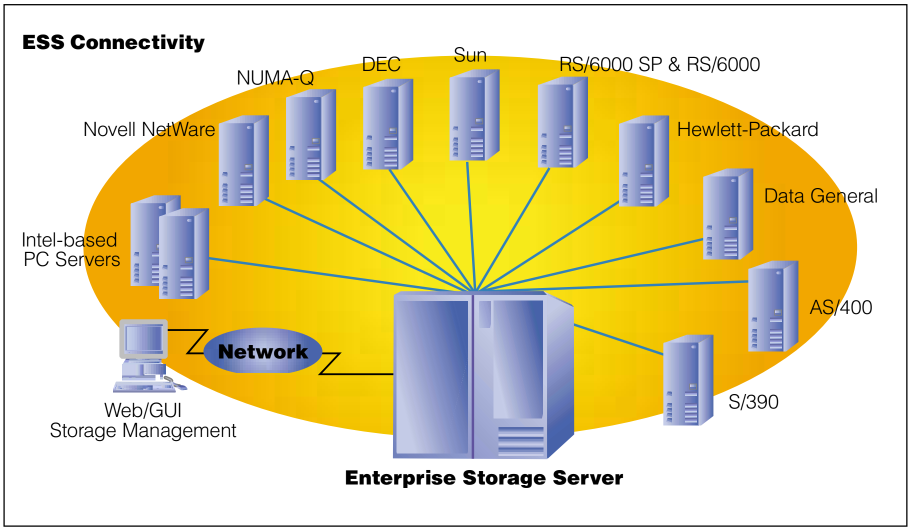
IBM Enterprise Storage Server

- *Peer-to-Peer Remote Copy* maintains a synchronous copy (always up-to-date with the primary copy) of data in a remote location. This backup copy of data can be used to quickly recover from a failure in the primary system without losing any transactions—an optional capability that can literally keep your e-business applications running.

- *Extended Remote Copy (XRC)* provides a copy of OS/390* data at a remote location (which can be connected using telecommunications lines at unlimited distances) to be used in case the primary storage system fails. The Enterprise Storage Server enhances XRC with full support for unplanned outages. In the event of a telecommunications link failure, this optional function enables the secondary remote copy to be resynchronized quickly—without requiring duplication of all data from the primary location—for full disaster recovery protection.

- *Custom volumes* enable volumes of various sizes to be defined for S/390 servers, enabling administrators to configure systems for optimal performance.

- *Storage partitioning* uses storage devices more efficiently by providing each server access to its own pool of storage capacity. Storage pools can be shared among multiple servers.



The IBM Enterprise Storage Server provides superior storage sharing for a range of servers.

High availability to safeguard data access

Support for 24x7 operations is built into the IBM Enterprise Storage Server. RAID 5 disk arrays help provide data protection while remote copy technologies allow fast data backup and disaster recovery. The IBM Enterprise Storage Server is a high-performance RAID 5 storage server featuring dual active processing clusters with fail-over switching, hot spares, hot-swappable disk drives, and nonvolatile fast write cache, and redundant power and cooling.

The IBM Enterprise Storage Server also contains integrated functions to help prevent storage server downtime by constantly monitoring system functions. If a potential problem is detected, the IBM Enterprise Storage Server automatically "calls home" to report the problem. A technician can be dispatched to make repairs, often before the problem is noticed by data center staff. Maintenance—including licensed internal code revisions—can typically be performed without interrupting operations.

Built-in flexibility

The IBM Enterprise Storage Server is a general-purpose disk system, providing outstanding flexibility with many options (IBM offers several predefined configurations to simplify ordering). The system consists of disk drives attached to a storage server via high-speed serial interfaces. A variety of host attachment options (UltraSCSI, ESCON, FICON, and Fibre Channel) enable the system to be optimized for the specific requirements of each computing environment.

Scalability for fast-growing environments

The IBM Enterprise Storage Server is especially designed for e-business and other applications with unpredictable growth requirements. It provides unprecedented scalability (up to 11 TB) while maintaining excellent performance. Disk drives for the IBM Enterprise Storage Server are provided as integrated packages of eight disk drives (known as eight-packs). Three disk drive capacities are available: 9 GB, 18 GB, and 36 GB. The server's base frame can hold a maximum of 16 eight-packs which, when used with 36 GB disks, yields a total capacity of nearly 3.4 TB. An add-on expansion enclosure is the same size as the base frame and can contain twice as many eight-packs—up to 256 hard disk drives—to deliver a maximum capacity of more than 11.2 TB.

Built-in investment protection

The IBM Enterprise Storage Server helps protect existing investments in IBM storage devices. For example, disk capacity from IBM Versatile Storage Server* frames and IBM 7133 Serial Disk System drawers (Models 020 and D40) can be attached to the IBM Enterprise Storage Server. Furthermore, first-generation (Models E10 and E20) Enterprise Storage Servers may be upgraded to the F-models, yielding up to 100% improvement in throughput. This upgrade protects customers' investments in ESS technology and enhances the scalability of installed Enterprise Storage Servers.

Performance enhancements for S/390 servers

Building on the capabilities of the IBM Versatile Storage Server and the RAMAC* Virtual Array family, the IBM Enterprise Storage Server improves function and performance for S/390 servers:

- *Multiple Allegiance:* This feature enables different operating systems to perform multiple, concurrent I/Os to the same logical volume—reducing queuing and significantly increasing performance. By enabling the Enterprise Storage Server to process more I/Os in parallel, Multiple Allegiance and optional Parallel Access Volumes can dramatically improve performance and enable more effective use of larger volumes. The result is simplified storage management at a reduced cost.

- *Parallel Access Volumes:* Previous S/390 systems allowed only one I/O operation per logical volume at a time. Now, performance is improved by enabling multiple I/Os from any operating system to access the same volume at the same time.

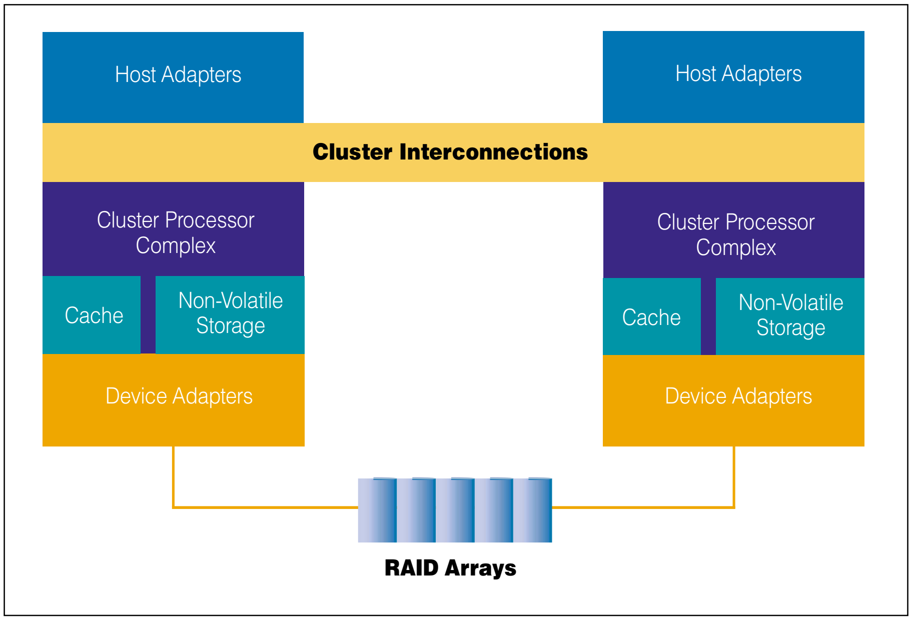
- *Priority I/O Queuing:* The storage server can ensure that important jobs have priority access to storage resources. With Priority I/O Queuing, the Enterprise Storage Server uses information provided by the OS/390 Workload Manager to manage the sequence in which I/Os are processed—matching I/O priority to your application priorities.

Enterprise-wide storage management to maximize productivity

IBM StorWatch* Enterprise Storage Server Specialist* is an integrated storage management tool that enables storage administrators to centrally monitor and manage IBM Enterprise Storage Servers. Using commonly available browser software, users can access the StorWatch Enterprise Storage Server Specialist tool from work, home, or on the road via a secure network connection. This enables additional control and flexibility in managing storage assets.

For more information

For more information, contact your IBM representative or IBM Business Partner or visit www.ibm.com/storage/ess.



The IBM Enterprise Storage Server logical structure provides outstanding flexibility to meet different price/performance requirements.

IBM Enterprise Storage Server at a glance

Characteristics	***Enterprise Storage Server 2105-F10***	***Enterprise Storage Server 2105-F20***
Disk storage capacity (baseframe)	420 GB to 1.68 TB	420 GB to 11.2 TB
Cache size	8 or 16 GB	8 or 16 GB
Host server attachments	Up to 32 SCSI or ESCON ports, up to 16 Fibre Channel ports, and intermix configurations	Up to 32 SCSI or ESCON ports, up to 16 Fibre Channel ports, and intermix configurations
Physical characteristics		
Dimensions	75.25" H x 54.50" W x 35.75" D (1913 mm x 1383 mm x 909 mm)	75.25" H x 54.50" W x 35.75" D (1913 mm x 1383 mm x 909 mm)
Weight	2160 lb. (980 kg)	2590 lb. (1175 kg)
Operating environment		
Temperature	60 to 90° F (16 to 32° C)	60 to 90° F (16 to 32° C)
Relative humidity	20 to 80%	20 to 80%
Wet bulb maximum	73° F (23° C)	73° F (23° C)
Caloric value	11,000 BTU/hr	16,000 BTU/hr
Power supply	Single phase 50/60 Hz	Three phase 50/60 Hz
Electrical power	3.5 kVA	5.0 kVA
Supported systems[1]		
S/390; AS/400 (9406 models); Data General; DEC; Hewlett-Packard (9000); Intel**-based PC servers; Novell NetWare; RS/6000*; RS/6000 SP; Sun**; NUMA-Q; Compaq		

[1] For more details on supported servers, visit www.ibm.com/storage/ess.

www.ibm.com/storage

IBM Storage Subsystems Division
5600 Cottle Road
San Jose, CA 95193

Produced in the United States
3-00

IBM hardware products are manufactured from new parts, or new and used parts. In some cases, the hardware product may not be new and may have been previously installed. Regardless, IBM warranty terms apply.

References in this publication to IBM products, programs, or services do not imply that IBM intends to make them available in all countries in which IBM operates.

GB equals one billion bytes when referring to hard drive capacity; accessible capacity may be less.

Product data is accurate as of initial publication and is subject to change without notice.

For Position Only

APPENDIX.D.2.1.4

IBM ESS Storage Performance White Paper

IBM Enterprise Storage Server

Performance White Paper

Version 1.0

IBM Storage Systems Division 24 September 1999

Authors: John Aschoff, Ruth Azevedo, Stu Goodgold, Joe Hyde, Lee LaFrese, Bruce McNutt, John Ponder, Brian Smith

General Editor: John Ponder

Table of Contents

Introduction . . . . - 3 -
Open Systems Performance with ESS . . . . - 4 -
OS/390 Performance with ESS . . . . **-** 15 -
PAVs and Multiple Allegiance . . . . - 22 -
DB2 Performance with ESS . . . . -33 -
TPF Performance with ESS . . . . - 38 -
Performance During Hardware Outage . . . . - 42 -
ESS Configuration Guidelines . . . . - 46 -
ESS Performance Monitoring . . . . - 54 -
Performance Marketing Tools . . . . - 56 -
Disclaimers, Acknowledgments, Trademarks . . . . - 58 -

Introduction

This paper provides performance information for the IBM Enterprise Storage Server (ESS), the 2105 disk storage subsystem. It is intended for use by IBM field personnel and their customers in making performance and capacity planning decisions regarding disk server solutions for their applications.

The ESS provides outstanding thruput (up to 230 MB/sec) and the greatest scalability (up to 11 TB) of any disk subsystem available at the initial publication of this paper. IBM incorporated the latest technology “horsepower” in order to achieve maximum thruput and performance -- PCI buses, advanced RISC microprocessors, SSA160, and 10K RPM disks. Superior "intelligence" - PAVs, Multiple Allegiance, and advanced cache algorithms -- deliver performance to the application when it’s needed, where it’s needed, virtually eliminating contention between systems or users for the same data resource. State-of-the-art RAID-5 design (with RAID-3 format writes) provides highest levels of data availability while exploiting the full bandwidth of the RAID rank.

ESS Performance Highlights:

- 32,000 ops/sec for 100% 4K read hits, open systems.
- 12,000 ops/sec for standard caching workload (70/30/50), open systems.
- 24,100 ops/sec for 100% 4K read hits, S/390.
- 11,300 ops/sec for cache standard workload, S/390.
- 230 MB/sec for sequential read operations from cache, 185 MB/sec for sequential reads from disk, and 145 MB/sec for sequential writes to disk.
- 800 MB/sec internal bus bandwidth.
- Eight device adapters, each of which can deliver up to 85 MB/sec sequential bandwidth.
- Up to 16 MB/sec bandwidth per ESCON channel.
- IOSQ and pend time sharply reduced or eliminated with PAVs and Multiple Allegiance.
  - PAVs allow up to 43 MB/sec data rate for a single logical volume (QSAM reads, 27K blocks), exploiting the bandwidth of the entire RAID rank.
  - RAID-5 and PAVs totally eliminate the "hotspots" associated with RAID-1 -- no need for balancing workload across logical volumes. This is a cost-of-ownership benefit, because fewer person-months will be spent on workload balancing to manage hotspots!
  - PAVs allow use of large logical volume images without creating hotspots or queuing issues, thus freeing up addresses and making space management easier.

With these stunning levels of performance and industry-leading features, the ESS is the most significant disk subsystem to be offered for enterprise computing since the advent of cached disk controllers in the early 80’s.

Chapter 1. Open Systems Performance with ESS

The following section describes the performance of ESS in open systems environments. The major comparisons include:

- Comparisons to the IBM Versatile Storage Server (VSS). ESS provides major architectural improvements beyond the VSS, resulting in overall improved performance. The improvements and comparisons are outlined in this section.
- Comparisons to the IBM 7133 and Serial Storage Architecture. ESS will provide better performance for most general purpose workloads, compared to SSA. However, there are some high-bandwidth applications for which native-attached SSA are particularly well suited, and some care should be taken in planning the right performance solution.

Comparisons to Versatile Storage Server

The IBM Enterprise Storage Server provides substantial performance improvements over the IBM Versatile Storage Server. Although these products share similar technology, many of the underlying hardware components have changed, incorporating the latest available technology. As a result, the performance of the ESS disk system is vastly better. The majority of improvements in ESS performance come from the following improvements in underlying architecture.

- ***Bus architecture.*** The ESS uses 6 PCI buses (compared to 4 PCI buses on VSS), to handle data transfers between host adapters, cache, and device adapters. This provides a total of approximately 800 MB/sec of internal bandwidth. In addition, efficiency in use of the buses has been tuned and improved. As a result, the sequential throughput has increased more than two times compared to VSS.
- ***Non-volatile memory.*** The amount of non-volatile memory has increased from 32 MB to 384 MB, allowing the disk system to better handle write-intensive workloads. In addition, the NVS is directly connected via PCI buses to custom-designed host interface adapters, allowing better parallelism in transferring two copies of cached data (one to the volatile cache and one to the non-volatile cache). Two copies of data are always maintained in cache, to guarantee the safety and integrity of the customer’s data. This results in cutting in half the service time of writes to cache compared to VSS (from about 1.8 ms down to 1 ms for a 4 KB record.)
- ***Disk.*** The latest technology 10K RPM disks and SSA160 disk interconnection technology is used. This improves the disk bandwidth, and reduces response time and increases throughput on cache-unfriendly workloads.
- ***SCSI adapters.*** Specially-designed host interface adapters allow higher efficiency use of the SCSI bus, and optimize writing multiple copies of data to cache and non-volatile storage. In addition, the number of UltraSCSI host interfaces has doubled (from 16 to 32), allowing increased host connectivity and throughput improvements for workloads constrained by the number of SCSI ports.

The results of these improvements can be seen in Table Open-1, showing comparisons of maximum throughputs achievable with the two products. This data was obtained directly from laboratory measurements. These measurements were obtained with a highly-tuned configuration.

Workload	VSS 6 GB cache 32 MB NVS 8 device adapters 128 7200 RPM disk 16 UltraSCSI	ESS 6 GB cache 384 MB NVS 8 device adapters 128 10K RPM disk 32 UltraSCSI
Sequential reads from disk - 64 KB	80 MB/sec	160-185 MB/sec
Sequential writes to disk - 64 KB	75 MB/sec	145 MB/sec
100% cache read hits - 4 KB records	26,000 SIO/sec	32,000 SIO/sec
Random reads - 4 KB records - no cache hits	8,000 SIO/sec	9,500 SIO/sec
Random 70/30 read/write ratio- 50% read cache hits - 4 KB records	9,500 SIO/sec	12,000 SIO/sec
Random writes to disk- 4 KB records	2,000 SIO/sec	3,000 SIO/sec
Writes - 100% cache hits - 4 KB	9,800 SIO/sec	12,000 SIO/sec

Table Open-1[1]

Figure Open-1 The following graphs shows performance comparisons of ESS to VSS for a standard open system workload.

Read intensive cache unfriendly workload. This workload is characterized by very random 4 KB reads. The accesses are extremely random, such that virtually no cache hits occur in the external cache. This might be representative of some decision support or business intelligence applications, where virtually all of the cache hits are absorbed in host memory buffers.
Standard workload. This workload is characterized by random access of 4 KB records, with a mix of 70% reads and 30% writes. This is also characterized by moderate read hit ratios in the disk system cache (approximately 50%.). This workload might be representative of a variety of online applications (e.g., SAP R/3 applications, Oracle, file servers, etc.)

This graph illustrates the general improvements in both response time and throughput for ESS.

[1] All throughput measurements stated in terms of MB/sec assume 1 MB = 1,000,000 bytes.

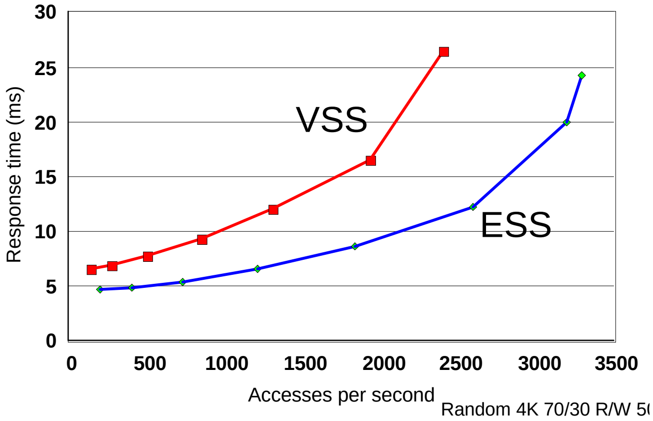


Figure Open-1

Performance Considerations for 7133 customers

Which provides higher performance: Serial Storage Architecture (SSA) with IBM 7133 or Enterprise Storage Server?

This question is asked frequently by open-systems users, and is one of the most difficult for which to provide a standard answer. SSA technology has continued to evolve and lead the industry in performance, and has unique performance attributes that in some environments can be difficult to match. There are many variables that determine the answer:

- Many users are not constrained at all by the performance of their disk system, and are perhaps more affected by processor or network performance. These users might see no difference in performance, regardless of the disk system. These users should certainly consider the many other factors that differentiate the products, and not worry about performance.
- Realizing the full potential of SSA technology often requires good use of facilities in the operating system, file systems, and data base management systems. For example, open

systems users can traditionally achieve tremendous benefit from the use of data caching in the host, either through the file system, data base management system, or logical volume manager (e.g.., caching of Virtual Shared Disks on RS/6000-SP). Likewise, logical volume managers often provide mechanisms for avoiding disk hot spots by striping data across multiple disks. For example, most AIX users take advantage of physical partition striping in AIX.

- ESS will provide equal or better performance for most workloads relative to typical SSA configurations. However, there are some workloads that have particularly high bandwidth requirements, transferring very large quantities of data. These are typically decision support and data warehousing applications. This type of workload may require more careful planning with use of ESS.
- ESS's robust design (including read and write cache, striped RAID-5 for balancing disk load, and SSA disk technology) can reduce much of the manual tuning, system administration, and use of host software features that are needed with the simpler native disk attachment design of 7133. This may be particularly important for organizations who must support mixtures of IBM and non-IBM system platforms.

When will ESS provide improved performance relative to SSA and 7133?

ESS would likely perform better than 7133 in the following circumstances:

- *Cache-constrained environments*. With use of 7133 (or any other storage device, for that matter), liberal use of host memory for caching data can be the most effective means of improving disk performance. This may mean allocating large buffers to a DBMS, large memory in the processor, and use of caching with VSDs, etc. However, there may be reasons that constrain a customer's use of inboard caching, ranging from not having sufficient memory in the processor, software and addressing constraints, or simply neglecting to allocate sufficiently large data base buffers. In these cases, the ESS cache can generate very good cache hit ratios, improving disk response time and throughput.
- *Write cache* can be especially important for some applications, such as data base loads, some batch applications, and data base logging. The primary benefits of write cache are reduction in response time for writes, and throughput improvement for applications that write short records sequentially. This may also be especially beneficial for AIX applications that must use mirror-write-consistency (MWC).
- *Non-optimized disk layout.* Many performance bottlenecks are the result of contention on one or a handful of disk drives. RAID-5 striping in ESS automatically balances load across all of the disks in a RAID array. Although AIX provides the ability to load balance with PP striping, not everyone uses the function. Those users could see substantial performance benefit for some disk-constrained workloads.
- *Older SSA technology.* SSA technology has continued to improve, with faster link speeds and adapters (SSA160), faster disks (10K RPM), larger write cache, and RAID.

The following graphs show relative performance of SSA compared to ESS for two representative random-access workloads: the same ***read-intensive cache-unfriendly*** workload and ***standard*** workload described above.

- **Standard workload.** This workload is characterized by random access of 4 KB records, with a mix of 70% reads and 30% writes. This is also characterized by moderate read hit ratios in the disk system cache (approximately 50%.). This workload might be representative of a variety of online applications: SAP R/3 applications, Oracle, file servers, etc.
- ***Read intensive cache unfriendly workload.*** This workload is characterized by very random 4 KB reads. The accesses are extremely random, such that virtually no cache hits occur in the external cache. This might be representative of some decision support or business intelligence applications, where virtually all of the cache hits are absorbed in host memory buffers.

Standard workload

The following shows SSA technology compared to ESS for a standard open systems workload. This workload might be representative of a variety of online applications: SAP R/3 applications, Oracle, file servers, etc. This workload has the following characteristics:

- Random 4K operations
- 70% reads and 30% writes
- For ESS, approximately 50% of the reads are hits in the ESS cache.
- For SSA and 7133, since there is no read cache, all of the read operations are disk accesses. All of the SSA measurements are using AIX mirroring without mirror write consistency.

For this workload, the configurations compared are the following:

1. SSA80 - 4 adapters - 64 7200 RPM disks - (This curve is projected from a measurement result using 2 adapters and 32 disks.)
2. SSA160 - 2 adapters - 32 7200 RPM disks
3. SSA160 - 2 adapters - 32 10K RPM disks - using 32 MB of write cache in each adapter
4. ESS - using 4 RAID arrays (32 disks)

Observations

This shows that good use of cache with ESS provides benefits over both the older and current SSA disk offerings.

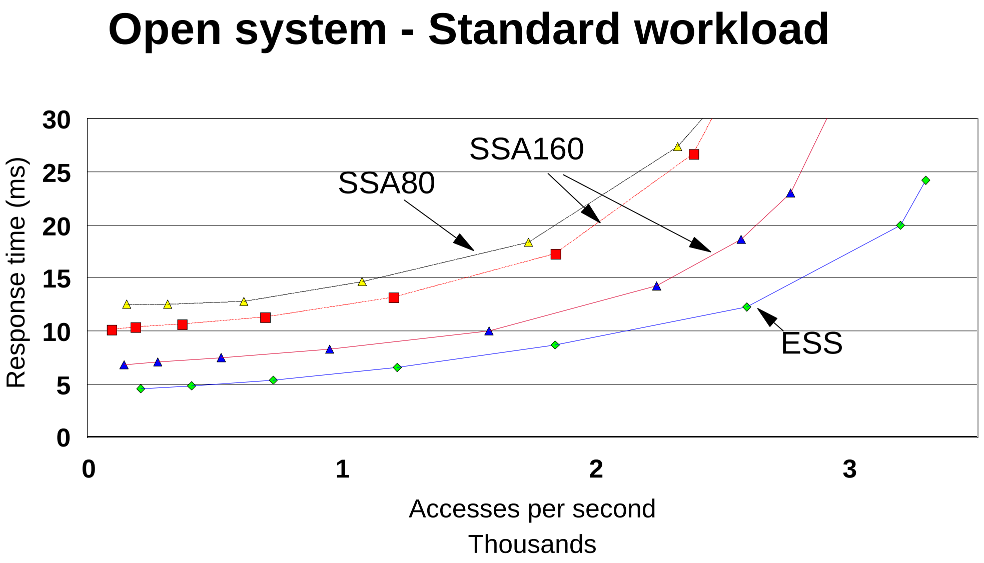


Figure Open-2

Read intensive cache unfriendly workload

Figure Open-3 The following graph shows comparisons of several SSA configurations compared to ESS. The following shows a comparison of two "older" SSA configurations for a very cache-unfriendly workload. This workload might be representative of some decision support applications in which caching is handled within the host processor. It is important to remember that this workload has **no** cache hits, sees no benefit from ESS caching, and is therefore a reasonable worst-case workload. One would ordinarily expect native disks to perform the best in this kind of environment.

The configurations shown are:

1. SSA80 with 32 disks. This uses AIX mirroring (RAID-1) with 2 SSA80 adapters, 16 disks on one adapter mirrored to a second adapter. It assumes that the workload is perfectly well-balanced across all of the disks, and therefore represents a very highly-tuned and optimal configuration.
2. SSA80 with 64 disks. This curve is an extrapolation of what would happen by increasing the number of SSA disks (64 disks and 4 SSA80 adapters).
3. ESS using 32 disks (4 RAID-5 arrays).

Observations

- This type of workload receives no real benefit from use of the cache in ESS. Despite that fact, the fast disks (10K RPM) and SSA160 technology provide better response time and throughput than the older SSA 32-disk configuration, disks and adapters up to approximately 2000 I/O per second, or around 500 I/O per second per array.

- For this type of workload, more disk arms helps improve throughput, as seen in the SSA 64 disk case. However, this relies on the ability to spread the workload evenly to all of those disks. This may be realistic only in some of the most highly-tuned configurations, as might be seen with parallel applications, such as DB2 UDB on RS/6000 SP environments. In this case, ESS provides better response time than the 64-disk SSA configuration up to approximately 2000 I/O per second, or roughly 80 I/O per second per disk.

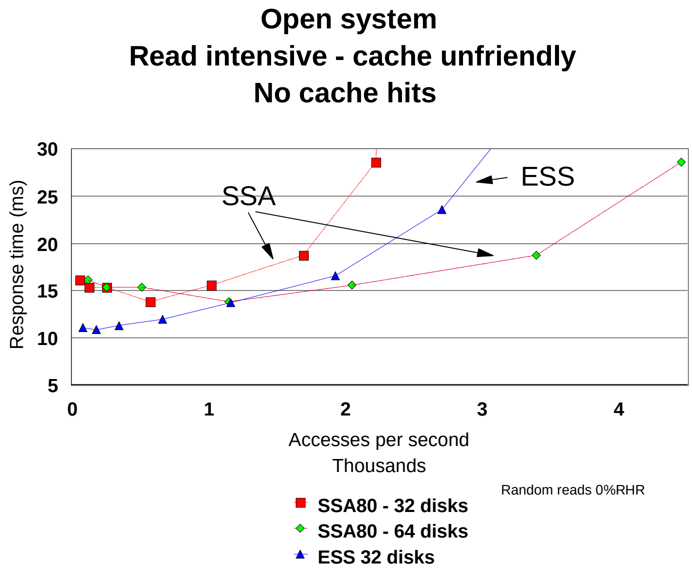


Figure Open-3

Figure Open-4 The following chart shows results for the same workload comparing the latest SSA disk technology to ESS. In this case, we are showing the following:

1. SSA160 with 32 7200 RPM disks. This uses AIX mirroring (RAID-1) with 2 SSA adapters, 16 disks on one adapter mirrored to a second adapter. It assumes that the workload is perfectly well-balanced across all of the disks, and therefore represents a very highly-tuned and optimal configuration.
2. SSA160 with 32 10K RPM disks.
3. ESS using 32 disks (4 RAID-5 arrays).

Observations

For this very cache unfriendly read-oriented workload, these results show slightly better performance for the latest SSA technology compared to ESS.

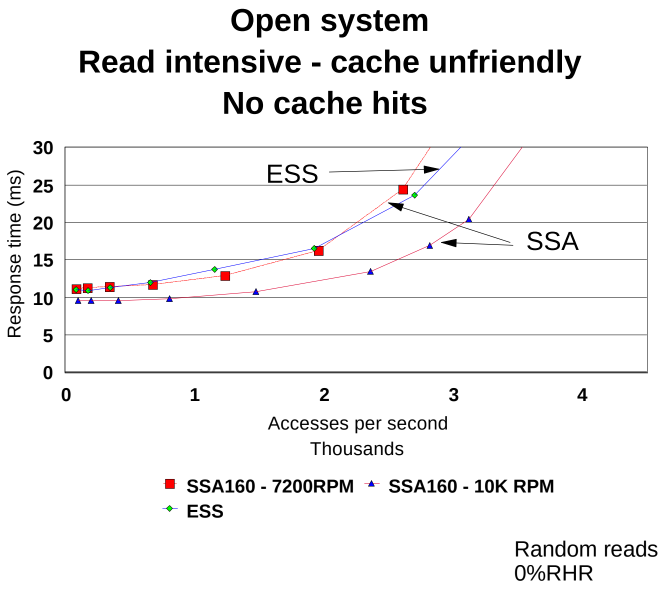


Figure Open-4

Use of SSA technology for high bandwidth applications

Some applications have particular requirements for high data rates, that is, very high sustained megabytes per second of data transferred. For example, **decision support applications** may generate parallel queries that scan data base tables in parallel, sending very large volumes of data to several host processors. An example of such an application might be implemented with a DB2 UDB and IBM RS/6000 SP configuration. For example, TPC-D benchmarks and user environments operating parallel queries can have sustained sequential read data rate requirements in excess of 1000 MB/sec.

Two SSA160 adapters operating on an RS/6000 F50-class processor with two PCI buses can achieve approximately 180 MB/sec of sustained sequential read throughput, or roughly equal to the sustained sequential read throughput capability of an ESS disk system. If an application requires sustained throughputs beyond that of a single ESS, then it is worth considering the choice of using 7133 and multiple SSA160 adapters. Most applications probably do not have requirements near these capabilities, except for some of the most demanding decision support and data mining applications.

Performance Measurements for Various Workloads

The following performance curves are intended to help determine the capability of a single ESS disk system to provide adequate performance for normal random-access type workloads. The following graphs and tables show how ESS performs for random workloads with various record sizes, read/write ratios, and cache hit ratios. These charts are based on actual laboratory measurements using the following configuration:

- 32 UltraSCSI ports
- 128 10K RPM disks
- 8 Device adapters

The workload profiles shown in these graphs are:

- 100% random reads, no cache hits
- 70% random reads, 30% writes, no read hits in cache
- 50% random reads, 50% writes, no read hits in cache
- 70% random reads, 30% writes, 50% read hits in cache

You can use these charts to help determine how many ESS disk systems are required to meet your performance needs.

- Select a workload profile that reasonably matches your application workload characteristics.
- Select a target I/O response time. (20 ms or less is usually a reasonable target.)
- Determine the throughput that a single ESS disk system can sustain.

These performance curves were obtained for a disk system with 128 disks. Configurations with 96 disks would achieve approximately 75% of the shown throughputs, and those with 64 disks would achieve approximately half the shown throughputs. You can use Disk Magic for evaluation of other workloads and configurations.

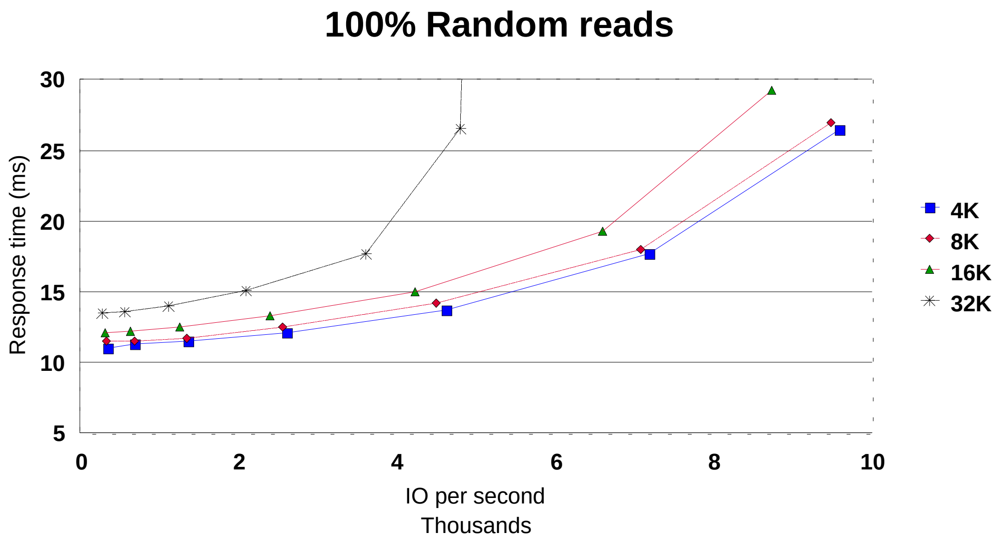


Figure Open-5

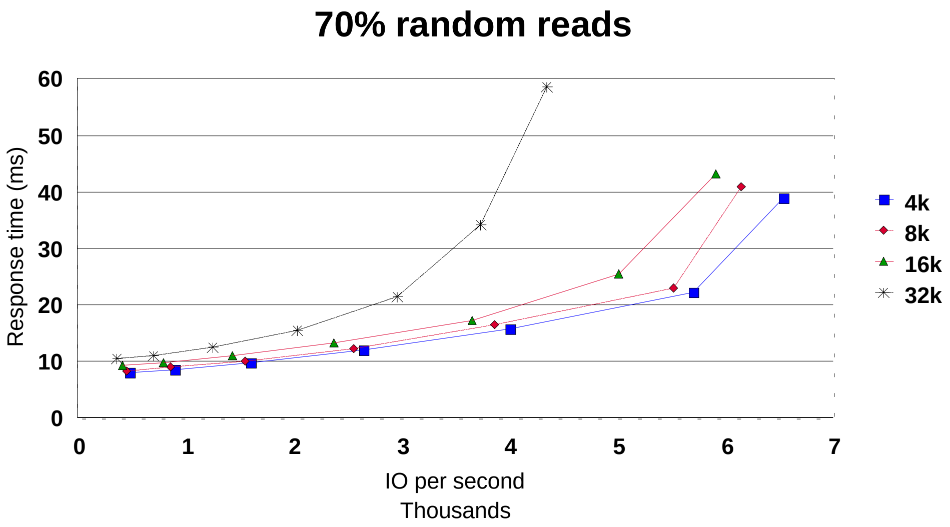


Figure Open-6

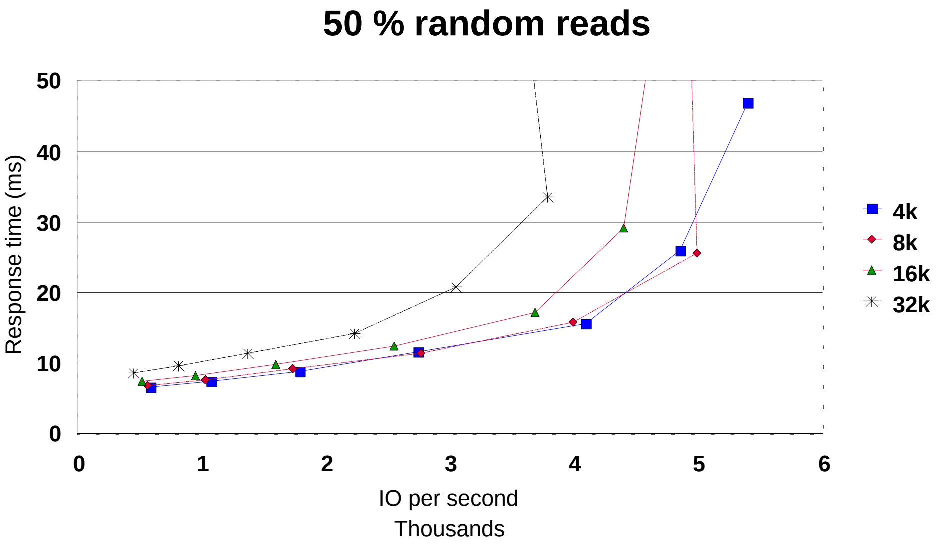


Figure Open-7

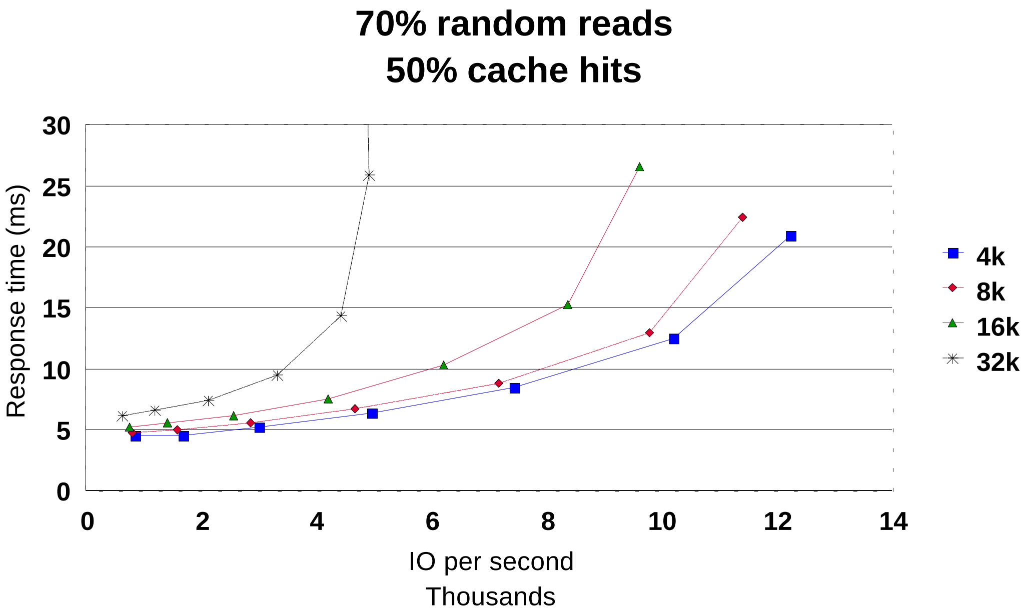


Figure Open-8

Chapter 2. OS/390 Performance with ESS

This section examines raw ESS performance in an OS/390 environment. We present the results of both random access and sequential benchmarks, including a range of hit ratios. In the benchmark tests of this section, increasing levels of I/O load are applied up to the maximum throughput of the subsystem.

No use of PAVs are made in the tests of the present section. Partly, this is because ESS does not require PAVs to achieve extraordinary levels of throughput. Also, the central advantage of PAVs is their ability to bring additional I/O processing resources to bear when these are needed by individual volumes. When *all* active volumes encounter stress loads, as in the tests of the present section, performance on one volume can only be boosted at the expense of another. For this reason, the benchmarks that are best suited to exploring the benefits of PAVs are those which examine specific, identified volumes or sets of volumes. Benchmarks of this type are presented below in the section "PAVs and Multiple Allegiance".

Database Performance

We begin with random access database performance, since the throughput for this type of workload is central to configuration planning in most OS/390 environments. Database workload conditions are tested using the Performance Assessment Workloads (PAWs) suite of database benchmarks. The test conditions are representative of a CICS data base environment, with a range of locality behavior from cache hostile (typical hit ratio 40 percent) to cache friendly (typical hit ratio 90 percent). The most representative level of cache locality is reflected by the PAWs "cache standard" test (typical hit ratio 70 percent in the tested ESS configurations). With one exception (the PAWs "uniform" test), a realistically severe skew is applied across the logical volumes tested. The transfer size for the database tests is 4K.

Configuration	Storage Capacity (GB)	ESCON Channels	Disk Drives (no. and type)	Cache Size (GB)
RVA T82	420	8	32 x 4.5 GB	4
ESS/8 ESCON	420	8	32 x 18 GB	3
ESS/16 ESCON	1260	16	128 x 9,18 GB	6
ESS/32 ESCON	840	32	128 x 9 GB	6
ESS/LRG DRV	1680	32	64 x 36 GB	6

Figure S390-1 *Tested configurations*

Figure S390-1 presents the tested configurations. The initial configuration of the table is the RVA Model T82. This extremely successful product, offered by IBM since August, 1997, is familiar to most IBM storage customers, and provides a useful base of comparison.

The ESS/8 ESCON configuration is well suited for comparison to the RVA Model T82. This configuration, which closely matches the RVA Model T82 in terms of ESCON channels, storage capacity, cache, and number of disk drives, consists of an ESS A-box, 25 percent populated with 18 gigabyte drives, and attached via eight ESCON channels.

The ESS/16 ESCON configuration is an ESS A-box, fully populated with an equal mix of 18 gigabyte and 9 gigabyte drives (128 drives altogether). Since both drives feature a rotation rate of 10,000 RPM, the performance of any single drive of one type is at an approximate par with a single drive of the other type. For this reason, the test results for the ESS/16 ESCON configuration reflect the performance which should be expected from any mix of 128 9 and 18 gigabyte drives, when configured with the same cache size and number of ESCON channels--for example, an 840 gigabyte configuration featuring 9 gigabyte drives, or a 1680 gigabyte configuration featuring 18 gigabyte drives. Based upon the results reported in the present chapter, the latter configuration (an ESS A-box with 128 18 gigabyte drives, 6 gigabytes of cache, and 16 ESCON channels) can be recommended as a reasonable "starting point" for configuration planning in most customer environments.

The ESS/32 ESCON configuration is included to demonstrate the maximum performance capability of ESS. It is an ESS A-box, fully populated with 9 gigabyte drives, and attached via 32 ESCON channels. For the same reasons as those just discussed in the previous paragraph, the test results for the ESS/32 ESCON configuration reflect the performance which should be expected from any mix of 128 9 and 18 gigabyte drives, when configured with the same cache size and number of ESCON channels. Thus, the central difference between the ESS/16 ESCON and ESS/32 ESCON configurations is the doubled number of ESCON channels offered by the latter.

Many customers find themselves in a position where they must be conservative in their use of ESCON attachments. It is anticipated that such customers, when running with a single processor, will prefer to attach the ESS with 8 or 16 (but not 32) channels. In more complex environments, 24 or 32 channel configurations may be attractive. For example, a 32 channel configuration can be attached to up to four processors without requiring the use of an ESCON port director. As another example, a 24 channel configuration might make it possible to set aside 8 channels for later use as PPRC links or as attachments to a planned XRC data mover.

The final row of Figure S390-1 shows the use of the 36 gigabyte drive option for configuring ESS. It is anticipated that many customers will be interested in this option due to its lower storage cost. The rotation rate of the 36 gigabyte drive differs, however, from that of the 9 or 18 gigabyte drives (7200 RPM versus 10,000 RPM). Our test results for the ESS/LRG DRV configuration allow customers to make an informed decision about the resulting tradeoff of storage cost versus performance. These data might best be compared with a configuration featuring the same cache size and number of ESCON channels, and a matching storage capacity

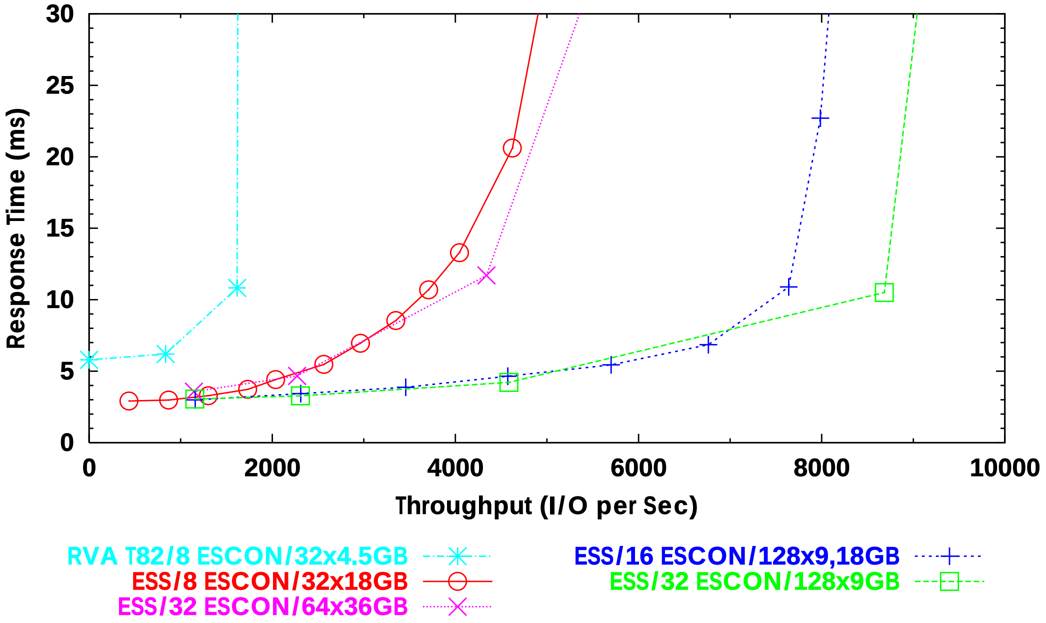


Figure S390-2 *Cache standard benchmark. Typical read hit ratio: 70 percent. R/W ratio: 3. Transfer size: 4K.*

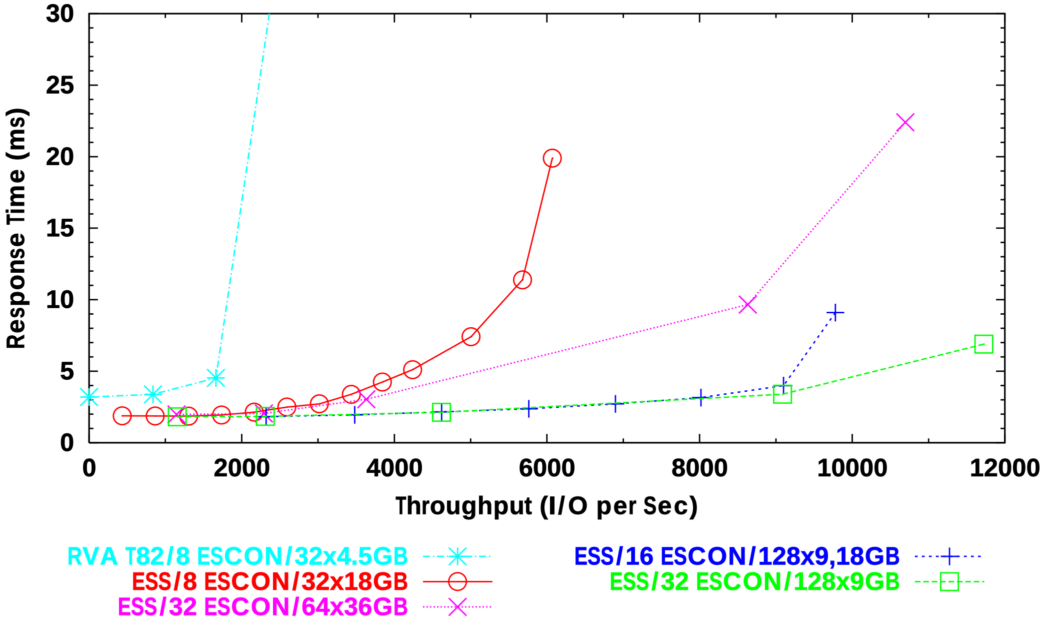


Figure S390-3 *Cache friendly benchmark. Typical read hit ratio: 90 percent. R/W ratio: 5. Transfer size: 4K.*

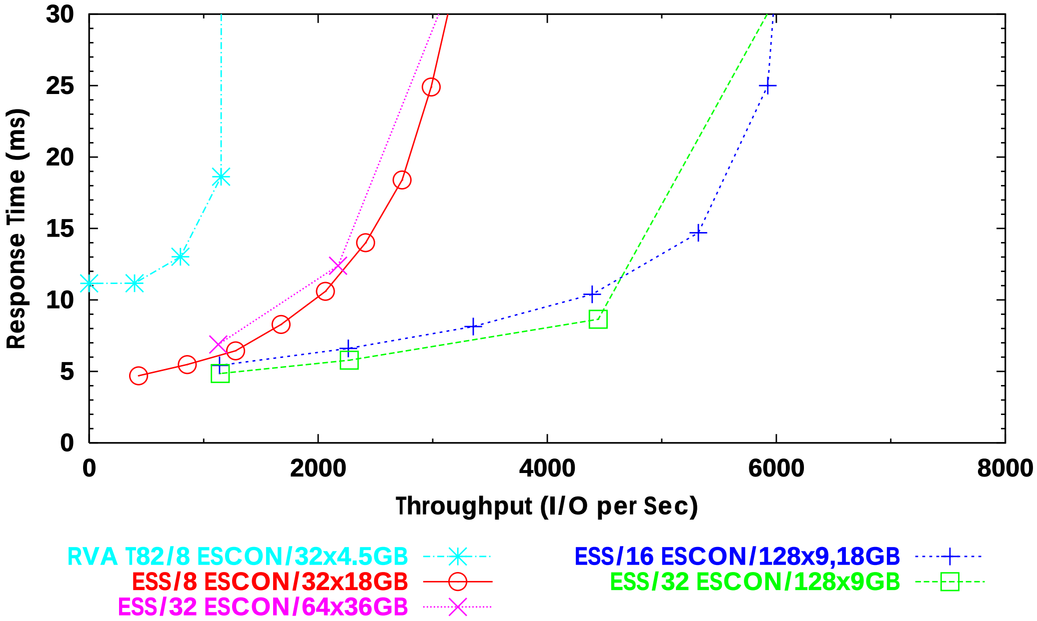


Figure S390-4 *Cache hostile benchmark. Typical read hit ratio: 40 percent. R/W ratio: 2. Transfer size: 4K*

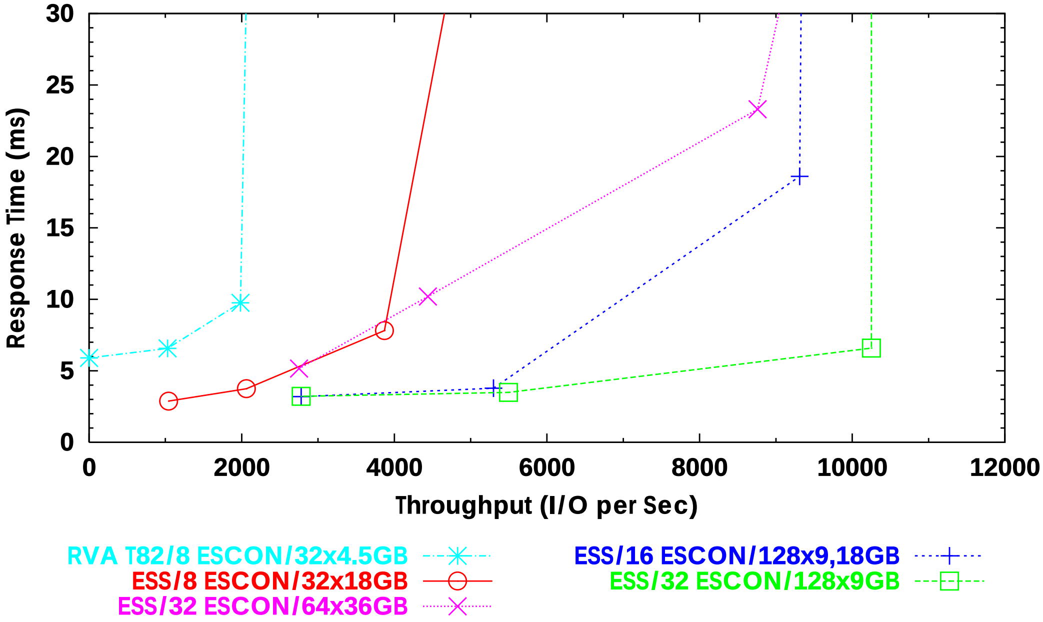


Figure S390-5 *Uniform benchmark. Typical read hit ratio: 80 percent. R/W ratio: 3.4. Transfer size: 4K.*

populated with 128 18 gigabyte drives. For the reasons just discussed in the preceding paragraphs, the results of such a comparison can be judged by referring to the ESS/32 channel configuration.

Figures S390-2 through S390-5 present the results of the database testing. In *every* database test, covering a wide range of workload conditions, the maximum throughput of the ESS/8 ESCON configuration exceeds that of the RVA Model T82 by a factor of two to three times. The maximum throughput of the ESS/16 ESCON configuration, in turn, exceeds that of ESS/8 ESCON by a factor of 1.5 to 2 times.

The results of the "cache standard" test should be noted particularly, since they best reflect a typical database environment. On this test, the ESS/8 ESCON configuration's maximum throughput exceeds that of the RVA Model T82 by a factor of three times. The ESS/16 ESCON configuration yields a further gain of 62 percent, or nearly *five times* the throughput of the RVA Model T82.

The ESS response times, at a given load level, are dramatically less than those of the RVA Model T82. Again using the cache standard workload as an example, the response time of the ESS/8 ESCON configuration is 3.0 milliseconds at a load level of 870 I/Os per second; the response time of the RVA Model T82 is more than twice as long at the same approximate load.

Sequential Performance

In many customer environments, sequential performance is critical due to the heavy use of sequential processing during the batch window. The types of sequential I/O requests that play an important role in batch processing cover a wide range. This section examines QSAM and VSAM sequential performance in detail. These specific forms of sequential processing are important in themselves, and also provide an indication of what level of performance might be expected for sequential processing in general.

Figures S390-6 and S390-7 present IBM's sequential test results. Each result shows a selected access method (QSAM or VSAM) in combination with a selected type of operation (read or write). Buffering was set so as to achieve transfer sizes per I/O of 2.5 tracks for QSAM, and 1 track for VSAM (5 buffers and 24 buffers respectively, with half of the VSAM buffers being used for any given I/O).

Figures S390-6 and S390-7 show benchmark results both for a single operation against a single data set (one "stream"), and also for multiple operations ("streams") running at the same time against multiple data sets. The case of a single stream is representative of a single batch job running in isolation; the case of eight streams is intended to reflect moderate-to-heavy batch window loads. The interpretation of the 16 stream case depends upon the number of ESCON channels. For a configuration with 8 channels (such as the RVA Model T82 or ESS/8 ESCON), the case of 16 streams represents a stress condition, under which the highest possible aggregate data rate is achieved at the cost of sharply elongating the elapsed time of any individual stream (for example, large-scale volume dumps might be performed in this manner). By contrast, an

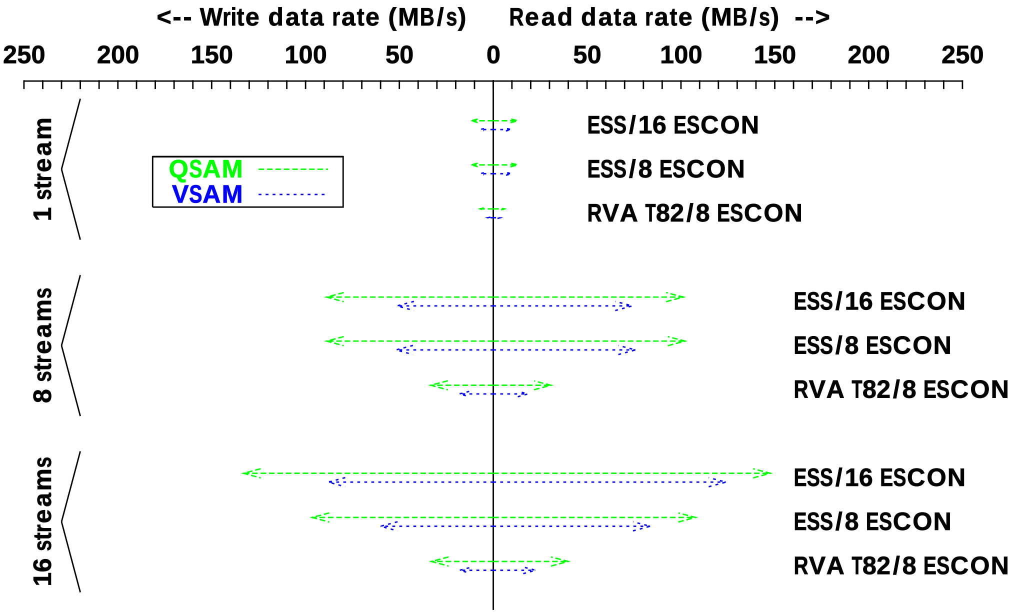


Figure S390-6 *Aggregate sequential throughput. Block size: 27K (QSAM), 4K (VSAM). Transfer size per I/O: 2.5 tracks (QSAM), 1 track (VSAM).*

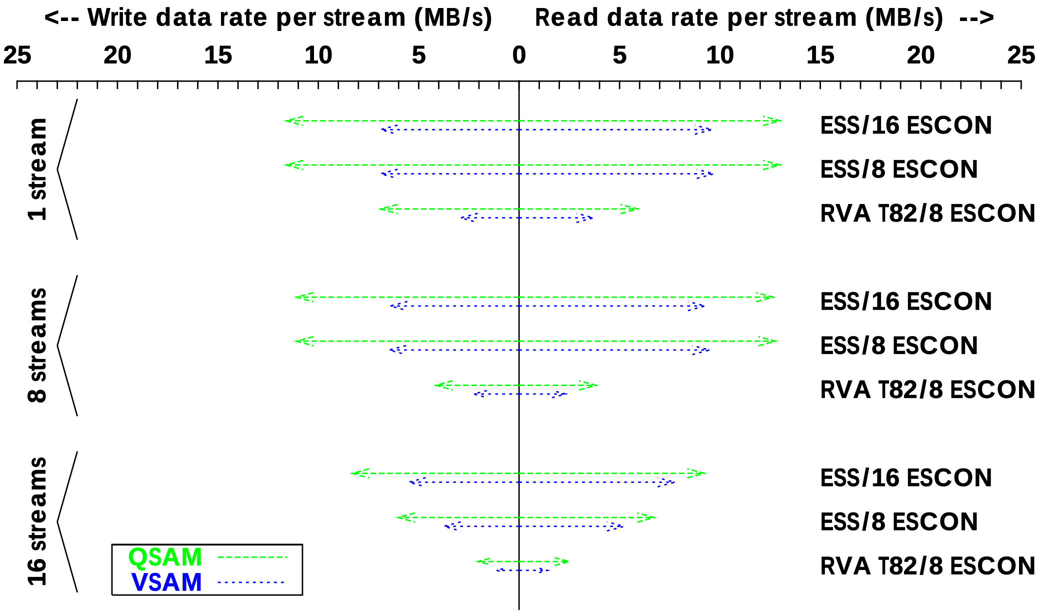


Figure S390-7 *Sequential throughput per stream. Block size: 27K (QSAM), 4K (VSAM). Transfer size per I/O: 2.5 tracks (QSAM), 1 track (VSAM).*

ESS configuration with 16 or 32 channels can accommodate 16 sequential streams while still delivering high throughput for each individual stream.

Each sequential test result is presented in two ways:

- Figure S390-6 shows the total data rate, taking into account all streams (usually the most important metric).
- Figure S390-7 shows the average data rate achieved by an individual stream. This metric may be of interest when examining the performance of individual batch jobs.

For every sequential measurement, with just one exception, the data rate of the ESS/8 ESCON configuration exceeds that of the RVA Model T82 by at least a factor of two. The exception is single-stream QSAM writes, where the ESS/8 ESCON configuration exceeds the data rate of the RVA Model T82 by two thirds (a factor of 1.67).

Test results show that ESS sequential performance depends primarily upon the nature of the sequential workload and the number of attached ESCON channels, as suggested by Figures S390-6 and S390-7. Sequential performance does not depend significantly upon the choice of 9, 18, or 36 gigabyte drives.

It should be noted that ESS offers new capabilities intended to support the optimization of sequential performance -- improved and more sensitive pre-fetching algorithms, and new channel commands which improve on-channel overheads and allow for increased bandwidth.. The tests results just reported do not reflect the new sequential optimizations. As one important example of their potential, ESS can deliver a data rate of 8.4 MB/second for the DB2 log (8.0 MB/second for a dual log). As another example, single-stream read and write data rates of as high as 15.8 and 14.5 MB/second, respectively, have been measured using extended format sequential data sets with half-track blocking. All DB2 and extended format I/O operations are performed by the Media Manager, and in this way benefit from the new ESS sequential optimization features.

Chapter 3. Parallel Access Volumes & Multiple Allegiance

Introduction

The UCB (Unit Control Block) has always been the point of control for access to devices in S/390. A request to do an I/O operation on a device must wait until the device is available, as indicated in the UCB. If the device is busy servicing another I/O, the new request must wait or queue on the UCB. The operating system samples the length of the queue on the UCB, and reports the average amount of wait as IOSQ in the RMF performance report.

For years, IOSQ time has been used as a significant indicator of performance at the device level. Applications that require the very best performance often tolerate little or no IOSQ time. Applications are designed and datasets are placed on logical volumes to minimize or eliminate IOSQ. Since the UCB is the point of access control to a device in S/390, this sometimes means that only a single important dataset is allocated on a device in order to guarantee that only a single known application will access the device, never having to wait on some other application or host to finish its I/O to the same device.

For multi-host systems, where two or more OS/390 hosts share the same disks, the UCB may show that a device is available, but when the I/O is attempted, the control unit for the device may respond with a “device busy” condition, indicating that the device is servicing an I/O for another host system. When the I/O finishes, the control unit tells the waiting host that the device is now free, and the I/O is attempted again. This type of wait (on an I/O from another system) shows up as PEND time. Some control units have the capability of accepting an I/O even when the device is busy on behalf of another system, queuing the request in the control unit until the device is available. This time is accumulated in the Control Unit Queuing field of the channel measurement block (CMB). The wait time in the control unit is reported separately as device busy delay time (AVG DB DLY) and is also part of PEND time in the RMF report.

Multiple Allegiance and Parallel Access Volumes

With the advent of OS/390 V1.3 (and later), and the ESS 2105 I/O subsystem from IBM, this state of affairs has dramatically changed, permitting multiple I/Os to be concurrently active on a given device at the same instant in time. In the case that the I/O come from different systems, this capability is called “multiple allegiance”, and when the I/O come from the same system, the feature is called “Parallel Access Volumes” or PAVs. PAVs are devices characterized by a base address and zero or more alias addresses. RMF rolls up the performance characteristics for all the addresses and summarizes the statistics on the base address

Multiple allegiance will allow two hosts, for example, to access two different datasets on the same logical volume at the same time. Similarly, PAVs will allow two different applications on the same system to access two datasets at the same time. There are some restrictions which apply. It is not possible for the two systems or applications to write the same records at the same time, and not advisable for one to read the same records that are being written by another application or system. The actual control and synchronization of these features relies on cooperative support in the operating system (OS/390), the access methods (VSAM or Media Manager, for example), and the I/O subsystem microcode.

The restrictions which apply are not stringent; the I/O operations that need to operate in parallel should involve different "extents" on the logical device. The idea of an extent in an I/O operation is a security feature in OS/390. Individual I/O operations are restricted to operate in a limited range on the logical device. Software in the operating system specifies the extent, while the microcode in the subsystem enforces the restriction. Some applications or access methods permit an I/O to operate anywhere within a dataset. Such a gross level of extent specification and enforcement means that two I/Os from these applications or access methods must operate on different datasets to achieve parallel service.

Putting it all together, OS/390 has been modified to allow multiple, concurrent I/O to a given device, relaxing the old use of UCB to enforce serialization to a device. The UCB is actually an artifact deriving from when a single I/O could only be serviced by a single actuator. Key applications and access methods have been modified to provide more precise "extent" information in each I/O operation, and the microcode in the I/O subsystem has been designed to enable and enforce the exploitation of multiple concurrent I/O to the same logical device, while still guaranteeing the same level of data consistency that OS/390 has always provided. Where applications have not been modified, the advantages still accrue if multiple datasets are involved.

The concept of PAVs is similar to that of "multiple exposures" as implemented for the IBM 2305 drum device, often used for paging and for job queue datasets in the days before cached control units. The cached 3880-21 also supported multiple exposures for paging. Cache technology and RAID-5 technology really leverage the feature for applications far beyond paging

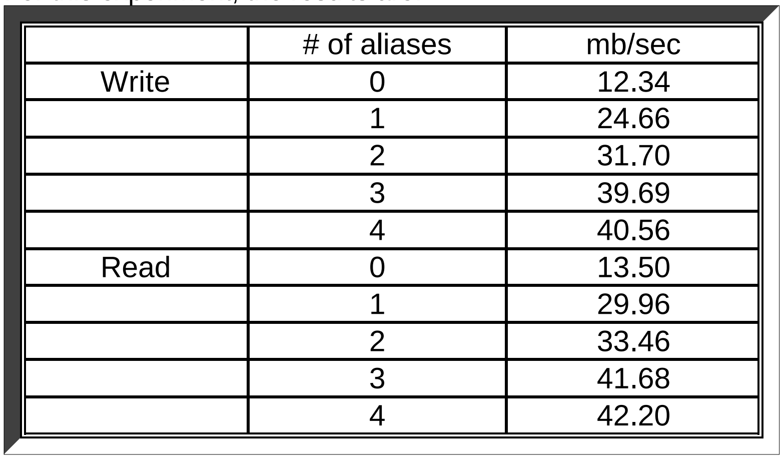
For example, on a single logical volume, the ESS subsystem can service as many concurrent hits as there are channel paths to the device (say eight). For cache misses, the ESS can service multiple misses concurrently, exploiting the multiple physical devices in a RAID-5 array. Perhaps the most common situation would be a cache miss, followed by an I/O to data already in cache. Prior to PAVs (and multiple allegiance), the second I/O must wait for the cache miss to be serviced, delaying its service by the 10 to 20 milliseconds typically required for a simple cache miss. With PAVs, the second I/O can be serviced while the control unit stages the miss from the RAID rank to the cache. So the short I/O is not queued behind the longer running cache miss. Similarly, a short I/O need not be queued behind a long running sequential transfer.

Who benefits?

In a word, wherever IOSQ is a problem, PAVs will be beneficial. One place where IOSQ time often shows up is on work volumes where many tasks access many different datasets on a volume. In modern RAID-5 I/O subsystems, sequential read jobs will trigger the pre-fetch of data from several of the physical devices in the RAID rank. This data is transferred from the devices to the cache at a rate that is characteristic of the attachment for the devices (e.g.. SSA for the ESS). If the jobs are forced to serialize on the UCB then (collectively) they will run no faster than the speed of one ESCON channel. With PAVs, some other resource in the data path will be the limiting factor.

Table PAV-1 depicts a QSAM experiment where multiple read or write streams are directed to different datasets on the same logical volume. There are six jobs reading or writing to datasets on a single logical volume. The number of aliases for the volume are varied from 0 (no PAVs) to 4. The jobs were designed to make sure the data is read and written all the way to the disks in the array group that contains the logical volume. For these sequential cases, we see the limiting sequential capability of a single RAID-5 rank at near 40MB/sec. When there are three or more aliases (plus the base address).

For this experiment, the results are:



	# of aliases	mb/sec
Write	0	12.34
	1	24.66
	2	31.70
	3	39.69
	4	40.56
Read	0	13.50
	1	29.96
	2	33.46
	3	41.68
	4	42.20

Table PAV-1 *Sequential throughput for reads and writes to a single volume with and without PAVs.*

Setup for this experiment

6 simultaneous jobs writing or reading to 6 different data sets on a single volume.

The same 6 jobs were then run against different volumes with a different numbers of aliases

Workload was 27K QSAM with bufno=5

Database workloads

On the page following (Figure PAV-1) is an example of a database workload running on many volumes spread across several arrays and SSA adapters. This is a standard, cache hostile workload we use in the lab. The I/O rate increases as additional users are added, each accessing their own “working set” in the large database stored across many logical devices. It shows how PAVs can improve performance by allowing cache hits to go ahead of misses, or by allowing multiple misses to be processed in parallel.

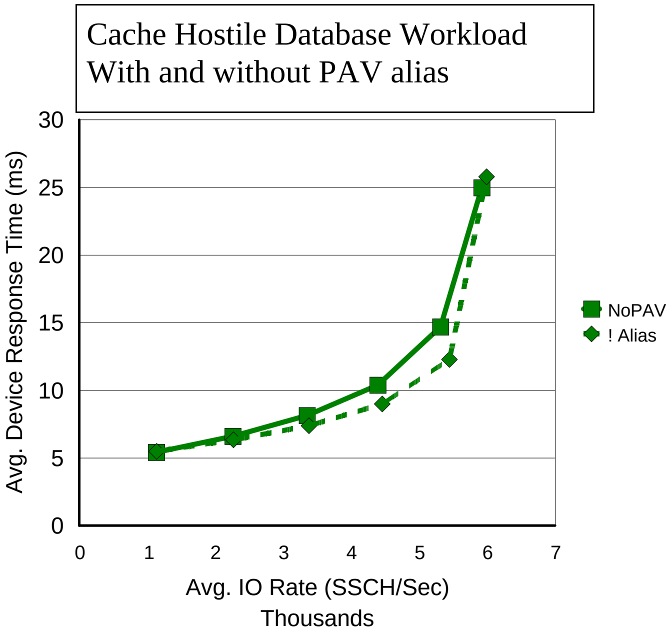


Figure PAV-1 *Cache Hostile Database Workload Performance with and without PAV alias.*

This graph points out a limit to the expectation for PAVs. PAVs cannot increase the aggregate maximum capabilities of an I/O subsystem. PAVs will improve the IOSQ and RESPONSE TIMES for volumes with aliases until some other resource in the subsystem saturates. The resource could be the set of channels, the number of aliases available, the internal bandwidth capabilities of other resources like the device adapters or RAID ranks. In RMF, the Response Time may increase in other components, like PEND or DISCONNECT. If the queue depths exceed the number of addresses (base plus aliases), IOSQ Time can reappear.

Nevertheless, aliases are extremely valuable as a way of giving preferential performance to some devices at the expense of others. The same workload (Cache Hostile) was configured to run on a much smaller configuration of four LCUs. Each LCU runs a copy of the same workload, but each has a different number of aliases. The first LCU has no aliases, the next LCU has one alias, Then two and three aliases. The next plot (Figure PAV-2 on the next page) shows the performance of each individual LCU as well as the average performance across all four LCUs.

One thing to notice from this graph is the benefit of having one alias over having none. Additional aliases matter less and less in this kind of workload. At higher and higher I/O rates for each LCU, the average queue depth increases, but by the time the average queue depth is sufficient to make great use of additional aliases, other resources, like channels, begin to have the dominant effects. Still, it is very common for a system to develop unexpectedly poor performance from just two jobs contending for the same volume. That first alias makes a profound difference.

How many aliases?

There are two ways to assign alias addresses to base addresses--static and dynamic. For stable, well-defined workloads, including most performance benchmarks, static assignment of aliases is attractive. For real workloads, for configurations where the number of device addresses is a problem, and for IS shops that want to exploit goal mode in the OS/390 Workload Manager, dynamic aliases will be attractive.

Suppose we consider an LCU built from two 6+P arrays of 18 Gigabyte HDDs--the maximum number of device addresses in an LCU is 256. If configured as 3390-3 volumes, there will be approximately 64 addresses consumed. Each base address could have three aliases, which would provide an optimum number of static aliases for a 3390-3 configuration. In laboratory experiments we generally observe that most of the increased throughput provided by PAVs accrues with three aliases defined. It depends on the workload, but at some point the increasing throughput saturates the RAID rank with increasing number of aliases (at about 43 MB/sec per rank), and you obtain no further benefit by adding aliases.

If the LCU were built of four 6+P arrays of 36 Gigabyte drives, you could expend all 256 addresses for the base addresses. If however, you define 3390-9 volumes, it is once again possible to have some alias addresses. But you can imagine configurations with 192 base addresses and only 64 extra addresses for aliases. This is where dynamic aliases save the day. The WLM will decide which devices profit most from aliases, and within the constraints of the

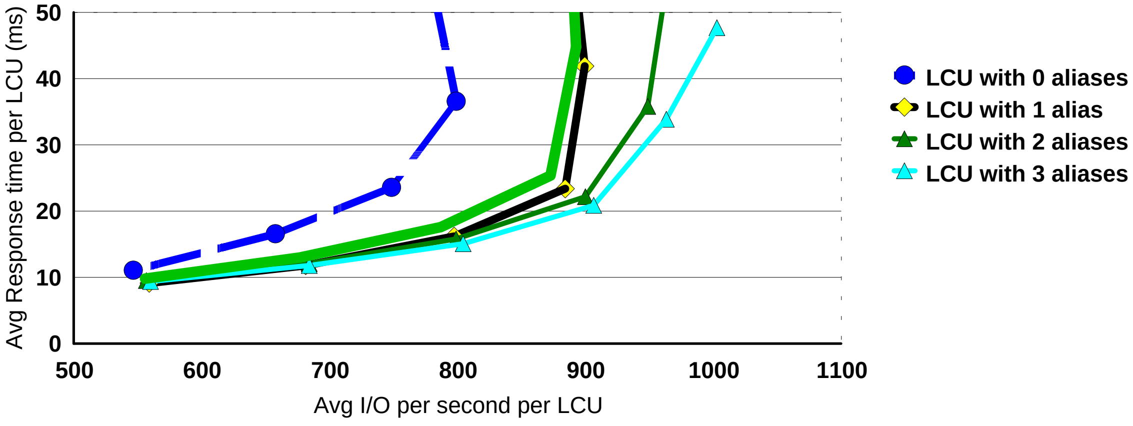


Figure PAV-2 *Cache Hostile Database Workload Performance varying number of aliases* goal mode performance policies, will dynamically move the aliases to where they do the most good.

It is difficult to design a convincing experiment, since one shop's goals may be totally different from another's. But the use of dynamic aliases and the special performance advantages they provide will make possible some data management options that are difficult today.
Where today, special, small volumes are sometimes defined for datasets which need special performance, we may see a trend toward larger volumes like 3390-9, with mulitple aliases or dynamically assigned aliases to provide the desired performance. Special sized, or "custom" volume will be used primarily to exploit every last gigabyte of space on a RAID array, but they will be less and less necessary for performance reasons.

Using PAVs with 3390-9 logical volumes

In an attempt to quantify the comparison of 3390-9 volumes with PAV aliases versus 3390-3 volumes, we designed the following experiment. Basically, we use one-half of an ESS configuration. That means we use eight ESCON channels connected to four LCUs (in an OS/390 environment, the terms LCU and LSS are interchangeable). Each LCU has 128 datasets spread across two 6+P RAID-5 arrays.

The workloads generate essentially the same I/O pattern to all four LCUs. They are our normal cache standard and cache hostile database workloads. These workloads generate more and more I/O by increasing the number of simulated "users" accessing each dataset. Thus, more users mean more I/O, poorer hit ratio in the cache, but most significant -- the potential for IOSQ on a logical volume. The charts plot the aggregate I/Os per second on the X axis against the response time behavior for each LCU on the Y axis. One obvious result is that having more aliases is better than having fewer -- a kind of priority scheme.

The first set of runs uses 3390-3 logical volumes with one dataset each. The LCUs are labeled A,B,C,D in these charts. LCU A has no aliases, LCU B has one alias per logical volume, LCU C has 2 aliases, and LCU D has 3 aliases per logical volume. Subsequently, we reconfigured the RAID-5 arrays as 3390-9 volumes, and put 3 datasets on each (well, 3 doesn't divide 128 exactly, but pretty close). The LCUs were further defined to have 0, 1, 3, and 6 aliases from A to D, respectively. In essence, each LCU is asked to do the same workload, whether configured as 3390-3 or as 3390-9. The number of PAV aliases affects the behavior of the LCU, but the limiting factors in the performance are channel utilization and RAID-5 array utilization.

If you study the next eight charts you will see cache standard and cache hostile charts for 3390-3, then 3390-9, and finally some charts that compare 3390-9 to 3390-3 where the 3390-9 have a few extra aliases to compensate for having three times as much data.

Here are initial observations about the behavior in these charts. For 3390-3 and 3390-9, the LCU with no aliases is severely impacted (by IOSQ). Furthermore, the addition of a single alias has profoundly beneficial effects. The addition of further aliases provide incremental improvements. Multiple aliases behave much like giving extra priority.

The main point of these experiments was to investigate the question whether a 3390-9 with enough aliases will perform as well as 3390-3 (with no or few aliases). The last two charts in this section show that 3390-9 with three aliases provide performance very similar to 3390-3 with one alias, at least for these workloads. And by the way, this is a fairly intuitive result! Take a look at the next eight charts, specifically the last two, where the 3390-9s are compared to 3390-3. Just remember, these large database workloads involving hundreds of datasets or volumes are far from the ideal environment to exploit PAVs. But they do provide enough information to suggest other experiments that need to be performed.

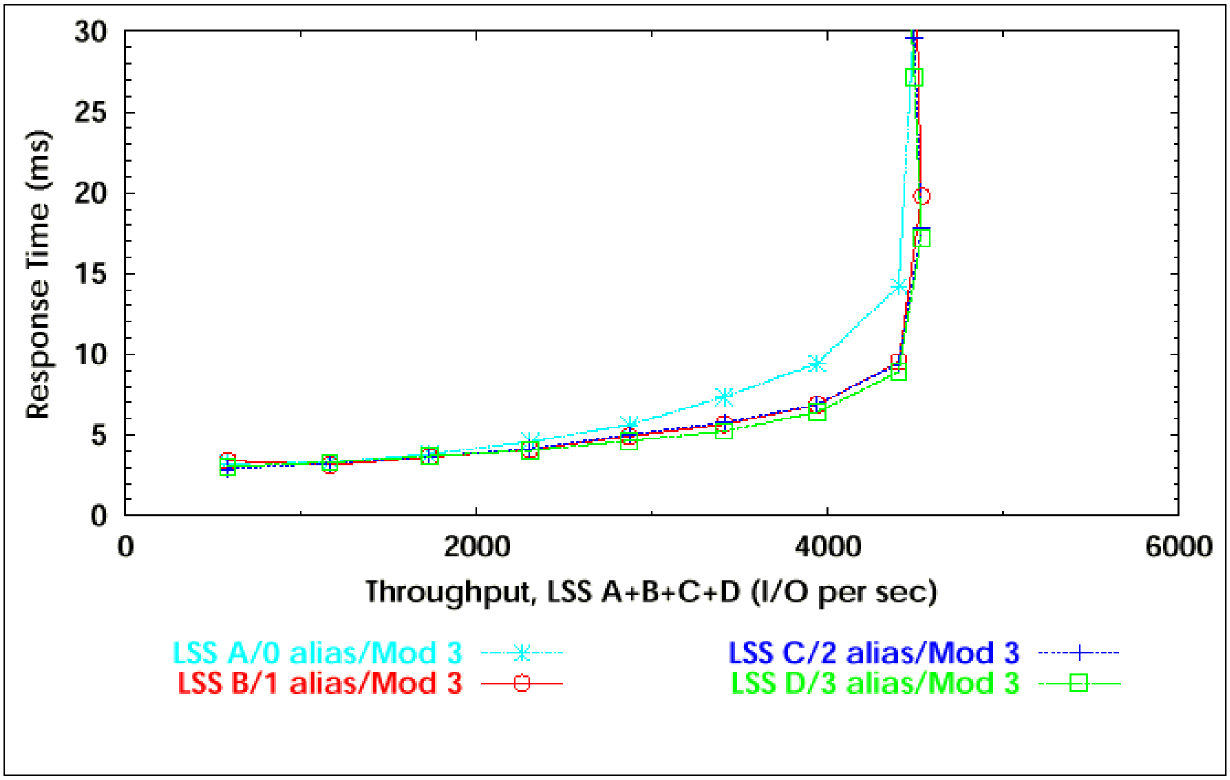


Figure PAV-3 *Database Cache Standard on Four LCUs of 3390-3*

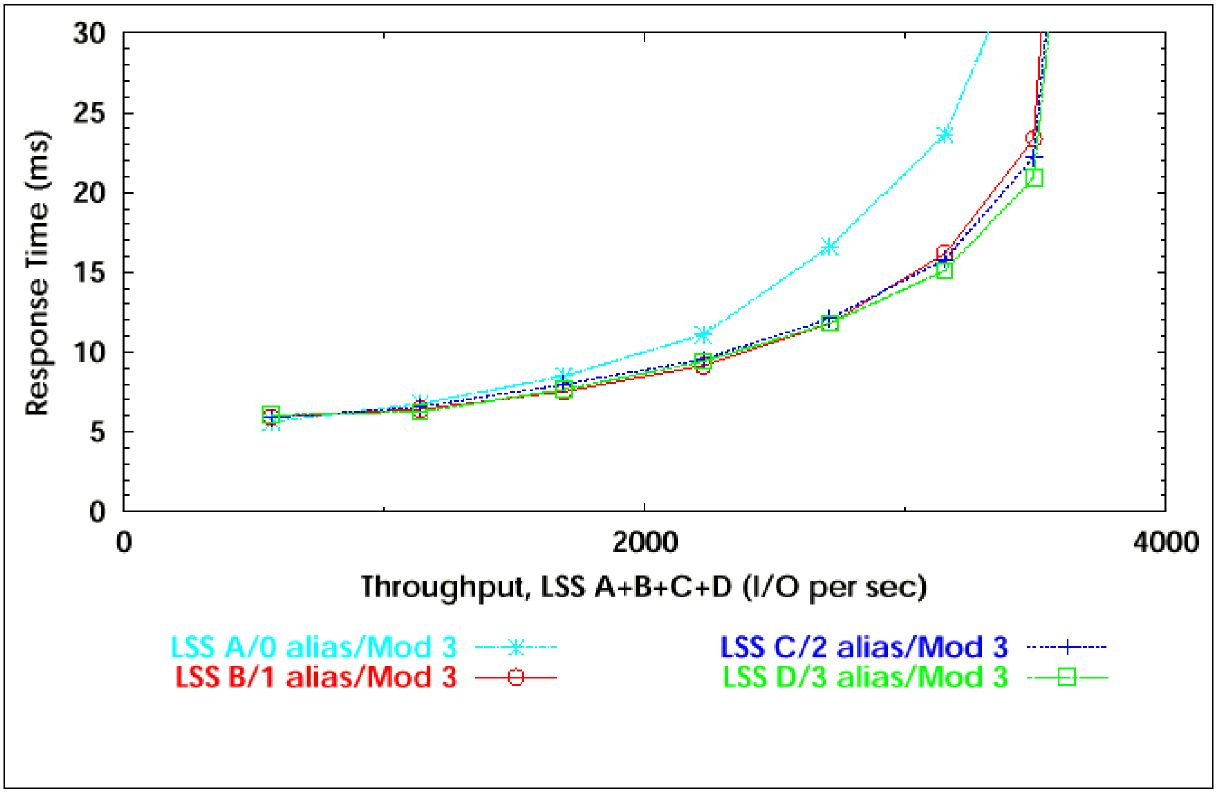


Figure PAV-4 *Database Cache Hostile on Four LCUs of 3390-3*

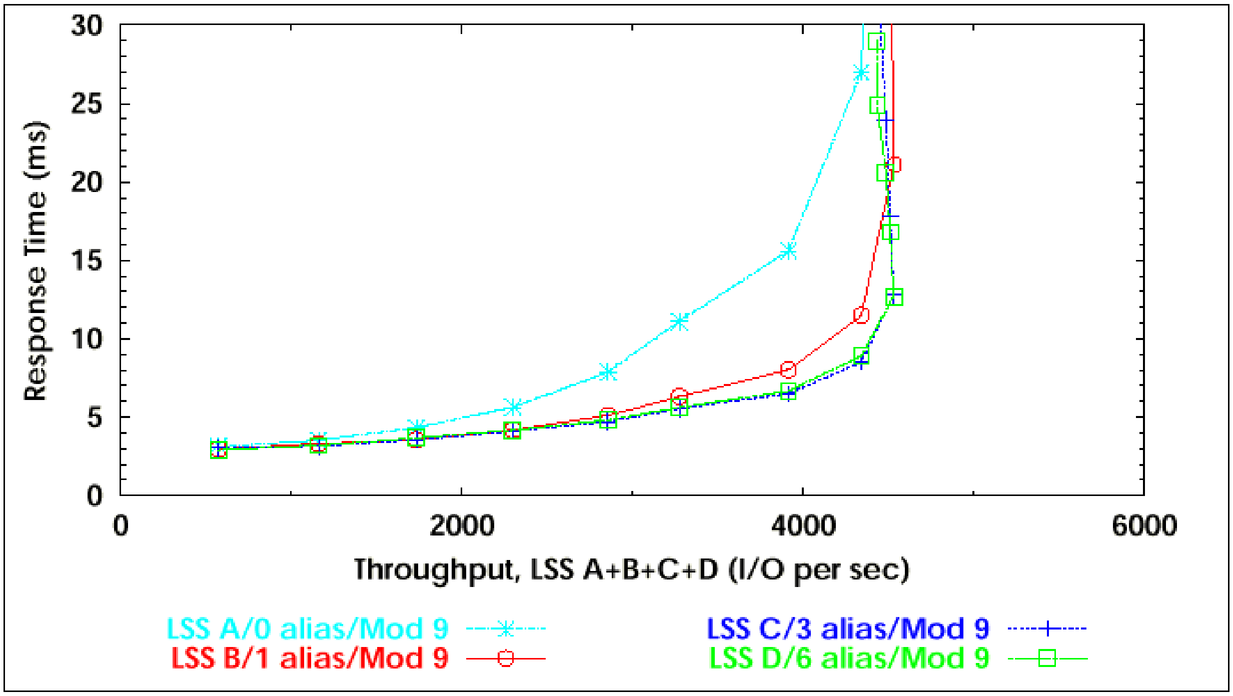


Figure PAV-5 *Database Cache Standard on Four LCUs of 3390-9*

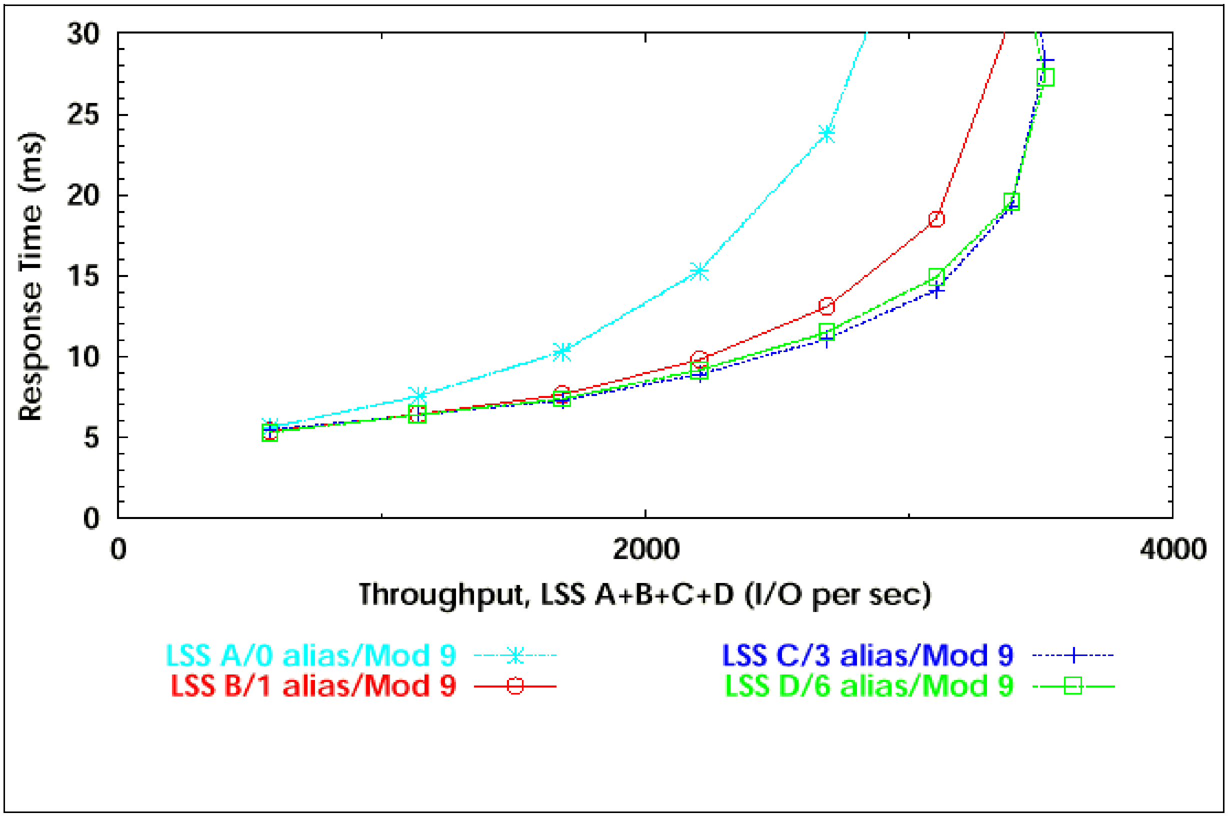


Figure PAV-6 *Database Cache Hostile of Four LCUs*

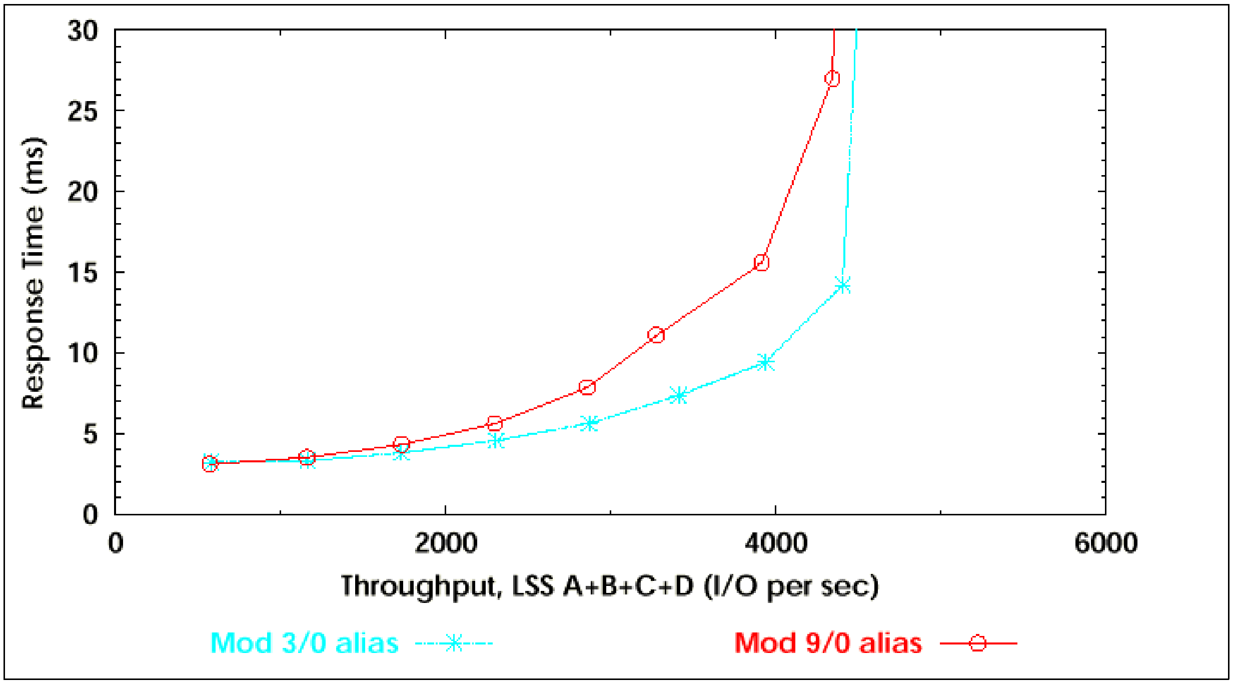


Figure PAV-7 *Cache Standard: 3390-9 vs. 3390-3 with no aliases*

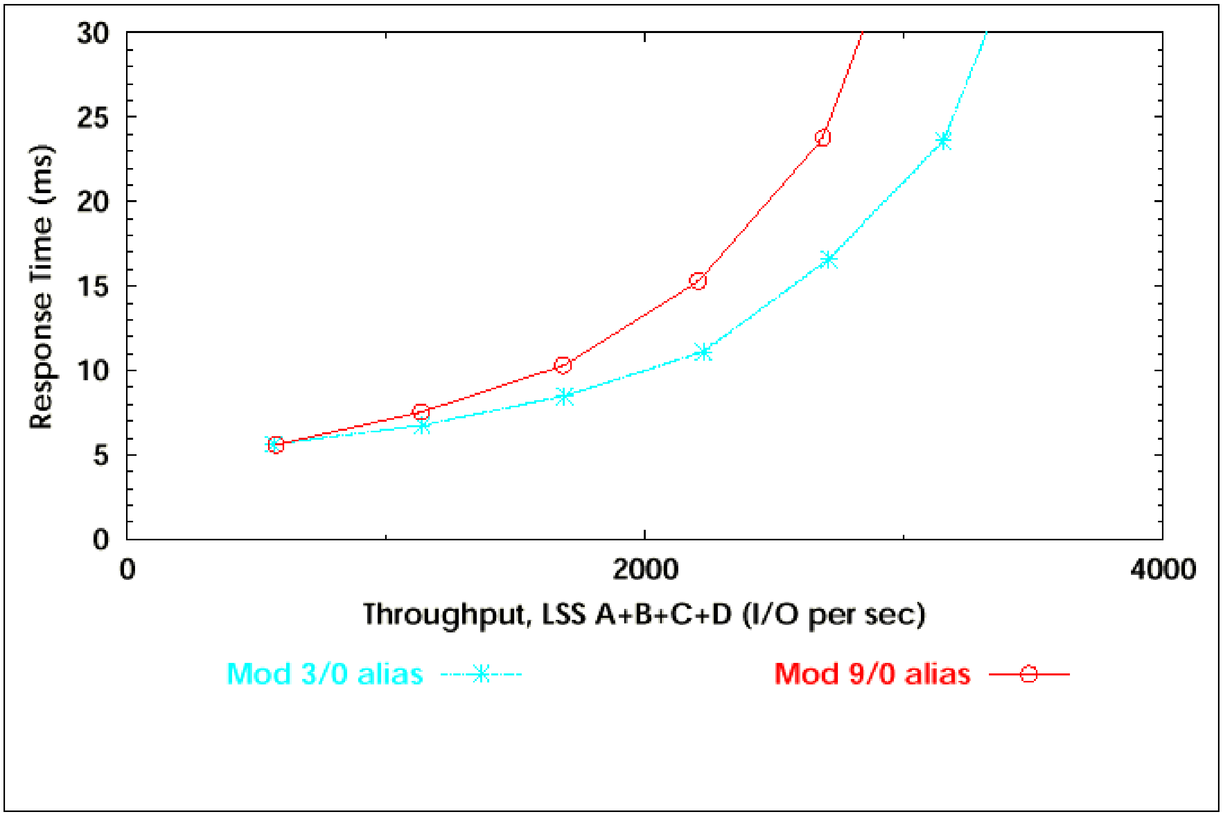


Figure PAV-8 *Cache Hostile: 3390-9 vs. 3390-3 with no aliases*

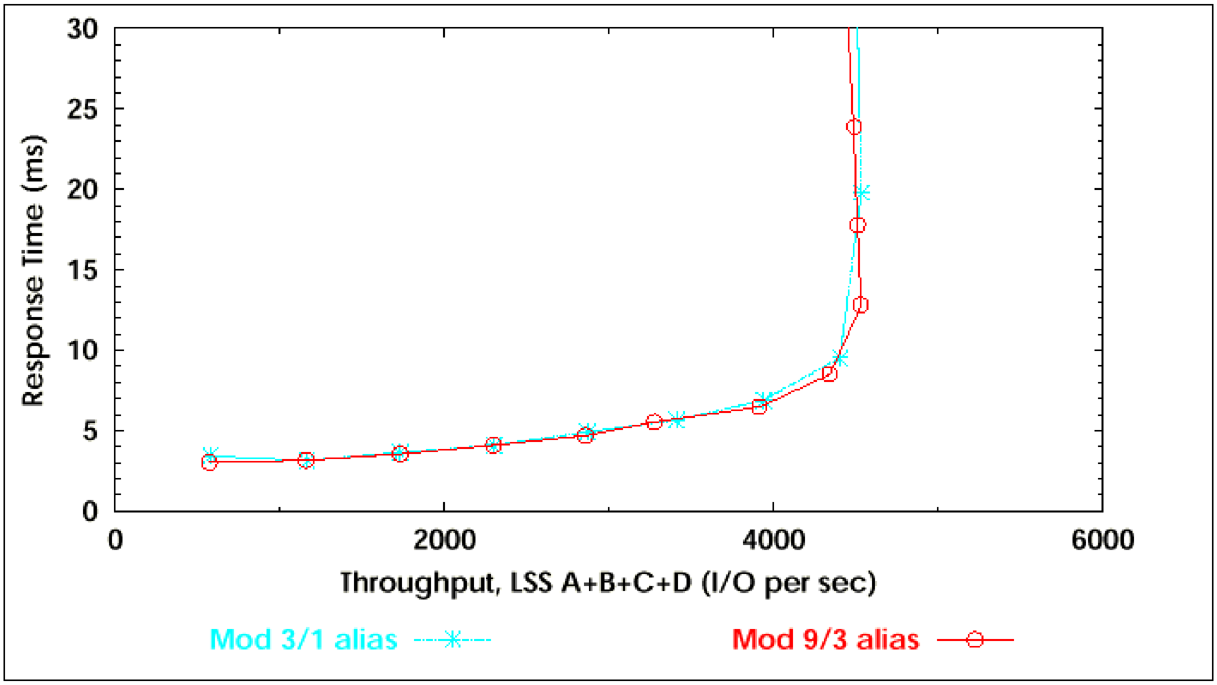


Figure PAV-9 *Cache Standard: 3390-9 with 3 aliases vs. 3390-3 with one alias*

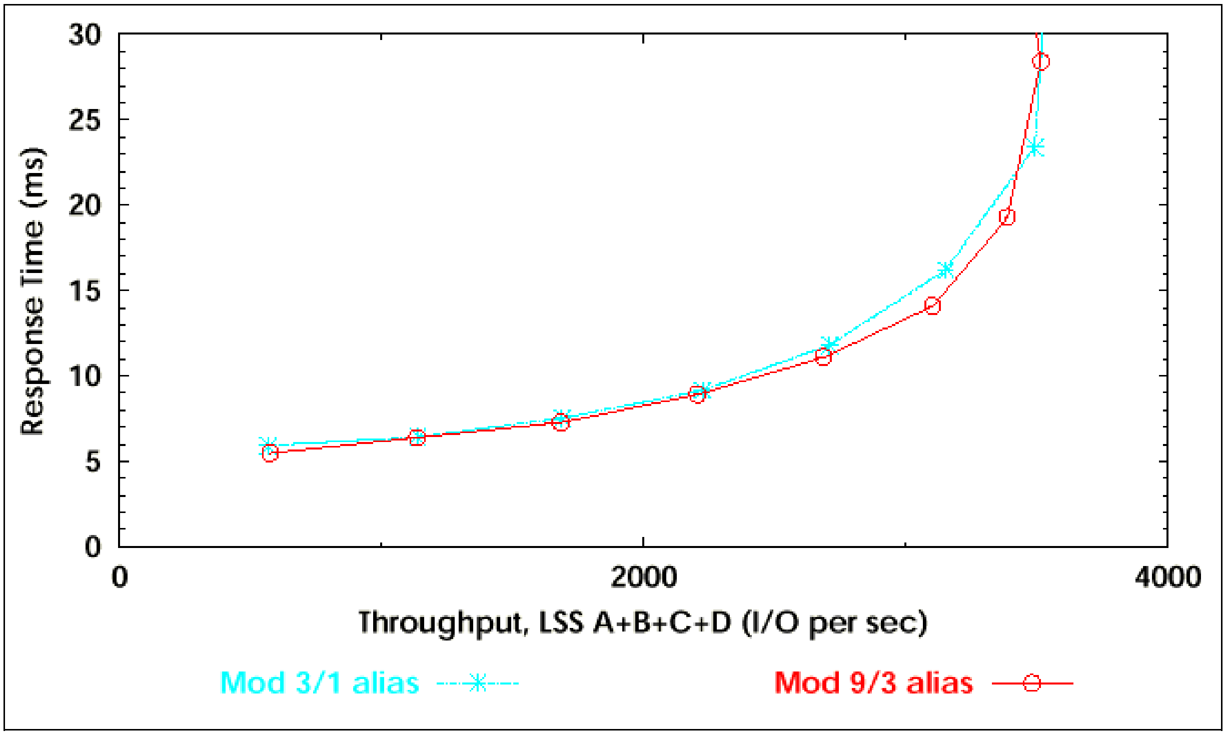


Figure PAV-10 *Cache Hostile: 3390-9 with 3 aliases vs. 3390-3 with one alias*

Chapter 4. DB2 Performance with ESS

DB2 Performance

Customers running DB2 on their OS/390 systems for mission-critical work will benefit from the improvements from the IBM Enterprise Storage Server and the software support added to ESS exploitation levels. The improvements will translate into fewer performance tuning issues, shorter durations for large queries, higher logging rates and faster elapsed times for DB2 utilities such as loads and re-orgs of DB2 tables.

The hardware and firmware design of the ESS provides the basis for the outstanding performance that can be achieved with DB2 workloads. Up to 32 ESCON channels can attach to a single ESS so that multiple OS/390 images could have their own paths into an ESS, if desired. The large number of channel interfaces, each of which can support concurrent I/O operations, provide industry leading configuration flexibility.

DB2 and multiple allegiance

In order to increase the concurrency of work performed by the ESS OS/390 base operating system and access methods have been enhanced to exploit Parallel Access Volumes. The ESS also provides an analogous feature, multiple allegiance, for allowing multiple OS/390 systems to access the same device concurrently. (See the PAV chapter in this document for more details on both features.) Multiple allegiance provides improved device response times for those DB2 V4 or later customers exploiting DB2's data sharing capability. Below is a table illustrating the benefits of multiple allegiance.

	System A "transaction" wkld single volume	System B "query" wkld single volume
Max ops/sec - isolation	767 SIOs/sec	55.1 SIOs/sec
Max ops/sec - no concurrency	59.3 SIOs/sec	54.5 SIOs/sec
Max ops/sec - full multiple allegiance	756 SIOs/sec	54.5 SIOs/sec

Table DB2-1 *Multiple allegiance benefits*

Without the concurrency of multiple allegiance, each system must wait on the other system's I/O to complete before its work can start. Since the service time of the query workload is larger than that of the transaction workload the transaction workload suffers much more than the queries. Most customers would prefer the other tradeoff. For System A, multiple allegiance removes the device utilization of the query workload from System B from its service time which allows it to run at about the same rate as with no query workload present.

DB2 and PAVs

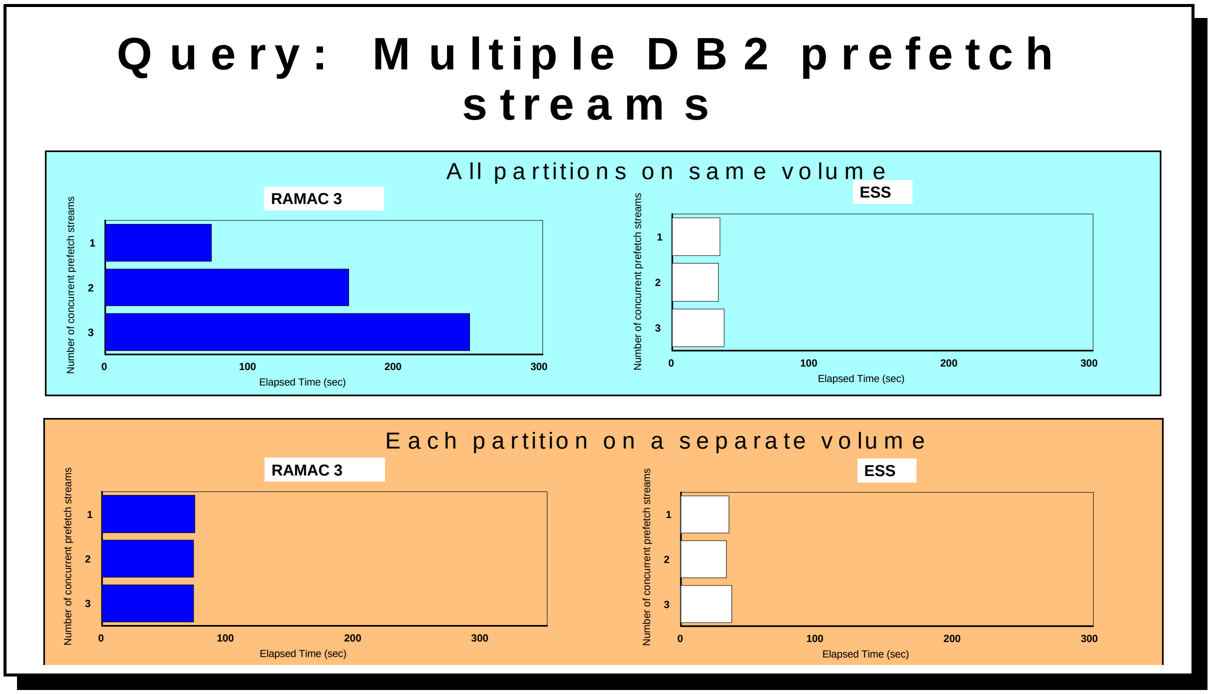
One of DB2's main advantages is the exploitation of a large buffer pool in processor storage. When managed properly the buffer pool can be used to avoid a good percentage of accesses to DASD. Depending on the application and the size of the buffer pool this can mean poor cache hit ratios for what in DB2 is called synchronous reads. Because ESS has a wide array size of at least 7 disks, PAVs can be used to increase the thruput to a device even if all accesses are read misses. Below is a table showing this PAVs advantage in a 100% read miss workload.

PAV status	Single volume 4K read miss
No aliases	145 SIOs/sec
1 alias	250 SIOs/sec
2 aliases	325 SIOs/sec
3 aliases	400 SIOs/sec

Table DB2-2 *Parallel access volume benefits for a 100% read miss workload*

Maximum parallelism in DB2 queries is equal to the number of partitions in the DB2 table space to be scanned. DB2 will issue concurrent I/O to partitions even if they reside on the same device. Below is a chart illustrating the advantages of PAVs for queries that occur on the same volume vs. different volumes. The multiple pre-fetch streams were generated by have separate partitions of a DB2 table space. The same result would be obtained by having multiple different queries to a single partition of a DB2 table space.

.



Query: Multiple DB2 prefetch streams

All partitions on same volume

Each partition on a separate volume

Figure DB2-3 *Multiple DB2 prefetch streams*

DB2 examples

Heavy update of DB2 databases may benefit from the following scenario. In some DB2 applications, IOSQ can be correlated to either DB2 deferred writes or checkpoint processing. DB2 accumulates changed data in its buffers until the deferred write threshold is reached, at which time a lot of random updates are written to the databases. These writes, even if (unlikely) hits in the write cache, must be serialized on the UCB. What is worse, if a new transaction needs to read data, even data in cache, it may be delayed by writes that are already queued on the UCB. These delays show up as IOSQ time in an RMF report. The combination of I/O Priority queuing (OS/390 V3.7) and PAVs will allow the read data to get to the head of the UCB queue, and PAVs will make it possible to start the read I/O on the first available alias, even if writes are taking a long time.

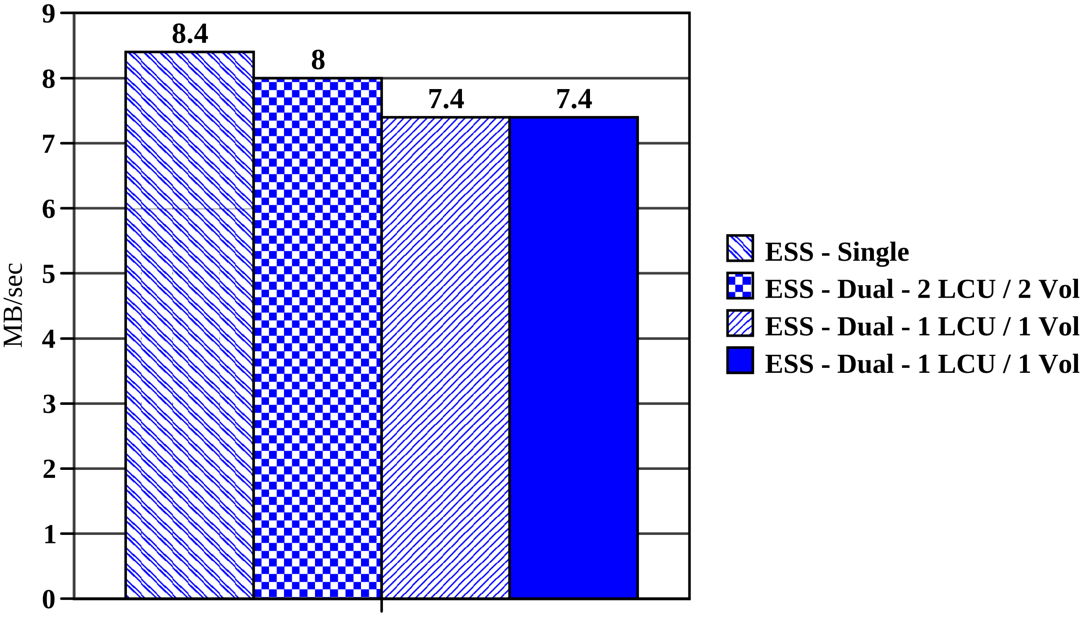
Sometimes two important datasets may accidentally get allocated to the same logical volume. In that case, PAVs may alleviate the IOSQ consequences. In one DB2 experiment, dual DB2 logs were allocated on the same logical volume; see results below.

A - Single logging. Log rate was 8.4 MB/sec.

B - Dual logging. Logs were on separate LCUs, rate was 8.0 MB/sec. This is recommended best layout for dual logging.

C - Dual logging. Logs were on the same LCU but different volumes. Logging rate went down about 7.5% compared to case B. This is not recommended for best layout, but an interesting test.

D - Dual logging. Logs were on the same volume (!). This is REALLY not recommended, but notice there is no degradation over case C. PAV really works, IOSQ time remains zero.



Large DB2 queries

The ESS architecture provides superior sequential bandwidth. For DB2 customers this means large queries can be accomplished much faster, given adequate CPU resource. Laboratory measurements of a DB2 table scan of a 60 partition, 30 GB table on a 24 ESCON ESS with 8 raid ranks was completed in under 200 seconds. The sustained data rate for this query was **155 MB/se**c!

DB2 Utilities

DB2 utilities such as DB2 loads, DB2 re-orgs, copies and recovers will have the advantage of the superior sequential bandwidth of the SSA arrays coupled with the concurrency of PAVs and a large number of ESCON channels. Early lab runs have shown 50% reductions in elapsed times for DB2 loads and more than 50% reductions for DB2 re-orgs compared with earlier RAMAC models. PAVs in particular will keep DB2 utilities which are run concurrently, such as copies and re-orgs, from having as large an adverse impact on transaction response time than with those subsystems that do not support PAVs.

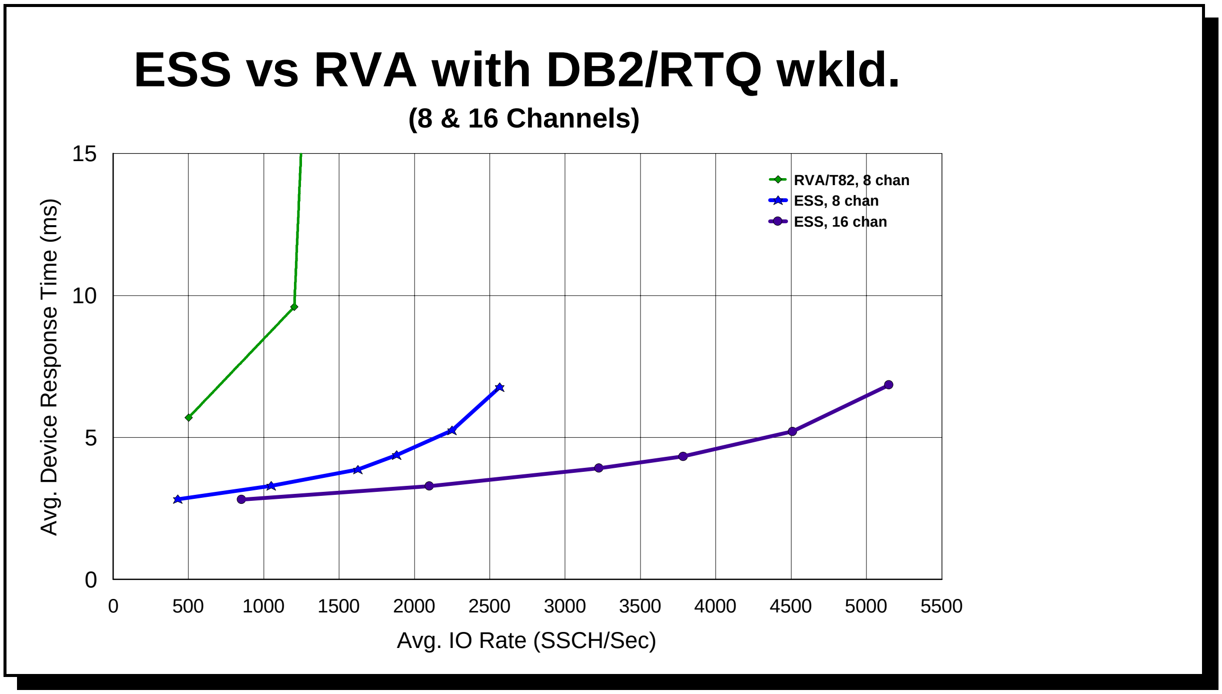


Table DB2-5 *DB2 Performance Comparison*

DB2 Transaction based workloads

Laboratory measurements of a DB2 transaction based workload were performed on ESS with 8 ESCON channels and 64 volumes on 4 RAID-5 arrays all residing on a single ESS cluster (see chart, below). Comparing this workload to RVA T82 you can see the superior thruput and response time the ESS provides. The workload was then run on 2 LPARs each with their own set of 8 channels and 64 volumes. The 16 channel run in the same chart represents this run. The resulting run shows a doubling of thruput compared with the 8 channel ESS run and 3.5 times as much as the RVA for this workload with better response times at all I/O rates.

Chapter 5. TPF Performance with ESS

ESS is the perfect choice for the high throughput and low response time required in today's demanding TPF environment. This section will discuss some of the ESS performance features that support this.

TPF 4.1 provides two levels of support for the ESS. The first level, is called transparency mode and does not require any TPF code changes. TPF was tested on the ESS at PUT08. We believe that ESS can work with TPF at earlier PUT levels, but we suggest that you contact the TPF support organization if you intend to use ESS at an earlier level to find out if any additional information is available. Please see the *contacts* section at the end of this section.

The second level of support, TPF Exploitation mode, requires PJ26692 for TPF41. This APAR is available through standard channels (TPF's APAR Disks) is included in the PUT11 shipment. If you are unfamiliar with these channels please contact your TPF Service Rep. This APAR will provide support for new and optimized CCWs within the ESS and as such will provide improved performance.

Please be aware that prior versions or releases of TPF do not support the ESS.

Performance Features

Multiple Allegiance

This allows two or more hosts to access the same logical volume at the same time provided there are no extent conflicts. It was first supported for TPF loosely coupled complexes with the IBM 3990-6 and 9390. With ESS, it is available for TPF and all other multi-host operating systems. Parallel Access Volumes are not currently supported for TPF.

I/O Queuing

Basic I/O queuing is supported for TPF systems, however priority queuing is not. Basic I/O queuing within the ESS improves performance by reducing the number of busy messages presented back to the I/O requester (TPF) when another I/O operation is in progress.

Improved CCW commands set

The ESS implements additional ECKD CCW commands for S/390 I/O. With PJ26511 for TPF41, a performance improvement can be realized by either uniprocessor or loosely coupled TPF41 systems. Full track operations, used primarily in TPF utility functions, will also benefit from the improvements in the ESS.

Track Emulation and FlexVolumes

The ESS will emulate both 3380 and 3390 format DASD. FlexVolumes allows for the definition of nontraditional sizes for either 3380 or 3390 emulated DASD. PJ23135, on PUT6, permits TPF to use disks with cylinder sizes above the physical limits of older devices for 3390s only.

Cache Efficiency

The ESS is typically more efficient in its cache utilization than an IBM 3990-6 or 9390. The ESS has smaller cache segments than these other storage controls which avoids wasted cache space for the small record sizes that are used with TPF.

JBOD Support

The ESS can be configured to run in a RAID-5, or non RAID-5 (JBOD) configuration. RAID-5 uses parity to protect against disk failures. IBM recommends that customers use RAID-5, which delivers very high performance. TPF customers already using RAID-5 have also found that it is efficient not to have to rebuild a module following a disk drive failure. If you believe that JBOD is preferable in your environment, please contact TPF support to discuss this.

Capacity

There are three sizes of disk drive modules (DDMs) available in the ESS: 9 GB, 18 GB and 36 GB. In general IBM recommends 9 GB DDMs be used for TPF. This is because of the high access density that is usually seen with TPF. For TPF workloads that are very cache friendly and have high destage efficiency, it may be possible to use 18 GB drives.

Although an ESS can support very high total capacities (up to 11+ TB), it is likely that relatively low capacity subsystems will be suggested for TPF. Total capacity of 420 to 840 GB should be considered typical. Again, the reason for this is high access density. For a rough estimate of how much capacity should be configured in an ESS, use the rule of 10,000. Estimate your current access density in SIO/sec/GB and divide this number into 10,000. The result is your capacity estimate in GB. For example, an access density of 20 SIO/sec/GB would result in an estimate of 500 GB per ESS. Your results may vary. A more accurate sizing can be done via performance modeling with Disk Magic.

Performance Expectations

TPF can benefit from the performance improvements designed into the Enterprise Storage Server. The actual performance of any one TPF system is very dependent upon the profile of the TPF workload and it is difficult to provide a single statement about the specifics of the expected performance, however ESS performance for TPF systems will be exceptional. As with all cache based DASD subsystems, read-to-write ratios, the cache-hit ratios, and destaging data will impact the actual response times.

ESS performance for TPF can be modeled using the latest version of Disk Magic for Windows. This is a marketing tool available for use by IBM employees and business partners. Disk Magic requires detailed knowledge of the caching attributes of the workload being modeled. TPF provides facilities to collect controller cache statistics for existing systems. The Cache Analysis Aide (CAA) is another marketing tool that may be useful to project cache statistics for various cache sizes and configurations. CAA requires a continuous file data reduction tape from each host. Disk Magic and CAA are available to available only for IBM storage marketing representatives through IBM internal distribution mechanisms.

Figures TPF-1 and TPF-2 show two examples of the modeled performance of the ESS. Note that these are only examples and you should not assume these models will reflect your system. Both examples are for loosely coupled systems.

In Figure TPF-1, the workload has a relatively high degree of read misses and destage operations. This means that many I/Os must access the physical disk. Thus the limiting resource is DDM. In these cases 9 GB drives will provide a substantial throughput benefit since the number of drives available is critical.

In Figure TPF-2, the workload is much more cache friendly. As a result, many more SIO/sec can be pushed through the ESS. However, at about 20,000 SIO/sec, the 32 ESCON channels become saturated. For these cases, 18 GB drives may be acceptable since most I/Os are satisfied in cache and the actual disks may not be overly busy.

In any case, performance modeling can help you understand your specific circumstances and make the right configuration choices.

Summary

The most noticeable attribute of the ESS for TPF systems is superior performance. Usually the largest individual component in the life of a TPF transaction is the time spent waiting on I/O to complete. Therefore performance improvements to your existing TPF system will have a noticeable and positive impact on the existence time of a transaction. Thus, depending on the storage platform you are currently running on, the performance attributes of the ESS may very well be the key to making substantial performance improvements.

Contacts for IBM Representatives and Customers

If your customer is in either AP or EMEA and you have additional questions regarding TPF support, please send your question through the TPFQA Notes id as with other support questions. If you are an IBM marketing representative or a customer and you reside in NA you may contact Bill Supon, Mary Kellner-Somers, or Bill Cohen directly. Generic ESS questions should be handled through your local Storage Specialists. Customers may contact their local TPF Support for more information.

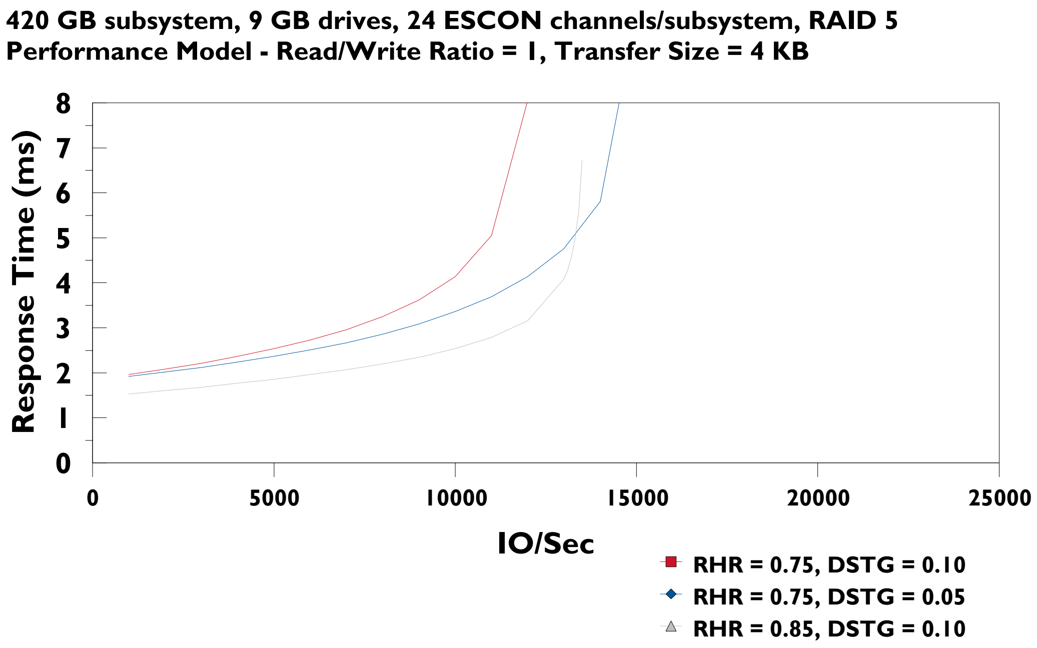


Figure TPF-1: *Performance Envelope for ESS DDM Limited Case*

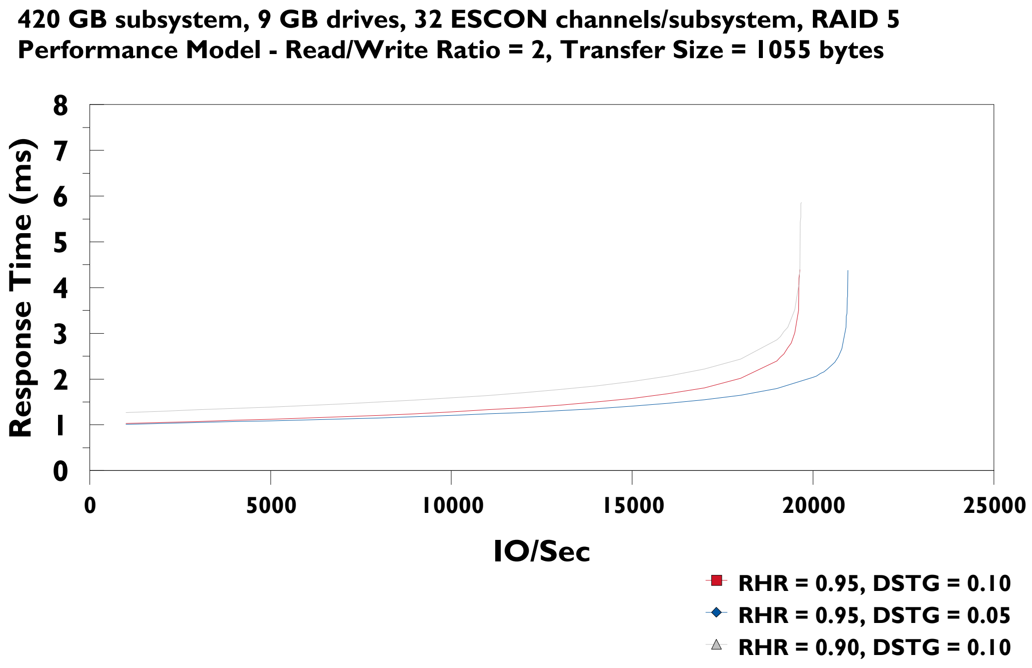


Figure TPF-2: *Performance Envelope for ESS SMP/Channel Limited Case*

Chapter 6. Performance during Hardware Outage

Types of Hardware Outages

The IBM Enterprise Storage Server has been designed to deliver 24 hour by 7 day data availability even when hardware failures occur. This design also allows for concurrent code load so microcode improvements can be made without loss of access to data. Expected performance during these outages can be understood by determining which resource is no longer available and what recovery action the ESS has taken to maintain 100% data availability. Below is a table of hardware resources that can be lost while the ESS maintains data availability:

Hardware Resource	ESS Actions
Disk Drive Module	Begin sparing operation. All other HW resources are still available.
Host Adapter	Fence adapter. All other HW resources are still available.
Host Adapter Bay	Fence adapter bay. Loses up to 4 of 16 host adapters.
Device Adapter	Control of RAID ranks belonging to failed adapter transferred to partner device adapter in other cluster through failover operation.
Cluster • mother board & components • NVS adapter • Device adapters • Common Platform Interconnect	Failover entire cluster. Remaining cluster assumes access to all data in the ESS. Loss of ½ cache, ½ of the NVS, ½ of the SMP processing power, ½ of the device adapters until the failing component is repaired.

Table Outage-1 *Hardware resource failure scenarios.*

Disk Drive Module Failure

Due to mechanical nature of the Disk Drive Module (DDM) their failures are the most likely type customers will encounter. DDM failures are detected by the Serial Storage Architecture (SSA) device adapters attached to them. The RAID-5 redundancy of an array is utilized to rebuild the failed DDM onto a spare DDM in the same SSA loop. (Note: the ESS cannot recover failed DDMs configured in JBOD mode. Normally, DDMs in JBOD mode would be used only by systems with redundancy built into the operating system such as TPF, AIX mirroring.) The device adapters are configured in two-way mode and both work in concert to rebuild the failed DDM. Once the rebuild is complete, performance returns to the same level as before the DDM failure.

Performance during RAID rebuild

The performance during RAID rebuild is close to the performance when the array is fully redundant. The reason for this is the large cache and NVS in the ESS reduces the accesses to the DDMs. It’s only when host accesses have saturated the array will you see noticeable performance loss during rebuild. Below is a table comparing the performance of an OS/390 workload at 80% the maximum rate for a 16 ESCON channel, 64 18 GB DDM configuration during normal operation and during a rebuild.

	total host I/O/sec	total host resp (ms)	array host I/O/sec	array host resp (ms)	array ops/s /DDM	rebuild array ops/s /DDM	array read resp (ms)	array write resp (ms)
Normal	8,844	7.11	1100	7.29	89	N/A	22.5	79.6
Rebuild	8,804	7.34	1058	8.32	86	13	25.0	121

Table Outage-2 *Performance During Rebuild*

The workload during this run had a 3.2 read-to-write ratio, .84 read hit ratio and 11% fast write destage (destages over total host I/Os).

Observations

There is a small increase in total host I/O response time (7.11 ms to 7.34 ms). This is due to having only 1 of the eight RAID-5 arrays affected by the rebuild. The host I/O directed to the array that is rebuild is affected more but the increased response time is reasonable considering the increased workload on the array. Notice the workload on the array before the rebuild is rather high at 89 "array ops/sec/DDM". These operations include host read misses and the four operations required of a RAID-5 array for a destage (write) of data from ESS cache to disk (read old data, read old parity, write new data, write new parity). A RAID-5 array can handle between 100 and 130 disk operations per second before it reaches saturation so the array is heavily utilized **before** the array rebuild begins. The ESS keeps track of read and write response times as observed at the device adapter level. This means read response time is host read miss response time and write response time is the time it takes a destage to complete (this includes multiple disk operations). During the rebuild the array activity increased from 89 ops/sec/DDM in the array to (86+13) ops/sec/DDM. This increased activity is reflected in the response time observed at the device adapter level.

Rebuild times

The duration of a rebuild is a function of host activity. If there is no host activity to the array, the rebuild operation will complete in roughly 40 minutes. With increasing load to the array, the completion of the rebuild operation will take longer, up to several hours, depending on the intensity and nature of the competing work. Fortunately, performance impact to production applications is very minimal, as described above. Open system arrays can take longer for rebuild operations due to how data is distributed across the array.

Host Adapter Failure

The host adapters in the ESS have a much lower intrinsic failure rate than DDMs. If a failure should occur the host adapter would be "fenced" and access to the ESS would continue on the other adapters. For OS/390 systems, the performance loss is akin to losing 25% of channel capability to a subset of volumes on the ESS if the configuration guidelines in this document are followed. Performance degradation would be noticeable as the pre-failure channel utilization approaches 60%. This is at the high end of normal operating ranges for DASD subsystems.

For Open Systems, the Data Path Optimizer is available for use with ESS to provide multiple paths from a host server to the same set of logical volumes (LUNS) on the ESS. With two paths to each LUN the maximum performance loss would be 50% for a subset of LUNS from the affected host server. With four paths to each LUN the maximum performance loss is just 25%.

Host Adapter Bay Failure

There are up to four host adapters configured in each of four host adapter bays in an ESS. If there is a host adapter bay failure or a bay must be placed in service mode to repair or change modify the configuration then the maximum performance loss of the ESS would be 25%. OS/390 workloads and Open systems configuration with four paths in a Data Path Optimizer won’t start noticing this performance loss until path utilization prior to the failure approaches 60%.

Device Adapter Failure

There are eight SSA device adapters in an ESS. The device adapters are configured in pairs so that if a device adapter fails, the other adapter in the pair can take over operation for all the RAID ranks in the two SSA loops belonging to the device adapter pair. The two device adapters that make up a pair reside in separate ESS clusters. For the takeover to occur, control of the Logical Subsystems (LSSs) in use by the failing adapter must be given to the other cluster. This occurs through a failover operation. The bandwidth of a single device adapter is very robust (up to 86 MB/sec), and thus only in rare cases will the remaining device adapter become a bottleneck for data under its control after the failure of its device adapter counterpart.

Cluster Failure

There are two clusters in an ESS. Each cluster has it’s own cache, NVS and device adapters. The clusters share the host adapters through a common platform interconnect (a PCI to PCI bridge). A failure of a component on the mother board (I/O planar) or an NVS adapter will cause the ESS to **failover**. A failover involves one cluster assuming control of all data in the ESS subsystem. Since the clusters are symmetric and each cluster has it’s own cache, NVS (albeit the NVS it owns is in the other cluster) and device adapters, losing a cluster will result in the loss of 50% of the maximum performance capability of the box.

In most cases, performance following cluster failover will be much better than might be anticipated. Figure Outage-1 shows the performance delta between an ESS running in normal mode (i.e., two clusters active) versus running in failover mode. With I/O rates less than 2,000 I/Os/SEC it’s difficult to detect a performance difference. As the I/O rate approaches 4,000 I/Os/SEC a noticeable response time difference of 30% exists. In failover mode the maximum I/O rate is just over 4,000 I/Os/SEC whereas the normal operating ESS can provide 5,000 to 6,000 I/Os/SEC depending on what you determine is an acceptable maximum average response time.

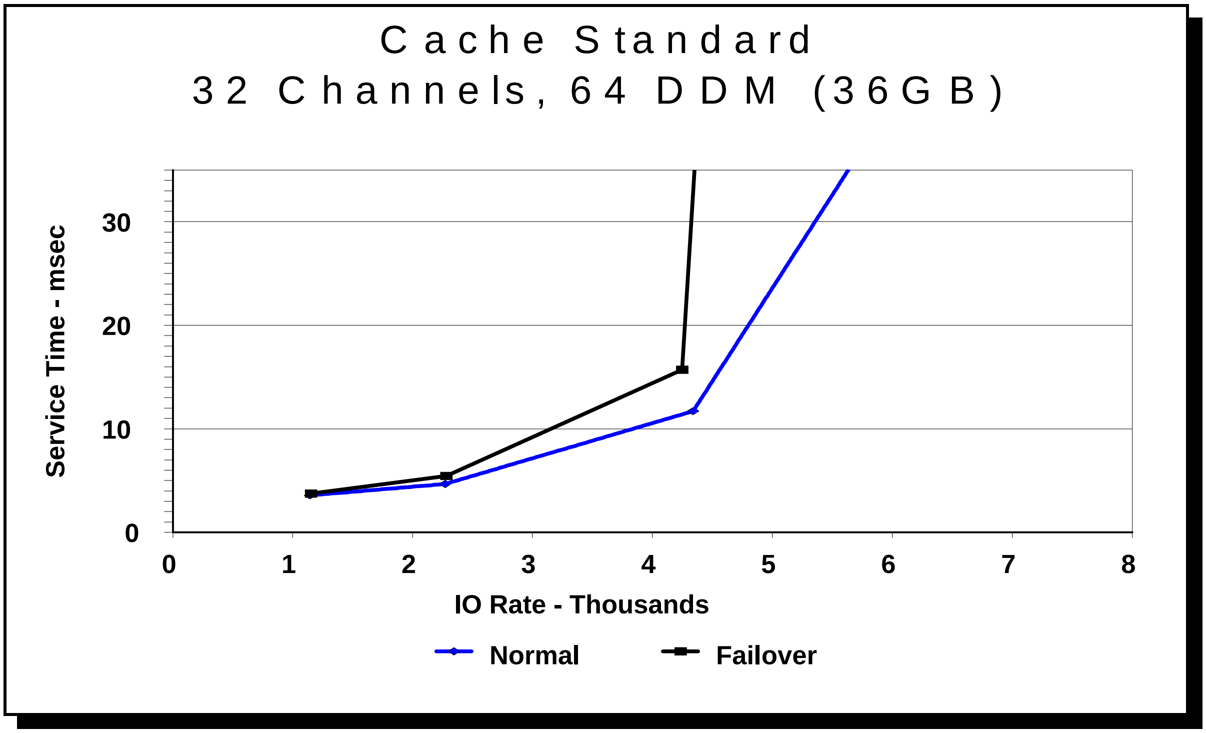


Figure Outage-1 *ESS performance in failover mode vs. normal mode*

Chapter 7. ESS Configuration and Tuning Guidelines

This section provides some general configuration guidelines and rules-of-thumb. Additional planning performance numbers are also provided. If a user requires additional detailed analysis, the Disk Magic modeling tool (see chapter 9) is available to IBM personnel and business partners to help conduct detailed analysis and disk system sizing. That tool can be used to help understand the performance effects of various configuration options: such as the number of ports and host adapters, disk drive capacity, number of disks, etc.

How many ESS disk systems?

One of the first tasks facing any user is determining how many ESS disk systems are required to meet the capacity and performance needs of the applications. A user might decide to implement more than a single ESS disk system for a variety of reasons:

- More storage capacity required than containable within a single system
- High performance requirements that exceed the capabilities of a single system
- Desire to separate different types of workloads onto different systems

As a general guideline, configurations of approximately 1.6TB or less of total useable capacity should be able to meet the performance requirements of most users without requiring significant detailed analysis. Since larger capacity configurations imply heavier user workload, IBM recommends some analysis of user workload and ESS performance capabilities for configurations significantly larger than 1.6TB.

Sequential workloads

If your workload is primarily sequential work, with a high demand for data bandwidth, you should estimate the following:

- Megabytes per second of data transfer
- Read/write mix

The ESS disk system has been measured with sustained sequential throughput capabilities as follows. These throughputs were measured under optimized conditions, with perfectly balanced workloads. You should generally use approximately 70% of these throughput values for normal workload planning purposes, depending upon your accuracy in estimating actual peak workloads, or ability to balance workload across the ESS disk system..

	Maximum Throughput (MB/sec)
Reads	160-185[2]
Writes	145

[2] Under optimal configuration conditions, ESS has been shown to provide even higher sequential read throughputs, approximately 185 MB/sec. However, for most customer configurations, 160 MB/sec is a better planning number.

Table Config-1 *ESS Throughput Capability*

Disk drive capacity

There are three choices available for disk drives used in ESS:

- 9.1 GB 10,000 RPM
- 18.2 GB 10,000 RPM
- 36.4 GB 7,200 RPM

By selecting a particular disk drive capacity for a given amount of disk storage, you are also selecting the number of disks available. Smaller disk capacities generally imply using more disks, fewer I/O per second per disk, fewer disk queuing delays, and generally better performance. The choice of disk drive capacity is usually one of meeting the required performance demands at the lowest possible cost per megabyte of storage.

The following are some general rules-of-thumb to help guide the choice of disk drive capacity.

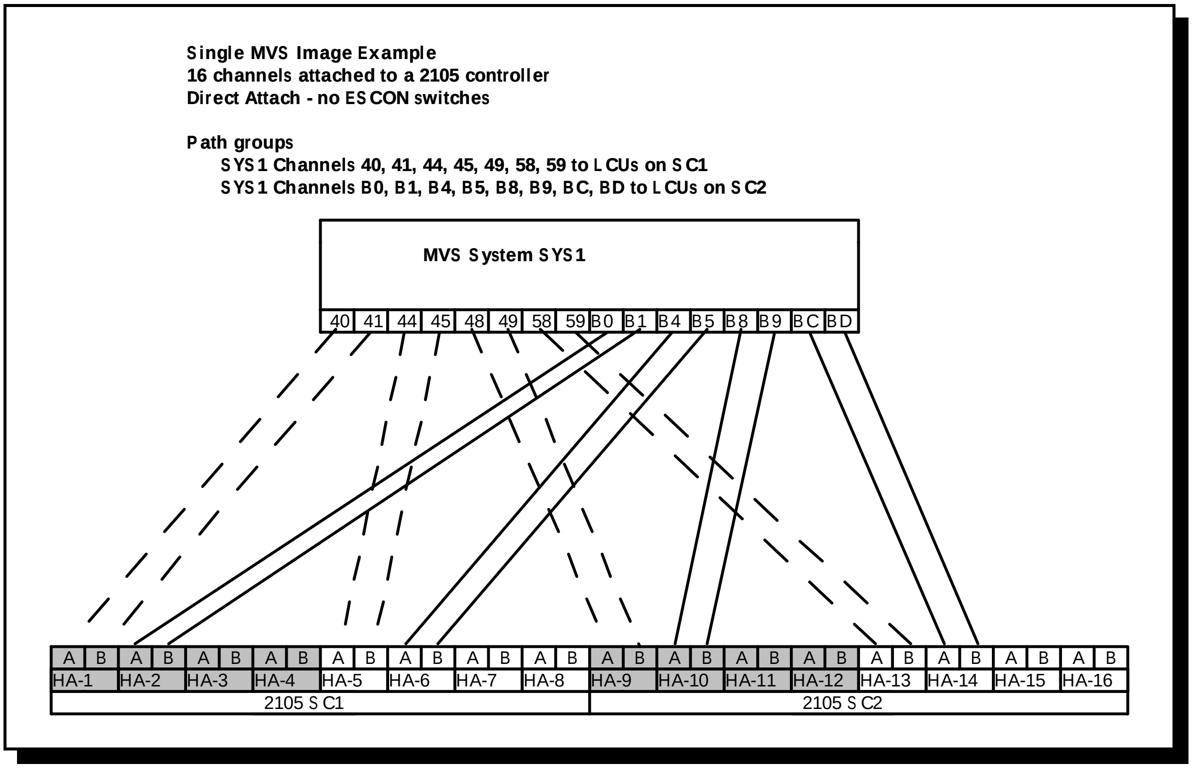
- 18 GB disks are most likely the correct choice for most applications. They should provide the proper balance of performance and cost for most users.
- 9 GB disks should be considered in the following circumstances:
  - Applications with very demanding performance requirements.
  - Workloads with very random access patterns or very high write content (sometimes referred to as “cache-hostile” workloads), in which there is likely to be very high disk activity.
  - Workloads that have very high access rates relative to the amount of total storage. For example, workloads with access rates in excess of 5 accesses per second per gigabyte of storage are likely candidates for 9 GB disks.
- 36 GB disks should only be considered in the following circumstances:
  - Applications that do not have demanding performance requirements.
  - Applications that make very good use of cache, with very high cache hit ratios (>95%).
  - Applications with very low access rates relative to the amount of storage (<1 access per second per gigabyte of storage).
  - **If you are planning on using 36 GB disks, you should consult with your IBM representative to model the effects with Disk Magic.**

Configuration of ESCON channels

These are some guidelines to configure ESCON channels to ESS to optimize performance:

- Always configure at least 16 channels per OS/390 image to the ESS
- Always use 8-path path groups. A path group is the set of channels defined for a particular LCU.
- Always plug channels for the 8-path path group into four adapters, one host adapter per bay
- Each 8-path path group access the LCU's on one cluster
- For maximum SIO/sec out of an 8-way path group, configure all 8 channels to 1 IOP. Splitting the 8 channels across 2 IOPs may degrade performance.

See below the simplest example of following the above recommendations. It depicts a single S/390 OS/390 system attached to eight ESS host adapters using 16 ESCON channels.


Single MVS Image Example
16 channels attached to a 2105 controller
Direct Attach - no ESCON switches
Path groups
SYS1 Channels 40, 41, 44, 45, 49, 58, 59 to LCUs on SC1
SYS1 Channels B0, B1, B4, B5, B8, B9, BC, BD to LCUs on SC2
MVS System SYS1
40
41
44
45
48
49
58
59
B0
B1
B4
B5
B8
B9
BC
BD
A
B
HA-1
HA-2
HA-3
HA-4
HA-5
HA-6
HA-7
HA-8
HA-9
HA-10
HA-11
HA-12
HA-13
HA-14
HA-15
HA-16
2105 SC1
2105 SC2

The second diagram following shows two S/390 OS/390 images each with 16 ESCON channels attaching to an ESS through two ESCON directors (channel switches). The 32 channels are mapped into 16 ESCON links on the back end of the switch and these links are attached to eight ESS host adapters. This example illustrates attaching more ESCON channels than host adapter ports using ESCON directors. It still follows the channel configuration rules outlined above.

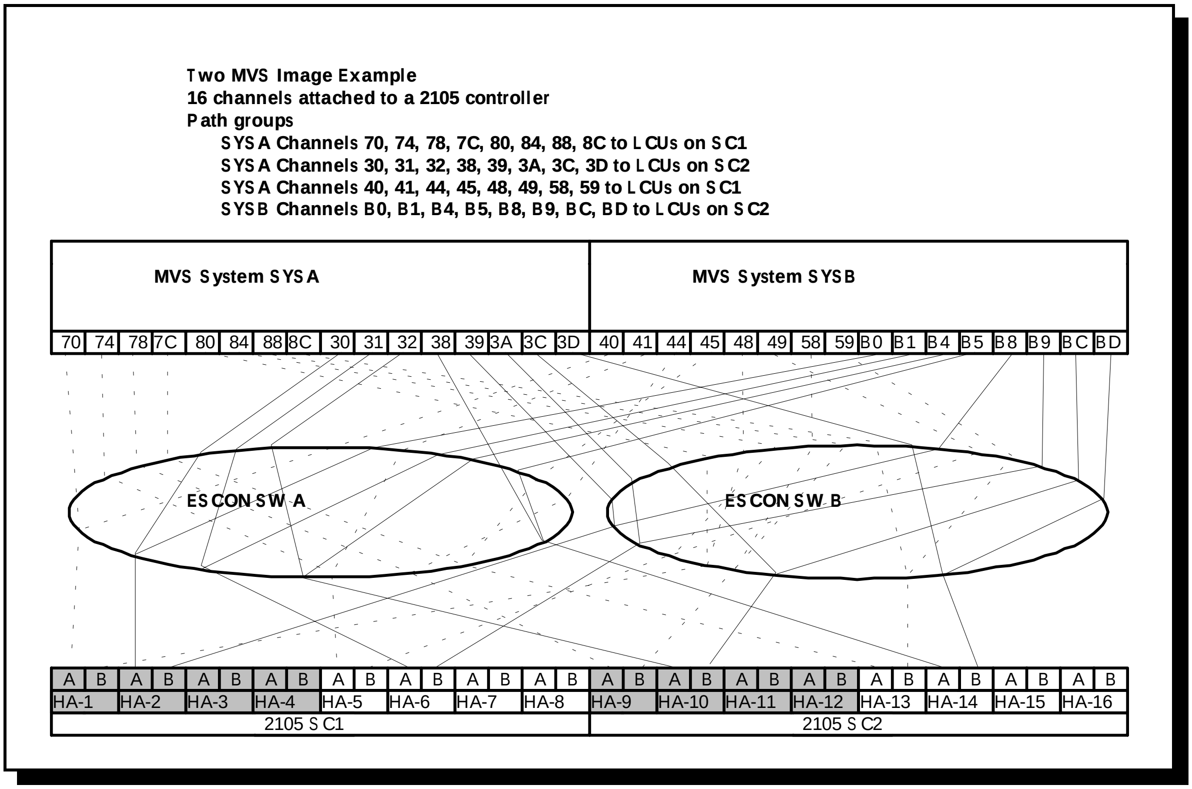

Two MVS Image Example
16 channels attached to a 2105 controller
Path groups
SYSA Channels 70, 74, 78, 7C, 80, 84, 88, 8C to LCUs on SC1
SYSA Channels 30, 31, 32, 38, 39, 3A, 3C, 3D to LCUs on SC2
SYSA Channels 40, 41, 44, 45, 48, 49, 58, 59 to LCUs on SC1
SYSB Channels B0, B1, B4, B5, B8, B9, BC, BD to LCUs on SC2
MVS System SYSA
MVS System SYSB
70 74 78 7C 80 84 88 8C 30 31 32 38 39 3A 3C 3D
40 41 44 45 48 49 58 59 B0 B1 B4 B5 B8 B9 BC BD
ESCON SW A
ESCON SW B
A B A B A B A B A B A B A B A B A B A B A B A B A B A B A B A B
HA-1 HA-2 HA-3 HA-4 HA-5 HA-6 HA-7 HA-8 HA-9 HA-10 HA-11 HA-12 HA-13 HA-14 HA-15 HA-16
2105 SC1
2105 SC2

Figure Config-2 *Example of ESS Channel Configuration with two hosts*

S/390 device type

ESS supports 3390-3, 3390-9 or 3390 with 3380 emulation device types. All device types are implemented the same way on ESS (i.e RAID-5 striping across several disks). There are no rotational delays or seek time penalties for the 3390-9. The choice of the device type depends on the customer needs, their migration paths and the application. Customers may want to consider larger capacity logical volumes (e.g. 3390-9) for one or more of the following reasons:

- Simpler storage administration with fewer total volumes
- Fewer “out-of-space” conditions with larger volumes
- UCB addressing relief

Customers who order the feature code for PAVs (Parallel Access Volumes) should consider the 3390-9 with multiple aliases as a choice for the device type. Here are some of the things to consider for the 3390-9 sized volumes:

- Four concurrent I/Os from a system can be active across 9 GB of capacity whereas with a 3390-3 configuration, there is a maximum of six concurrent I/Os per 9 GB.
- Specifying a 3390-9 capacity volume size means the data migration scheme to the subsystem does not rely solely on physical dump/restore.
- Physical dumps of 3390-9 volumes require a 3390-9 device for restore. It might make sense to do logical rather than physical dumps.
- Applications that are doing frequent Reserves may cause more serialization when using the larger capacity volumes.

Parallel Access Volumes (PAVs)

The Parallel Access Volumes (PAVs) feature of the ESS will provide significant performance improvements for S/390 workloads where IOSQ time is a problem. PAVs do not address problems caused by reserve/release, etc., which may also result in long IOSQ time. The standard configuration process will default to one alias per base 3390-3 and three aliases per 3390-9 base device but will allow the user to modify the default. Additional aliases can be configured using the ESS Specialist.

ESS and OS/390 software provide greater concurrency of I/O through the use of PAVs and multiple allegiance. Larger volume sizes allow for better exploitation of PAVs due to the 256 address limit per logical subsystem. Reserve/release reduces this concurrency and thus will hinder performance. In addition, using larger volumes means reserves affect more data since they're a volume level mechanism. Therefore, IBM strongly recommends that customers use Global Resource Serialization (GRS) or equivalent software to convert reserves into enqueues and so fully exploit the advantages of PAVs and multiple allegiance.

How many UltraSCSI ports for Open Systems?

For SCSI host attachments, the number of SCSI ports determine the ultimate capability of the subsystem to transfer data to the host. The following rule-of-thumb should help determine that an adequate number of ports have been defined. Note that this rule assumes that the ports are ordinarily active, that is, they are not just defined as an alternate path in case of failure.

- 120 GB of storage for each attached UltraSCSI port
- 80 GB of storage for each attach SCSI F/W port

These rules have been devised with the assumption that the user does not know specific attributes of the workload. If the user knows specific workload requirements, specifically how many MB/sec of data transfer (which for example can be obtained from an IOSTAT report on UNIX systems) required to meet their application needs, they can determine the correct number of ports by assuming the following sustained throughput capabilities per port:

- 30 MB/sec per UltraSCSI port
- 15 MB/sec per SCSI F/W port

It is best to have at least two ports available for any particular host, to be used in combination with IBM's Dynamic Path Optimizer software or other equivalent availablity-management software. This helps ensure availability of data in case a SCSI adapter or port failure (either within ESS or the host), reduce manual tuning and improve performance.

Size of ESS logical volumes (LUN size) for Open Systems

The size of the ESS logical volumes (LUN size) defined by the user generally does **not** have an impact on performance of the subsystem. ESS does not serialize I/O on the basis of logical devices. Most operating systems also support command tag queuing, allowing parallel operations to a single volume.The user can select the size based on operating system constraints or storage administration needs. The following are some general considerations:

- Some operating systems have specific requirements for LUN sizes. For example, Windows/NT may require large LUNs to hold a single data base in multi-host clustered environments.
- For those operating systems that do not have special LUN size requirements, many find that 16 GB LUNs may meet most requirements. This size may allow the best flexibility in being able to reassign storage from one host to another, and also not so small as to cause a proliferation in the number of LUNs seen by the host operating system.

Balancing load for maximum throughput

Generally speaking, spreading I/O load for performance-critical applications across the resources within a ESS subsystem (clusters, arrays, and device adapters) will maximize that application’s performance. When attempting to balance load within a ESS, placement of application data is the determining factor. The following are the resources most important to balance, roughly in order of importance:

- *Balance activity to the RAID arrays***.** Use as many RAID arrays as possible for the critical applications. Most performance bottlenecks occur because a few disks are saturated. Spreading an application across multiple RAID arrays ensures that as many disk drives as possible are available. **This is extremely important for Open Systems environments where cache hit ratios are usually low.** The following table specifies an I/O intensity rule of thumb to ensure that disk drives and RAID arrays are not overloaded, providing good application response time. It is based upon a conservative set of workload assumptions. It is intended to represent a planning estimate to avoid bottlenecks, not the maximum achievable throughput This rule of thumb applies to a random access workload using record sizes of 16K or less. When configuring a ESS configuration, define enough RAID arrays so that I/O per second per array does not exceed these values. This table can also be used to decide the number of arrays to use when configuring LUNs for a given application.

Read percentage	I/O per second per array
100% read	450
70% read	300
50% read	250

Table Config-2 *Raid array throughput capability*

- *Balance activity to the clusters***.** When selecting RAID arrays for a critical application, spread them across separate clusters. Since each cluster has separate memory buses and cache memory, this will maximize use of those resources.
- *Balance activity to the device adapters***.** When selecting RAID arrays within a cluster for a critical application, spread them across separate SSA device adapters.
- *Balance activity to the SCSI ports or ESCON channels.* For open systems, Use DPO (Storwatch Data Path Optimizer) or similar software for other platforms to balance I/O activity across SCSI ports. For S/390, load balancing is handled by the operating system and functions such as dynamic path reconnection.
- *Balance activity to the adapter bays.* When selecting SCSI ports or ESCON channels to assign to a given server or processor, spread them across the different adapter bays.
- *Balance across SCSI adapter cards.* For best performance, it is best to assign one port from each SCSI host adapter (there are two ports on the adapter), and then use the second port on each adapter card.

Considerations for mixing workloads

When using ESS, users may very likely be combining data and workloads from several different kinds of independent servers onto a single ESS. Examples of mixed workloads include:

- S/390 and UNIX
- UNIX from different vendors
- UNIX and Windows/NT
- Mission-critical and test

Sharing resource in ESS has advantages for storage administration and resource sharing, but does have some implications for workload planning. Resource sharing has the benefit that a larger resource pool (e.g.. disk drives or cache) is available for critical applications. However, some care should be taken to ensure that uncontrolled or unpredictable applications do not interfere with mission-critical work. This requires the same kind of workload planning the one would use when mixing those types or work on a server.

If a user has a workload that truly “mission-critical” (which could be either S/390 or UNIX-based), they may want to consider isolating it from other workloads, particularly if those other workloads are unimportant (from a performance perspective), or very unpredictable in their demands. There are several ways to isolate the workloads:

- Place the data on separate RAID arrays. Note that S/390 and open system data automatically are placed on separate arrays. This will reduce contention for use of disks.
- Place the data behind separate device adapters.
- Place the data behind separate ESS clusters. This will isolate use of memory buses, microprocessors, and cache resource. However, before doing that, make sure that a “half-ESS” provides sufficient performance resource to meet the needs of your important application. Note that Disk Magic provides a way to model the performance of a “half-ESS”.

Chapter 8. ESS Performance Monitoring

ESS Expert - Using the Expert for Performance Reporting

The ESS Expert allows you to examine the performance of the Enterprise Storage Subsystem from any location where you have network connections to the ESS Expert manager routine. The ESS Expert manager routine must run on a system that is, in turn, connected to the ESS Specialist, either over a private network, or through the Internet.

Once installed, you may have the Expert collect performance statistics on any number of ESS boxes, save the results in an internal database, and generate performance reports at various levels of detail. The Expert can be set up to discover any new hardware within its scope, and then periodically collect statistics from these ESS boxes.

Reports can be viewed for varying time periods, down to one hour. They can also be viewed at higher or lower levels of hardware details. This flexibility allows an analyst to quickly determine whether a performance problem exists or is brewing, then drill down to the appropriate level of detail, and assess where and when the problem is occurring.

There are reports for:
- Disk Utilization
- Disk - Cache Transfers
- Cache Usage

These reports may be viewed at the level of:
- Cluster
- Device adapter
- Disk Group
- Logical Volume

with the exception that Disk Utilization is not shown for Cluster or Logical Volume.

Some of the statistics provided are: disk utilization, disk to cache reads, cache to disk writes, cache hit ratios, and cache residency times.   Thresholds are provided to call attention to potential performance problems; they are reported in separate columns and generally show if there was a one-time anomaly or a recurring problem.

A skilled systems analyst can use these reports effectively. For example, let’s assume that the previous week’s cache summary reports show that I/O Requests per second have been abnormally high. By drilling down by day, then hour, you may find that this rate peaks at 10 am on Monday. Then by drilling down to the cluster, adapter, disk group, or logical volume, you can uncover exactly where the bulk of the I/Os are occurring. At this point, you can go to the configuration reports and determine which system(s) are using the volumes in question. From there, it is a matter of determining where the culprit is running on the system in question. In rare cases, it may be appropriate to load balance among RAID ranks or device adapter loops. Load balancing can be done at the system level or by reassigning the use of the logical volumes.

Detailed online help and useful suggestions are provided for all the Expert’s panels.

RMF

OS/390 customers will continue to have RMF direct access device activity and caching reports. This applies to ESS toleration, transparency & exploitation levels of OS/390. With ESS exploitation levels of software new information will be provided in the reports:

> MX - This field is in the Direct Access Device Activity report. It reports the current number of "paths" to a device. For Parallel Access Volumes, the number of paths is one plus the number of alias addresses assigned to the base address. Only the base address is reported and the statistics for this address is the aggregate of all activity for the address (base address plus all alias address activity). If the number of paths to the device has changed during the interval RMF flags the MX field with an asterisk (*). When calculating the aggregate values shown on the report, RMF considers the proportion of time aliases existed in the interval, if the number of aliases has changed during the interval.

A special note should be made of the RMF field "Device Busy Delay" in the Direct Access Device Activity report. Currently, RMF adds the control unit queuing delay for an I/O into this field. Control unit queuing is calculated by the control unit and is passed back to the host via an ESCON frame. Control unit queue occurs when there are extent conflicts for two or more I/Os that would ordinarily operate concurrently. These I/Os may be from the same system and arrive at the controller via different paths of the same PAV or they can come from different OS/390 systems. True device busy delay will only occur in normal circumstances with Device Reserve. RMF provides a measure of reserve activity. Without reserves, the device busy delay can be attributed to control unit queuing.

Chapter 9. Performance Marketing Tools

IBM SSD provides a number of **Performance Marketing Tools** that may be helpful for ESS performance analysis and capacity planning.   These tools are for use by IBM marketing personnel and business partners. Customers that are interested in understanding how the ESS will perform in their specific environment should contact their IBM storage marketing representative to discuss what type of analysis may be appropriate.

In order to use Performance Marketing Tools, customers must provide data on their existing storage configuration and current performance. For OS/390 customers, this data will usually be in the form of RMF and Cache summary reports. For Open Systems customers, IOSTAT or similar reports will usually be requested.   The accuracy of the analysis depends on the quality of the data provided.

Disk Magic for Windows is used to project disk performance attributes such as response time or throughput for new hardware. For S/390, most competitive and IBM disk hardware, including ESS are supported. For Open Systems, only the ESS and VSS are currently supported.   A typical Disk Magic study would have the following steps:

1. Collect data describing your current disk performance levels and workload. Select a time interval that represents peak levels of I/O activity
2. Input the current performance data and configuration into Disk Magic. Disk Magic will allow a “base” model of the existing configuration and workload to be created.
3. After a base is created, the workload may be migrated to an ESS or other new hardware configuration. Disk Magic will project the performance of the workload on the new hardware.
4. Disk Magic provides a connection to Lotus 1-2-3 which allows various graphs to be created which illustrate comparisons of performance projections to current results.

In general, Disk Magic provides highly accurate results. The user should be aware of some limitations that exist. Disk Magic uses an analytic queuing model of I/O subsystem performance. One of the important assumptions is that I/O arrives in a random manner. This is usually a good assumption for online transaction workloads but a bad one for sequential batch workloads. In general, if the percentage of sequential I/O exceeds 15%, it is a good idea to select a different workload interval to model.   Disk Magic models of intervals with high sequential content are often inaccurate. In these cases, you should consider using the Sequential Sizer (see description below).

Disk Magic for Windows may be useful in selecting the correct ESS configuration to meet your needs. For example, the following performance and configuration questions may be answered with the help of Disk Magic:

1. What disk drive size should be used?
2. How many channels should be configured?

3. How much disk capacity can the ESS support?
4. How sensitive is disk performance to hit ratio?
5. How much I/O growth can the ESS support?
6. How much performance improvement can ESS provide over current hardware?
7. What is the difference in performance between RAID and non-RAID?

In many cases, there is no other way to answer these questions. That is why Disk Magic is highly recommended when considering a migration to ESS.

In the past, Disk Magic and it's predecessor, DASD Magic for OS/2 were used for S/390 disk only. With the introduction of VSS and ESS, Disk Magic now supports Open Systems performance modeling as well. Since there are not tools such as RMF which provide detailed performance data for Open Systems, it may be difficult to obtain input data for modeling open systems. Under UNIX or AIX, the IOSTAT tool can provide some useful information. In many cases, you may need to approximate inputs such as hit ratios and transfer sizes from your knowledge of the application. It is recommended that you do sensitivity analysis on any input parameters which you must estimate.

Cache Analysis Aide (CAA) is a tool that simulates the behavior of cache in the ESS. A GTF CCW trace or I/O summary trace is required to run this tool. Since the ESS currently does not have multiple cache sizes available, using CAA is no longer as important as it was for prior disk subsystems. However, often you will consider merging several existing disk subsystems into a single ESS. If the original subsystems have large caches, the total cache of the existing subsystems may exceed the ESS cache size of 6 GB. In that case, CAA can be used to see what effect, if any, the smaller total cache size will have on hit ratios.

CAA is also useful for obtaining cache data needed to do Disk Magic studies. Although Disk Magic can do automatic cache modeling, this is less accurate than the cache simulation that CAA does.

CAA has been enhanced to provide limited support of AIX and UNIX. The latest CAA release includes a program which will convert the AIX I/O Trace to a GTF record format. The converted trace may either be run by itself or merged with S/390 GTF traces and run through the CAA. This allows the user to project Open systems and heterogeneous workload hit ratios.

Sequential Sizer is a set of Lotus 1-2-3 spreadsheets which may be used to project expected improvements to S/390 sequential batch performance when migrating to ESS. Sequential Sizer is not a model but uses rules of thumb for sequential disk performance analysis. It requires a Cache Subsystem Report from RMF for input. It can estimate percentage of improvement when moving various sequential workloads from IBM RAMAC or RVA disk subsystems to ESS.

Your IBM storage marketing representative can give you more details about these and other performance marketing tools for ESS.

DISCLAIMERS

The performance information contained in this document was derived under specific operating and environmental conditions. While the information has been reviewed by IBM for accuracy under the given conditions, the results obtained in your operating environments may vary significantly. Accordingly, IBM does not provide any representations, assurances, guarantees or warranties regarding performance. Please contact your IBM marketing representative for assistance in assessing the performance implications of the product in your specific environment.

Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

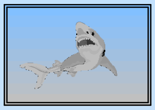
ACKNOWLEDGMENTS

This white paper reflects the work of many dedicated individuals in both San Jose and Tucson. Thanks go to Dr. Marc Goldfeder, Lee LaFrese, and Dr. Olga Yiparaki for their modeling work which has served as the measuring rod against which ESS performance has been measured. Much of the fundamental 390 measurements, algorithm verification, and microcode tuning work was accomplished by the Tucson team led by Larry Blount in partnership with Joe Hyde and Sonny Williams. The open systems measurements were mostly done by Neena Cherian in San Jose. Jim Sibley accomplished the 390 PAWS measurements plus the measurements which verified PAVs and other complex workload measurements. Thanks finally to Chris Saul, Brent Beardsley, and Mike Hartung, who provided careful review (alongside of the performance team!)

TRADEMARKS and COPYRIGHTS

APPENDIX.D.2.1.5

IBM ESS Storage Guide to Value



The

IBM

Enterprise Storage Server

□ Guide to Value and Capabilities for Business □

strategy/vision storage consolidation low total cost of ownership
breakthrough performance 24x365 availability ease-of-management
SANs investment protection host-to-host data transfer
and more....

The Enterprise Storage Server (ESS) is IBM’s next-generation leadership disk system.

The ESS represents the integration of multiple advanced technologies for safely storing, quickly accessing, and easily managing data. The ESS can be described as an all-in-one disk system for the entire enterprise that stands apart from the competition in significant ways.

Delivering innovations in design, performance, and function, the ESS meets - and generally exceeds - the stringent requirements that demanding customers place on large-scale disk storage subsystems. Satisfying needs ranging from easier management of storage resources, to 24x365 access, to higher levels of performance and lower cost of ownership, the ESS provides demonstrable business value. With concurrent support for a broad range of server families - such as NT servers, UNIX servers, and S/390 servers - the ESS is truly a storage solution for the entire enterprise. And the ESS is backed by IBM quality and support.

With the ESS, online data storage is not just a peripheral anymore.

This Guide identifies specific benefits the ESS provides today - and others planned for the future - that make a difference to your business. Together, these comprehensive benefits make a compelling case for the ESS.

Aligning storage technology with business needs

Notices

Published January 12, 2000

The information in this paper is provided by IBM on an "AS IS" basis without any warranty, guarantee or assurance of any kind. IBM also does not provide any warranty, guarantee or assurance that the information in this paper is free from errors or omissions.

Trademarks
AIX, AS/400, DFSMSdss, Enterprise Storage Server, Enterprise Storage Server Specialist, ESCON, FICON, FlashCopy, IBM, InfoSpeed, OS/390, RMF, RS/6000, Seascape, StorWatch, System/390, S/390, and Versatile Storage Server are trademarks of International Business Machines Corporation.

Other company, product, and service names may be trademarks or registered trademarks of their respective companies.

Please send comments via e-mail to: djsacks@us.ibm.com

> “Enterprise storage solutions such as IBM’s Enterprise Storage Server create an information delivery platform that provide[s] the needed flexibility to adapt in times of rapid change...Through use of an enterprise storage solution, companies can more effectively utilize their existing staff and — through the intelligence of the new solutions — free their staff to focus on the strategic issues of application value rather than the mundane and repetitive issues of data availability, capacity planning, and performance management.”
>
> -- John McArthur, International Data Corporation (IDC), November, 1999

IBM Enterprise Storage Server

Contents

Value Area **Page**

Enterprise Storage Server - Value Highlights 5
Strategy / Vision 6
Comprehensive Storage Consolidation Platform 7
Breakthrough Performance 8
Round-the-clock Access to Data 10
Ease-of-Management / Operations / Maintenance 12
Storage Area Network (SAN) Support 14
Attractive Cost of Ownership 15

This Guide discusses each value area from two perspectives. *Value to you* identifies specific business/ownership benefits. *How the ESS does it* identifies associated ESS attributes and facilities that deliver those benefits through technology.

Notes:

- Not every identifiable *Value to you* item or associated *How the ESS does it* item are necessarily cited in this guide.
- All items apply to all supported processors and operating systems unless specifically qualified.
- Virtual Architecture in the ESS is an IBM Statement of Direction.
- Attachment to S/390 processor FICON channels initially requires an IBM 9032 ESCON Director bridge card. Plans for direct FICON attachment for the ESS have been publicly previewed by IBM.
- Attachment to processor Fibre Channel adapter cards initially requires an IBM 2108 SAN Data Gateway. Plans for direct point-to-point, FC-AL, and switched Fibre Channel connections to an ESS have been publicly previewed by IBM.
- The term “midrange” is used throughout this guide to refer to non-mainframe server platforms such as AS/400, Novell NetWare, UNIX-based servers, and Windows NT.

For additional information on IBM storage please contact your IBM storage sales representative. Or visit the IBM storage Internet site at: www.ibm.com/storage. The specific home page for the Enterprise Storage Server is currently: www.ibm.com/storage/ess.

Enterprise Storage Server - Value Highlights

This table summarizes some of the most important benefits and How-the-ESS-does-it items on one page. Of course, one page can't identify every benefit of this comprehensive disk system. Please refer to the detailed itemized lists in this paper for a more comprehensive guide to ESS business and technical values.

ESS - Value Highlights
Comprehensive Storage Consolidation Platform
Scalable to over 11TB usable RAID-protected capacity
ESCON, Fibre Channel, UltraSCSI, FICON (preview) connections
Up to 32 direct host connections; hundreds possible in a SAN
Choice of disk capacities, most at 10K RPM
Breakthrough Performance
Large cache
Internal bus bandwidth near 800MB/second
Up to 64 internal disk paths / all can be transferring data concurrently
Up to 32 direct host connections / all can be transferring data concurrently
Data is striped for balanced disk utilization and parallel access
Specialized S/390 performance accelerators
Round-the-Clock Access to Data
No hardware single points of failure
Concurrent maintenance, repairs, microcode changes, upgrades supported
Hot-pluggable / hot-swappable components
Ease-of-Management / Operations / Maintenance
Web-based management interface
Logical volumes presented to hosts in customized capacities
Auto-call home for service
Remote problem analysis / diagnostics
Support for system management solutions via SNMP messages
Storage Area Network (SAN) Support
Fibre Channel attachment, including switch, hub, and gateway support
Attractive Cost of Ownership
Many popular features are standard (e.g., spare disks, management interface)
High performance RAID 5 lowers cost
Added-value features available & priced based on ESS capacity
Three year warranty standard
Added-Value Capabilities
Near-instantaneous volume copy (FlashCopy)
Remote Copy (remote mirroring) options (PPRC and S/390 XRC)
Subsystem-independent data transfer (InfoSpeed)
Virtual disk architecture (data compression and more)

Note: PPRC, FlashCopy, and InfoSpeed are not supported for the AS/400.

Strategy / Vision

Selecting a large-scale disk system is not always an easy process. Customers sometimes wish they could combine the performance of product A, the availability of product B, and the function of product C - all at the price of product D.

The ESS combines and integrates multiple, proven, world-class storage technologies into a single disk system. By integrating these multiple advanced technologies together, the ESS delivers capabilities not otherwise available in a single disk storage system. In this way, the ESS reduces barriers to efficient I/S operations by reducing technology constraints between users and the data they need.

Added-value functions, such as remote mirroring (peer-to-peer remote copy) already available in IBM's popular S/390 RAMAC disk system family, are extended in the ESS to support platforms such as NetWare, NT and UNIX. The "under the covers" use of the latest generation of serial disk technology supports data movement inside the ESS at unprecedented speeds. With its planned incorporation of virtual disk architecture and other enhancements, the ESS is poised for future total cost of ownership benefits such as data compression and dramatically reduced capacity requirements for internal data replication.

In this way, the ESS brings you closer than ever to the vision: any data / at any time / from any place / really, really fast / easily managed.

Value to you:

Integrated, proven storage architectures
State-of-the-art design
Conformance to industry standards
Stable, proven technologies
Positioned for future enhancements

How the ESS does it:

- Next generation Seascape Architecture disk system. Seascape products implement the principles of universal data access, operating-system based storage servers, and snap-in building blocks. Seascape's value is proven in the successful Virtual Tape Server and Network Storage Manager solutions from IBM.
- Next generation ANSI standard Serial Storage Architecture. (Internal SSA paths at 160 MB/second provide higher bandwidth than UltraSCSI at 40MB/second or Fibre Channel at 100MB/second.)
- Next generation enterprise-wide disk system for both midrange and S/390 servers: the best from previous subsystems plus leadership enhancements
- Next generation virtual disk architecture (Statement of Direction)
- Integrated industry standards including: SCSI-3, Serial Storage Architecture, Fibre Channel, ESCON, FICON, PCI buses, PPRC, XRC, Java, Ethernet, SNMP, TCP/IP, HTML, and Secure Sockets Layer (SSL).

Comprehensive Storage Consolidation Platform

Distributed computing has often led to islands of disks and data locked inside isolated processors. This not only makes enterprise storage management more complex, but also raises costs due to the inability to share storage resources.

The ESS enables installations to replace islands of isolated disks with a smaller number of shared subsystems. Some immediate benefits include reduced operational complexity, and capacity that can be easily distributed to processors that need it. Moreover, while many midrange systems' isolated disk storage is historically low-function, nearly all of the ESS's sophisticated capabilities are accessible to all attached servers, improving the overall quality of service the I/S organization delivers to its customers.

Drawing on IBM's years of experience in attaching storage systems to both IBM and non-IBM platforms, a single ESS supports concurrent attachment to a wide range of processors. Dozens of processors can be connected to a single ESS at the same time - with the ability to grow this number into the hundreds as SAN implementations continue to evolve.

Value to you:

Improved service to internal users and external customers
Ability to shift disk capacity to the host that needs it
Better control of assets/resources
Centralized management
Reduced operational complexity
Reduced staff/skills to manage storage
Enhanced data security
Shared (pooled) resources
Shared functions / common procedures

How the ESS does it:

- A wide variety of servers and operating systems can concurrently connect to the same ESS
- Flexible combinations of SCSI, Fibre Channel, ESCON, and FICON processor attachments
- Scalable growth to double-digit terabytes of usable space, in one system / one footprint
- Availability and performance attributes make consolidation practical
- Capacity is a pool that is flexibly partitioned for distribution to attached processors
- Capacity can be reassigned by the customer among processors while the ESS is online
- Capacity is customizable as a combination of standard (JBOD) and RAID-protected disks
- Capacity is customizable as logical volumes of flexible sizes to fit the needs of different servers
- Logical volumes can be assigned to multiple processors/paths for server clusters and data sharing
- Internal data replication (FlashCopy) supports nearly all processor environments
- Remote mirroring (Peer-to-Peer Remote Copy) supports nearly all processor environments
- No application program changes are required to migrate existing data to an ESS
- Dual internal clusters, multiple RAID arrays, and separate internal paths to groups of disks allow workload separation to minimize resource contention
- Single (Web-based) management interface
- Multiple SCSI-attached servers can optionally be daisy-chained to a single ESS SCSI port
- Fibre Channel, ESCON, and FICON allow processors to be located at extended distances from an ESS
- Fibre Channel and FICON allow a reduced number of host connections compared to SCSI and ESCON
- Disks in existing IBM VSS and 7133 subsystems can be consolidated into an ESS

Breakthrough Performance

The time it takes to perform I/O operations is often the single largest contributor to response time, particularly in commercial environments. The two primary dimensions of disk storage I/O performance are: I/O response time and I/O throughput. In simple terms, "how fast" and "how many."

The ESS incorporates numerous design elements from top to bottom that deliver unprecedented levels of performance in a high-function disk system, regardless of processor platform. For the OS/390 platform in particular, the ESS introduces innovative I/O performance accelerators that further strengthen the ESS's position as the performance leader among enterprise disk offerings.

Value to you:

Faster and more consistent service to users and customers
Improved productivity
Batch work completed in less time
Better handling of peak or unpredictable I/O workloads
Reduced I/O bottleneck
Reduced need for performance tuning
Minimized application delays waiting for backup tasks
Increased flexibility to consolidate distributed storage into a single disk system
Increased flexibility to mix applications together in one disk system
Reduced need to replicate data for performance reasons
Reduced application development time
Extended processor life

How the ESS does it:

- High performance design for both cache-friendly and cache-unfriendly workloads
- Large read/write cache for satisfying many I/O requests at electronic speeds. Multiple algorithms optimize cache efficiency. (Examples: 1) a read miss initiates staging into cache the optimum amount of data based on ESS analysis of current I/O patterns. 2) sequential read-ahead plus accelerated discarding of such data from cache after it is read by applications makes more cache available for other users.)
- All host connections can transfer data concurrently
- 64 data transfers at 40MB/second each can concurrently be in process across internal disk adapters. Every single disk can be in the process of transferring data in an interleaved manner
- All RAID-protected data is *striped* across multiple disks providing: parallel cache-miss processing even for a single logical volume, reduction of disk “hot spots” by naturally balancing utilization across disks, faster sequential throughput due to the ESS automatically processing large read and write requests in parallel across multiple disks, and, as a result of these benefits, the additional benefit of reduced manual tuning
- Disks have high-performance characteristics, including up to 10,000 RPM
- Disk capacity selection (4.5GB, 9.1GB, 18.2GB, 36.4GB) allows optimized price/performance
- High disk media transfer rates are supported by the high speed of paths between disks and cache
- Dual active 4-way SMP RISC processor clusters manage system activities
- Floating spare design eliminates the overhead of moving data back from a spare disk to a replaced disk

(Item list continued on next page)

Breakthrough Performance (page 2 of 2)

- Data (file) transfer between OS/390 and selected midrange hosts is performed at channel speeds and offloaded from both the network and disk system, using the IBM InfoSpeed solution.
- Internal aggregate bus bandwidth near 800MB/second
- Near-instantaneous internal data replication (FlashCopy), with an innovative option to reduce data movement overhead. The replicated copy can be used as the source of backup processing while production applications continue to use the original production volumes. Additionally, FlashCopy provides application developers a means to quickly create copies and backups of data for ad hoc testing with easy fallback.
- RAID rebuild, invoked in the rare case of a disk failure, favors preserving performance of production I/O
- RAID management is offloaded to the internal disk adapters
- RAID write penalties are minimized or eliminated in many cases. (For example, applications are not delayed for disk parity updates because writes are processed at cache speed.)
- Remote mirroring performance optimizations
- Dual internal clusters, multiple RAID arrays, and separate internal paths to groups of disks allow workload separation to minimize resource contention
- Fast I/O performance can offset delays caused by higher processor utilization
- The ESS can provide a high-speed replacement for traditional tape: no manual or robotic delays, and large total capacity that is planned to be further increased through data compression provided by virtual architecture.

Platform-specific items:

- AIX and NT: The optional Data Path Optimizer balances I/O traffic over multiple paths to the same volumes. (Support for additional selected servers is planned.)

- Midrange platforms: JBOD (non-RAID disk) support provides for a) eliminating RAID processing altogether for volumes where RAID protection isn't needed, and b) host-based mirroring in the rare case where that provides even higher levels of performance
- Midrange platforms: SCSI Command Tag Queuing (supports increased I/O parallelism)

- S/390: FICON channel support at 100MB/second (previewed)
- S/390: significantly increased I/O parallelism for OS/390 via the Parallel Access Volume (PAV) feature, a major advance over conventional subsystems that only support one I/O to one logical volume at a time from a given system image
- S/390: multiple system increased I/O parallelism via the Multiple Allegiance feature, a major advance over conventional subsystems that only support one I/O to one logical volume at a time from multiple system images
- S/390: application I/O priority support for OS/390 via the I/O Priority Queuing feature, a major advance over conventional subsystems that process I/Os only in first-come first-served order
- S/390: performance improvements through ESS support for new channel commands
- S/390: asynchronous remote mirroring to minimize application delays (OS/390 Extended Remote Copy)
- S/390: intra-job FlashCopy backups eliminate tape usage/delays
- S/390: fair access (a faster processor can't delay I/Os from a slower processor)
- S/390: support for the sequential-notification bit set in channel programs by operating systems (in addition to ESS sequential-detect algorithms); this is more efficient than disk systems that rely only on sequential-detect.

Round-the-clock Access to Data

Stored data can be converted to useful information only when applications can access it. Therefore, a key storage system value is its facilities that contribute to continuous data availability.

The ESS incorporates a comprehensive fault-tolerant hardware design, meaning that the failure of an individual hardware component should not prevent applications from accessing data. ESS fault tolerance begins with support for redundant connections to attached processors, extends to two copies of data written to the system maintained in two separate caches, includes storing data on RAID-protected disks, and more.

But the ESS's comprehensive availability protection goes well beyond fault tolerance. The ESS lets applications continue to access data while a failed component is being replaced. Additional disk capacity can be added dynamically. Software running on the internal, redundant processor clusters can generally be changed while applications continue to run. Further, innovative intra-subsystem data copy facilities can reduce application outages for backups from hours to seconds, or can keep applications running while a point-in-time copy of data is being sent to another location. And, inter-subsystem (remote) copy facilities support continuing business operations even across extended planned or unplanned outages.

Value to you:

Consistent service to internal users and external customers
Stable/reliable/dependable storage environment
Greater assurance of meeting production schedules
Protection against both planned and unplanned outages
More practical to consolidate distributed storage into a single disk system
Comprehensive protection against loss of data or loss of data integrity
Support for 24x365 business operations

How the ESS does it:

- Data integrity is protected in multiple ways. Examples: ECC (error correction) for data in cache and data on disks, Longitudinal Redundancy Checking (LRC: additional error checking) for data as it moves through the ESS, the use of dual write caches to protect against cache failures, and more.
- Predictive Failure Analysis can predict/prevent disk failures before they occur. Examples: internal periodic measurements of signal quality and head flying height.
- FlashCopy data replication minimizes application downtime for backups, data transfers, creating test copies, etc.
- High performance design reduces the need to take data offline for performance tuning
- Licensed Internal Code (internal software) can generally be changed while the ESS remains online
- Built-in spare disks support automated restoration of RAID protection after a disk failure
- Internal batteries (two redundant batteries) insure clean system shutdown in case of total loss of external power; if power loss is only transient, the ESS continues with normal operations.

(Item list continued on next page)

Round-the-clock Access to Data (page 2 of 2)

- Redundant AC and redundant DC power supplies
- Redundant cooling fans
- Redundant power cords
- Redundant, dual active processor clusters with automatic failover/failback
- Redundancy for data on disk via RAID protection, ensuring data remains accessible even if a disk drive fails
- Redundant internal adapters and data paths for every disk
- Repair/replace actions ("hot plugging") performed while the ESS remains online
- Upgrades done while the ESS remains online. (Example: adding new disks to increase capacity.)
- Internal hard disks (one per cluster) hold both the current and previous level of internal software, for quick fallback if necessary
- Remote mirroring (Peer-to-Peer Remote Copy) provides for continuing business I/S operations without data loss across an extended planned or unplanned outage
- Specialized data integrity protection for remote mirroring (Peer-to-Peer Remote Copy)

Platform specific items:

- AIX and NT: Data Path Optimizer product manages I/Os over multiple paths to the same volumes; in case of loss of a path, I/Os are automatically routed over other paths. (Note that some operating systems such as OS/390 include their own multi-path facilities.)

- S/390: Changes to parallel I/O processing can be made while the ESS remains online; for example, Parallel Access Volume (PAV) aliases can be added dynamically
- S/390: Concurrent Copy minimizes application down time during copy and dump operations
- S/390: Extended Remote Copy (XRC) provides asynchronous remote mirroring for optimized application performance in a remote copy solution
- S/390: Extended Remote Copy (XRC) in the ESS provides enhanced suspend/resume support
- S/390: Peer-to-Peer Remote Copy supports Geographically Dispersed Parallel Sysplex
- S/390: Virtual architecture minimizes application downtime for full volume DFSMSdss "defrags"

Ease-of-Management / Operations / Maintenance

A high-capacity high-function disk system is generally mission-critical in day-to-day I/S operations. A disk system needs to be easily and efficiently manageable to allow a business to fully realize the benefits of its technology.

The ESS pursues this goal from multiple perspectives. As a single, large-scale storage system, the ESS provides a single point of management control. The ESS provides a familiar, consistent management interface by taking advantage of its internal operating system-based intelligence to run a Web server under-the-covers. A thoughtful maintenance philosophy minimizes customer awareness of maintenance activity. ESS capacity is defined by the customer as logical volumes of different sizes, independent of the capacities of the physical disks installed.

Value to you:

Simplified operations / reduced staff
Any-time / any-location view and control
Secure access to management facilities
Support of enterprise-wide systems management solutions
System maintenance performed without customer intervention
Logical configuration tailorable to the I/S environment
Easy migration of data from conventional products to an ESS
Centralized backup of midrange data to OS/390 tape
Indirect improvement of access to other I/S resources

How the ESS does it:

- ESS StorWatch management interfaces are Web-based, providing ease-of-use and any-location management
- The StorWatch Specialist provides configuration customization under customer control
- The StorWatch Expert reports on the ESS internal (global) view of performance and capacity
- Access to ESS management facilities is secured by password, SSL (Secure Sockets Layer) protocol, and/or private network
- E-mail and pager notifications of system events can be automatically sent to designated personnel
- SNMP messages can be sent to designated addresses such as systems management applications
- Built-in spare disks support automated restoration of RAID protection after a disk failure
- Data Copy Services can be combined (example: you can make a FlashCopy of a remote-copied volume for input to backup processes)
- Data Copy Services procedures can be managed as named tasks
- FlashCopy: once a FlashCopy of a given volume is complete, another can be started to the same or different target volume; this can be repeated without any limit to the number of outstanding copies.
- FlashCopy: target volumes are specified dynamically by the customer; dedicated volumes are not required
- FlashCopy: tape drive use for disk backups can be reduced or rescheduled. (Once the FlashCopy is made, production applications can be restarted without waiting for disk data to first be written to tape. Further, multiple backups can potentially be batched onto one tape reducing tape cartridge requirements.)
- Reduced manual tuning effort due to high levels of performance

(Item list continued on next page)

Ease-of-Management / Operations / Maintenance (page 2 of 2)

- Flexible intermix of RAID-5 and JBOD (non-RAID) storage under customer control
- Logical volumes: up to 8192 can be created in one ESS (4096 for midrange, and 4096 for S/390)
- Logical volumes are configurable over a wide range of capacities (from .5GB to the capacity of a RAID array).
- “Phone home” automatically notifies IBM if the ESS needs service
- IBM product specialists can run system diagnostics from a remote support center and download any needed software changes; access to the ESS is password-protected.
- High performance, high scalability, high availability, and added-value function attributes make it practical to consolidate multiple disk subsystems into a single ESS, simplifying storage operations and management
- The ESS may be a potential tape replacement due to its large capacities and other benefits (such as RAID protection and elimination of manual and robotic delays)
- Space management simplifications accrue due to planned virtual architecture (e.g., reduced wasted space)
- Small floor space requirement

Platform Specific Items:

- Midrange systems: Data migration to an ESS can be via standard operating system commands or mirroring/split mirroring facilities. Optional IBM services are also available.

- OS/390 ⇔ Midrange file transfer: The InfoSpeed product provides high-speed data transfer between selected midrange and OS/390 servers. The implementation is subsystem-independent. Data transfer overhead is offloaded from both the network and the disk system. One-step transfer of database data is supported, eliminating the need for intermediate temporary flat files required by conventional data transfer facilities.
- OS/390 ⇔ Midrange backup/restore: The InfoSpeed product provides high-speed backup of midrange data to OS/390 tapes/libraries, offloading the network. The Tivoli Storage Manager (formerly ADSM) supports this high-speed solution.

- S/390: FlashCopy is detected and automatically invoked by the standard DFSMSdss copy utility - no JCL changes are needed
- S/390: RMF and the IDCAMS LISTDATA command report on ESS performance statistics
- S/390: Multiple data migration methods are available: Examples: IBM services (which employs a nondisruptive migration solution), or using the ESS as an XRC target from a source disk system that supports XRC.
- S/390: Because of its unprecedented performance (high performance design augmented by S/390 accelerators), it is practical to define ESS logical volumes as large capacity 3390 model 9 devices, reducing the total number of volumes to manage, reducing processor memory requirements, and conserving disk addresses.
- S/390: Parallel Access Volume (PAV) tuning can be automated by the OS/390 Workload Manager
- S/390: P/DAS (Peer-to-Peer/Dynamic Address Switching) provides nondisruptive volume movement, and is enhanced in the ESS

- AS/400: Selecting a logical volume size of 36GB allows for logical volume reassignment among AS/400 and other midrange platforms

Storage Area Network (SAN) Support

Storage Area Networks (SANs) hold the promise of significantly improving the overall storage environment - particularly in those installations with multiple heterogeneous midrange processors and storage systems.

IBM’s Enterprise SAN initiatives include connectivity/components, management, applications (e.g., capacity pooling and LAN-less backup), and service offerings. The ESS, as a disk system, participates as a SAN component that connects to Fibre Channel-based SANs which can include hubs and switches. This ability offers immediate benefits such as increased distance, performance, and access to storage. As SAN applications emerge, additional benefits will accrue.

Through its operating-system based intelligence, the ESS is well-positioned to provide additional SAN support and exploitation as industry standards evolve.

Please visit the IBM SAN Web site at www.ibm.com/san for details on the IBM Enterprise SAN.

Value to you:

Separates storage from processor dependencies
Offloads I/O traffic from the LAN/WAN network to a storage-optimized network
Supports increased connectivity of many servers to one storage system
Storage capacity and function shareable by multiple processors
Positions installations to support emerging SAN-based applications

How the ESS does it:

- Flexible intermix of UltraSCSI, Fibre Channel (FC) , ESCON, and (previewed) FICON host connections
- The IBM SAN Data Gateway allows attaching host FC connections to ESS SCSI ports while preserving FC benefits such as distance and performance
- The ESS supports the IBM Fibre Channel Hub and the Netfinity Fibre Channel Hub, providing distances of up to 11km between an ESS and attached servers
- The ESS supports the IBM SAN Fibre Channel Switch for improved performance, reliability, connectivity, and management over hubs
- IBM’s commitment to open SANs, demonstrated by its SAN Interoperability Lab, maximizes enterprise flexibility and SAN component choice
- The ESS is positioned to support applications such as LAN-less and Server-less data movement. IBM’s Tivoli subsidiary has published a road map to deliver these SAN functions and more. Already, IBM’s InfoSpeed facility provides specialized LAN-less file transfer between OS/390 and selected midrange servers, and LAN-less backup of midrange servers to OS/390, with subsystem-independent facilities.
- StorWatch management is provided for the ESS, IBM SAN Data Gateway, and IBM SAN Fibre Channel Switch
- IBM Global Services can assist in SAN planning, design, implementation, and testing
- Seascape storage server provides operating system-based intelligence and flexibility for future enhancements

Attractive Cost of Ownership

With the rapid pace of change in storage technology, customers are increasingly concerned that any investment they make may too soon be obsolete. Or, that they may unnecessarily pay more than they need to for a disk system and added-value features. The ESS addresses these concerns in multiple ways.

For example, the newer ESS design can lower costs compared to conventional disk systems with older designs. Several ESS advanced features are optional, and priced based on system capacity. And, ESS capacity can be configured in multiple ways to meet individual price/performance requirements.

The ESS preserves the investment many organizations have already made in selected IBM disk subsystems. SSA disk drawers (whether in IBM 7133 or VSS subsystems, subject to configuration rules) can be redeployed as ESS disk capacity, preserving an existing investment in disk drives. Such investment protection/enhancement is rare in the industry.

The ESS is also designed to incorporate "releases" of enhancements over time, so that even the first ESS ever delivered to a customer can evolve to keep pace with advances in technology.

Value to you:

Attractive upfront costs
Attractive ongoing costs
Attractive upgrade costs
Support for budget constraints
Ability to affordably acquire new technology
Investment justification
Investment protection
Long asset life
Potential elongation of processor life
Minimized technical/business risk

How the ESS does it:

- Multiple spare disks are standard at no extra charge
- StorWatch Specialist management interface is standard at no extra charge
- Three year warranty is standard
- Design efficiencies minimize upfront hardware costs. (Examples: Efficient cache algorithms and fast internal data paths mean less cache may be required to achieve higher performance than conventional subsystems. A high-performance RAID 5 design significantly reduces the number of disks needed compared to a mirroring design.)
- All supported processors may be intermixed on one ESS without additional charge
- Upgrade options (e.g., additional disks, host attachments, and optional features) allow you to buy what you need when you need it
- Design for upgrade-in-place with planned future enhancements
- Optional added-value functions are incrementally priced based on system capacity
- ESS software (Licensed Internal Code) and purchased features are owned by the customer for ease of resale

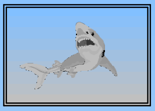
(Item list continued on next page)

Attractive Cost of Ownership (Page 2 of 2)

- No monthly maintenance charges for added-value features (S/390 Parallel Access Volumes, FlashCopy, remote copy)
- FlashCopy: target volumes are specified dynamically under user control; pre-purchased, dedicated volumes are not required
- ESS functions and facilities are shareable across nearly all server platforms
- High ESS performance may extend processor life / allow processor upgrades to be delayed. (The simple reason: processor response time degrades at higher utilizations - this may be offset by the speed of the ESS.)
- Independent scalablity of the number of disks and the number of host connections
- Quality IBM technology, service, and support reduces risks
- RAID and non-RAID (JBOD) disks are intermixable for optimized price/capacity/availability
- An ESS is flexibly redeployable across different servers in the enterprise, increasing asset life
- Existing 7133 and VSS disk assets can be re-deployed in an ESS; in particular, usable capacity is multiplied when migrating mirrored disks to ESS RAID 5, providing an attractive return-on-investment
- Scalability / consolidation facilitate cost savings (compared to buying additional subsystems)
- Storage capacity is reassignable across hosts under customer control, while the ESS remains online. This maximizes the use of existing resources without vendor involvement or special fees.
- Two ESS footprints, base frame + optional expansion frame, scale floor space with capacity
- Virtual architecture data compression multiplies usable disk capacity
- Virtual disk copies of data require minimal or no additional physical capacity, effectively increasing usable system capacity. (This demonstrates synergy between FlashCopy and virtual architecture.)
- Virtual disk minimizes wasted space. An example is a logical volume that is only partly filled. Another example is allocated-but-unused space within a logical volume. With virtual disk the unused space remains in a pool of available space shared by the entire disk system.

Platform Specific Items:

- Selected S/390 performance accelerators are standard at no extra charge (I/O Priority Queuing, Multiple Allegiance, support for new optimized I/O commands)



The

IBM

Enterprise Storage Server

□ Value for Business □

Aligning storage technology with business needs