IBM Storwize family storage systems with SAS workloads on IBM Power Systems servers

Storage recommendation

IBM Systems and Technology Group ISV Enablement
March 2013

IBM

Table of contents

Abstract

This paper describes the IBM Storwize family storage systems along with configuration and implementation best practice considerations for the IBM Storwize V7000 to help optimize I/O performance for SAS workloads on IBM Power Systems servers.

The primary audience for this paper is technical such as SAS server, network, and storage administrators, and technical implementation architects.

Introduction

SAS® Business Analytics provide an integrated environment for predictive and descriptive modeling, data mining, text analytics, forecasting, optimization, simulation, experimental design and more. Together with IBM® infrastructures consisting of IBM Power Systems™ and IBM System Storage® components, the resulting SAS and IBM solutions form a powerful and robust platform optimized for client analytic workloads.

The IBM Power Systems family of servers includes proven workload consolidation platforms that help clients to control costs while improving overall performance, availability, and energy efficiency. With these servers and IBM PowerVM® virtualization solutions, an organization can consolidate large numbers of applications and servers, fully virtualize it's system resources, and provide a more flexible and dynamic IT infrastructure.

IBM Power Systems, combined with IBM Storwize® family storage systems, provide an integrated analytical environment that can support the business analytics needs of demanding organizations.

The IBM System Storage SAN Volume Controller (SVC) and IBM Storwize® family disk systems provide enterprise-grade reliability, storage efficiency, and ease of management for midsize data storage, combined with support for advanced features.

One of the primary advantages of SAN Volume Controller and Storwize family disk systems is using the storage virtualization features. Storage virtualization allows an organization to implement pools of storage across physically separate disk systems (including systems from different vendors). Storage can then be deployed from these pools and can be migrated between pools without any outage to the attached host systems.

Storage virtualization yields numerous benefits for storage administration and management that includes:

- Combining storage capacity from multiple heterogeneous disk systems into a single reservoir that can be managed as a business resource rather than as separate boxes
- Increasing storage utilization by providing host applications with more flexible access to capacity
- Improving productivity for storage administrators by enabling management of heterogeneous storage systems through a common interface
- Improving application availability by insulating host applications from changes to the underlying physical storage infrastructure
- Enabling a tiered storage environment where the cost of storage can be matched to the value of data and easily migrated between those tiers

This technical white paper describes some of the features and capabilities that set apart the Storwize family storage products and how these capabilities can be used for I/O performance needed by SAS workloads.

This paper also focuses on the IBM Storwize V7000 storage system. Storwize V7000 is a mid-range disk storage system designed to provide ease of use along with solid performance for clients' demanding application workloads, such as virtualization, analytics, and cloud computing. For these reasons, this paper focuses on best practice implementations using Storwize V7000 configurations that help optimize SAS analytical workloads with the understanding that (in general) these concepts apply to the whole Storwize family of storage offerings.

This paper extends the detailed joint SAS / IBM white paper, *Storage best practices: SAS 9 with IBM System Storage and IBM Power Systems* published at:
http://www.sas.com/partners/directory/ibm/SAS_IBM_Storage_Best_Practices0311.pdf

Refer to the detailed paper for additional SAS I/O characteristics, and for detailed storage-related configuration and tuning guidelines.

IBM Storwize family storage systems

The IBM Storwize family provides robust midrange storage solutions well-suited for many SAS customer environments. Although some of the new SAS Business Analytics applications perform some random data access, for the most part, SAS workloads can be characterized as predominately large-block sequential I/O requests with high volumes of data. Specific Storwize family I/O configurations and tuning possibilities for SAS are examined in the following sections of this paper.

IBM has a broad set of midrange storage offerings running the same common SAN Volume Controller code base. In this paper, **Storwize family** refers to these IBM offerings (as of April 1, 2013):

- IBM Storwize V3500 – available only in GCG (China)
- IBM Storwize V3700
- IBM Storwize V7000 (block only) and IBM Storwize V7000 Unified (block and file)
- IBM Flex System™ V7000 Storage Node
- IBM System Storage SAN Volume Controller

Storage can be provided for the SAS application by many different methods, such as using internal server storage, direct-attached storage systems (SAS or iSCSI), or Fibre Channel (FC) switch attached storage area network (SAN) storage. This paper focuses on the block-based I/O Fibre Channel switch attached SAN storage products for SAS, and specifically on Storwize family storage systems. This paper also assumes that the reader has prior knowledge of IBM PureSystems™, IBM PureFlex™ System and the IBM Flex System. You can find additional information regarding the IBM PureSystems product family, including the IBM PureFlex and IBM Flex System at: **ibm.com**/ibm/puresystems/us/en/pf_overview.html.

Storwize family storage virtualization

The Storwize family of storage systems supports the virtualization of external storage arrays from a wide range of leading storage vendors.

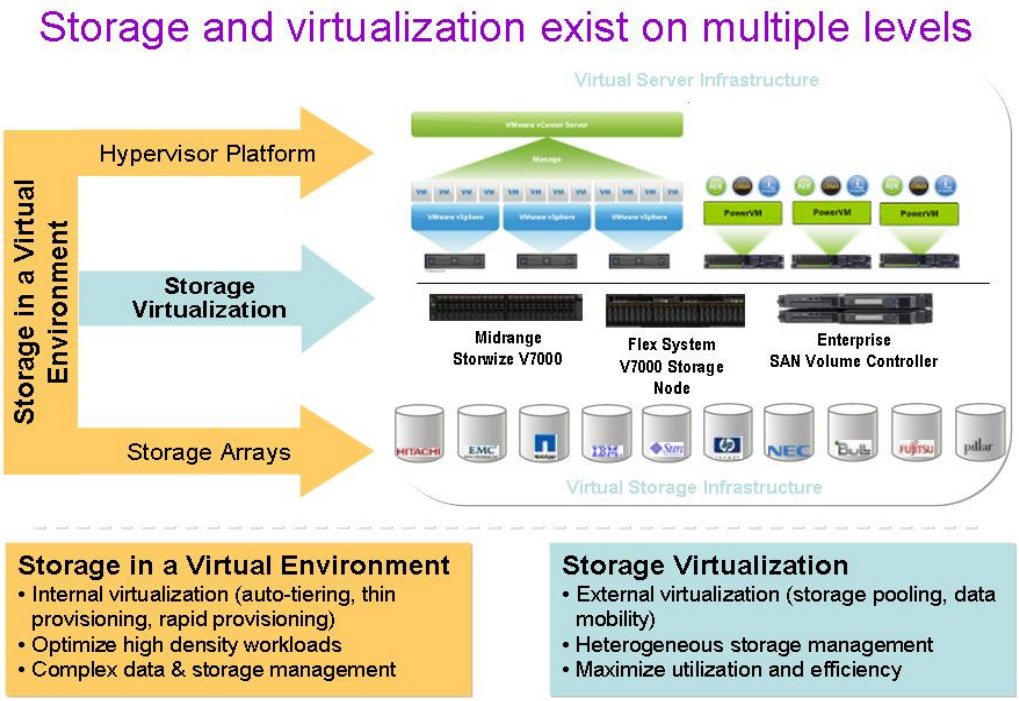

Figure 1: Storwize family storage virtualization

IBM SAN Volume Controller, IBM Storwize V7000, and IBM Flex System V7000 Storage Nodes provide modular and scalable storage that includes the capability to virtualize external SAN-attached storage and their own internal storage. Refer to Figure 2. The SAN Volume Controller can have up to four internal solid-state drives (SSD) that can be used in addition to external storage for virtualization. The IBM Storwize V3700 system is positioned for smaller environments and provides virtualization for its internal storage only.

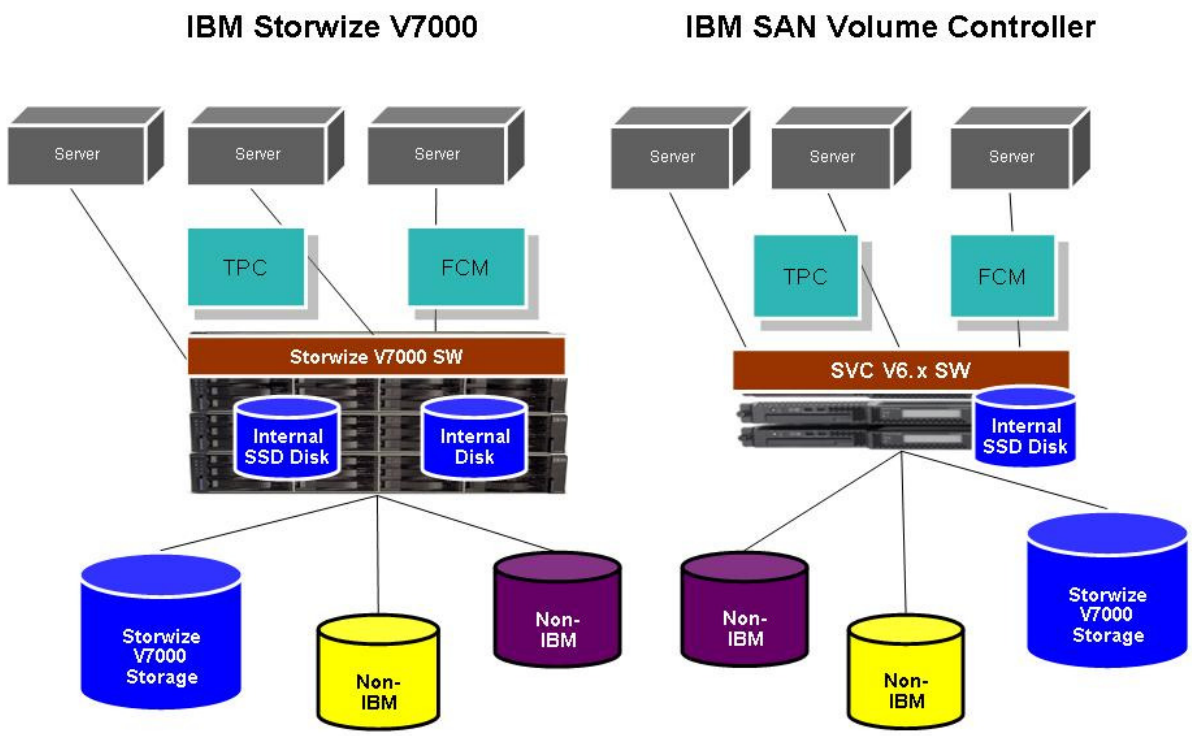

Figure 2: Storwize V7000 and SVC

Note: You can refer to the SVC, Storwize V7000, and Flex System V7000 documentation in the "Resources" section for storage virtualization best practices, hardware support matrix, and design-specific guidelines.

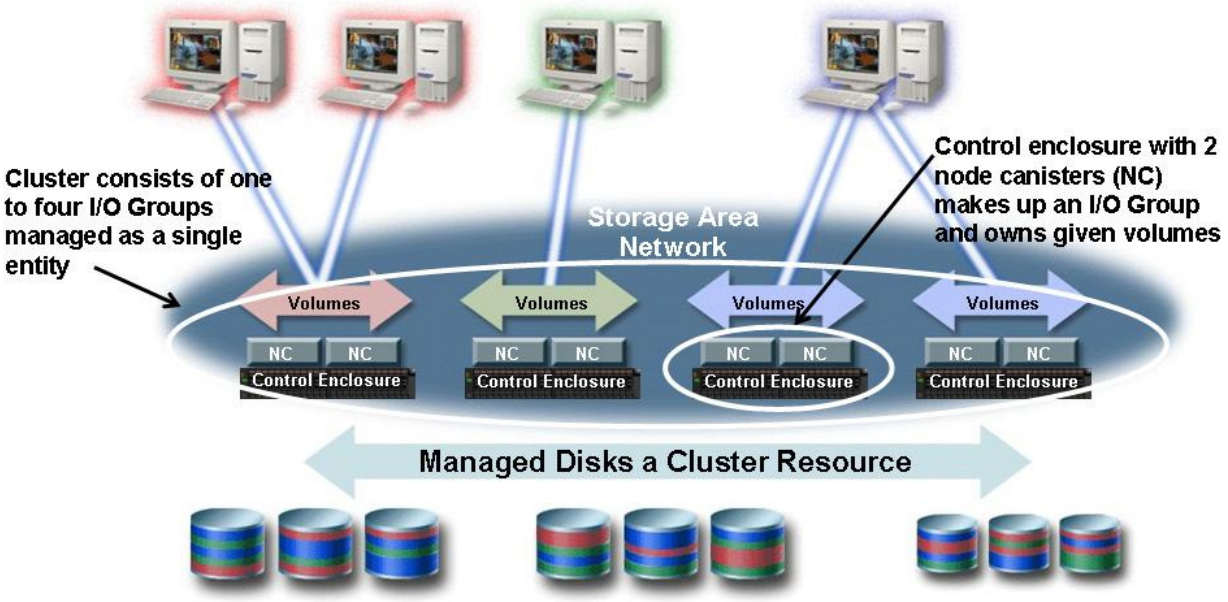

Figure 3: Storwize family I/O groups and virtualization

The Storwize V7000, Flex System V7000, and Storwize V3700 systems consist of a pair of redundant controllers, called nodes, and expansion enclosures that can support SSD, serial-attached SCSI (SAS), and near-line SAS drive types. As mentioned, the SAN Volume Controller does not support expansion enclosures or internal disk storage. The SAN Volume Controller virtualizes external disk storage only. The pair of redundant nodes forms an I/O group. Clustering allows multiple I/O groups to be managed as a single storage system. The SAN Volume Controller, Storwize V7000, or Flex System V7000 cluster can have up to four I/O groups, while only a single I/O group is allowed in a Storwize V3700 cluster.

Storwize family virtualization components

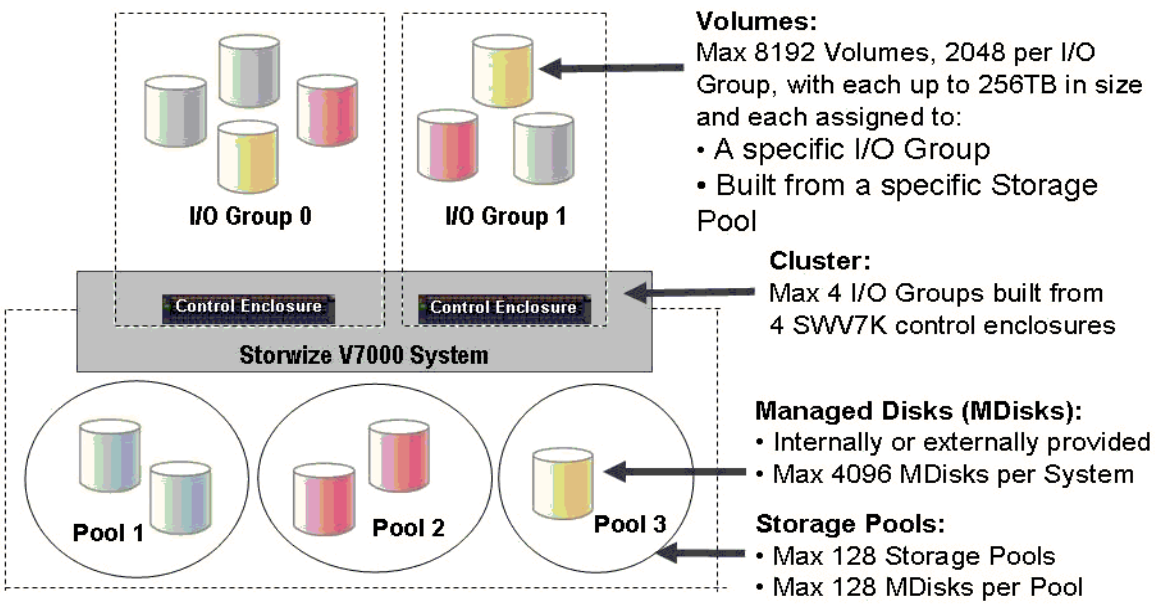

Figure 4: Storwize V7000 virtualization big picture

Managed disks

The Storwize family storage systems use a basic storage unit, called managed disk (MDisk). MDisks are collected into one or more storage pools. An MDisk must be protected by Redundant Array of Independent Disks (RAID) to prevent loss of the entire storage pool. These storage pools provide the physical capacity to create volumes for use by hosts. The Storwize family allows the following two different types of managed disks:

- Internal array MDisk
  - The internal RAID implementation inside the system takes drives and builds a RAID array with protection against drive failures.
- External SAN-attached MDisk
  - An external storage system provides the RAID function and presents a logical unit (LU) to the Storwize family storage system.

Storage pools

A storage pool is a collection of managed disks that can be allocated into volumes or virtual disks for the SAS application file system use.

- The primary property of a storage pool is the extent size which by default when using the GUI is 256 MB (see *Table 1: Storage pool extent size and volume size*).
  - Each MDisk is divided into segments of equal size called extents
  - This extent size is the smallest unit of allocation from the pool.
- When adding managed disks to a pool they should have similar performance characteristics:
  - Same RAID level
  - Roughly the same number of drives per array
  - Same drive type (SAS, NL SAS, SSD except if using Easy Tier. See Easy Tier section in this document for additional information.)
  - Similar performance characteristics for external storage system Mdisks
- Data from each volume will be spread across all Mdisks in the pool, so the volume will perform approximately at the speed of the slowest MDisk in the pool.
  - The exception to this rule is that if using Easy Tier you can have 2 different tiers of storage in the same pool, but the Mdisks within the tiers should still have the same performance characteristics.

Extent size	Maximum volume capacity for normal volumes (in GB)	Maximum volume capacity for thin-provisioned volumes (in GB)
16	2,048 (2 TB)	2,000
32	4,096 (4 TB)	4,000
64	8,192 (8 TB)	8,000
128	16,384 (16 TB)	16,000
256	32,768 (32 TB)	32,000
512	65,536 (64 TB)	65,000
1024	131,072 (128 TB)	130,000
2048	262,144 (256 TB)	260,000
4096	528,288 (512 TB)	520,000
8192	1,05,576 (1,024 TB)	1,040,000

Table 1: Storage pool extent size and maximum volume size

Internal storage RAID

The Storwize family storage system with internal drives will initially contain drive objects that are not associated with storage arrays (MDisks) or storage pools. The drives must be formatted members of storage pools to be used for volumes and file systems by the SAS application. The unformatted drive objects cannot be directly added to storage pools. These drives are managed by the same software component with SVC which manages SAN-attached managed disks. The drives need to be included in a RAID to provide protection against the failure of individual drives.

Storwize family supported RAID levels are:

- RAID 0 (striping, no redundancy)
- RAID 1 (mirroring between two drives)
- RAID 5 (striping, can survive one drive fault)
- RAID 6 (striping, can survive two drive faults)
- RAID 10 (RAID 0 on top of RAID 1)

The Storwize family storage systems GUI also provides templates for commonly used and recommended input parameters for:

- Volume creation
- IBM FlashCopy® creation
- RAID array creation

You can find additional information about RAID presets in the IBM Redbooks® entitled, *Implementing the IBM Storwize V7000*,(SG24-7938) at **ibm.com**/redbooks/redbooks/pdfs/sg247938.pdf.

More detail about what the RAID levels actually mean is available at http://en.wikipedia.org/wiki/RAID.

Write/Read combine

When accessing MDisks, the Storwize family storage system issues write or read I/O requests based on:

- Strip size for array read operations which is by default 256 KB
- Stripe size for array write operations which depends on the array width
- 256 KB for SAN-attached external managed disks

In this example, strip and stripe are defined as follows:

- Strip – In a striped RAID array, a strip is the number of consecutively addressed blocks on each drive.
- Stripe – An array stripe is a portion of the array data starting with one strip on the first array disk member and proceeding through a corresponding strip on each of the other disk members.

Why write/read combine matters?

- In general, coalescing or combining I/O adds less strain on external storage systems virtualized by the Storwize family storage system.
    - There is less chance to reach I/O operations per second (IOPS) limits using virtualization on slower disk technology (for example, SATA drives or more commonly, ports on the external controllers).
- Full stripe write operations increase performance for RAID5/RAID6 arrays, as this allows parity to be calculated without any reads.

Interoperability within the Storwize family

Because one Storwize family storage system may be configured to act as backend storage to another Storwize family system, SVC and Storwize V7000 use a layered concept to define how two products running the same virtualization software interact.

- Replication layer (SVC, Storwize V7000, Flex System V7000)
- Storage layer (Storwize V7000, Flex System V7000, Storwize V3700)

They interoperate as shown in Table 2.

Layer	Replication	Storage
Replication	Replication layer will perform remote copy between systems including SVC	Replication layer will treat the storage layer as a back-end storage system to virtualize.
Storage	Replication layer will treat the storage layer as a back-end storage system to virtualize	Will perform remote copy between systems except SVC

Table 2: Storwize family interoperability

Note: Provisioning Storwize V7000, Storwize V3700, or Flex System V7000 as the storage layer to systems configured as the replication layer requires that all systems are running V6.4.x or later software as shown in the example in Figure 5.

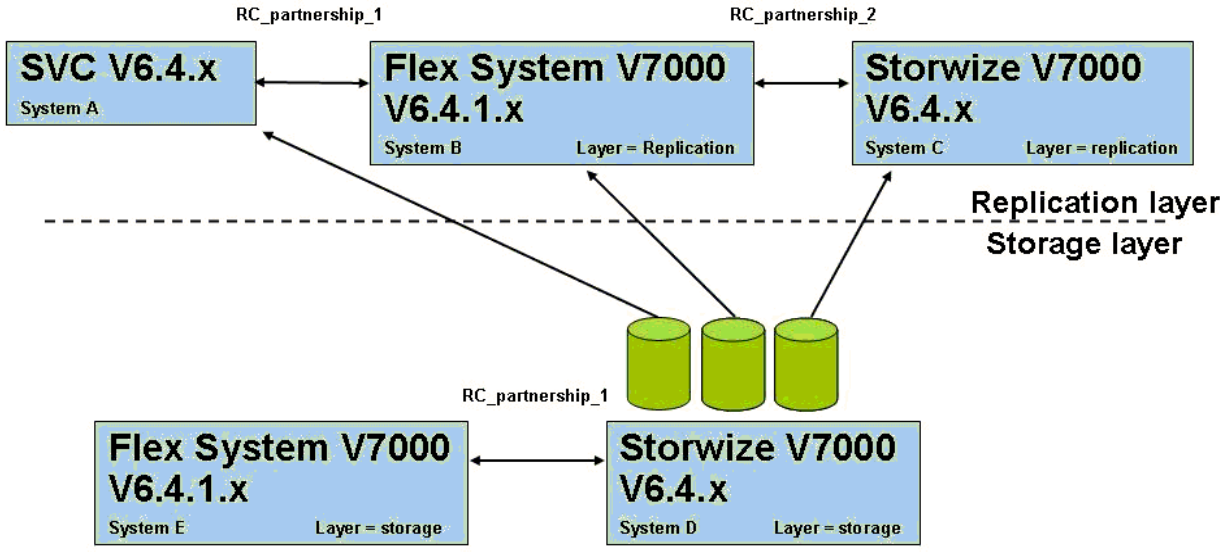

Figure 5: Remote copy / virtualization – configuration Example

For additional information regarding external storage virtualization, refer to *Implementing the IBM System Storage San Volume Controller* or *Implementing the IBM Storwize V7000* Redbooks listed in the "Resources" section of this paper.

Other Storwize family features

Powerful GUI: Another feature of the Storwize family is a simple and easy-to-use GUI designed to allow rapid and efficient storage deployment.

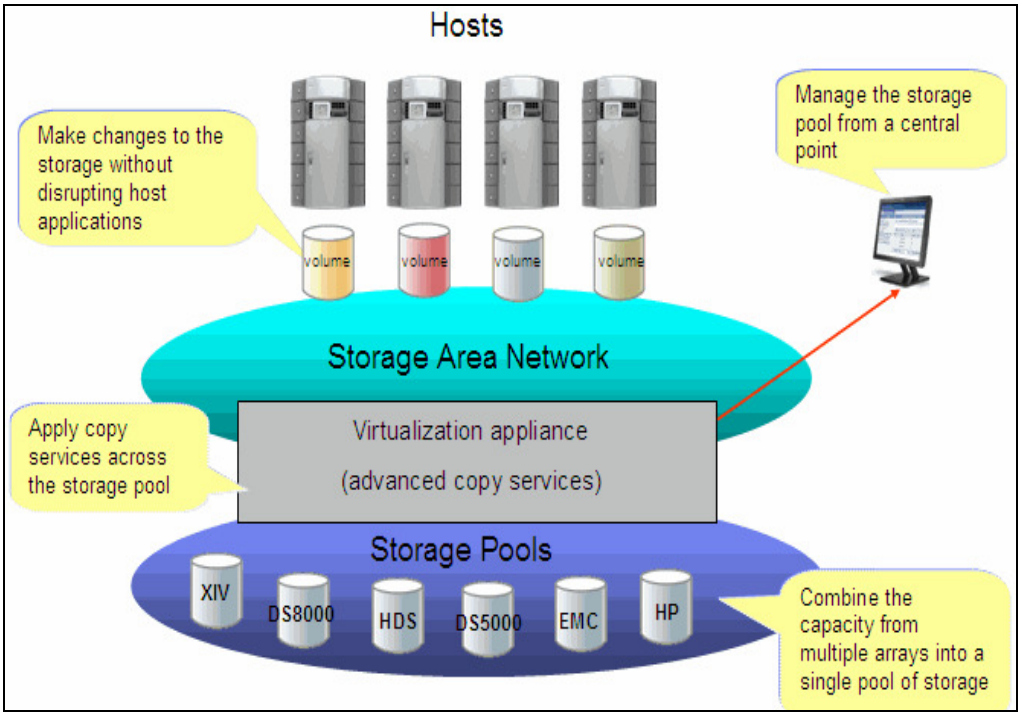

Figure 6: Features of the Storwize family web-based GUI

The GUI runs through a web browser window on the SAN Volume Controller and Storwize family systems. The GUI runs optionally through the IBM Flex System Manager™ (FSM) on the Flex System V7000 Storage Node. There is no requirement for a separate console.

Physically, the SAN Volume Controller and Storwize family storage system hardware include flexible host connectivity options with support for 8 Gb Fibre Channel, 1 Gb iSCSI, and 10 Gb iSCSI host connections. In addition, the Storwize family also contains a full array of advanced software features that include:

- Seamless data migration
- Thin provisioning
- Volume mirroring
- Global Mirror and Metro Mirror replication (except Storwize V3700)
- FlashCopy – 256 targets, cascaded, incremental, space efficient (thin provisioned)
- Integration with IBM Tivoli® Productivity Center
- IBM System Storage Easy Tier® that provides a mechanism to seamlessly migrate hot spots to a higher performing storage pool.

Benefits of using IBM Storwize V7000 with SAS workloads

This section of the paper examines the general strengths and features of the Storwize V7000 system for most customer environments. Later sections of this paper focus on the general guidelines for setting up and tuning the IBM Storwize family system storage environment for the specific needs of SAS workloads.

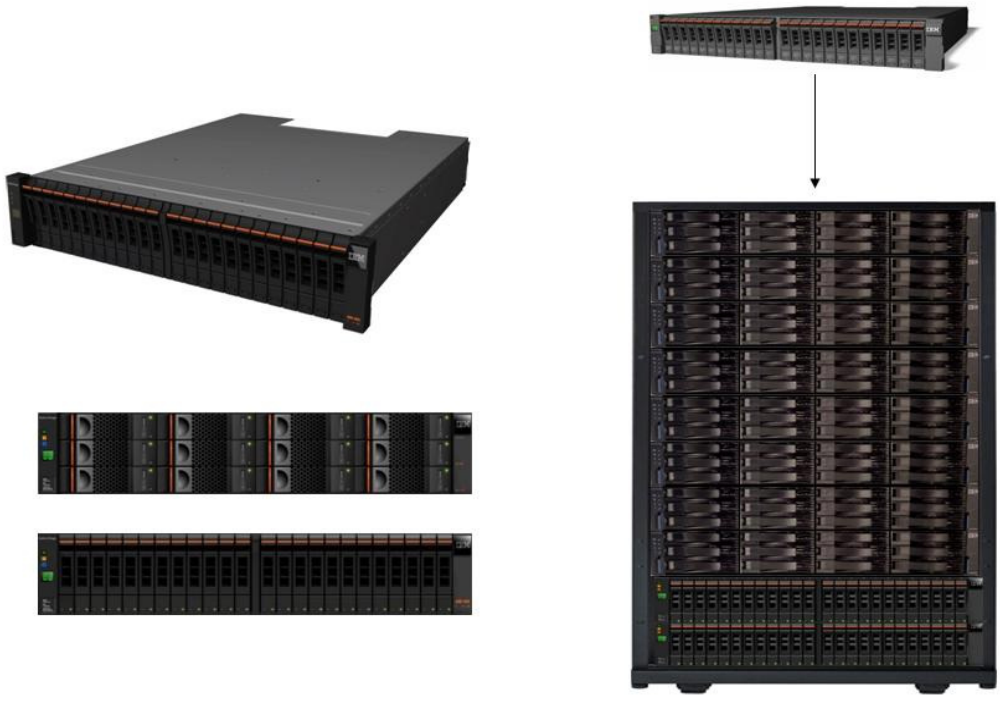

Figure 7: Storwize V7000 disk system

The IBM Storwize V7000 system is a midrange storage solution designed to deliver enterprise-level performance, scalability, and ease of management in many application environments such as SAS. The following points list some of the beneficial characteristics and features of the Storwize V7000, and the potentially beneficial additional products to use with the Storwize V7000.

- Modular, midrange disk storage that grows as your needs grow
  - Scales to 960 HDDs (forty 24 bay enclosures)
  - Based upon proven IBM System Storage SAN Volume Controller software technology.
- Support for 10 Gbps iSCSI server attach
  - Up to seven times improvement in iSCSI throughput compared to 1 Gbps iSCSI
  - Upgradable from existing Storwize V7000 systems
- Enterprise class capabilities in a midrange offering
  - Built in virtualization, thin provisioning, FlashCopy, and Easy Tier
  - Sophisticated local and remote mirroring

- High performance in a midrange disk system
  - Uses clustered nodes, controller cache, back-end I/O coalescing, and balanced disk I/O.
  - Supports up to almost 1M IOPS
  - Supports up to 82,000 IOPS for *database-like* 70/30 read / write workloads
  - Integrated Tivoli Storage Productivity Center (TPC) for Disk Midrange Edition Performance Optimization technology helps reduce service times of resource constrained applications by an average of 48% up to a maximum of 90%, depending upon the application requirements and other factors.
- Easy to set up and manage
  - Innovative, intuitive GUI eliminates complexity
    - Simplified tasks such as storage configuration, storage provisioning, storage tiering, and capacity upgrades help increase administrator productivity.
    - Preinstalled software, simplified provisioning, online data migration, simplified remote mirroring, Easy Tier, and event monitoring make it easy to add functions as needed.
    - Virtualization can help increase storage administrator productivity by up to twice as much as with using non-virtualized storage.
  - Non-disruptive data migration differentiates a product in this class.
  - TPC for Replication improves the productivity of administrators by automating storage replication.
- Storage efficiency
  - The Storwize V7000 system is designed to deliver enterprise-level performance.
  - Automated tiering can improve performance by up to 300% by moving as little as 10% of data to solid-state storage without requiring any administrator intervention.
    - **Note:** These figures are dependent upon many factors such as the I/O block size, and the type and mixture of each workload such as random or sequential I/O. Refer to the SAS notes in the “Easy Tier SSD usage for SAS workloads” section of this paper.
  - Storage virtualization makes it easy to resolve performance bottlenecks.
  - Storage virtualization minimizes downtime due to storage migrations.
  - Virtualization can increase disk utilization and reduce storage growth.
  - Save space using space-efficient FlashCopy.
  - Storage resource management with Tivoli Productivity Center can increase disk utilization.
  - Thin provisioning can dramatically reduce disk storage needs. Refer to the SAS notes in the “Thin provisioning and Real-time Compression” section of this paper.
- Investment protection
  - Virtualize external storage arrays and extend asset life.
- Cost effective – dramatic savings of money and space compared to competitive offerings providing total cost of ownership (TCO) savings.

Storwize family disk storage performance tuning with SAS I/O

From a high level, the IBM AIX® OS I/O stack contains several layers that an I/O must traverse. At each layer, AIX keeps track of the I/O. Some of the layers offer specific queues that can be tuned to help optimize performance. The I/O stack layers are:

- Disk
- Disk subsystem
- Interconnect to the disk (SAN)
- Adapter device driver
- hdisk device driver
- Subsystem Device Driver (SDD) or Subsystem Device Driver Path Control Module (SDDPCM), if used
- Logical Volume Manager (LVM), which is optional
- File system (optional)
- Application

SAS I/O characteristics

SAS includes solutions and applications for data management, quality, and analysis. These solutions and applications support SAS reporting and analysis, analytics and advanced statistical modeling, data mining, online analytical processing (OLAP), business intelligence, and large data computing. Using the SAS data set structure, the SAS Scalable Performance Data Server data structure and native access engines for most third-party databases on almost every operating system (OS), SAS covers the gamut of data stores. Offering vertical solutions in practically every industry, such as Financial and Insurance, Retail, Manufacturing, Customer Management, Business Performance, and many more, SAS has a broad business reach. There can be hundreds of individual SAS processes running in a large environment simultaneously.

The dominant access pattern for most SAS processes is large-block sequential I/O. This is true of data management and exploitation. The type of reporting, analysis, and statistical modeling performed by SAS customers typically involves large data stores accessed sequentially. Given the data sizes of the stores (row lengths and widths), heavy sequential I/O is performed. Operations are favored by the underlying I/O subsystems that are optimized for large-block retrieval of large data stores. This is often diametric to many business applications that rely on random access of smaller data units from database structures that are optimized for that.

Although the dominant SAS I/O patterns are large-block sequential, there are SAS operations that invoke small-block random processing. There are at least three typical scenarios in which a SAS process might exhibit a random I/O characteristic. The first scenario is traversing heavily indexed files for the random retrieval of records against B-tree or hybrid bitmap indexes. The OLAP-structure is traversing with the SAS OLAP Server in the second scenario. In the third scenario, there can be data set traversal, retrieval, update techniques that rely on POINT= type processing, or random access of heavily indexed datasets. Segregation of large-block sequential I/O, and random I/O applications data should be considered. Refer to the “Appendix A: Additional SAS I/O and storage reference material” section for white papers offering details about SAS I/O characteristics and workloads, and general advice on architecting storage for SAS.

Disk and disk subsystem

This section provides some general best practices for MDisk, storage pool, I/O groups and clusters, and volume disk implementations.

MDisk / LUN (RAID array volumes) best practices

- In general, use the same RAID formats across MDisks.
  - This depends on whether the drives are SAS HDDs or SSDs and the number of each drive type available.
- When adding MDisks, consider creating new storage pools rather than adding to existing pools.
- Create each logical unit number (LUN)/MDisk to use the entire capacity of a RAID group or array.
- MDisks/LUNs should have the same stripe width as the SAS bufsize. This improves physical I/O transfer efficiency to the underlying storage array by aligning the SAS application I/O size to match a full stripe I/O to storage. SAS environments may benefit from increasing the bufsize. Any permanent changes should be based upon first testing which higher multiple of bufsize is beneficial. SVC / Storwize V7000 default is 256 KB for MDisks.
  - Recommendation is to align and set the MDisk/LUN stripe size from 64 KB to 256 KB to align with your SAS application data size requirements.

Storage pool (managed disk groups) best practices

- Use MDisks from one external storage system per storage pool.
- Each storage system should provide MDisks to a single SVC cluster.
- Each externally virtualized RAID array group must be included in only one storage pool.
- For non-XIV, limit storage pool to no more than approximately 10 MDisks.
- Implement striped volumes for all workloads.
- Select an extent size that balances cluster capacity and volume granularity (such as 256 MB extents). Keep extents as a multiple of 64 KB. For SAS workloads, set the extent size to 256 MB unless customer testing shows a different multiple to be beneficial.
- Use at least eight MDisks to use all eight ports in an I/O group (round robin placement of assigning preferred paths to LUNs).

I/O groups and clusters best practices

- Volumes (virtual disks): Maximum 8192 volumes in total (2048 per I/O group) and up to 256 TB in size
- Each volume is assigned to:
  - Specific node-pair (I/O group)
  - Specific storage pool
- Cluster:
  - Maximum of four node-pairs (eight nodes in total) or I/O groups
  - Large environments may have multiple clusters
- Managed disks (MDisks):
  - LUNs (MDisks) from up to 64 physical disk subsystems
  - Maximum of 128 storage pools (MDG)
  - Maximum of 128 MDisks per pool
  - Maximum 4096 MDisks per cluster

- Add or remove from pool

Volume disk best practices

- Note that new volumes go to the least utilized storage pool within a cluster.
- Configure the volume size that is appropriate for hosts. Fewer and larger volumes work better from a performance and management perspective. Volumes should not consume more than 10% of the storage pool capacity and not more than 20% of performance bandwidth of storage pool.
- Use striped volumes in most cases, even if application or host has its own striping as long as stripe sizes are dissimilar. Fine-grained host LVM striping can be beneficial for performance.
- For thin-provisioned volumes, set the cache mode to read/write to cache metadata.
- Set the grain size for non-FlashCopy thin volumes to 256 KB. FlashCopy volumes set to 64Kb (old default was 32Kb)
- Set the value of thin disk to *autoexpand*.
- Use I/O governing.
- If using Volume Mirroring, consider balancing the primary setting of all the volumes across A/B storage pools for read I/O balance.

General best practices

- The SAN Volume Controller (not Storwize V7000) allows for one to four internal SSDs per node, and a maximum of 8 nodes for a maximum of 32 SSDs per cluster.
- The Storwize V7000 (and Flex System V7000) supports up to 960 internal drives.
  - Plan for a maximum of 24 to 32 SSDs (depending on workload) per control enclosure. Testing shows less can be as effective with the use of Easy Tier.
  - Ensure drives are split across SAS disk drive chains for optimal SSD performance. With later versions of SVC this is not an issue.
- As of SVC v6.4 storage pool extent sizes can be from 16 MB to 8 GB.
  - All MDisks in a storage pool will have the same extent size.
  - Same extent size can be used for all storage pools in a cluster.
  - The default extent size is 256 MB.
  - Large SVC implementations might use 512 MB or larger extent size to maximize the capacity to be virtualized.
- SVC and Storwize V7000 can have Easy Tier active on selected volumes in a multi-tiered storage pool.
  - Provides the ability to *pin* a volume onto SSDs or exclude volume(s) from SSDs.
- Easy Tier can manage striped mode volumes, but not sequential mode or image mode volumes.
- You must have free extents for Easy Tier to function.
- Plan on having approximately 10 free extents per MDisk in the pool.
- At SVC V6.1, variable block sizes of up to 256 KB are allowed compared to the 32 KB pre SVC V6.1.
  - Handled automatically by the SVC system without requiring user control.
  - Might increase I/O throughput to the back-end storage and put stress on the front-end adapters of the managed storage systems.

SAN interconnect to the disk

The following best practices sections demonstrate the common use of SVC code base by the Storwize family of products. The term SVC is used interchangeably with the Storwize V7000 system to indicate the Storwize family of products including Storwize V7000, Flex System V7000 Storage Node, SVC, and other family members. These best practices sections are abbreviated for quick reference. For additional best practice details, refer to the SVC and Storwize V7000 implementation Redbooks reference links in the "Resources" section of this paper.

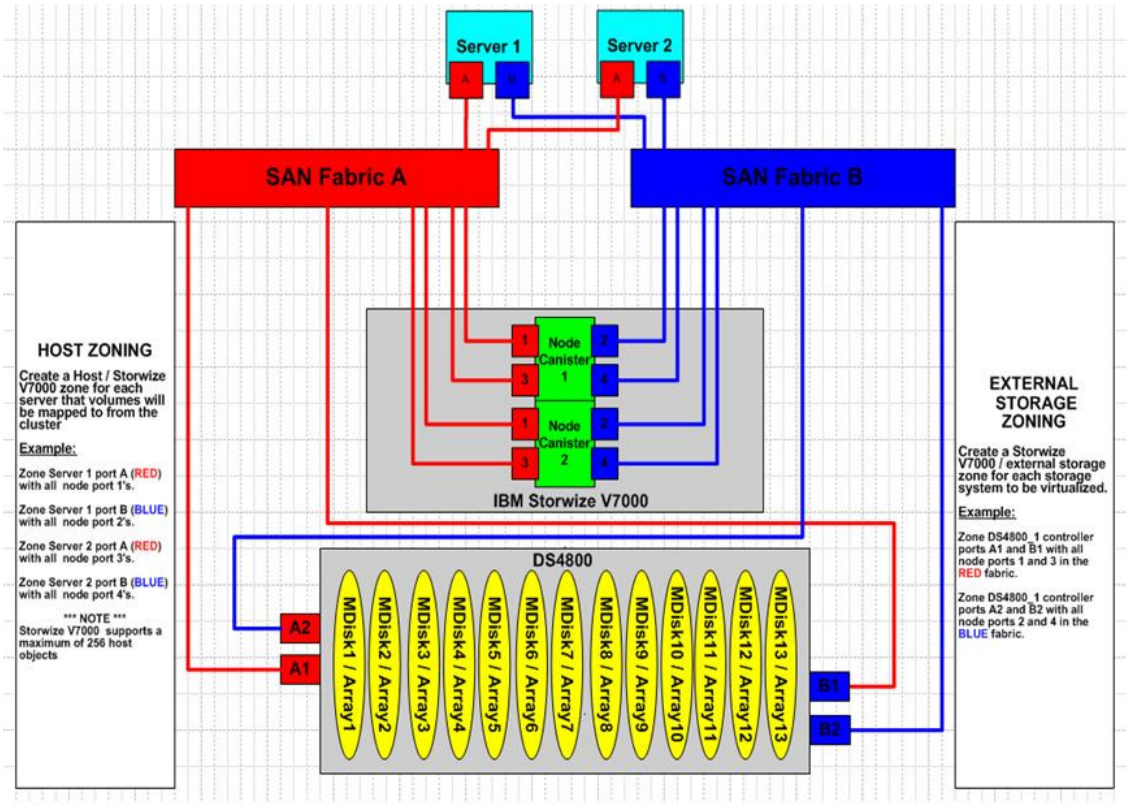

Figure 8: SAN configuration example

SAN fabric best practices

Note on ports: IBM Storwize V7000 supports a maximum of 16 ports or WWPNs from a given external storage system that will be virtualized.

- Zone configuration considerations:
  - Create a SVC cluster zone for intra-node communication.
    SVC to storage zones – all SVC nodes should be zoned to all storage ports.
    SVC to host zones – each host should only be zoned to one specific I/O group.

- Have physical separation of dual fabrics. Refer to the fabric and zoning example in Figure 8.
  - In this example a Storwize V7000 system virtualizes external storage.
  - In each fabric, a zone is created for the four IBM Storwize V7000 worldwide port names (WWPNs), two from each node canister, along with up to a maximum of eight WWPNs from each external storage system.
- Avoid inter-switch link (ISL) congestion (70% utilization).
- Ensure that all SVC node ports are connected to the same SAN switches as storage devices they manage. For example, storage traffic or intra-node traffic should never traverse across ISLs.
- Place high bandwidth servers on the same switch as SVC nodes and storage.
- Never allow more than three hops between the host and SVC cluster.
- Use the same port speed for the SVC nodes.
- Ensure that a single host is not zoned to both SVC and the SVC-managed storage controller.

Server/Host side connectivity best practices

- Use the multipathing software to configure two to four paths to an SVC I/O group.
  - In SVC (pre R6.3), ensure that the multipath I/O (MPIO) software is aware of the preferred path.
    - SVC uses simple multipathing from each node in the cluster to the target storage allowing the SVC to communicate with many disk controllers.
  - In SVC R6.3, I/O is submitted using one path per target port per managed disk per node.
    - I/O to a managed disk progress in a *round robin* fashion.
    - I/O is spread across multiple storage system ports.
    - Paths are chosen according to port groups presented by the storage system.
- Put no more than three ISL switch jumps between host and SVC cluster.
- Limit host connects to four paths optimal, eight acceptable.
- Do not share host bus adapter (HBA) between disk and tape.
- Use a single I/O group per host.
  - SVC uses a preferred node scheme to load balance by I/O group so the volumes are assigned automatically to a SVC node or controller within that I/O group.
- Ensure that when booting from SAN, the LUN ID is the lowest Small Computer System Interface (SCSI) ID (zero in most cases).
- Configure Fibre Channel adapter Q depth appropriately for the SAS workload to not over-run SVC node.
- Note that SAS is not software cluster aware, but can use clustered file systems, such as IBM General Parallel File System (IBM GPFS™) across multiple logical partitions (LPARs) and SAN connectivity. Considerations for clustered file systems are the same as for non-shared file systems.

XIV connectivity to Storwize family

- Zone two ports (one per fabric) from each interface module with the SVC Ports
- If not using XIV remote mirroring, then change role of port 4 from *initiator* to *target* on all interface modules. Use port 1 and port 3 from every interface module for SVC attachment. Only single-rack XIV configurations are supported.

Storwize V7000 considerations

- The Storwize V7000 can present storage to be virtualized by the SVC. SVC V6.3 introduced the cluster property called *layer* for use with remote copy cluster replication. With SVC V6.4, a Storwize V7000 system can also present storage to be virtualized by another Storwize V7000 system using the cluster property *layer*. A Storwize V7000 system can be virtualized with layer=storage behind another Storwize V7000 with layer=replication.
- A zone should include at least 1 port per storage cluster.

PowerVM, VIOS, and NPIV

IBM Power Systems with IBM PowerVM virtualization offerings enable businesses to consolidate servers and applications, virtualize system resources, and provide a more flexible, dynamic IT infrastructure.

IBM POWER Hypervisor™ enables the hardware to be divided into multiple LPARs, and ensures isolation between them. The hypervisor orchestrates and manages system virtualization, including the creation of LPARs and dynamically moving resources across multiple operating environments. The POWER Hypervisor also enforces partition security and can provide inter-partition communication that enables the Virtual I/O Server (VIOS) virtual SCSI, virtual Fibre Channel, and virtual Ethernet functions.

PowerVM Editions deliver advanced virtualization functions for AIX, IBM i, and Linux® clients such as IBM Micro-Partitioning® technology, Virtual I/O Server, Integrated Virtualization Manager (IVM), and Live Partition Mobility (LPM). PowerVM features provide the ability to virtualize processor, memory, and I/O resources to increase asset utilization and reduce infrastructure costs. PowerVM also allows server resources to be dynamically adjusted to meet changing workload demands, without a server shutdown.

PowerVM Editions provide Virtual I/O Server technology to facilitate consolidation of local area network (LAN) and disk I/O resources and minimizes the number of required physical adapters in a Power System server. The VIOS actually owns the resources that are shared with clients. A physical adapter assigned to the VIOS partition can be shared by one or more other partitions. The VIOS can use both virtualized storage and network adapters, making use of the virtual SCSI, virtual Fibre Channel with N-Port ID Virtualization (NPIV), and virtual Ethernet facilities.

You can achieve continuous availability of virtual I/O by deploying multiple VIOS partitions (dual VIOS) on an Hardware Management Console (HMC)-managed system to provide highly available virtual services to client partitions.

Virtual SCSI

The Virtual I/O Server allows virtualization of physical storage resources, accessed as standard SCSI-compliant LUNs by the client partition, through virtual SCSI adapters. Virtual SCSI allows client LPARs to share disk storage and tape or optical devices that are assigned to the VIOS partition. The functionality for virtual SCSI is provided by the POWER Hypervisor. Virtual SCSI allows secure communications between partitions and a VIOS that provides storage backing devices. The VIOS is capable of exporting a pool of heterogeneous physical storage as a homogeneous pool of block storage in the form of SCSI disks.

Virtual Fibre Channel is most likely the best choice for accessing application volumes, but virtual SCSI is often used to configure the root volume group for AIX client partitions by configuring SAN disk that can be accessed from a VIOS partition, and use native MPIO installed on the VIOS.

For details for configuring virtual SCSI in a dual-VIOS environment, refer to chapter 4 of *IBM PowerVM Virtualization Introduction and Configuration* in IBM Redbooks at:
ibm.com/redbooks/redbooks/pdfs/sg247940.pdf.

Virtual Fibre Channel

Because of ease-of-use in SAN storage management, most administrators will want to consider setting up virtual FC adapters using NPIV technology instead of using virtual SCSI to access the application volumes in production LPARs.

NPIV is a technology that allows multiple LPARs to access independent physical storage through the same physical FC adapter. Each partition is identified by a pair of unique WWPNs, enabling you to connect each host or client partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk. The VIOS partition acts only as a pass-through, managing the data transfer through the IBM POWER Hypervisor. The primary advantages of using NPIV are that storage device configuration is less complicated, and the client LPAR (host node) recognizes the disk volumes as if they had dedicated FC adapters directly attached. When using NPIV, well established SAN storage management techniques, such as zone mapping and masking, are valid without the need to additionally map volumes as backing devices in the VIOS partition.

To implement NPIV technology to access Storwize V7000 volumes, verify that the Fibre Channel HBAs, switches, and directors are NPIV-capable, and that the versions of AIX, VIOS, and HMC support it. The virtual Fibre Channel adapters for both the VIOS partition and the client partition will need to be configured in the partition profile using the HMC. Finally, a virtual Fibre channel adapter must be created on a client LPAR that was assigned to the single physical Fibre Channel adapter port in the VIOS partition.

For details on a recent VIOS and NPIV highly available environment configuration and test exercise, refer to *Disaster recovery using IBM Storwize family storage with IBM PowerHA SystemMirror Enterprise Edition 7.1* in IBM Techdocs at the following URL:
ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102245

Subsystem, hdisk, and adapter device drivers

In this section, the focus is on tuning the middle layers consisting of SDD (or SDDPCM), hdisk, and adapter device drivers. The goal is to improve simultaneous I/O capability and realize efficient queue handling. Refer to for some of the parameters that can affect disk and adapter performance. In general, SAS applications can benefit from careful consideration and tuning of these parameters.

Both the disk and adapter have maximum transfer parameters that can be adjusted to handle larger I/O, reduce I/O splitting, and coalesce I/O as it moves up and down the stack. In addition, both have I/O queues that can be adjusted to accept additional I/Os.

AIX SDD and Data Path Optimizer

Generally SDDPCM is used for the IBM Storwize family of products with IBM Power® hosts. There can be reasons to choose SDD or SDD Device Specific Module (SDDDSM for Microsoft® Windows® hosts, not covered in this paper). The following section explains the tuning of both the SDD and how this relates to the SDDPCM tuning parameters.

Before implementing the SAS storage multipathing environment, go to the following website for SDD/SDDPCM technical support and for the most current SDD documentation and support information: **ibm.com**/servers/storage/support/software/sdd/.

If SDDPCM is used with the storage environment, a single multipath disk device is presented to the OS. SDDPCM manages the multipathing. The hdisk device driver tunables for queue management are used. Refer to disk device tunable parameters below.

If SDD is used with the storage environment, then the Data Path Optimizer (DPO) device I/O queue should be evaluated. SDD provides a virtual path to the storage subsystem LUN / logical disk and provides several hdisk devices through the physical paths (such as FC adapters). So, with SDD, you can issue queue_depth times the number of paths to LUN.

However, when the dpo device queue is enabled (default is *yes*), any excess I/O that can not be serviced in the disk queues go into the single wait queue of the dpo device. The benefit of this is that the dpo device is designed to provide fault tolerant error handling. This might be preferred for high availability applications but for other applications there are advantages of disabling the dpo device queue and using multiple hdisk wait queues for each SDD vpath device. Note that this is not an exhaustive discussion and does not detail any possible AIX limitations for total number of I/O. Also, the queue parameters should be carefully evaluated before implementing any change. For tuning guides specific to a particular IBM storage system, such as the IBM System Storage DS8000® or IBM Storwize V7000, refer to the additional information section. Next, you can look at the disk and adapter I/O tuning parameters, and the SDD- and SDDPCM-related parameters.

Disk and adapter I/O tuning parameters

Parameter	Description
max_xfer_size	FC adapter maximum I/O that will be issued.
max_transfer	Disk maximum I/O that will be issued.
queue_depth	Disk maximum number of simultaneous I/Os. The default is 20 but can be set as high as 256 for Storwize V7000.
num_cmd_elems	FC adapter maximum number of simultaneous I/O. The default is 200 per adapter but can be set up to 2048.
qdepth_enable	SDD data path optimizer device queuing parameter. The default is *yes*. A setting of *no* disables SDD queuing. Use this with Storwize V7000 (that uses SDDPCM and not SDD) and DS8000 storage.
lg_term_dma	Long-term Direct Memory Access (DMA) is the memory area that the FC adapter uses to store I/O commands and data.

Parameter	Description
LTG	AIX volume group Logical Track Group parameter. LTG specifies the largest I/O that the LVM can issue to the device driver. In AIX 5.3, the LTG dynamically matches the disk maximum transfer parameter.

Table 3: Disk and adapter I/O tuning parameters

Note: It is important to understand the I/O characteristics of the application in order to properly tune within the I/O stack layers. If the SAS application is of predominantly large I/O, then the application performance can benefit from adjusting the maximum transfer sizes, long-term DMA, and the LTG. The recommended starting values for a large I/O and highly sequential workload are *lg_term_dma=0x800000* and *max_xfer_size=0x200000*.

Queue information can be monitored in AIX 5.3 and later with the *iostat –D* command. For AIX 5.1 and AIX 5.2, SAR can be used. It is recommended that *qdepth_enable=no* to use the hdisk wait queue rather than the dpo device wait queue.

It is recommended to increase the *num_cmd_elems* value for the FC adapter from the default (initially start at 400). Some of these parameters require a system reboot to take effect. For additional guidelines, refer to the tuning guide links found in the additional information section of this paper.

Use the following commands to display and modify disk and adapter parameters and settings.

Disk/SDDPCM disk driver – max_transfer, queue_depth

- 'lquerypv –M hdisk#' displays maximum I/O size a disk supports.
- 'lsattr –El hdisk#' displays current disk values.
- 'lsattr –R -l hdisk# -a max_transfer hdisk" displays allowable values.
- 'chdev –l hdisk# -a max_transfer=value –P' modifies current disk values

Note: The device should be in an offline/disabled state before changing any parameters. Then cfgmgr will need to be issued.

FC adapter – max_xfer_size, lg_term_dma, num_cmd_elems

- 'lsattr –El fcs#' displays current value.
- 'chdev –l fcs# -a max_xfer_size=value -P' modifies current value.

Note: The device should be in an offline / disabled state before changing any parameters. Then cfgmgr will need to be issued.

SDD only, not SDDPCM – qdepth_enable

- 'lsattr –El dpo' displays current value.
- Use the *datapath* command to change if at SDD 1.6 or greater. Otherwise, the *chdev* command can be used. Example: 'datapath set qdepth disable'

AIX MPIO and SDDPCM

The AIX MPIO disk driver that is used for multipath support for the Storwize V7000 is the IBM SDDPCM. SDDPCM provides multipath and load-balancing management for the redundant FC paths defined in the environment.

Also note that when using VIOS and NPIV with the SAS environment SDDPCM must be loaded on the client LPAR. If you decide to use virtual SCSI to access the SAS application volumes, multipath software should be loaded on the VIOS partitions.

You can download the latest SDDPCM software for the Storwize V7000 system with AIX at the following website:

ibm.com/support/docview.wss?rs=540&context=ST52G7&uid=ssg1S4000201&loc=en_US&cs=utf-8&lang=en+en#SVC.

You will also be required to download and install the specific host device or *Host Attachment for SDDPCM on AIX* from the following website: **ibm.com**/support/docview.wss?uid=ssg1S4000203.

You can find installation directions for both packages in the SDDPCM user's guide: **ibm.com**/support/docview.wss?rs=540&context=ST52G7&q=ssg1*&uid=ssg1S7000303&loc=en_US&cs.

Go to the following Web site for SDD/SDDPCM technical support and for the most current SDD documentation and support information: **ibm.com**/servers/storage/support/software/sdd/.

Go to the following SAN Volume Controller web site for latest SDD/SDDPCM support version for SAN Volume Controller: **ibm.com**/systems/storage/software/virtualization/svc/interop.html.

Go to the following System Storage Interoperation center (SSIC) web site for latest SDD/SDDPCM support version for all other supported IBM storage systems: **ibm.com**/systems/support/storage/config/ssic/displayesssearchwithoutjs.wss.

After the multipath software install is complete, check the multipath status through **smit**, and using the **lspath** command.

aixhost1> smitty mpio --> *MPIO Device Management --> Change/Show MPIO Device Characteristics --> select device name.*

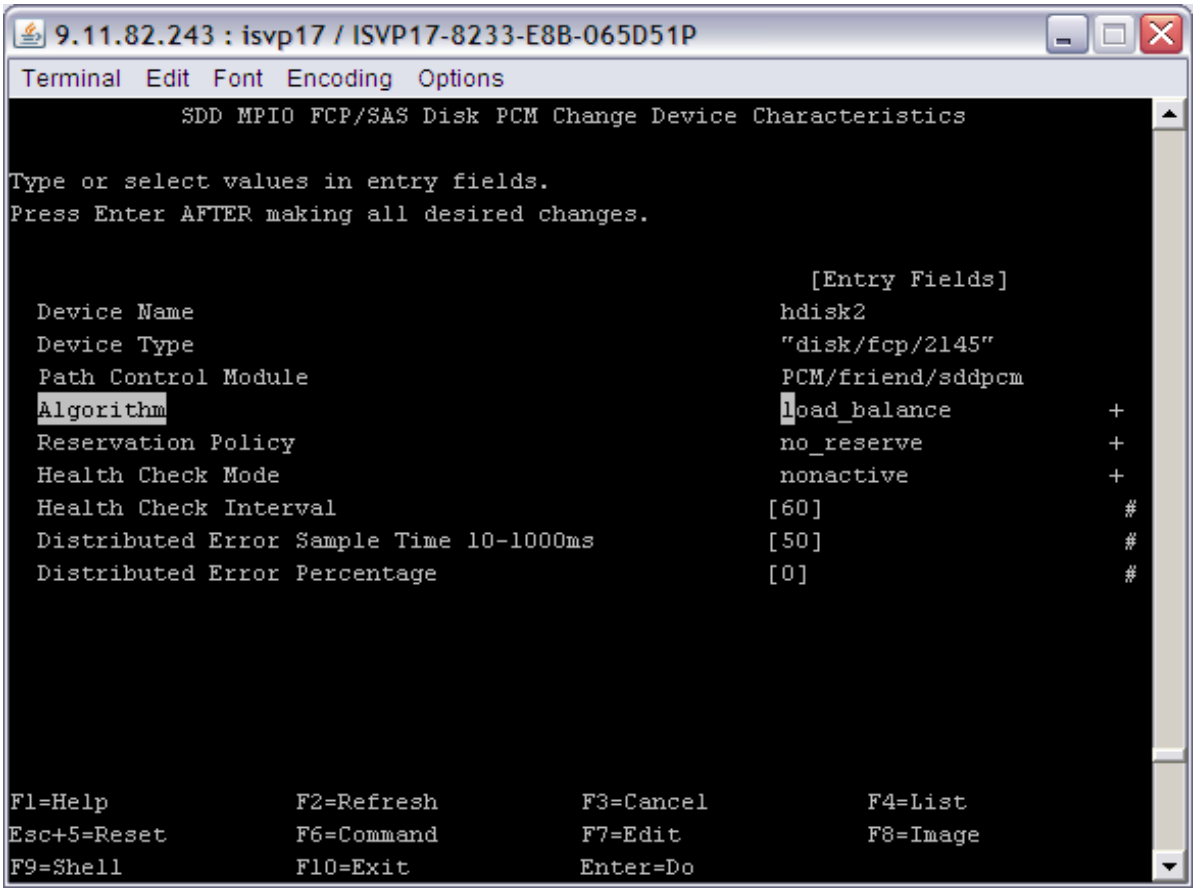

Figure 9: Sample smit screen showing MPIO characteristics

aixhost1> lspath -l hdisk2
Enabled hdisk2 fscsi0
Enabled hdisk2 fscsi0
Enabled hdisk2 fscsi1
Enabled hdisk2 fscsi1

AIX, LVM, and file system

During the deployment of SAS, specific operating system tunable parameters must be considered for optimal storage performance in addition to general SAS tuning parameters. In the following sections AIX storage tunable parameters specific to SAS 9 are examined. These AIX tunable parameters are also available in the *SAS AIX 5L, AIX 6 and AIX 7 Tuning Guide* found at: **ibm.com**/support/techdocs/atsmastr.nsf/WebIndex/WP101529.

There are many non-storage related SAS tunable parameters (not mentioned in this paper) that should be examined for optimal SAS application performance. You can find more information about these AIX non-storage and storage-related tunable parameters in the *SAS AIX 5L, AIX 6, and AIX 7 Tuning Guide*. The following discussion includes storage configuration guidelines directly from the previously mentioned AIX tuning guide white paper.

AIX general storage guidelines

- Use enhanced journaled file system (JFS2) on a 64-bit AIX kernel.
  - With the introduction of the IBM AIX 5L™ operating system, IBM introduced a new file system referred to as JFS2 that provides greater scalability than journaled file system (JFS). JFS2 is designed and optimized for a 64-bit kernel environment taking full advantage of the 64-bit functionality. JFS2 is the default file system for a 64-bit kernel.
  - In general, SAS requires large rates of sequential disk I/O. The AIX file system named JFS2 can detect and use the read-ahead and write-behind characteristics of the application under normal file-caching policy.
  - You can choose **3. Enable 64-bit kernel** and **4. Create JFS2 File Systems** on **Install Option** screen during AIX installation, for example, from CD.

- It is strongly recommended that SAS WORK or UTILLOC point to a more robust file system than /var.
  - AIX instructions use /var for various purposes, such as storing temporary files, mail spool files, and all security logging information. Running out of /var space might cause SAS processes to terminate abnormally.

- Configure paging space at least with the following suggestions:
  - Place paging spaces on dedicated disk(s), like Storwize family storage volumes, to eliminate I/O contention.
  - Use multiple paging spaces spread over multiple disks.
  - Make the primary paging space *hd6* a little bigger than the secondary paging spaces.
  - Insure that the paging space is sufficient to support the number of concurrent SAS process as the number of SAS process can be dynamic depending upon the application workload.
  - Set *nokilluid=10* with *vmo.*

- Determine the application's I/O access patterns, which is important for I/O layout and tuning.
  - To achieve the best I/O performance, the access patterns and storage configuration should be compatible. If the application's I/O patterns are not known, then additional data can be gathered to determine dominant patterns. For example, in the test experiments, AIX trace indicates that the SAS Revenue Optimization application drives traditional large sequential I/O characteristics but it also contains a fair amount of random I/O. Thus, optimization for different I/O access patterns (dominant and non-dominant) is recommended.

- Ensure tuning from a system-wide perspective [for example, virtual memory manager (VMM), LVM, Fibre Channel (FC), disk storage] for the SAS workload.

- Use the appropriate number of HBAs from the storage to the host server to provide the required front-end application bandwidth.
  - Many SAS I/O workload patterns can be throughput-intensive. However, this is not always the case for all SAS applications or necessarily true while running the entire SAS application.

- High performance storage channels, such as Fibre Channel technology, must be considered over slower mediums.
- Use dynamic multipathing, if possible, to spread the I/O load over multiple adapters. Care needs to be exercised when locating SAS data libraries on mount points.

- Spread the I/O workload across many physical disk spindles rather than fewer, larger capacity disks. Storwize family already provides many of the disk management and performance benefits listed below.
  - Provide better I/O performance by sizing for quantity of disks instead of capacity of disks. **Note1**: The Storwize family provides a balance of disk performance versus capacity of storage.
  - Implement storage system RAID striping across multiple physical disks. See note1 above.
  - In general testing, it has been observed that there is a slight performance advantage to using RAID10 over RAID5 for SAS temp space file systems. This is not necessarily the case for other SAS file systems. Use RAID10 or RAID5 depending on the level of redundancy and total capacity compared to usable capacity that is required for each type of file system.
  - Use LVM striping instead of concatenation. **Note2**: This recommendation is for older non-Storwize family disk systems. The Storwize family disk volumes provide robust performance and storage contention reduction benefits through the use of internal RAID and storage pool extent striping, and storage virtualization without requiring LVM striping or separate physical disks for SAS temp and SAS data file systems.

- Minimize disk contention between SAS temp space and data spaces.
  - Avoid disk contention by placing SAS temp space file systems and SAS data file systems on physically separate disks. This recommendation is for older disk systems and is not specific to the Storwize family. This is not applicable to the Storwize family. See note1 and note2 above.
  - Use multiple storage server controllers to further separate and isolate the I/O traffic between SAS temp and data spaces. This also provides a more robust disk back end to handle I/O.
  - Use multiple mount points for SAS file systems. Place system O/S, SAS, user, SAS temp, and SAS data file systems on separate physical disk. This is not applicable to the Storwize family. See note2 above.
  - Separate the single SAS temp space file system (SAS WORK) into separate SAS temp file systems with physically separate disk **if** multiple users otherwise would have to share the SAS temp space (SAS WORK) **and** sharing the disk or file system increases disk contention beyond acceptable response levels. The physically separate disk requirement is not necessary for the Storwize family. See note2 above.
  - Create separate JFS2 log files on separate physical disk for each SAS file system. This does not apply to the Storwize family. See note2 above.

- Isolate SAS I/O from non-SAS workloads.

- In general SAS applications can be highly sequential large I/O workloads. Disk contention between SAS applications and other non-SAS small I/O random IOPS applications will increase service times of all applications and decrease I/O performance.

- Use the AIX scalable volume group or big volume group with the *mklv –T 0* option to avoid the logical volume control block reserve of the first 4 K of space.
  - With the logical volume control block (LVCB) present, the first data block will start with a 4 K offset.
  - When logical volume control blocks exist on a logical volume, they can cause I/O to span multiple physical volumes due to this offset.

- Be mindful that AIX file systems are aligned on a 16 K boundary when choosing the disk stripe or segment size or array stripe size.
  - A strip is the size of data to be written to each physical disk in the array. A stripe is the size of the full write across all the physical disks in the array.
    Example: strip size x number of disks = stripe size.
  - Note that the AIX LVM stripe size that can be selected from the *smit lv* create panel is actually the single strip size (not stripe) or size of data to be written to each of the array disks and not the full stripe size across all the physical disks.

- Synchronize SAS BUFSIZE with the storage system stripe size and the AIX LVM stripe size (if using LVM striping), and VMM read-ahead increments.
  - Synchronizing of I/O sizes results in more efficient I/O, while reducing the total number of I/O requests to the storage subsystem.
  - Note: LVM striping may or may not provide better performance depending on the SAS application or the storage subsystem configuration. Testing your specific application is recommended.

AIX LVM tuning

The LVM provides an abstract logical view of the underlying physical disk devices. Logical volumes are employed to contain paging spaces and dump areas, but most often they underlie file systems. LVM uses a construct called **pbuf** to control a pending disk I/O. A single pbuf is used for each I/O request. The application generating large amount of I/O requests or striping and mirroring environment usually requires more pbufs to satisfy the system requirements. Running out of pbufs can degrade the performance as the I/O initiating process is suspended until pbufs are available again.

The parameter *pv_pbuf_count*, used to control the number of pbufs available to the LVM device driver, can be set for each logical volume using the **lvmo** command.

The AIX file system (FS) is called journaled file system or enhanced journaled file system. FS presents a logical view of files and directories linked together to form a hierarchical tree structure.

AIX JFS and JFS2 tuning

In general, SAS applications have a great deal of large sequential read and write disk I/O. If the workload has many large I/O requests to a file system (for example large sequential I/O to JFS2), I/O requests can be bottlenecked at the file system level while waiting for a construct called *bufstructs*. The bufstructs for

JFS2 is dynamic and the number of bufstructs per file system can be increased. The file system must be remounted for the new value to take effect.

The I/O characteristics of SAS usually creates the situation where VMM read-ahead, and write-behind algorithm can be used to improve the performance of sequential file access. The parameters listed in Table 4 can be tuned using the *ioo* command.

Most frequently used AIX file system tuning parameters

Parameter	Description
j2_dynamicBufferPreallocation	This tunable (16 by default) specifies the number of 16 k chunks to preallocate when the file system is running low of bufstructs.
j2_nBufferPerPagerDevice	This tunable (512 by default) specifies the number of bufstructs that start on the paging device. JFS2 will allocate more bufstructs dynamically. It may be appropriate to change this value if j2_dynamicBufferPreallocation tuning has already been attempted and the number of external pager file system I/O requests blocked due to no fsbuf increases rapidly.
j2_maxPageReadAhead	This tunable (128 by default) specifies the upper limit for AIX JFS2 prefetching. It affects efficiently when doing large I/O.
j2_nPagesPerWriteBehindCluster	This tunable controls the gathering I/O requests for sequential write behind. The default is 32.

Table 4: Most frequently used AIX FS tuning parameters

Release-behind mechanism for JFS and enhanced JFS

Release-behind is another suggested tuning mechanism for SAS. This feature allows the file system to release the file pages from file system buffer cache as soon as an application has read or written the file pages. This feature helps the performance when an application performs a great deal of sequential read or write operations and most often, if accessed one time, these file pages will not be accessed again in the near future.

If release-behind is not used, it might cause threads to wait on page replacement to supply sufficient free frames to handle file read or write operations. In the worst case, the page replacement activity might cause paging. When writing a large file without using release-behind, write operations will happen very fast whenever there are available pages on the free list. When the number of pages drops to minfree, VMM uses the least recently used (LRU) algorithm to find candidate pages for eviction.

A trade-off of using the release-behind mechanism is that application can experience an increase in processor utilization for the same read or write throughput rate (as compared to not using release-behind). This is because of the work required to free pages, which is normally handled at a later time by the LRU daemon. Also, note that all file page accesses result in disk I/O as file data is not cached by VMM. However, applications (especially long-running applications) with the release-behind mechanism applied, still performs more optimally and with more stability.

This feature can be configured on a file system basis. When using the *mount* command, enable release-behind by specifying one of the following three flags:

- Release-behind sequential read flag (-rbr)
- Release-behind sequential write flag (-rbw)
- Release-behind sequential read and write flag (-rbrw)

GPFS tuning for distributed SAS workloads

GPFS tuning is specific to the workload type. Here are some GPFS tuning recommendations in a distributed SAS 9 environment. These parameters, as listed in Table 5, are in addition to the *subsystem, hdisk, and adapter driver* storage tuning recommendations section found earlier in this paper.

GPFS parameters

Parameter	Description
File System Blocksize -B *blocksize*	The GPFS file system block size should be set based on the specific SAS solution I/O requirements. **Example:** For a mixed workload of sequential and random I/O in a specific SAS markdown optimization test, setting the GPFS block size to 256 KB rather than 2 MB improved the overall file system performance by 22%. **Note: Real world performance might vary and is dependent on specific test environments.**
pagepool	Set pagepool to 8 GB if there is available memory and the workload includes random file access. More cache will help the random portion of SAS workloads. Currently 8 GB is the maximum size allowed for pagepool (GPFS 3.1). The next version of GPFS might support a larger value of pagepool.
maxFilesToCache	The default value for maxFilesToCache is 1000. Increasing this value to 20,000 or greater in general improves the performance for user-interactive tasks. **Example:** The SAS markdown optimization solution tested uses thousands of files and benefited from increasing this value.
Disk bandwidth	In a distributed environment, the amount of available memory might be smaller on each system. In this case, to further improve the performance of the solution, add additional disk bandwidth to GPFS. This means adding additional physical disks to the GPFS storage pools.

Table 5: GPFS parameters

Refer to the *IBM GPFS tuning guidelines for deploying SAS on IBM Power servers* paper for additional GPFS tuning parameters of a mixed SAS workload published at:
ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102255.

Easy Tier SSD usage for SAS workloads

Easy Tier is a built-in dynamic data relocation feature available on the IBM System Storage DS8700, SVC, Flex System V7000, Storwize V7000, and (optionally) Storwize V3700 that allows host transparent movement of data among the storage subsystem resources. Easy Tier is a no-charge feature that automates the placement of data among different storage tiers. Easy Tier can automatically migrate data at the sub-LUN/sub-volume level to the most appropriate storage tier. This includes the ability to automatically and non-disruptively relocate logical volume extents with high activity to storage media with higher performance characteristics, while extents with low activity are migrated to storage media with lower performance characteristics. This capability achieves the best available storage performance for your workload in your environment.

In general, Easy Tier is not recommended for large, highly sequential workloads. To confirm or disprove this statement, future SAS application testing by the joint SAS/IBM team can look at the possible benefits of Easy Tier on the Storwize V7000 storage system. The Easy Tier hot extent algorithm might prove useful for base SAS workloads, and further testing is planned. However, typical workload characteristics that are usually a best fit for effective SSD usage are random I/O, a high read to write ratio, and smaller block I/O (less than 8 KB). Although the base SAS workload is expected to be predominantly large-block (greater than 512 KB) sequential I/O, with a read-write ratio of around 50 to 50, the test team found it to still benefit from the SSD usage. Write operations do not perform quite well as read operations, but still typically outperform conventional storage. Read operations perform much better. Both randomized and sequential SAS workloads can benefit from Easy Tier, when used on appropriate tier levels (non-SATA disks). I/O performance might improve with Easy Tier on mixed workload systems, and when the array must be shared with other applications.

Thin provisioning and Real-time Compression

The current recommendation is to not use thin provisioning or IBM Real-time Compression™ with Storwize V7000 for very large, base SAS workloads. Future SAS and IBM joint testing might explore possible benefits compared to penalties using Real-time Compression.

Thin provisioning non-SAS general considerations

- In general I/O very large intensive workloads such as SAS usually are not good candidates for thin provisioning. Thin provisioning also adds a slight additional processing workload on a disk subsystem. Also, note that there may be a storage capacity compared against the number of disk spindles trade-off. Thin pools on large drives might not yield sufficient disk spindles to aggregate throughput if the thin provisioning places multiple workloads on too few disks. This scenario might result in multiple workloads concurrently accessing the thin-provisioned volume at the same time causing performance degradation. Smaller workloads, or workloads that are not consistently heavy, might better utilize thin-provisioning benefits, provided over-subscription does not take place, large pools that are striped across many disks are used (for example, striped everything), and administrators monitor that adequate numbers of physical disk spindles are actually available underneath the thin-provisioned volume to balance real capacity with required throughput.
- Use a grain size of 256 KB for thin-provisioned volumes in storage pools whether using Easy Tier or not. In older versions, the grain size defaulted to 32 KB, which triggered Easy Tier algorithm to consider all I/O requests to the thin-provisioned volumes. At a 32 KB grain size, even

large sequential I/O requests from the host will be broken up into 32 K I/O requests, resulting in odd Easy Tier behavior and performance issues. Refer to the following URL for more details: **ibm.com**/support/docview.wss?uid=ssg1S1003982.

SAS deployment examples

This section describes a sample SAS grid deployment on IBM Power servers with SAN Volume Controller.

SAS Grid on Power servers with SAN Volume Controller

The following example gives a brief overview of a SAS grid deployment on IBM Power servers with the IBM System Storage SAN Volume Controller providing virtualized cloud storage with the GPFS file system. Alternatively, other storage products from the Storwize family, such as Storwize V7000 or Flex System V7000 can be used for a combination of internal and external virtualized storage instead of the SVC storage system (refer Figure 10).

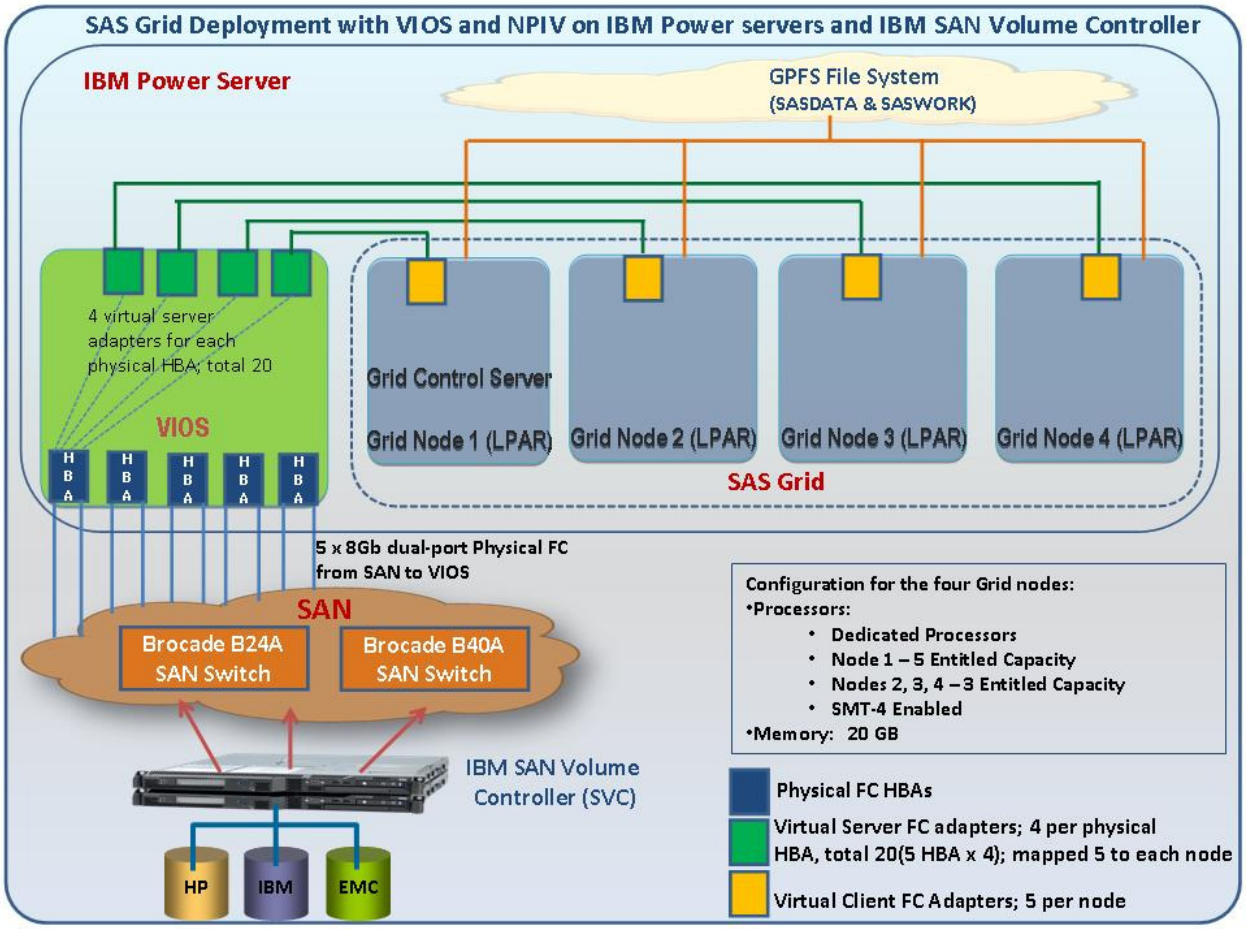

Figure 10: SAS Grid deployment architecture on IBM Power servers with SAN Volume Controller system storage

The SAS Grid deployment shown in Figure 10 illustrates strong virtualization capabilities of the IBM server, storage, and software products. An IBM Power server provides four AIX LPARs. The PowerVM VIOS servers virtualize the five dual-port HBAs, providing four virtual server adapters for each HBA. Each grid node LPAR receives five virtualized Fibre Channel adapters, as well as virtualized Ethernet adapters from the VIOS server. The required cloud storage is provided to the SAS Grid nodes through the use of GPFS. The underlying virtualized IBM and non-IBM SAN storage is provided by the SAN Volume Controller.

Summary

SAS 9 is a robust and rich business analytics solution with specific I/O patterns driven by specific application and environment variables and the associated server and storage hardware. The required I/O throughput and IOPS can be achieved by carefully factoring in the application I/O variables and requirements into each server and storage hardware design, selecting the appropriate high-performing IBM server and storage systems for the performance requirements, and adapting the provided I/O rules of thumb thoughtfully into the specific application environment through OS, application, and hardware tuning.

The IBM Storwize family storage systems with their capabilities and features, combined with the use of general storage planning and storage tuning parameters, provide SAS environments with robust enterprise-grade storage performance.

Appendix A: Additional SAS I/O and storage reference material

SAS and the New Virtual Storage Systems, Paper 487-2013 by Tony Brown and Margaret Crevar, SAS Institute Inc., 2012
http://support.sas.com/resources/papers/proceedings09/487-2013.pdf2

Best Practices for Configuring your IO Subsystem for SAS 9 Applications Updated: August 2011 by Margaret A. Crevar, SAS Institute Inc., and Tony Brown, SAS Institute Inc.
http://support.sas.com/rnd/papers/sgf07/sgf2007-iosubsystem.pdf

Guidelines for Preparing your Computer Systems for SAS, Paper 363-2012 by Margaret Crevar, and Tony Brown, SAS Institute Inc.
http://support.sas.com/resources/papers/proceedings12/363-2012.pdf

How to Maintain Happy SAS 9 Users Paper 310-2009 by Margaret Crevar, SAS Institute Inc.
http://support.sas.com/resources/papers/proceedings09/310-2009.pdf

For additional background information about IBM SAS I/O performance considerations and setup concepts refer to the IBM SAS ICC Information Brief entitled *Deploying SAS Enterprise Business Intelligence in an AIX virtual environment – Guide for installing in a medium-scale POWER6 environment* found at http://www.sas.com/partners/directory/ibm/EBIinAIXvirtual_enviro.pdf

Storage Best Practices: SAS 9 with IBM System Storage and IBM Power Systems – Considerations for optimal storage layout
http://www.sas.com/partners/directory/ibm/SAS_IBM_Storage_Best_Practices0311.pdf

SAS 9 on AIX 5L, AIX 6, and AIX 7 Tuning Guide (I/O specific recommendations and additional SAS tuning tips)
ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101529

Grid Computing in SAS 9.3
http://support.sas.com/documentation/cdl/en/gridref/64808/PDF/default/gridref.pdf

SAS Grid Computing
http://www.sas.com/resources/factsheet/sas-grid-computing-factsheet.pdf

Additional SAS information can be obtained at http://www.sas.com.

Appendix B: Performance monitoring tools and techniques

Monitoring scope

Monitoring scope includes processor and memory utilization, disk I/O, network I/O, chip and memory subsystem, application, logical partition/logical processor and system environment/configuration. It monitors each of those areas from overall to detailed perspectives.

When to monitor

- When the run reflects a representative time slice of application workload
- When facing a performance bottleneck
- Anytime you prefer to see the details of insight

Suggested performance tools for monitoring

Tool	Description
Vmstat	A tool that can monitor overall system performance in the areas such as processor, virtual memory manager (VMM) activity, and I/O.
Tprof	A global and micro-profiling tool and it is used to check the *hot spots*.
Curt	A tool that can produce detailed processor utilization for process/thread/pthread activity.
Trace	The trace can be post-processed to check events, such as krlock contention, workload access pattern, inode contention, and so on.
Svmon	The tool can monitor the detailed memory consumption on real and virtual memory.
Ps	A tool to monitor process/thread status and memory consumption as well.
Iostat	A tool to monitor overall I/O stats including disks loads or adapters, and system throughput.
Sar	A tool to report the per-processor, disk, run queue statistics.
Filemon	A magnifying glass tool, and it is used for detailed file I/O activity (for example, hot lv, pv).
Netstat	A tool to report network and adapter statistics.
Netpmon	A tool to report detailed statistics on network I/O and network-related processor usage, data rates, and response time.
Hpmcount	A tool that programs the on-chip and memory subsystem's Performance Monitor facilities to count a set of events.
Lparstat	A tool to report logical partition related information. For example, partition configuration, hypervisor call, and processor utilization statistics.
Mpstat	A tool to report logical processor information in logical partition. For example, simultaneous multithread (SMT) utilization, detailed interrupts, detailed memory affinity and migration statistics for AIX threads, and dispatching statistics for logical processors.
topas_nmon or topas	A tool to report the local system's statistics, including: processor, network, I/O, processes, and workload management classes utilization.

Nmon	A commonly used freeware tool for capturing AIX performance data. Use this tool together with nmon analyzer, which loads the nmon output file and automatically creates dozens of graphs reflecting key system performance characteristics. Refer to SAS Performance Monitoring – A Deeper Discussion (at the URL: http://www2.sas.com/proceedings/forum2008/387-2008.pdf) SAS Global Forum 2008 paper for the procedure on collecting nmon trace.

Table 6: Suggested performance tools for monitoring

Monitoring example - processor utilization monitoring

- Suggested monitoring tools: vmstat, iostat, ps, sar, tprof.
- Overall processor utilization monitor:

  A system is probably processor-bound if the system processor utilization (usr+sys) is always greater than 80 percent. *iostat*, *vmstat*, and *sar* can help determine whether a system is processor bound. Here, *vmstat* is used to demonstrate the processor monitoring methodology.

  The four processor utilization groups such as us, sys, wa, idle in *vmstat* report indicates processor spent in user mode, system mode, idle, or I/O wait. The first group, *kernel thread* of two columns **r** and **b** represents statistics about thread queues. It is suggested to check these two columns first.

% us:	Percentage of processor time spent in the application code (that is, SAS). In order to maximize the throughput, ideally this value should be as high as possible.
% sys:	Percentage of processor time spent in the system calls and kernel code. Ideally, system time should be as low as possible. High percentage of system time needs to be investigated.
% wa:	Percentage of processor time spent waiting for an I/O (disk read/write, network, and so on) to be completed. Ideally, this value should be zero. If not, it means that there is some opportunity to improve system throughput by either tuning disk or network or memory configuration.
“r”:	Average number of runable kernel threads during the sampling interval. The run queue is used to display the number of active tasks that are currently waiting for processor resources. The higher the value in **r**, the more the amount of processor work there is to do, which is an indication of processor bottleneck.
“b”:	Average number of kernel threads in the wait queue during the sampling interval. If threads are consistently being forced to wait, the processor performance will get degraded.

- Detailed processor utilization analysis:

  If you decide that the system is processor bound, *tprof* can be used to check which process or program is dominating the processor usage. *ps* can also be used but profiler is a better method. After identifying the high utilization process, you can decide if this behavior is normal and then tune as needed. Conducting further analysis before just adding more processing power to the server is always recommended.

  - Step 1 – Profiling entire system: In order to have a better understanding of SAS workload characteristics on IBM Power, establishing a baseline profile is the first step. You can get

familiar with the workload pattern by checking if there are outstanding routines based on the profiling data. An outstanding routine means that the processor spends more time in this routine compared to others for example, 25% compared to 3%). Further evaluation is required for the outstanding routine.

- Step 2 – Micro-profiling SAS user application: Micro-profiling can focus on where the processor spent the maximum time in the application. For instance, vmstat reports that the processor utilization was mainly on user, and micro-profiling of the application is reasonable.

Understanding processor utilization on Power Systems - AIX

Traditionally, users have been accustomed to use processor utilization as the primary metric to understand the performance of a system running a workload, to do capacity planning and to do charge back. Processor technology has undergone tremendous changes in the past decade and this has called for a change in the way that processor utilization is computed and correctly interpreted. To better understand how processor utilization is computed in AIX and what changes it has undergone in the past decade in synchronization with the IBM POWER® processor-based technology changes, refer to the IBM article, *Understanding Processor Utilization on Power Systems – AIX* at: **ibm.com**/developerworks/mydeveloperworks/wikis/home?lang=en#/wiki/Power%20Systems/page/Understanding%20CPU%20utilization%20on%20AIX

Appendix C: Additional IBM System Storage product information

The IBM System Storage disk systems products and offerings provide storage solutions with superior value for all levels of business (from small and medium business (SMB) to high-end enterprise systems).

IBM System Storage DS8000 series offers high-performance, high-capacity, and secure storage systems that are designed to deliver resiliency and total value for the most demanding, heterogeneous storage environments.

IBM XIV® Storage System is a ground-breaking, high-end, open disk system designed to support business requirements for a highly available information infrastructure. The XIV architecture is a grid of standard Intel®/Linux components connected in the any-to-any topology by means of massively paralleled, non-blocking Gigabit Ethernet, providing outstanding enterprise-class reliability, performance, and scalability.

IBM System Storage DS5000 and DS3000 series are entry and midrange and storage system offering in IBM System Storage DS® series.

For a complete and current list of IBM System Storage offerings, visit the IBM System Storage website at: **ibm.com**/systems/storage/disk/index.html

Appendix D: Additional GPFS information

IBM General Parallel File System is a critical component for SAS Grid solution deployments. GPFS is a high-performance shared-disk cluster file system that provides file system services to parallel and serial applications. GPFS allows parallel applications simultaneous access to a single file or multiple files from any node in the GPFS cluster while managing a high level of control over file system operations. GPFS is particularly appropriate in an environment where the need for data bandwidth exceeds the capability of a distributed file system server. In addition to high-speed parallel file access, GPFS provides fault tolerance, including automatic recovery from disk and node failures.

GPFS provides a flexible virtual storage option for distributed SAS applications that require high performance data access. GPFS currently powers many of the world's largest scientific supercomputers and commercial applications requiring high-speed access to large volumes of data.

GPFS allows users shared file access within a single GPFS cluster and across multiple GPFS clusters. A GPFS cluster consists of:

- AIX nodes, Linux nodes, or a combination of both. A node may be:
  - An individual operating system image on a single computer within a cluster.
  - A system partition containing an operating system.
- One or more shared disks that are defined in GPFS as Network Shared Disks (NSDs)
- A network for GPFS communications, allowing a single network view of the configuration. A single network is used for GPFS communication.

GPFS software can be added to existing analytic servers and coexists with local file systems. This allows migration to GPFS and the use of multiple file system types for SAS solution.

This paper does not cover the installation, implementation, or administration of GPFS, the creation and management of storage pools, or definition of file placement policies. For these topics and tasks, refer to the GPFS product documentation at: http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html or GPFS Redbooks at: **ibm.com**/redbooks/cgi-bin/searchsite.cgi?query=GPFS.

Appendix E: Resources

These web resources provide useful references to supplement the information contained in the paper.

IBM Redbooks

- IBM System Storage SAN Volume Controller Best Practices and Performance Guidelines, SG247521
  ibm.com/redbooks/redbooks/pdfs/sg247521.pdf

- Implementing the IBM Storwize V7000, SG24-7938
  ibm.com/redbooks/redbooks/pdfs/sg247938.pdf

- Implementing the IBM System Storage SAN Volume Controller V6.3, SG24-7933
  ibm.com/redbooks/redbooks/pdfs/sg247933.pdf

- IBM Flex System V7000 Storage Node Introduction and Implementation Guide, SG248068
  ibm.com/redbooks/Redbooks.nsf/RedpieceAbstracts/sg248068.html

- IBM PureFlex System and IBM Flex System Products and Technology, SG247984
  ibm.com/redbooks/Redbooks.nsf/RedbookAbstracts/sg247984.html

- IBM System Storage Solutions Handbook, SG24-5250
  ibm.com/redbooks/redbooks/pdfs/sg245250.pdf

- IBM Midrange System Storage Implementation and Best Practices Guide, SG24-6363
  ibm.com/redbooks/redbooks/pdfs/sg246363.pdf

- IBM XIV Storage System: Architecture, Implementation, and Usage, SG247659
  ibm.com/redbooks/abstracts/sg247659.html

- IBM System Storage DS8000 Series: Performance Monitoring and Tuning, SG24-8013
  ibm.com/redbooks/redbooks/pdfs/sg248013.pdf

- Implementing an IBM b-type SAN with 8 Gbps Directors and Switches, SG24-6116
  ibm.com/redbooks/redbooks/pdfs/sg246116.pdf

- IBM Tivoli Storage Productivity Center V5.1 Technical Guide, SG248053
  ibm.com/redbooks/redpieces/pdfs/sg248053.pdf

- SAN Storage Performance Management Using Tivoli Storage Productivity Center, SG247364
  ibm.com/redbooks/redbooks/pdfs/sg247364.pdf

- Deployment Guide Series: Tivoli Storage Productivity Center for Data (2009), SG24-7140
  ibm.com/redbooks/redpieces/abstracts/sg247140.html?Open

- Using IBM Tivoli Storage Productivity Center for Disk to Monitor the SVC (2005), REDP-3961
  ibm.com/redbooks/abstracts/redp3961.html?Open

Product information and support

- IBM System Storage SAN Volume Controller:
  ibm.com/systems/storage/software/virtualization/svc/

- General Parallel File System - Document Library:
  http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html

- General Parallel File System FAQs (GPFS FAQs):
  http://publib.boulder.ibm.com/infocenter/clresctr/topic/com.ibm.cluster.gpfs.doc/gpfs_faqs/gpfs_faqs.html

- IBM Tivoli Storage Productivity Center, IBM System Storage SAN Volume Controller, and other IBM Storage products:
  ibm.com/systems/storage/index.html

- IBM Tivoli Storage Productivity Center V4.2.2 Hints and Tips (Updated)
  ibm.com/support/docview.wss?rs=1133&context=SS8JB5&context=SSWQP2&dc=DA4A10&uid=swg27008254&loc=en_US&cs=utf-8&lang=en

- IBM Power Systems Information Center
  http://publib.boulder.ibm.com/infocenter/pseries/index.jsp

- Power Systems on IBM PartnerWorld®
  ibm.com/partnerworld/systems/p

- AIX on IBM PartnerWorld
  ibm.com/partnerworld/aix

Acknowledgements

Thanks to the many contributing SAS and IBM team members who assisted in the planning, direction, editing, and reviewing of this paper and those who contributed to previous white papers referenced in this paper.

In particular, special thanks to the following team members:

- Margaret Crevar, SAS, Sr. Manager, Performance Validation, Research and Development
- Leigh Ihnen, SAS, Advisory SAS Solutions Architect, SAS Enterprise Excellence Center
- Frank Battaglia, IBM Certified IT Specialist, Systems and Technology Group
- Ling Pong, IBM Certified IT Specialist, Advanced Technical Skills Solutions
- Narayana Pattipati, IBM Senior Technical Consultant, Systems and Technology Group

About the authors

- Brian Porter, IBM Senior IT Specialist, Systems and Technology Group
- Tony Brown, SAS, Principal Software Developer, Performance Validation, Research and Development
- Harry Seifert, IBM Senior Certified IT Specialist, Advanced Technical Skills Solutions

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

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

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in IBM product announcements. The information is presented here to communicate IBM's current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O

configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.