

Designing and Optimizing

Dell/EMC SAN Configurations **Part 1**

Dell/EMC storage area networks offer storage architects and administrators a wide range of design options and optimization settings. In the first of a two-part series, members of the Dell™ Server and Storage Performance team present collective best practices for designing logical storage units and optimizing storage processor cache settings.

BY **ARRIAN MEHIS AND SCOTT STANFORD**

Dell/EMC storage area networks (SANs) support various workload demands—from transactional databases and file sharing to media streaming and backup servers. This article examines how Dell/EMC SANs can benefit from optimized, well-planned storage processor (SP) and logical storage unit (LUN) designs.

Settings and configurations for SPs and LUNs can be interdependent or independent of each other and often affect back-end performance, depending on the type of workload, optimization, and design configurations. This close symbiotic relationship between SP and LUN configurations can quickly become complex, so storage architects who strive to design their SANs for maximum application or database performance will benefit from understanding this relationship.

SAN performance is affected by the workload read/write type, size, and activity; RAID group design and LUN allocation; SP cache settings; LUN binding and metaLUN design; and host bus adapter (HBA) performance tuning. This two-part series explores detailed concepts and processes for each of the preceding factors, and explains both the theoretical and practical performance advantages that can be realized by implementing properly tuned and optimized SAN configurations. Part 1

focuses on best practices for designing LUNs and optimizing SP cache settings.

Designing LUNs: Best practices

RAID storage offers large capacity, failover protection, high performance, or various combinations of all three. RAID groups can be subdivided into one or more LUNs; RAID groups represent the physical layer with which the SAN hardware communicates, while LUNs represent the logical layer with which the operating system (OS) communicates.

RAID groups

Figure 1 shows how individual physical disks are incorporated into a single RAID group. EMC® Navisphere® Management Suite—the storage management software for Dell/EMC SANs—supports up to 128 LUNs per RAID group¹ and the following RAID types:

- **RAID-5 (individual access array):** Provides data integrity using parity information that is stored on each disk in the LUN. This RAID type is well suited for multiple applications that transfer different amounts of data in most I/O operations.

¹ EMC Navisphere Management Suite version 12 or higher; for more information about EMC Navisphere, visit http://www.emc.com/products/storage_management/navisphere.jsp.

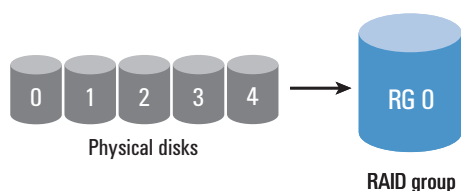


Figure 1. Incorporating five physical disks within one RAID group

- **RAID-3 (parallel access array):** Provides data integrity using parity information that is stored on one disk in the LUN. This RAID type is well suited for single-task applications—such as video storage—that transfer large amounts of data in most I/O operations.
- **RAID-1 (mirrored array):** Provides data integrity by mirroring (copying) its data onto another disk in the LUN. This RAID type provides the greatest data integrity at the highest cost in disk space; it is ideal for an OS disk.
- **RAID-0 (individual access array without parity):** Provides the same individual access features as RAID-5, but does not have parity information. As a result, if a disk in the LUN fails, the information on the LUN is lost. Also, RAID-0 is not technically RAID because the setup is not redundant.
- **RAID-10 (mirrored individual access array without parity):** Provides the same individual access features as RAID-5, but with higher data integrity. This RAID type is well suited for the same applications as RAID-5, but should be used when data integrity is more important than the cost of disk space.
- **Disk (individual disk):** Functions just like a standard single disk and, as such, does not have the data integrity provided by parity or mirrored data. This RAID type is well suited for temporary directories that are not critical.
- **Hot spare (global spare):** Serves as a temporary replacement for a failed disk in a RAID-5, RAID-3, RAID-1, or RAID-10 LUN. Data from the failed disk is reconstructed automatically on the hot spare—either from the parity or the mirrored data on the working disks in the LUN—so data on the LUN is always accessible.

LUN-to-RAID ratio

Although Dell/EMC SANs can support up to 128 LUNs per RAID group, dedicated RAID groups can help maximize performance. Dedicated RAID groups have a one-to-one relationship with a LUN; that is, only one LUN is allocated from the RAID group. The reason for this design is that, for each I/O operation on the LUN, the read or write still must be executed on the physical layer.

Understanding this concept is particularly important for applications such as Microsoft® Exchange Server 2003, because each

client-side change (read or write) usually requires or expects a corresponding return acknowledgment. Delays in I/Os resolving to the physical layer or LUN level can cause upstream delays and message queues to build. For example, if three LUNs are bound to one RAID group, depending on the workload imposed on each LUN, all three LUNs could contend with one another because they are tied to the same physical resource. However, if each LUN were bound to its own RAID group, the LUNs would not compete for physical resources. Figures 2 and 3 illustrate how a RAID group can be allocated to one or multiple LUNs.

Binding multiple LUNs to a single RAID group is a common practice in file-sharing scenarios, where slices of capacity from the RAID group are divided among different users or groups that want to share information. File-sharing environments typically receive random accesses slightly similar to database workloads. However, file shares do not deploy or rely upon the delayed-write methodology at the application level, which is used in transactional database implementations.² In contrast, database workloads experience random I/O behavior and need fast, efficient flushes, or *commits*, from transaction log buffers to the database volumes. Therefore, storage architects should configure one LUN per RAID group for database workloads to reduce the possibility of LUN contention.

LUN and back-end loop symmetry

Just as implementing the optimal binding ratios for RAID groups and LUNs is critical to preventing LUN contention, so too is the concept of maintaining symmetry between LUNs and SP ownership. When binding LUNs using EMC Navisphere software, administrators have several options for assigning LUN ownership. LUNs can be automatically assigned to an SP or designated to a specific SP.

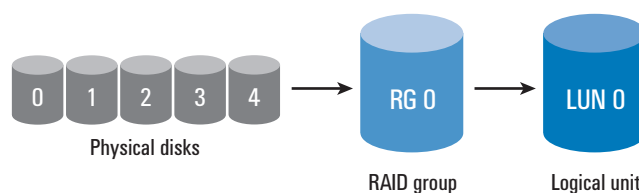


Figure 2. Incorporating physical disks into a RAID group that is mapped to a single LUN (dedicated RAID group)

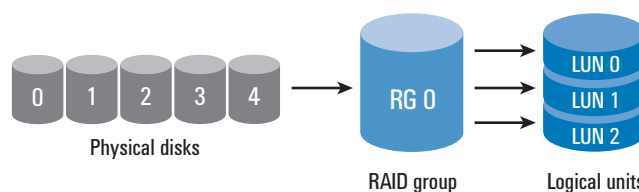


Figure 3. Incorporating physical disks into a RAID group that is mapped to multiple LUNs

² In file-sharing implementations, most network operating systems use *lazy writes*, in which writes are committed to the file system on the storage medium based upon predetermined OS settings.

Dell/EMC CX series arrays	Back-end loops (2 Gbps)	Drives per array
Entry-level	CX200	1
	CX300	1
Mid-level	CX400	2
	CX500	2
Enterprise-level	CX600	2
	CX700	4

Figure 4. Back-end loop and drive quantity matrix for Dell/EMC CX series storage arrays

Regardless of what method is used, administrators should ensure that symmetry between LUNs and SP ownership is maintained.

A typical technique used to formulate LUN symmetry is balancing by workload, which helps spread the workload evenly across the SPs and therefore reduce SP latency.³ If LUN symmetry is not balanced and LUNs are heavily skewed on one SP, client actions originating from a mailbox housed on one SP and destined for a mailbox on a different SP can experience lengthy delays. The ultimate result of such delays can be higher SP cache misses.

All Dell/EMC CX Fibre Channel storage systems include dual SPs and either one or two back-end loops, depending on the model. LUNs should be designed for the best possible balance across the SPs and the back-end loops to help avoid or significantly minimize performance lag caused by overloading the SPs or back-end loops.

Back-end loops are external extensions from the SPs to the disk array enclosures that operate at 2 Gbps bandwidth—the more back-end ports, the more drive support, bandwidth, and I/O capability. Because each SP can be viewed as having two virtual connections to each individual disk on the system, the loops can help to balance workloads across the SPs. For example, the entry-level CX200 and CX300 each have one back-end loop to support additional drives. The mid-level CX400 and CX500 each have two back-end loops. The enterprise-level CX600 also has two back-end loops, but as shown in Figure 4, the next-generation CX700 has four back-end loops supporting the same amount of drives as the CX600.

Sample scenario: Creating symmetry in the SAN

Figure 5 depicts a hypothetical scenario in which a small business deploys a Dell/EMC CX400 storage array. The LUN layout shows an imbalanced symmetry for the SPs and back-end loops. Small, medium, and large workloads are represented respective to the LUN size.

In Figure 5, Storage Processor B (SPb), including both back-end loops (BE0 and BE1), is heavily overloaded with one large workload and two medium workloads. Storage Processor A (SPa)

manages three light workloads. Figure 6 shows a more effective balance of LUN and loop symmetry, given the weight and type of workloads for this small business scenario. Although in Figures 5 and 6 each storage processor has two paths to every disk in the enclosure for failover purposes, these are hypothetical scenarios; the LUNs could just as easily span more than one back-end loop, depending on which loop the comprising disks reside.

An existing LUN can be easily shifted to a different SP using the trespass operation in EMC Navisphere. However, because the physical disk permanently resides on its own specific loop, reassigning LUNs to different loops requires destroying the LUN and recreating RAID groups based on physical disk location. To design more effective SP workload symmetry, this process is necessary.

Optimizing storage processor cache settings: Best practices

Cache settings are integral to SP and LUN configurations, specifically cache page size, cache flush watermarks, and cache allocation. Just as reading from or writing to memory is much faster than reading from or writing to disk, cache can help significantly reduce read and write latencies from the SP to and from the disk.

SP cache is subdivided into two types: read cache and write cache. The cache is present in, and also shared by, the SP memory; read and write caches can be enabled or disabled per SP. LUNs can be allowed or disallowed to use SP read and write caches. This capability is helpful when a workload targeted for a specific LUN will not benefit from cache—either read cache, write cache, or both—while the other LUNs owned by the SP will benefit from cache. So, when designing LUNs for optimum response times and low latency, administrators can disable cache on certain LUNs to make more cache available to LUNs that will benefit as well as to reduce unnecessary cache thrashing.

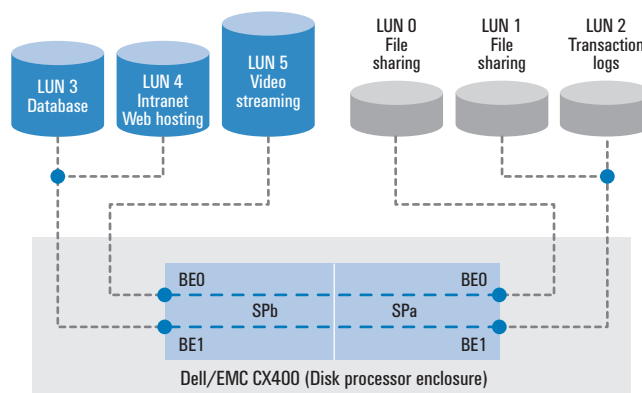


Figure 5. Imbalanced SP workload distribution for hypothetical small-business scenario

³ EMC PowerPath® software can be used in combination with Navisphere to help provide load balancing among SPs. However, to fully leverage the capabilities of a Dell/EMC storage array, administrators should carefully maintain proper SP-to-LUN symmetry.

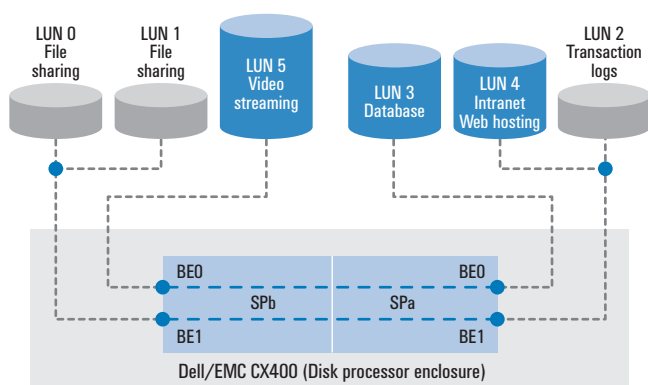


Figure 6. Balanced SP workload distribution for hypothetical small-business scenario

Cache page size

Page size sets the number of kilobytes stored in one cache page. A page is a portion of memory reserved for a specific block of I/O, which is the actual data being sent to or read from the LUN. Unlike physical disk I/O logic controllers, the SPs manage the read and write caches by pages instead of sectors—the larger the page size, the more continuous sectors the cache stores in a single page. EMC recommends the following page sizes:

- **General file-server applications:** 8 KB or 16 KB
- **Database applications:** 2 KB or 4 KB

Cache flush watermarks

A cache watermark determines when a processor flushes its write cache—also known as watermark processing. When an SP flushes its write cache, it writes dirty pages to disk. A *dirty page* is a write cache page with modified data that has not yet been written to disk. The high watermark is the percentage of dirty pages in the write cache; when the high-watermark threshold is reached, the SP begins flushing its write cache. The low watermark also is the percentage of dirty pages in the write cache; when this threshold is reached, the SP stops flushing its write cache. EMC recommends that the high watermark be set at 80 percent and the low watermark be set at 60 percent. Although watermark settings can be adjusted above and below the default recommendations, administrators should fully understand the implications of changing these settings and how those changes can affect cache miss rates or cache thrashing.

Read cache

Read cache relies heavily on prediction and the type of workload to which the LUNs are exposed. Read cache hits are extremely high during sequential reads of contiguous data streams. This type of scenario is typical in file-sharing environments where large amounts of sequential data may stream back and forth from the client to the storage system. In such an environment, administrators should

allocate a large portion of the read cache to anticipate a high percentage of cache hits based on the workload.

Cache prefetching

The EMC read-prediction algorithm—also known as read-ahead caching or prefetching—is adaptive in nature. The SP will prefetch data (assuming prefetching is enabled for the corresponding LUN) and fill the read cache only after two sequential reads that share spatial locality. By using this algorithm, the SP assumes that if there were two reads from the same location, there is a high probability that it will need that data from the next sequential location again. The SP will first check the read cache for the data. If the data is not present, it will then check the disk. This algorithm can help reduce SP-to-LUN latency by filling the read cache with prefetched data before it is actually needed.

EMC Navisphere offers three types of cache prefetching:

- **None:** Disables all prefetch properties.
- **Constant:** Prefetches data of a constant length. This type of prefetching is recommended if the read data size is unvarying and regular in length. If selected, only *prefetch size* and *segment size* options are available. Prefetch size is the number of blocks of data to prefetch for one host read request. Segment size is the number of blocks of data to prefetch in one read operation from the LUN. An SP reads one segment at a time from the LUN, because smaller prefetch requests interfere less with other host requests.
- **Variable:** Prefetches data of variable length. If selected, only *prefetch multiplier*, *segment multiplier*, and *maximum prefetch* options are available. The prefetch multiplier determines the amount (in disk blocks) of data to prefetch. For example, if the prefetch multiplier is set to 4 and the amount of data requested is 2 KB, then the variable prefetch size is 8 KB (16 disk blocks). The segment multiplier determines the size (in disk blocks) of the segments that make up the prefetch operation. This option allows the variable prefetch size to be broken into smaller chunks of data, because smaller prefetch requests interfere less with other host requests. For example, if the segment multiplier is set to 2 and the amount of data requested is 2 KB, then the variable segment size is 4 KB (8 disk blocks). The maximum prefetch is the maximum number of disk blocks to prefetch; the default setting is 4,096.

In an environment where reads are sequential, such as a file server, administrators should select variable prefetching. Although Navisphere defaults to a prefetch multiplier of 4 and a segment multiplier of 4, the multipliers can be further tuned for specific needs. For example, if reads are very sequential and constant, such as with media streaming or backup servers, it may be best to select

constant prefetching. Depending on the typical read size, regardless of whether variable or constant prefetching is used, the prefetch size may be increased to further bridge the gap between the SP and the cache. This will allow the cache to be loaded with more read data before it is needed by the SP, ultimately decreasing read latency.

In an environment where reads are random, administrators should select variable prefetching, with a prefetch multiplier of 1 and a segment multiplier of 1. That way, the prefetch data traffic will be reduced to avoid interfering with other host requests, but it will still allow for occasional read cache hits with minimal unused prefetched blocks.

In an environment where reads occur rarely (*very* random) or never, such as a database transaction log LUN in which only writes occur, administrators should set the prefetch type to none. Because no reads occur and the cache will never be used by the respective LUN, enabling any type of read-ahead logic to fill the cache is unnecessary. Ultimately, administrators may want to disable read cache at the SP level if read cache hits are seldom to none.

To use any of the read prefetch features, read cache must be enabled on the LUNs before they can benefit from prefetching and SP read cache.

Write cache

The Dell/EMC CX series Fibre Channel storage arrays introduce an additional step to the storage-system writing process, in which writes are performed in the SP write cache before the destination LUN. In a typical database application, new transactions or updates bound for a transaction log file housed on the transaction log LUNs must first be written to the SP write cache. However, the use of write cache applies to all write types, not just random database writes, so any write that takes place on the storage system will write to cache first. Specifically for typical database applications, when the cache is considered dirty, it is flushed to a transaction log file (buffer). Then, when the log file reaches a preset size, all writes are committed to the database.

High-availability (HA) cache vaulting is available in Dell/EMC CX series storage systems. HA cache vaulting determines the availability of storage-system write caching; if enabled, it will disable write caching when a single vault disk fails. Before disabling HA cache vaulting, administrators should consider the resiliency requirements of the application and storage array. Disabling cache vaulting simply allows write caching to continue on a failed drive. HA cache vaulting will have no effect on performance unless a drive fails, in which case the cache *image* will be quickly dumped to disk in an effort to save cached data.

Note: Write cache must be enabled on the LUNs before the respective LUN can benefit from SP write cache.


Advanced cache optimization

After using the previously discussed settings—read and write cache, cache page size, cache watermarks, and prefetching—as a starting

point for tuning, administrators can further tune Dell/EMC CX series storage systems using EMC Navisphere Analyzer. Key Navisphere Analyzer counters such as read hit percentage; write hit percentage; write cache flush ratio; dirty pages percentage; throughput and bandwidth; and SP, LUN, and disk utilization provide solid indicators of the effect of each setting on I/O operations or effective bandwidth utilization.

Building a better SAN

LUN design methods and the interrelationships among LUN, RAID, and back-end loop symmetry can have a significant impact on achieving optimal I/O operations with low latencies. Storage processor cache optimizations further reveal how advanced storage settings can be utilized to support typical application workloads.

Performance, capacity, and redundancy are key considerations when determining the optimal storage solution for specific application workloads. EMC default performance settings can provide a solid foundation that can help administrators achieve optimal overall SAN performance. However, maximum performance can be obtained only through a thorough analysis of workload requirements and an understanding of the effects of performance settings and design concepts. 

References

EMC Corporation. “EMC CLARiON Fibre Channel Storage Fundamentals.” EMC Engineering White Paper, October 31, 2003. http://www.emc.com/products/systems/pdf/H1049_emc_clariion_fibre_channel_storage_fundamentals_ldv.pdf.

EMC Corporation. *EMC Navisphere R13 Help Tutorial*.

Arrian Mehis (arrian_mehis@dell.com) is a systems engineer on the Server and Storage Performance team in the Dell™ Enterprise Product Group. His responsibilities include Microsoft Exchange Server single-node performance analysis on Dell server, SCSI, SAN, and RAID solutions. Arrian has a B.S. in Computer Engineering with a minor in Information Systems from the Georgia Institute of Technology.

Scott Stanford (scott_stanford@dell.com) is a systems engineer on the Server and Storage Performance team in the Dell Enterprise Product Group. His current work focuses on Microsoft Exchange Server cluster benchmarking and server/storage performance analysis. He has an M.S. in Community and Regional Planning from The University of Texas at Austin and a B.S. from Texas A&M University.

FOR MORE INFORMATION

Dell/EMC:
<http://www1.us.dell.com/content/products/category.aspx/storage?c=us&cs=555&l=en&s=biz>