penguinpunk.net
storage design principles
revision 0.2

synopsis

The following is a simple document providing information and background on a number of principles that should be followed when designing storage solutions using EMC midrange systems such as the CLARiiON, Celerra and VNX. It is intended that this document provide a high-level view of the potential issues facing the storage designer, and that further, detailed design decisions be undertaken with the aid of the VNX Configuration Guidelines document and support from EMC.

table of contents

table of figures

tables

introduction

The purpose of this document is to provide configuration guidance for the deployment of storage solutions. It covers a number of storage design considerations and caveats, and provides a framework upon which can be built robust, available and scalable solutions. This document does not provide a specific detailed design document, but rather a set of design principles that will drive detailed design activities. The document “EMC VNX Configuration Guidelines – Version 0.01” (released XX.12.2011) provides further information on specific configuration scenarios and best practices for the EMC VNX midrange array.

executive summary

There are a number of principles that should be observed when designing storage solutions.

Design for Performance and Capacity

Understanding the performance workload as well as the space required is critical to a successful design.

RAID 5 vs RAID 6 vs RAID 1/0

Knowing what RAID is and understanding the different types available is important. RAID 1/0 isn't always a waste of time, but it's not always the answer either.

EFD vs SAS / FC vs NL-SAS / SATA-II

All disks were not created equal. Particularly not "flash" ones.

Traditional LUNs vs MetaLUNs

Sometimes your LUN design needs to be a bit fancy.

RAID Groups vs FAST VP Pools

RAID Groups are solid, FAST VP Pools are not always great, but they are convenient.

Block vs File

Maybe you just want something to access a bunch of storage over the network. There's nothing wrong with that.

Storage Virtualisation

Doing more with less, more or less.

Native Multipathing vs Vendor-specific tools

Sometimes it's worth paying a little extra.

design for performance and capacity

overview

Good storage design should not simply revolve around capacity requirements. While these are fairly important to the overall success of a deployment, the performance requirements of solutions are frequently overlooked. This generally leads to much angst and gnashing of teeth on the part of the customer, operational support staff and the solution designer. The following section provides a high-level view of a number of considerations that should be made when designing for performance-oriented storage solutions. This is by no means an exhaustive list of considerations, but can be used as a starting point for further discussion. While it is not always possible to understand the workload requirements of a given solution (particularly when that solution is new), it is important to attempt to understand the workload before solutions are delivered.

Note that this document focuses primarily on midrange EMC storage, such as that found in the CLARiiON and VNX range. Design principles for VMAX or other vendors' arrays will differ slightly, although the general concepts are similar.

understanding your I/O profile

Any discussion of a device or interface being able to support x Gbs of bandwidth, and y IOPS of throughput is meaningless without knowing the size of the I/O.

A key point to understand is the ratio of writes to reads, that is, on EMC midrange systems, a write will generally consume more resources than reads do. If you are looking at deploying a write-intensive application, the locality of the workload in relation to other workloads may not be as simple to achieve on consolidated storage.

When comparing sequential vs random vs mixed workloads, keep in mind that small, random I/Os will use more storage system resources than large sequential I/Os. It is also important to note that, generally speaking, sequential workloads have better bandwidth than random or mixed I/O.

When deciding whether you're dealing with large-block or small-block, a rule of thumb is that small is considered to be 16KB or less, while large is considered to be 64KB or greater.

Consider whether the workload will be steady or bursty. I/O traffic can be steady or can vary widely over a short period of time (bursty). Bursty behaviour results in spikes in traffic. This can cause some amount of consternation for operational support staff if they are not forewarned.

The following table provides some generic workload categories, and the approximate workload types and characteristics that can be associated with them for planning purposes. Note that this is no replacement for actual workload performance analysis, but it can be used as a general guide for performance sizing activities.

Workload	**I/O Type**		**Access Type**		**I/O Size**		**I/O Flow**		**Descriptive Metric**	
	Random	***Sequential***	***Reads***	***Writes***	***Small***	***Large***	***Steady***	***Bursty***	***Throughput***	***Bandwidth***
OLTP	X		X	X	X			X	X	
Messaging	X		X	X	X	X		X	X	
File serving	X		X	X	X	X		X	X	
DSS	X	X	X			X	X			X
Backup-to-Disk		X		X		X	X			X
Rich Media		X	X			X	X			X

Table 1 - Workload Profiles

Finally, consider the spindle RPM to latency relationship

Spindle RPM	Average Latency (ms)
5400	5.6
7200	4.2
10000	3.0
15000	2.0

Table 2 - Spindle RPM to Latency

RAID 5 vs RAID 6 vs RAID 1/0

general

A number of people do not like to use RAID 1/0, as they feel that RAID 5 generally provides performance that is "good enough". Sometimes, however, RAID 5 won't provide the bandwidth required for the proposed workload, and you'll need to look at using RAID 1/0. RAID 6 is generally used where spindle sizes are greater than 1TB and the risk of losing 2 disks during a RAID group rebuild is higher due to the time it takes to rebuild these larger RAID Groups.

raid 1/0

RAID 1/0 combines RAID 0 stripes with RAID 1 mirrors. A series of RAID 1 mirrors is configured in a RAID 0 stripe. One disk in each mirror can be lost while still providing data redundancy. RAID 1/0 receives a performance benefit from striping. RAID 1/0 includes some optimization of reads that takes advantage of two drives with identical data. It has very good random read and write performance. It also has good sequential read and write performance.

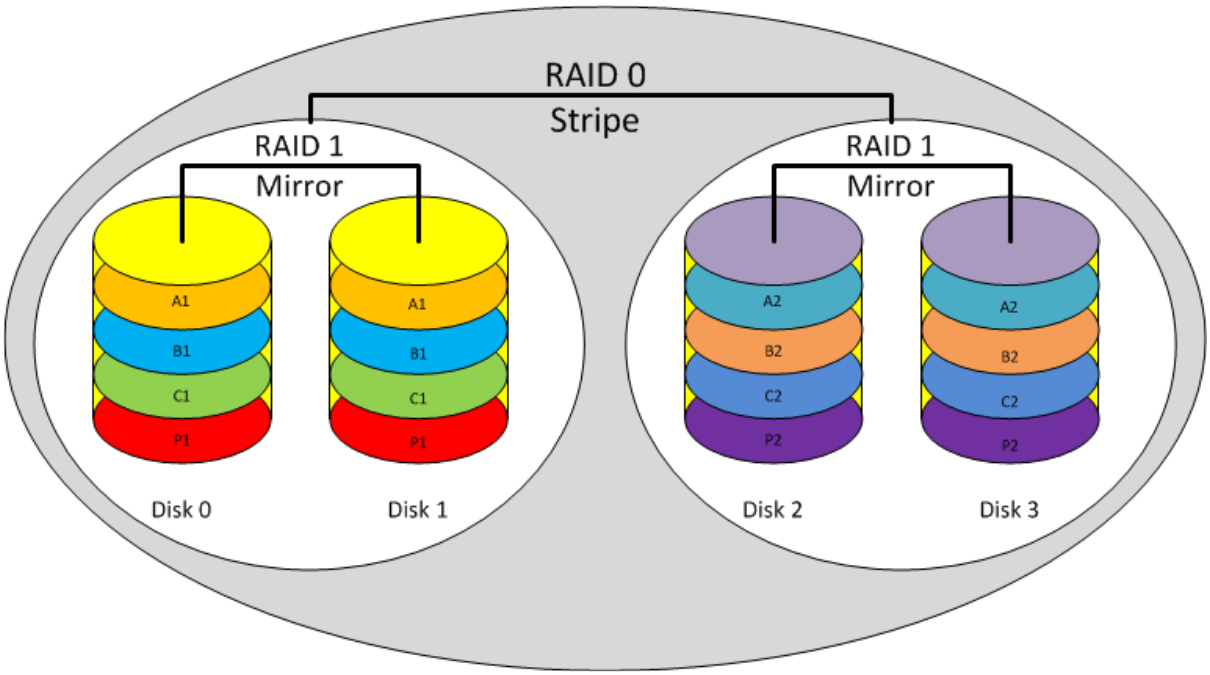

Figure 1 - RAID 1/0

Note that, when you're calculating the IOPS that can be delivered by a RAID 1/0 configuration, a read is one back-end operation and a write is two. On a VNX, RAID 1/0 is commonly configured in a 4 + 4 configuration.

raid 5

RAID 5 provides striping of data with parity. The equivalent of one disk is used for parity data, striped across all disks in the group. One disk in the stripe can be lost while still providing data redundancy. RAID 5 has excellent random read performance. Performance improves with

increasing numbers of disks in the RAID Group. Random write performance is fair, due to the parity calculation. RAID 5 random write performance is better than RAID 3, because there is no dedicated parity disk to cause a bottleneck. Sequential read performance is good. Sequential write performance is good.

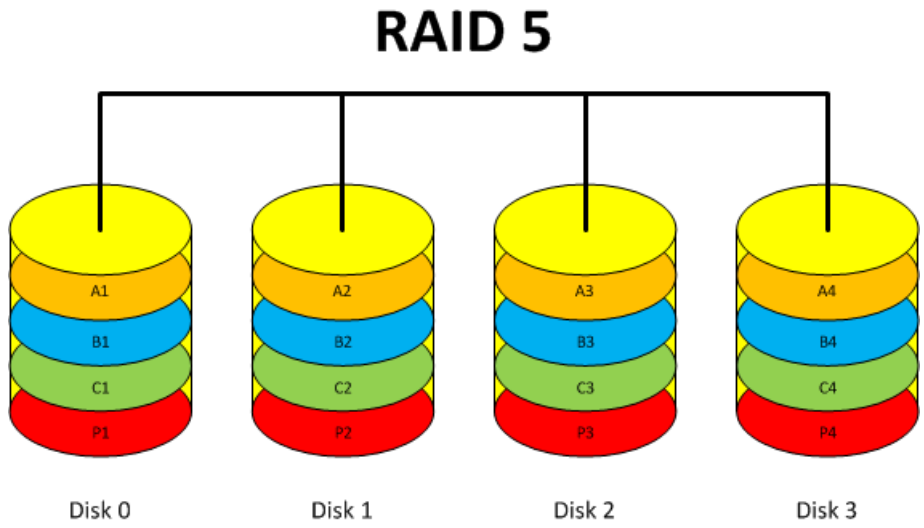

Figure 2 - RAID 5

In RAID 5, a read is one back-end operation, and a write is one for a cached full stripe and up to four for a single small I/O less than a stripe. On a VNX, RAID 5 is commonly configured in a 4 + 1 configuration.

raid 6

RAID 6 provides striping of data with dual-parity. The equivalent of two disks is used for parity data, striped across all disks in the group. Up to two disks in the stripe can be lost while still providing data redundancy. RAID 6 has similar performance to RAID 5. Where RAID 6 suffers in comparison is in the requirement for the additional parity calculation. It has the lowest random-write performance (equal user drive count) of any RAID level. It has excellent random read performance. Sequential read performance is good. Performance improves with smaller stripe widths. Sequential write performance is fair.

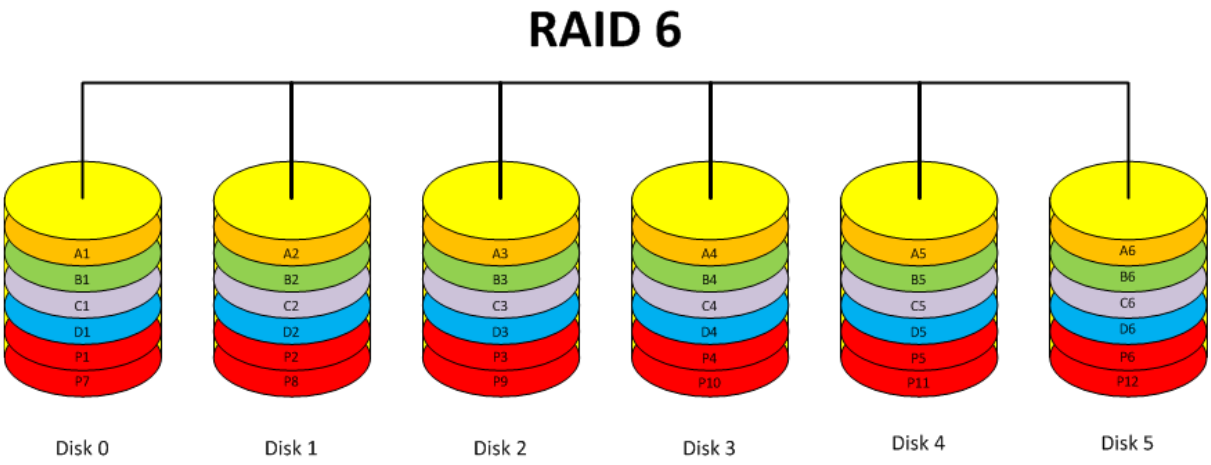

Figure 3 - RAID 6

In RAID 6, a read is one back-end operation, and a write is one for a full cached stripe and six for a single small I/O less than a stripe. On a VNX, RAID 6 is commonly configured in a 6 + 2 configuration.

other raid types

RAID 0 stripes data across all disks with no parity. RAID 0 is sometimes used by gamers who know no better, but RAID 0 is never an option if you care about your data. RAID 3 provides striping similar to RAID 5, but with a dedicated parity disk. RAID 3 was popular for about 5 minutes with backup-to-disk and streaming media applications, but most people have found that RAID 5 provides comparable performance.

The following table provides a summary of the strengths and weaknesses of the various RAID types.

RAID	Config	Random Read	Random Write	Sequential Read	Sequential Write	Parity Penalty - Performance	Parity Penalty - Capacity
5	4 + 1	Excellent	Fair	Good	Good	R=1, W=1 or 4	1 disk
6	6 + 2	Excellent	Bad	Good	Fair	R=1, W=1 or 6	2 disks
1/0	4 + 4	Very Good	Very Good	Good	Good	R=1, W=2	4 disks

Table 3 - RAID Strengths and Weaknesses

efd vs sas / fc vs nl-sas / sata-ii

enterprise flash drives

An SSD is a data storage device that uses solid-state memory to store persistent data with the intention of providing access in the same manner of a traditional block I/O hard disk drive[1]. An EFD is an “enterprise-grade” implementation of an SSD, using single-level cell flash memory. Broadly speaking, EFD drives provide Tier 0 performance, particularly with sustained, read-type workloads. It does not necessarily perform well with random I/O.

serial attached scsi / fibre channel

Serial Attached SCSI (SAS) is a point-to-point serial SCSI interface that replaces Parallel SCSI. The latest iterations of this interface run at 6Gbps. It is commonly a replacement for Fibre Channel, which in turn was the replacement enterprise version of Parallel SCSI. Fibre Channel drives can typically run at 2Gbps or 4Gbps, and are still used in EMC CLARiiON CX4 midrange arrays.

nearline-sas / sata-ii

Nearline-SAS (NL-SAS) drives are enterprise SATA drives that use a SAS interface. They are generally considered slightly faster than SATA drives, but shouldn’t be treated as a high-capacity performance drive.

performance by drive type

The following table provides information (based on figures provided by EMC) on the approximate IOPS available per disk. Fibre Channel and SATA figures are included to accommodate legacy CX4 configurations. It is recommended that a figure of 70% of the listed IOPS figure be used to ensure that designed workloads can accommodate bursts in throughput requirements.

Drive Type	IOPS
SAS 15K RPM	180
SAS 10K RPM	150
NL-SAS 7.2K RPM	90
EFD	3500
Fibre Channel 15K RPM	180
Fibre Channel 10K RPM	140
SATA 7.2K RPM	80

[1] http://en.wikipedia.org/wiki/Solid-state_drive

SATA 5.4K RPM	40

Table 4 - IOPS by Drive Type

The following table provides average seek time figures based on particular drive types.

	EFD	SAS		NL-SAS
Rotational Speed (RPM)	N/A	15000	10000	7200
Attachment Speed (Gb/s)	6			
Typical Average Seek Time (ms)[2]	0.03	3.1	3.7	9.0

Table 5 - Average Seek Times by Drive Type

traditional luns vs metaluns

A Logical Unit Number (LUN) is the mechanism by which volumes are presented to hosts. LUNs can be configured as Thick (fully provisioned) or Thin (blocks consumed as required by the filesystem).

MetaLUNs are an EMC technology where LUNs are created from 2 or more traditional LUNs. MetaLUNs are one solution to LUNs requiring capacity that exceeds the capacity of a single RAID group or LUNs with greater maximum capacity than available within a Virtual Pool. MetaLUNs can also be used to increase the performance of a standard LUN by striping components across multiple RAID Groups (thus providing more than 16 disks worth of performance).

There are two methods of creating and expanding MetaLUNs: Concatenation and Striping. Concatenated expansion simply adds capacity to the base entity.

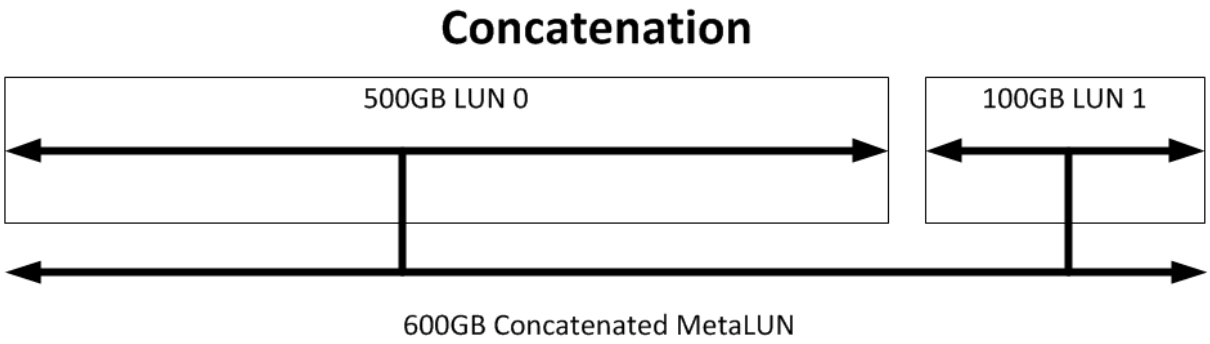

Figure 4 – Concatenated MetaLUN

Striped expansion restripes the base LUN's data across the base LUN and component LUNs being added. A requirement of striped expansion is that the additional extent shares the same size and performance characteristics as the original MetaLUN components.

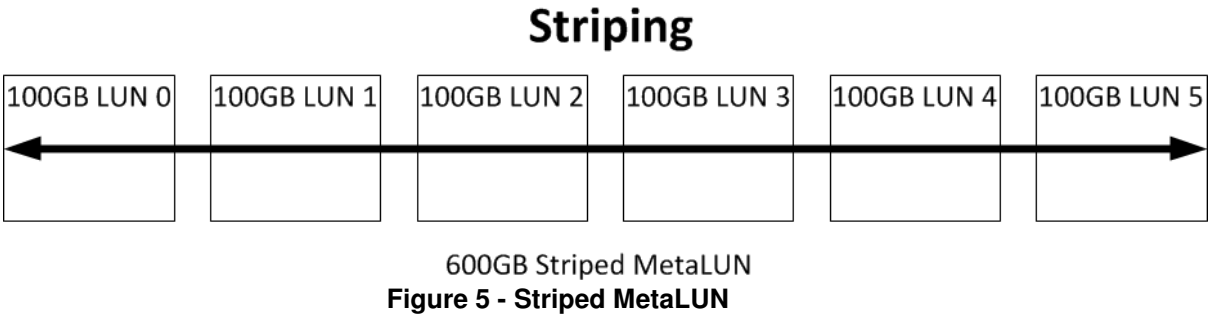

Figure 5 - Striped MetaLUN

[2] Read / write for EFD

raid groups vs fast vp pools

RAID Groups are the mechanism by which various disks are grouped together to provide a platform for LUN configuration. On EMC midrange systems, the maximum number of disks in a RAID group is 16.

FAST VP Pools are similar to RAID Groups, but they use internal, private RAID Groups to get around the traditional RAID Group limitation. It should be noted, however, that various midrange EMC systems have limitations on the number of disks that can be included in a pool. The core idea of pools is that a number of different disk types (EFD, SAS and NL-SAS) can be included in the one pool and EMC's Fully Automated Storage Tiering (FAST) can move slices of data between tiers based on workloads. It's a great idea but not technologically mature enough to consider it as an automatic deployment option.

For example, if you have a 15 disk pool, and then add 5 disks to the pool, new LUNs will be placed on these 5 disks until their usage is equal to the other disks. Once this is done, data is then striped (not re-striped) across all disks. This can create hotspots and reduce performance significantly in a scenario where you increase the number of disks in a pool in an incremental fashion. The solution to this is to fully provision the pool from the start, but occasionally this won't be possible, or feasible.

It is suggested that FAST VP Pools shouldn't be positioned for performance-sensitive applications that require low millisecond response times. EMC are working on this though, so stay tuned.

One other, minor, point is that a given LUN's queue depth equals (14 * (the number of data drives in the LUN)) + 32. So for a LUN that resides on a 4 + 1 RAID 5 RAID Group, the queue depth would be 88. To get around this limitation, you can use MetaLUNs in combination with traditional RAID Groups to provide an increased queue depth for the LUN. You cannot do this with LUNs in a FAST VP pool. It should also be noted that the practical maximum number of concurrent host requests at the front-end port on a CLARiiON is around 1600 – so you need to plan your high workload host access accordingly. A symptom of this threshold being exceeded is a high number of QFULL errors being encountered on the array. This is bad for both the array and the hosts attached to it.

block vs file

The mechanism by which storage can be presented from the array to the host can be as critical as the back-end storage itself. There are two types of storage presentation – block and file. Block storage reads and writes blocks of data using logical block addresses (LBAs), which are translated into sector addresses on drives. This is similar to having an internal, raw device presented to the host's disk controller. File storage adds a filesystem to the block storage.

Block presentation of storage can be via Fibre Channel, iSCSI or FCOE. File presentation can be done via NFS or CIFS. There are a number of pros and cons for each type of connection mechanism. FC has traditionally been preferred for high bandwidth applications, as 8Gbps was seen as a much better option than 1Gbps iSCSI or NFS. However, with the advent of 10Gbps IP networking, the decision is not as simple as it could be. The disadvantage of FC has been the high cost of host bus adapters (HBAs) and associated fabric infrastructure. People have thus positioned iSCSI and NFS as cheaper alternatives, able to leverage existing infrastructure. The issue with this is that, unless you use QoS or dedicated IP storage infrastructure, the impact on production IP workloads can be dramatic.

storage virtualisation

At a high level, storage virtualisation is used to provide a layer of abstraction between physical and logical storage devices. This allows for:

- The deployment of heterogeneous storage devices;
- Replication of storage to remote sites;
- Dynamic re-allocation of workload to under-utilised devices;

- The pooling of devices and addition of caching to increase performance; and
- Non-disruptive data migration.

The following diagram provides a conceptual view of storage virtualisation.

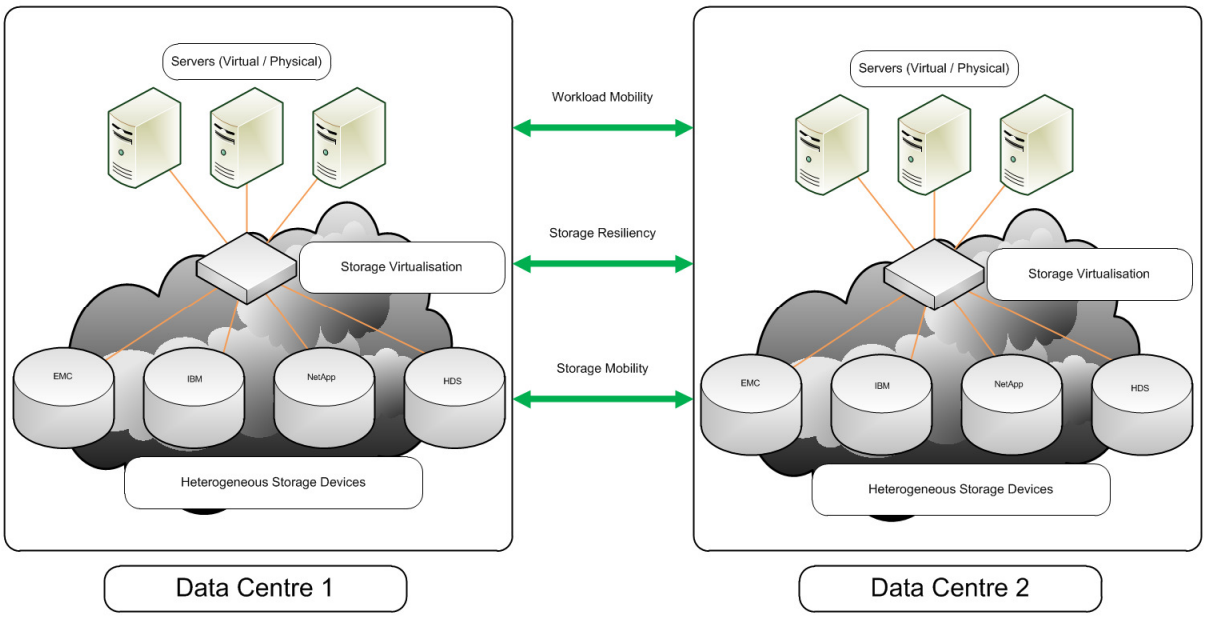

Figure 6 - Storage Virtualisation - Conceptual

Depending on the use case, storage virtualisation may introduce unnecessary complexity into the storage environment.

native multipathing vs vendor-specific tools

general

The following section provides information on a number of storage multipathing options available from Microsoft, VMware, and EMC. The decision to use built-in multipathing software versus vendor-supplied will often depend on the solution's available budget and the performance requirements.

Generally speaking, when a vendor-supplied multipathing solution is available and affordable, it is recommended that this be used.

microsoft mpio

New MPIO features in Windows Server 2008 include a Device Specific Module (DSM) designed to work with storage arrays that support the asymmetric logical unit access (ALUA) controller model (as defined in SPC-3), as well as storage arrays that follow the Active/Active controller model.

The Microsoft DSM provides the following load balancing policies. Note that load balance policies are generally dependent on the controller model (ALUA or true Active/Active) of the storage array attached to Windows-based computers.

- **Failover** No load balancing is performed. The application specifies a primary path and a set of standby paths. The primary path is used for processing device requests. If the primary path fails, one of the standby paths is used. Standby paths must be listed in decreasing order of preference (the most preferred path first).
- **Failback** Failback is the ability to dedicate I/O to a preferred path whenever it is functioning. If the preferred path fails, I/O is directed to an alternate path until function is restored to the preferred path, but I/O automatically switches back to the preferred path when function is restored.

- **Round-robin** The DSM uses all available paths for I/O in a balanced, round-robin fashion.
- **Round-robin with a subset of paths** The application specifies a set of paths to be used in a round-robin fashion, and a set of standby paths. The DSM uses paths from primary pool of paths for processing requests, as long as at least one of the paths is available. The DSM uses a standby path only when all primary paths fail. Standby paths must be listed in decreasing order of preference (most preferred path first). If one or more of the primary paths become available, DSM uses the standby paths in their order of preference. For example, given 4 paths — A, B, C, and D — A, B, and C are listed as primary paths, and D is standby path. The DSM chooses a path from A, B, and C in round-robin fashion as long as at least one of them is available.

  If all three fail, the DSM uses D, the standby path. If A, B, or C become available, DSM stops using D and switches to available paths among A, B, and C.
- **Dynamic Least Queue Depth** The DSM routes I/O to the path with the least number of outstanding requests.
- **Weighted Path** The application assigns weights to each path; the weight indicates the relative priority of a given path. The larger the number, the lower the priority. The DSM chooses the path that has the least weight from among the available paths.

The Microsoft DSM preserves load balance settings even after the computer is restarted. When no policy has been set by a management application, the default policy that is used by the DSM is either Round Robin, when the storage controller follows the true Active/Active model, or simple failover in the case of storage controllers that support the SPC-3 ALUA model. With simple Failover, any one of the available paths can be used as the primary path, and remaining paths are used as standby paths.[3]

vmware native

These multipathing policies can be used with VMware ESX/ESXi 4.x and ESXi 5.x:

- Most Recently Used (MRU) — Selects the first working path, discovered at system boot time. If this path becomes unavailable, the ESX/ESXi host switches to an alternative path and continues to use the new path while it is available. This is the default policy for Logical Unit Numbers (LUNs) presented from an Active/Passive array. ESX/ESXi does not return to the previous path when if, or when, it returns; it remains on the working path until it, for any reason, fails.
- Fixed (Fixed) — Uses the designated preferred path flag, if it has been configured. Otherwise, it uses the first working path discovered at system boot time. If the ESX/ESXi host cannot use the preferred path or it becomes unavailable, ESX/ESXi selects an alternative available path. The host automatically returns to the previously-defined preferred path as soon as it becomes available again. This is the default policy for LUNs presented from an Active/Active storage array.
- Round Robin (RR) — Uses an automatic path selection rotating through all available paths, enabling the distribution of load across the configured paths.
- For Active/Passive storage arrays, only the paths to the active controller will used in the Round Robin policy.
- For Active/Active storage arrays, all paths will used in the Round Robin policy. **Note**: This policy is not currently supported for Logical Units that are part of a Microsoft Cluster Service (MSCS) virtual machine.
- Fixed path with Array Preference — The VMW_PSP_FIXED_AP policy was introduced in ESX/ESXi 4.1. It works for both Active/Active and Active/Passive storage arrays that support ALUA. This policy queries the storage array for the preferred path based on the arrays preference. If no preferred path is specified by the user, the storage array selects the preferred path based on specific criteria. **Note**: The VMW_PSP_FIXED_AP policy has been removed from the ESXi 5.0 release and VMW_PSP_MRU became the default PSP for all ALUA devices

[3] http://technet.microsoft.com/en-us/library/cc725907.aspx

The Round Robin (RR) multipathing policies have configurable options that can be modified at the command-line interface. Some of these options include:

- Number of bytes to send along one path for this device before the PSP switches to the next path.
- Number of I/O operations to send along one path for this device before the PSP switches to the next path.

For more information, see *Round Robin Operations with esxcli nmp roundrobin* in the vSphere Command-Line Interface Installation and Reference Guide for the appropriate version of VMware product. [4]

ibm system storage multipath subsystem device driver (sdd)

A good overview of the functionality of the SDD is provided in the "IBM System Storage - Multipath Subsystem Device Driver User's Guide"[5]. IBM states that "[t]he SDD is a software solution to support the multipath configuration environments in supported storage devices. It resides in a host system with the native disk device driver and provides the following functions:

- Enhanced data availability
- Dynamic input/output (I/O) load-balancing across multiple paths
- Automatic path failover protection
- Concurrent download of licensed machine code."

Here's a picture of what an SDD implementation looks like.

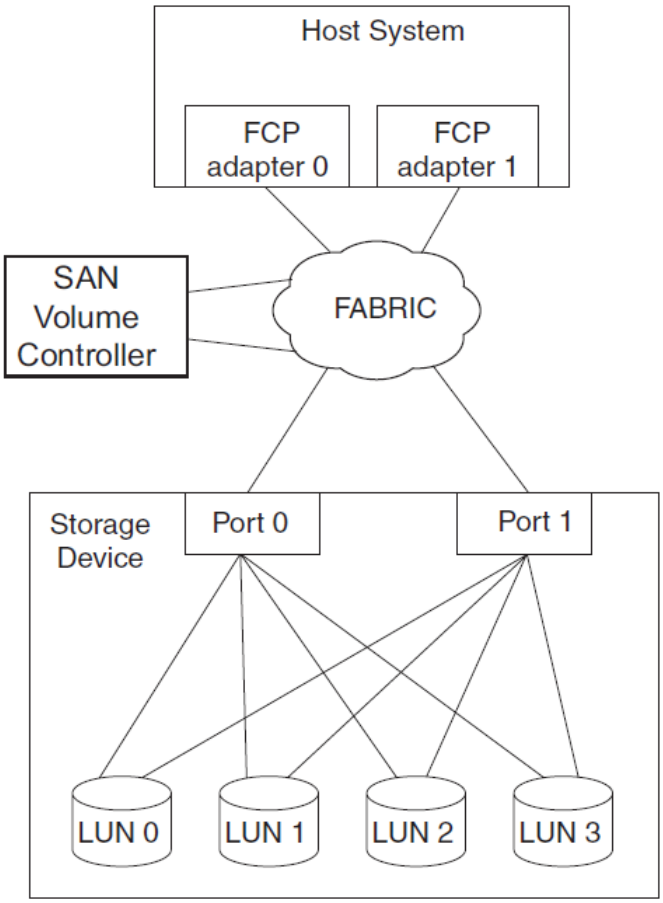

Figure 7 - IBM SDD Overview

[4] http://kb.vmware.com/kb/1011340
[5] ftp://ftp.software.ibm.com/storage/subsystem/UG/1.8--3.0/SDD_1.8--3.0_User_Guide_English_version.pdf

emc powerpath and powerpath/ve

EMC PowerPath and PowerPath/VE provide the following advantages:

- Failover from port to port on the same SP (minimizes LUN trespassing);
- Port load balancing across SP ports and host HBAs;
- Higher bandwidth from host to storage system.

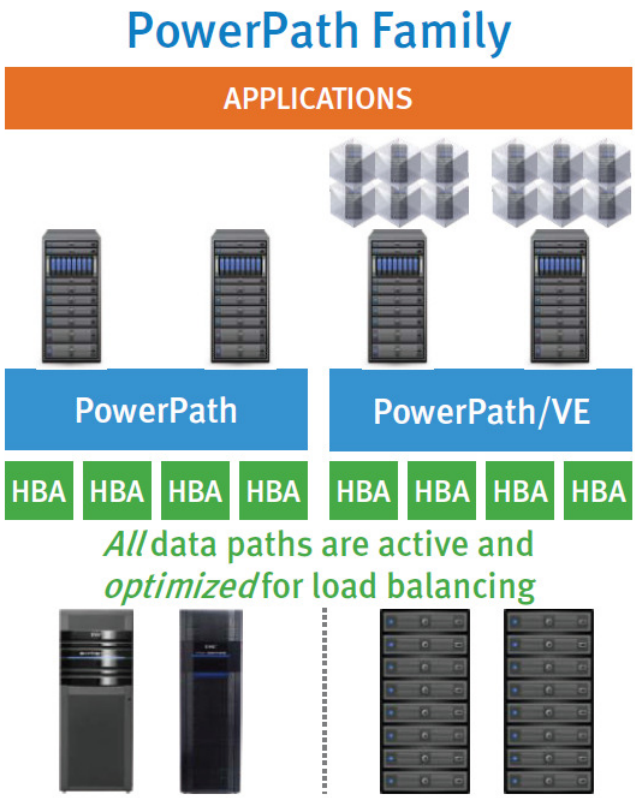

Figure 8 - PowerPath Family[6]

There is, however, a cost associated with balancing storage workloads across active paths. These include:

- Some host CPU resources are used during both normal operations, as well as during failover;
- Every active and passive path from the host requires an initiator record (this is limited by the system);

Active paths increase time to fail over in some situations (tries several paths before attempting to trespass the LUN).

conclusion

There are a few things that you need to be mindful of when designing and deploying EMC midrange storage. Hopefully this document provides a solid foundation upon which you can build performance-focused storage solutions for the midrange.

[6] EMC PowerPath Family: PowerPath and PowerPath/VE Multipathing. EMC Data Sheet. EMC P/N L751.27. August 2011.