penguinpunk.net
vnx5700 configuration guidelines
revision 0.1

synopsis

The following is a simple document providing information and background on a number of configuration guidelines specific to EMC's VNX5700 (Block). It is intended that this document be read in conjunction with the high-level "Storage Design Principles" document, as well as supporting documentation from EMC and elsewhere.

table of contents

table of figures

tables

introduction

The purpose of this document is to provide configuration guidance for the deployment of block storage, with particular focus on the VNX5700. It covers a number of configuration scenarios and caveats, and provides a framework upon which can be built robust, available and scalable solutions.

This document does not refer to a specific detailed design document, but rather a process overview that will drive detailed design activities.

Most of the information in this guide is available in "EMC Unified Storage Best Practices for Performance and Availability Common Platform and Block Storage 31.0 - Applied Best Practices" - revision 23/06/2011 (h8268_VNX_Block_best_practices.pdf), available on EMC Powerlink.

vnx5700 hardware overview

The following documents provide further information on the VNX5700:

- EMC VNX5700 Hardware Information Guide (July 2011);
- EMC VNX Family Data Sheet, H8520.2 (October 2011);
- EMC VNX Series Unified Storage Systems Specification Sheet, H8514.4 (October 2011);
- Global Services Product Support Bulletin VNX5700 and VNX7500 – VNX-PSB-05 (08.03.2011).

With the VNX drive types can be inter-mixed in any DAE. So if you have workloads that don't necessarily conform to the 15 or 25-disk DAE this is no longer a problem as SAS, EFD, or NL-SAS can all be in the same DAE. Note that the new 60-drive DAEs do not support drives with speeds greater than 10K RPM.

While the initial release of Block OE doesn’t allow connectivity to another domain, the latest Block OE 05.31.000.5.502 resolves this issue.

If you are using only 2 backend ports on a VNX5700, use ports 0 and 2, or 1 and 3 for the best performance.

VNX Model	SAS Backend Ports per Storage System
VNX5100	2
VNX5300	2
VNX5500	2
VNX5700	4
VNX7500	4 or 8

Table 1 - SAS Backend Ports

vnx5700 configuration maximums

It should be noted that some of these numbers will change with VNX OE code updates. Here are some useful numbers to know when considering a VNX5700 deployment:

- Maximum no. of disks - 500;
- Maximum RAID Groups - 500;
- Maximum drives per RAID Group - 16;
- Minimum drives per RAID Group - R1/0 - 2, R5 - 3, R6 - 4;
- Stripe Size R1/0 (4+4) and R5 (4+1) - 256KB, R6 (6+2) - 384KB;
- Maximum LUNs (this includes private LUNs) - 4096;
- Maximum LUNs per Pool / all Pools - 2048;
- Maximum LUNs per RAID Group - 256;
- Maximum MetaLUNs per System - 1024;
- Maximum Pool LUN size (thick or thin) - 16TB;
- Maximum traditional LUN size = largest, highest capacity RAID Group;
- Maximum components per MetaLUN - 512;
- EMC still recommends 1 Global Hot Spare per 30 drives.

It is recommended that SP CPU utilization be kept at 50% or less per SP, with bursts of up to 70%. The logic here is that, during an NDU activity, one SP will have sufficient processing to manage the load of the entire array for the time that the peer SP is off-line. If processor utilization is constantly higher than 80% there is a problem.

configuration options

raid groups

When you add drives to a Storage Pool or RAID Group they are zeroed out - this is a background process but can take some time. New drives shipped from EMC are pre-zeroed and won't be "re-zeroed"; the drives you bought off ebay are not. To pre-zero drives prior to adding them to a Storage Pool or RAID Group run the following commands with ***naviseccli***:

```
naviseccli zerodisk -messner <disk-id> <disk-id> <disk-id> start
naviseccli zerodisk -messner <disk-id> <disk-id> <disk-id> status
```

By default, RAID groups should be provisioned in a single DAE. You can theoretically provision across backend ports for increased performance, but oftentimes you'll just end up with crap everywhere. Storage Pools obviously change this, but you still don't want to bind the Private RAID Groups across DAEs. But if you did, for example, want to bind a RAID 1/0 RAID Group across two backend ports - for performance and resiliency - you could do it thusly:

```
naviseccli -h <sp-ip> createrg 77 0_1_0 1_1_0 0_1_1 1_1_1
```

(Where the numbers refer to the standard format **Bus_Enclosure_Disk)**.

RAID Groups and binding between disks on the DPE / DAE-OS and other DAEs. It's a minor point, but something people tend to forget when looking at disk layouts. Ever since the days of Data General, the CLARiiON has used Vault drives in the first shelf. For reasons that are probably already evident, these drives, and the storage processors, are normally protected by a Standby Power Supply (SPS) or two. The SPS provides enough battery power in a power failure scenario such that outstanding writes can be copied to the Vault disks and data won't be lost.

The thing to keep in mind with this, however, is that the other DAEs in the array aren't protected by this SPS. Instead, you plug them into UPS-protected power in your data centre. So when you lose power with those, they go down. This can cause "major dramas" with Background Verify operations when the array is rebooted. This is a sub-optimal situation to be in. As EMC have said for some time, you should bind RAID groups across disks that are either contained in that first DAE, or exclusive to that DAE.
Now, if you really must do it, there are some additional recommendations:

- Don't split RAID 1 groups between the DPE and another DAE;
- For RAID 5, ensure that at least 2 drives are outside the DPE;
- For RAID 6, ensure that at least 3 drives are outside the DPE;
- For RAID 1/0 - don't do it, you'll go blind.

It's a minor design consideration, but something I've witnessed in the field when people have either a) tried to be tricky on smaller systems, or b) have been undersold on their requirements and have needed to be creative. As an aside, it is also recommended that you don't include drives from the DPE / DAE-OS in Storage Pools. This may or may not have an impact on your Pool design.

Finally, you can't defragment RAID 6 RAID Groups - so pay attention when you're putting LUNs in those RAID Groups.

When it comes to fancy RAID Group configurations, EMC recommend that a single DAE should be the default method for RAID Group provisioning. If you use vertical provisioning make sure that: for RAID 5, at least 2 drives per port are in the same DAE; for RAID 6, 3 drives are in the same DAE; and for RAID 1/0, both drives of a mirrored pair are on separate Backend ports. It should be noted that parity RAID Groups of 10 drives or more can benefit from binding across 2 Backend ports - this reduces rebuild times when you pop a disk.

metaluns

MetaLUNs can be quite useful where short-stroking for performance is a requirement. MetaLUNs can't be created in Storage Pools. If you are using pools and require MetaLUN-like functionality, host-based volume striping will be required. When creating multiple MetaLUNs ensure that the primary or first component of the MetaLUN does not share a RAID group with other primary components.

storage pools

Refer to Vijay's blog post (http://virtualeverything.wordpress.com/2011/03/05/emc-storage-pool-deep-dive-design-considerations-caveats/) for a good discussion on storage pool design considerations.

The primary reason for using Storage Pools is to take advantage of FAST Virtual Provisioning (FAST VP). A secondary reason is to provide a simple mechanism for capacity expansion without the need for MetaLUNs.

A Pool LUN's performance will be adversely affected after a trespass. It is recommended that you avoid doing this, except for NDU or break-fix situations.

The maximum number of Storage Pools you can configure is 40. It is recommended that a pool should contain a minimum of 4 private RAID groups.

VNX Model	Maximum Storage Pools per Storage System
VNX5100	10
VNX5300	20
VNX5500	40
VNX5700	
VNX7500	60

Table 2 - Maximum Storage Pools

While it is tempting to just make the whole thing one big pool, you will find that segregating LUNs into different pools may still be useful for FAST cache performance, availability, etc. The mixing of drives with different performance characteristics in a homogenous pool is also contra-indicated. When you create a Storage Pool the following Private RAID Group configurations are considered optimal (depending on the RAID type of the Pool):

- RAID 5 - 4+1
- RAID 1/0 - 4+4
- RAID 6 - 6 + 2

Pay attention to this, because you should always ensure that a Pool's private RAID groups align with traditional RAID Group best practices, while sticking to these numbers. So don't design a 48 spindle RAID 5 Pool. This would be considered non-optimal.

With current revisions of FLARE 30 and 31, data is not re-striped when the Pool is expanded. It's also important to understand that preference is given to using the new capacity rather than the original storage until all drives in the Pool are at the same level of capacity. So if you have data on a 30-spindle Pool, and then add another 15 spindles to the Pool, the data goes to the new spindles first to even up the capacity. For RAID 1/0, avoid private RAID Groups of 2 drives.

A Storage Pool on the VNX5700 can be created with or expanded by 120 drives at a time, and you should keep the increments the same.

VNX Model	Maximum pool drive incremental increases
VNX5100	20
VNX5300	40
VNX5500	80
VNX5700	120
VNX7500	180

Table 3 - Maximum Pool Drive Increases

When FAST VP is working with Pools, remember that you're limited to one type of RAID in a pool. So if you want to get fancy with different RAID Types and tiers, you'll need to consider using additional Pools to accommodate this. It is, however, possible to mix thick and thin LUNs in the same Pool. It's also important to remember that the consumed capacity for Pool LUNs = (User Consumed Capacity * 1.02) + 3GB. This can have an impact as capacity requirements increase.

A LUN's tiering policy can be changed after the initial allocation of the LUN. FAST VP has the following data placement options:

- Lowest;
- Highest;
- Auto;
- No movement.

This can present some problems if you want to create a 3-tier Pool. The only workaround is to create the Pool with 2 tiers and place LUNs at highest and lowest. Then add the third tier and place those highest tier LUNs on the highest tier and change the middle tier LUNs to No Movement. What would be a better solution is to create the Pool with the tiers you want, put all of your LUNs on Auto placement, and let FAST VP sort it out for you. But if you have a lot of LUNs, this can take time.

It is also important to expand a FAST VP pool by the same amount of disks each time. So if you've got a 15-disk pool to start with, you should be expanding a pool with another 15 disks. This can get unwieldy if you have a 60-disk pool and then only need 5 more disks.

While you can have up to 10 DAEs (15 disk DAE) on a bus or a maximum of 250 disks (using the new 25 disk DAE), legacy design best practices with the VNX can still be used: RAID1/0 across two backend ports, and FAST cache provisioned vertically across all available backend ports.

Note that Pool-based LUNs, EFD-based LUNs, FAST VP LUNs, and FAST Cached LUNs do not benefit from file system defragmentation in the way traditional LUNs do. This might require a bit of education on the part of the system administrators.

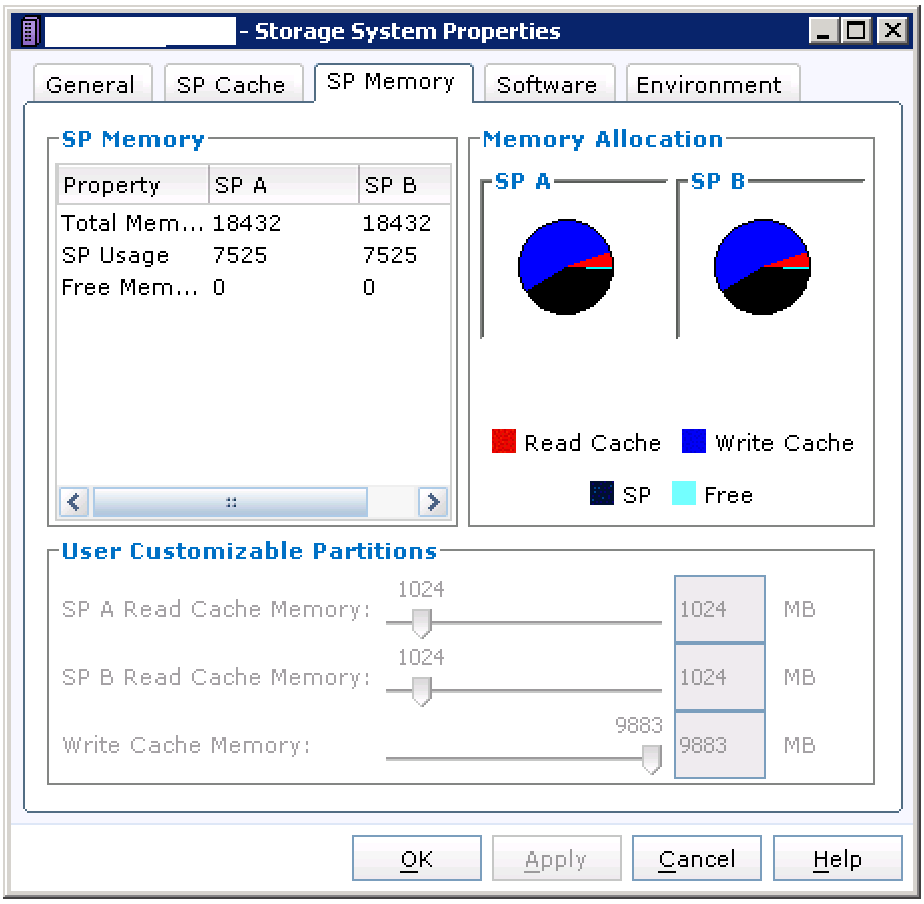
Finally, it should be noted that you can't use Vault drives in a FAST VP pool. I still prefer to not use them for anything.

thin luns and compression

For thin NTFS LUNs - use Microsoft's sdelete to zero free space. When using LUN Compression - Private LUNs (Meta Components, Snapshots, RLP) cannot be compressed. EMC recommends that compression only be used for archival data that is infrequently accessed.

cache configuration

Different models of VNX have different Storage Processor memory configurations. It should be noted that a certain amount of SP memory is also occupied by operating code, so don’t assume that a VNX5700 has a full 18GB per SP available. On the VNX5700, for example, approximately 7.5GB of SP memory is unavailable for Cache use.



SP Memory

Property	SP A	SP B
Total Mem...	18432	18432
SP Usage	7525	7525
Free Mem...	0	0

User Customizable Partitions

SP A Read Cache Memory: 1024 MB

SP B Read Cache Memory: 1024 MB

Write Cache Memory: 9883 MB

Figure 1 - SP Memory Usage

The following table provides figures for the various VNX models and their Cache settings.

Per SP Memory	VNX5100	VNX5300	VNX5500	VNX5700	VNX7500
System Memory (MB)	4000	8000	12000	18000	24000
Maximum Read / Write Cache (MB)	801	3997	6988	10906	14250
Recommended initial Read Cache (MB)	100	400	700	1024	1024

Table 4 - Cache Configuration Maximums

It should also be noted that the recommended Cache watermark levels have changed from the CX4 to the VNX. The following table provides guidance across the various flavours of VNX. It is important to note that these figures should be revisited regularly to ensure performance requirements are being met.

VNX Model	VNX Recommended Write Cache Watermarks	
	High	Low
VNX7500 VNX5700 VNX5500	60	50
VNX5300 VNX5100		40

Table 5 - Recommended Write Cache Watermarks

migration tools and notes

The LUN Migration tool provided by EMC has saved my bacon a number of times. If you need to know how long a LUN migration will take, you can use the following formula:

LUN Migration duration = (Source LUN (GB) * (1/Migration Rate)) + ((Dest LUN Capacity - Source LUN Capacity) * (1/Initialization Rate)).

The Migration rates are - Low = 1.4, Medium = 13, High = 44, ASAP = 85 (in MB/s). Up to 2 ASAP migrations can be performed at the same time per SP. Keep in mind that this will belt the SPs though, so, you know, be careful.

	Maximum LUN Compressions per SP	**Maximum LUN Migrations per Storage System**
VNX5100	N/A	8
VNX5300	5	8
VNX5500	5	16
VNX5700	8	24
VNX7500	10	24

Table 6 - Maximum LUN Migrations

fast cache

EMC recommend that if you're going to spend money on EFDs, you should do it on FAST Cache before making use of the EFD Tier.

When configuring FAST Cache on an array, it is important to locate the primary and secondary drives of the RAID 1 pair on different Backend ports. Ideally this would be done using vertical provisioning. The order the drives are added into FAST Cache is the order in which they are bound. This is most easily done using ***naviseccli***. Run the following switches after the standard **naviseccli -h sp-ip-address**

cache -fast -create - this creates FAST Cache.
cache -fast -destroy - this destroys FAST Cache.
cache -fast -info - this displays FAST Cache information.

When you create FAST Cache, you have the following options:
cache -fast -create -disks disksList [-rtype raidtype] [-mode ro|rw] [-o]

Here is what the options mean:

-disks disksList - You need to specify what disks you're adding, or it no worky. Also, pay close attention to the order in which you bind the disks.
-mode ro|rw - The ro is read only mode and rw is readwrite mode.
-rtype raidtype - I don't know why this is in here, but valid RAID types are disk and r_1.
-o - Just do it and stop asking questions!

naviseccli cache -fast -create -disks 0_1_6 1_1_6 -mode rw -rtype r_1

In this example I've used disks on Bus 0, Enclosure 1, Disk 6 and Bus 1, Enclosure 1, Disk 6.

The disabling of FAST Caching of Private LUNs is recommended (these include the WIL, Clone private LUNs and Reserved LUN Pool LUNs). However, you shouldn't disable FAST Cache for MetaLUN components.

You should consider using FAST Cache under the following circumstances:

- SP utilization is under 70%;
- Regular write cache flushing;
- Majority I/O block under 64K;
- RAID group utilization consistently greater than 70%;
- Majority read versus write I/O;
- High percentage read cache misses;

Higher than acceptable host response time.

disk types

If you're using EFDs for "Tier 0", you'll get good performance with up to 12 EFDs per Backend port. But if you're after the highest throughput, it is recommended that this number be kept to 5.

It is recommended that you use RAID 6 with NL-SAS drives of 1TB or greater. This has some interesting implications for FAST VP heterogeneous Pool configurations and the use of 15 vs. 25-disk DAEs. Refer to Section 0 for further discussion on this.

When architecting for optimal response time, limit throughput to about 70% of the following values:

Drive Type	IOPS
SAS 15K RPM	180
SAS 10K RPM	150
NL-SAS 7.2K RPM	90
EFD	3500

Table 7 - IOPS per Drive type

It is considered prudent to plan for 2/3 of IOPS for normal use - this will give you some margin for burst and degraded mode operation.

EFD can provide performance benefits when:

- Drive utilization >70%;
- Queue length >12;
- Average response times >10ms;
- I/O read-to-write ratio of 60% or greater;
- I/O block-size from 4KB to 16KB.

SAS 10K vs. 15K

- Sequential Reads – either speed consistent performance for 8KB and larger sequential reads and writes
- Random reads – 15K provides for better service time
- Sequential writes – similar performance
- Random writes – 15K slightly better than 10K

NL-SAS

- Sequential reads – approximates 10K SAS
- Random reads – half the performance of 15K SAS
- Sequential writes – comparable but lower than 15K SAS
- Random writes - half the performance of 15K SAS. Decreases with increased queue depth

Vault drives use 192GB per disk for system files. Don’t put Reserved LUN Pool LUNs or Mirrors on there. It is recommended that nothing other than static archive data is stored on Vault drives. Ideally, however, the Vault drives would not be used for user data.

The following table provides recommended IOPS and throughput maximums for any load configured on Vault drives.

Maximum Vault Drive Host I/O Loading		
System Drive Type	**Maximum IOPS**	**Maximum Bandwidth (MBs)**
15K RPM SAS	150	10
10K RPM SAS	100	10

Table 8 - Maximum Vault Drive IOPS

RAID 6, NL-SAS and Disk Layout Options

EMC strongly recommends using RAID 6 when you're using SATA-II / NL-SAS drives that are 1TB or greater. However the current implementation of FAST VP uses Storage Pools that require homogeneous RAID types. So you need multiple tools if you want to run both RAID 1/0 and RAID 6. If you want a pool that can leverage FAST to move slices between EFD, SAS, and NL-SAS, it all needs to be RAID 6. There are a couple of issues with this. Firstly, given the price of EFDs, a RAID 6 (6+2) of EFDs is going to feel like a lot of money down the drain. Secondly, if you stick with the default RAID 6 implementation for Storage Pools, you'll be using 6+2 in the private RAID groups. And then you'll find yourself putting private RAID groups across backend ports. This isn't as big an issue as it was with the CX4, but it is considered non-optimal.

The VNX will create non-standard sized RAID 6 private RAID groups. If you create a pool with 10 spindles in RAID 6, it will create a private RAID groups in an 8+2 configuration. This seems to be the magic number at the moment. If you add 12 disks to the pool it will create 2 4+2 private RAID groups, and if you use 14 disks it will do a 6+2 and a 4+2 RAID group. You can theoretically split the 8+2 across two DAEs in a 5+5. In this fashion, you can increase the rebuild times slightly in the event of a disk failure, and you can also draw some sensible designs that fit well in a traditional, 15-disk DAE. Expanding pools in increments of 10 disks is going to be a pain, particularly for larger Storage Pools, and particularly as there is no re-striping of data done after a pool expansion. The downside to all this, of course, is that you're going to suffer a capacity and, to a lesser extent, performance penalty by using RAID 6 across the board. In this instance you need to consider whether FAST VP is going to give you the edge over split RAID Storage Pools or traditional RAID groups.

The two figures below illustrate both an optimal and non-optimal deployment of a mixed, 32 - 40 disk RAID 6 Storage Pool (using FC and SATA-II). In the first illustration, a standard pool is configured with all 32 disks to begin with. The VNX creates 2 RAID 6 private RAID groups in a 6+2 configuration. However the remaining 4 disks of each type cannot be used, as RAID 6 has a minimum requirement of 6 disks.

The second illustration uses RAID 6 as well, but by creating the Pool in increments of 10, the VNX will create 8+2 private RAID groups instead of 6+2. If this method is used, the remaining 5 slots in each DAE could also be used to add storage to the pool (assuming that host spares are located elsewhere on the array). In addition to this, the parity penalty is not as great per RAID group. In a 40 disk, 2TB RAID 6 pool, the difference in useable capacity is approximately 3.6TB. Additionally, the 5 slots at the end of each DAE can be used for either traditional RAID 5 RAID groups, or RAID 5 Storage Pools using private RAID groups in a 4+1 configuration.

Slot 0	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7	Slot 8	Slot 9	Slot 10	Slot 11	Slot 12	Slot 13	Slot 14	
🟦 3200	🟦 3201	🟦 3202	🟦 3203	🟦 3204	🟦 3205	🟦 3206	🟦 3207	3208	3209	3210	3211	3212	3213	3214	B3E2
🟦 2200	🟦 2201	🟦 2202	🟦 2203	🟦 2204	🟦 2205	🟦 2206	🟦 2207	2208	2209	2210	2211	2212	2213	2214	B2E2
🟩 1200	🟩 1201	🟩 1202	🟩 1203	🟩 1204	🟩 1205	🟩 1206	🟩 1207	1208	1209	1210	1211	1212	1213	1214	B1E2
🟩 0200	🟩 0201	🟩 0202	🟩 0203	🟩 0204	🟩 0205	🟩 0206	🟩 0207	0208	0209	0210	0211	0212	0213	0214	B0E2

Table 9 - Example VNX RAID 6 Configuration 1

Slot 0	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7	Slot 8	Slot 9	Slot 10	Slot 11	Slot 12	Slot 13	Slot 14	
🟦 3200	🟦 3201	🟦 3202	🟦 3203	🟦 3204	🟩 3205	🟩 3206	🟩 3207	🟩 3208	🟩 3209	3210	3211	3212	3213	3214	B3E2
🟦 2200	🟦 2201	🟦 2202	🟦 2203	🟦 2204	🟩 2205	🟩 2206	🟩 2207	🟩 2208	🟩 2209	2210	2211	2212	2213	2214	B2E2
🟩 1200	🟩 1201	🟩 1202	🟩 1203	🟩 1204	🟦 1205	🟦 1206	🟦 1207	🟦 1208	🟦 1209	1210	1211	1212	1213	1214	B1E2
🟩 0200	🟩 0201	🟩 0202	🟩 0203	🟩 0204	🟦 0205	🟦 0206	🟦 0207	🟦 0208	🟦 0209	0210	0211	0212	0213	0214	B0E2

Table 10 - Example VNX RAID 6 Configuration 2

conclusion

There are a few things that you need to be mindful of when designing and deploying the EMC VNX5700 (Block). Hopefully this document provides a solid foundation upon which you can build reliable, available and scalable solutions based on this platform.