EVA Best Practices and Other Goodies

1

EVA Services Information

- Know Your Environment
- Utilize a "Site Management Guide" to contain as much site-specific information as possible

- Hardware Platform
  - Processor Type and Speed
  - Memory size
  - Internal Disk Drive Information
- O.S. version and patch levels
- File System Information
- All HBA Types and Driver versions
- All Network Information
- SAN Switch Information
  - Firmware Revision
  - Port Information

- Storage Array Information
  - Model
  - Subsystem Name
  - WWN
  - Node
  - Firmware Version
  - Disk Group Information
    - Number of disks
    - Raw Capacity
    - RPM
  - Additional information

Contact Information

- "Company Name" Contacts:
  - SAN Administrator - Primary:
    - Office number:
    - Cell number:
    - Email address:
    - Shift
    - Location
  - SAN Administrator - Secondary
  - OS Administrator - Primary
  - OS Administrator - Secondary
  - SAN Architect
  - IT Manager
  - CIO
- HP Contacts
  - Local HP Services Technician – Primary
  - Local HP ASE - Primary
  - Local HP Services Manager
  - HP Gold Support Technical Account Manager
  - HP Services Sales Representative
  - HP Business Critical Systems Sales Representative
  - HP BCS Solutions Architect
  - HP Storage Sales Representative
  - HP Storage Solutions Architect
  - HP Account Executive
  - HP Corporate Business Manager3

Living History

- Service Level Descriptions and Definition
- Change Management
- Trouble History
- Drawing Graphics and Connection Tables
- Call Flow

EVA Best Practice Items

- ***Cost of ownership:*** Do not mix disk sizes in a single disk group.
- ***Cost of ownership:*** To minimize the amount of "reserved" capacity, use a single disk group. But, availability should be considered and is described in the next section.
- ***Cost of ownership:*** Fill the EVA with as many disk drives as possible.
- ***Cost of ownership:*** Use lower performance, larger capacity disks wherever possible.
- ***Availability:*** For critical database applications, consider placing data and log files in separate disk groups. This means that in a majority of situations, two (2), possible three (3) disk groups, depending on the database design, is advisable.

Best Practice Items

- ***VRaid5 availability:***
  - There should be a minimum of 8 shelves in a configuration for VRaid5
  - All disks should be arranged in a vertical fashion, i.e. distribute the disks among the shelves such that the same bay in each shelf has a disk.
  - The total number of disks in a disk group should be an integer multiple of eight.
  - When creating a disk group, let the EVA choose which disks to place in the group.
- ***VRaid1 availability:***
  - If the array contains less than 8 disk shelves, use VRAID1.
  - All disks should be arranged in a vertical fashion, i.e., distribute the disks among the shelves such that the same slot in each shelf has a disk.
  - When creating a disk group, let the EVA choose which disks to place in the group.

Best Practice Items

- ***VRaid0 availability:***

  VRaid0 offers the best performance of all the raid levels. It also supplies the lowest level of data availability and should only be used in instances where possible data loss can be tolerated.

- ***Availability - disk replacement:***

  - Wait for the sparing to be completed. This will be signaled by an entry in the event log (VCS versions 2.002 and above).
  - Remove the failed disk from the shelf and replace with a new one into the same slot as the failed one.
  - Add the new disk into the original disk group.
  - After inserting a disk drive into a shelf, wait 60 seconds before inserting another disk.

Best Practice Items

- ***Performance:*** Fill the EVA with as many disk drives as possible.
- ***Performance:*** Keep the number of disk groups a low a possible.
- ***Performance:*** Using 15K RPM disks is generally best, but carefully consider cost and quantity tradeoffs between 10K and 15K RPM disks.
- ***Performance:*** Under certain, carefully considered circumstances, disabling write cache mirroring will result in significantly increased write performance.
- ***Performance:*** Always leave read caching enabled on a LUN.

Best Practice Items

- ***Performance:*** Always attempt to balance LUNs between the two controllers on an EVA based on the I/O load.
- ***Performance:*** It is acceptable, but not best practice to include drives with different performance characteristics in the same disk group. It is possible that the result might be lower subsystem performance than if they were in separate disk groups.
- ***Performance:*** It is acceptable, but not best practice to include drives with different capacity in the same disk group and could result in higher subsystem performance than if they were in separate disk groups.

9

Summary

	Cost	Availability	Performance
Mixed disk capacities in a disk group	No	No	Acceptable
Number of disk groups	1	1-2-3	1-2
Number of disks in a group	Maximum	Multiple of 8	Maximum
Total number of disks	Maximum	Multiple of 8	Maximum
Higher performance disks	No	-	Probably
Write cache mirroring	-	Yes	Maybe
Mixed disk speeds in a disk group	-	-	Marginal
Read cache	-	-	Enabled
LUN balancing -	-	Yes	Yes

10

EVA Goodies and Gotchas

11

Squeezing More from an EVA

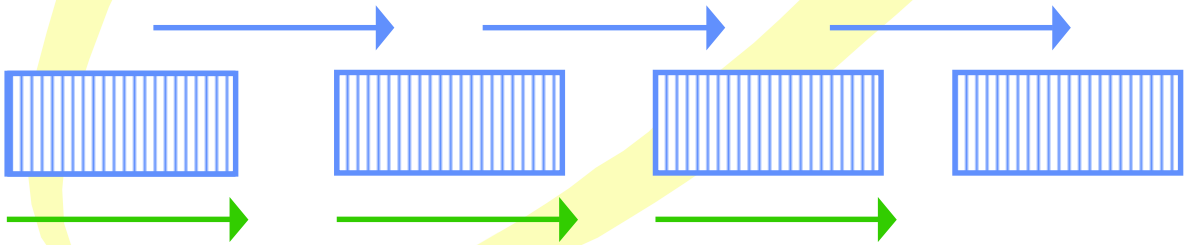
Vraid5 sequential write performance under Windows 2000 can be greatly improved by aligning Partition starting block with the Windows' DISKPAR utility.

-- See CUSTOMER ADVISORY OI040301_CW02:

"Windows applications may experience performance issues with EVA Virtual Disks during a heavy write load"

-- Advisory located at:

http://h20000.www2.hp.com/bizsupport/TechSupport/Document.jsp?objectID=PSD_OI040301_CW02



Squeezing More from an EVA

One person's testing:

- Vraid 1 LUNS are not affected by the offset issue.

- If you try to diskpar a dynamic disk, it is reverted to a basic disk

- Dynamic disk performance also suffers from the offset issue (so don't use dynamic disks)

- Only the first partition on a given LUN must be offset, subsequent partitions are aligned by offseting the first partition

- Sequential write performance deltas:

  -1K I/O's Same

  -2k I/O's 35% increase

  -4k I/O's 55% increase

  -64k I/O's 95% increase

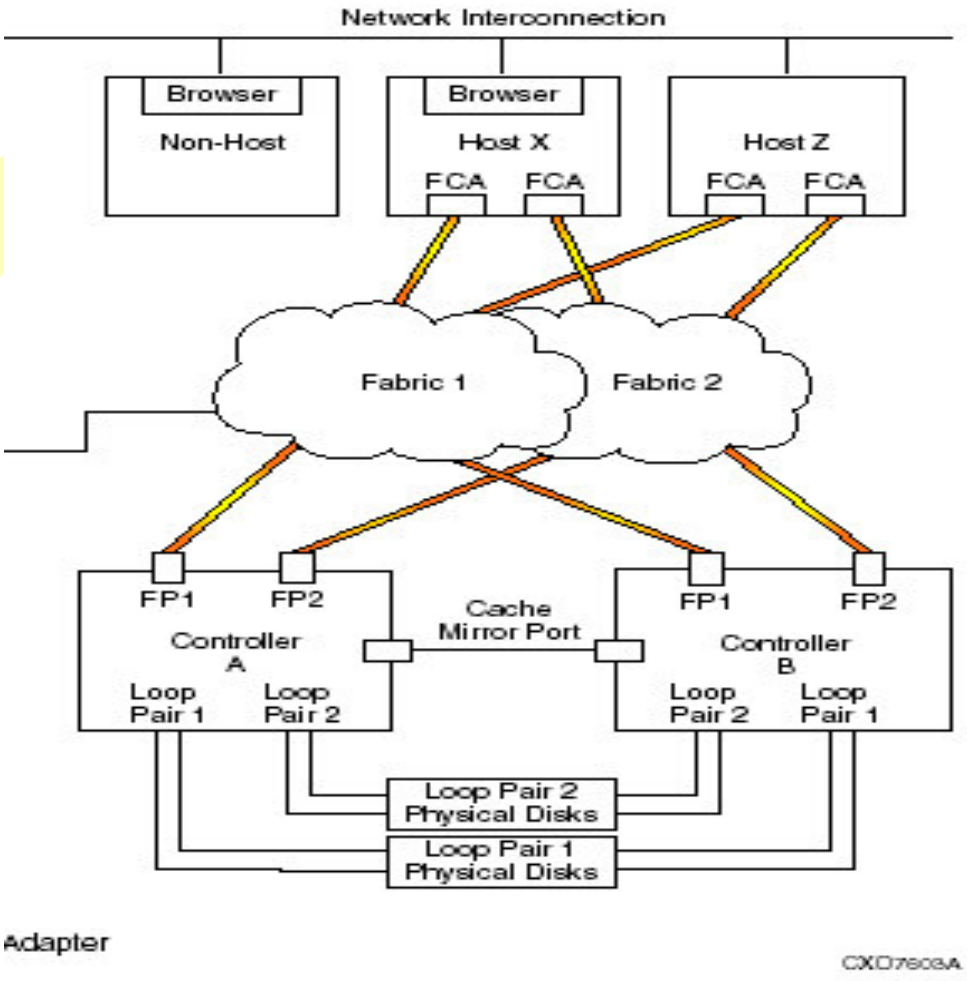
- Random write performance didn't seem to be heavily affected during my limited testing. Interestingly the offset partition held a pretty straight line in perfmon throughout the test while the non-offset partition jumped up and down a lot.

13

Squeezing More from an EVA

Load Balancing

- balancing load across all available paths
- Dividing load across EVA controllers with Secure Path and Command View
- Dividing load across HBAs with Secure Path
- Dividing load across EVA host ports with Secure Path



k diagram of the storage system's connections

Squeezing More from an EVA

Benchmarking tricks and issues

- prewrite data to avoid 1st write penalties
- Avoid reading unwritten blocks unless you want to truly impress your audience
- Increase HBA Queudepths if you can't saturate your driver CPU
- Use oldest Virtual Disks to minimize head movement, if possible
- Make sure the EVA is not doing maintenance such as leveling or zeroing

Squeezing More from an EVA

- Hardware versus software capacities
  - Physical 1000 000 000Bytes = 1GB
  - Software 1073 741 824Bytes = 1.07GB Physical (2^{30})
- System metadata overhead
  - System metadata
  - MLD—HSV Element Manager metadata
  - Virtual Disk metadata
- Vraid overhead
  - Vraid0 - 0% (1 block for every 1 block usable)
  - Vraid1 - 50% (2 blocks for every 1 block usable)
  - Vraid5 - 20% (1.25 blocks for every 1 block usable)
- Snapshot working space
  - Snap — depends on rate of change of original data
  - Snapclone — same physical capacity as virtual disk
- Spare capacity
  - 2 X physical capacity of the largest physical disk X protection selected

~ 7% Variance → 1GB Physical = 0.93GB Software

0.2% system overhead

0.0% -> 50% raid overhead

0.0% -> 100% snap and clone

50% to less than 1% sparing overhead

16

Squeezing More from an EVA

Performance best practice: Using 15K RPM disks is generally best, but carefully consider cost and quantity tradeoffs between 10K and 15K RPM disks.

Disk type	IOPS	Cost	$/IOP	Disk count at $100K budget	Total IOPS at $100K budget
72 GB 10K RPM	110	$2,300	$20.91	43	4,730
72 GB 15K RPM	150	$3,400	$22.67	29	4,350
Improvement	15K = 36%				10K = 9%
Best choice	**15K**	**10K**	**10K**		**10K**

Table 1 – Disk RPM Tradeoffs

Squeezing More from an EVA

Growing file systems

	File System Type / Version	Comments
Windows 2000	NTFS/ Basic Disks	SVG or Diskpart
Windows 2003	NTFS/ Basic Disks	SVG or Diskpart
Redhat		
SuSe		
Solaris	UFS with growfs/ requires dismount	VxFs with Veritas VxVM 4.0 unsupport
AIX		
OVMS	Files-11, V7.3-2	$set volume/limit
Tru64	?	
HP-UX	Lun concatenation only	

18

When Good Hardware Fails

- Always set-up monitoring and event forwarding (ISEE/PRS)

19

When Replicating

- Undo your careful balancing act
  - DR Groups must be accessed from a single HBA
  - By implication – no Secure Path HBA load balancing
- Try to keep your inter-site load balanced across ISL's
  - Use a combination of preference settings in Secure Path and Command View on the source and target to try to get traffic flowing across both links
- Don't saturate your links
- Keep your distance to a minimum
- Avoid applications that require more than 8 luns in a consistency group (DR Group)
- Monitor Design and Operations Guide for updates

20

Leveling for the EVA

What is "Leveling"?

- The process used by the EVA to distribute physical storage among the disks in an array.
- Its purpose is to distribute the physical allocation of storage for the collection of logical disks created by the user, such that the usage for a given logical disk on a given physical volume is proportional to the contribution of that physical volume to the total amount of physical storage available for allocation to a given logical disk.
- In other words, if a given physical volume contributes 10% of the total storage, 10% of each logical disk will be allocated on that volume.

Leveling for the EVA

EVA Storage Allocation - In order to understand leveling, it is necessary to understand some basic storage space allocation concepts.

- Each physical volume (disk) is segmented into small units of storage called PSEGs (Physical SEGment).
- Physical volumes are grouped by the controller and/or management agent to form a Logical Disk Allocation Domain (LDAD).
- An LDAD is also known as a Disk Group in the Management Agent.
- Each logical disk created by the user is associated with a particular LDAD, and the physical storage of a logical disk is chosen only from the set of physical volumes in that LDAD.
- Each LDAD can be further subdivided by the array controller into Redundant Storage Sets (RSS) in order to improve the fault tolerance characteristics of the LDAD.
- An RSS may be a subset of the volumes that make up an LDAD, depending on the size of an LDAD.
- Another way of stating this is that an LDAD may contain one or more RSSs.
- RSSs are further subdivided into allocation units called RSTOREs.
- RSTOREs are assigned and allocated based on the RAID type of the logical disk being created.
- The operation of moving disks from one RSS to another is referred to as a RSS migration, and is typically done as the result of a split or merge.
- If an RSS drops below a minimum membership, it is typically merged with other disks to keep the RSS greater or equal to its minimum and less than its maximum.
- If an RSS expands to its maximum limit, it is split into two RSSs.

22

Leveling for the EVA

Leveling Methodolgy – to ensure that the physical storage used for a given Logical Disk is allocated proportionally across the RSSs in the LDAD containing the Logical Disk, as well as proportionally across the volumes within each RSS.

In other words, if a given RSS contains 15% of the storage in the LDAD, 15% of the Logical Disk will be allocated from that RSS and if a given volume in the RSS contains 10% of the storage within the RSS, 10% of the Logical Disk space allocated in that RSS will be allocated on that volume.

- Differing RAID levels will have different calculations for physical storage.
- For example, in VRAID1, since a pair of drives is necessary for full redundancy, the smaller of the two drives is used to compute the available physical storage.

Leveling for the EVA

Leveling Triggers:

- The triggers for a leveling event are divided into three groups.
  - The first group occurs in response to a configuration change due to either drive failure or drive appearance. These may be due to hardware errors or user action. The second group occurs inresponse to an action by the user, and the last group occurs as a result of restart ormaster controller failover.
- **First Group:**
  - When an RSS migration request is made (includes marry operations)
  - When a pending RSS migration request is sent to the level wait queue
  - When an RSS migration completes
  - When the number of normal members in an LDAD transitions above a minimum
  - Following a successful merge (joining of two RSSs) operation
  - Following any reconstruction (data recovery through redundancy) or when a failed drive is replaced (successful or not)

Leveling for the EVA

Leveling Triggers:

- **Second Group:**
  - When a snapclone completes
  - Following a capacity change
  - Following a snap creation, if leveling was in progress at the start of the snap creation
  - In VCS v2.003 and beyond, via a maintenance command
  - In VCS v3.x and beyond, when a logical disk is deleted
- **Last Group:**
  - Following a restart or master transition (controller reset, or mastership moving from one controller to the other)

Leveling for the EVA

Should the Customer Worry About Leveling?

- Leveling is a consequence of both normal and abnormal events that occur in the life of a cell.
- Most of the events that trigger the leveling process are not within the customer's control.
- The one event that is within the customer's control is when disks are inserted. In this case, HP suggests that all disks are added at the same time (one at a time with a few seconds delay in between drive insertions), unless the array is at or near maximum capacity. If the latter scenario is the case, then only two disks should be added initially and leveling should be allowed to finish. After leveling is complete, the rest of the disks should be added.
- If the user has a choice it would also be a good practice to add the disks in a given disk group such that they are all of same size as the existing drives in that disk group. The user should never worry about leveling. It may, however, be useful to be aware of it, as there are side effects such as temporary loss of capacity while leveling is in progress. Users can also run into problems when the amount of free capacity is so low that leveling cannot complete, because there is no space in which to move the data. By following the suggested guidelines around capacity and always ensuring at least single protection, these problems should be eliminated.

26

Leveling for the EVA

Reducing the Impacts of Leveling:

- In general, to reduce the impacts of leveling, the customer should always be at the latest version of EVA firmware, set a minimum protection of "single" on all disk groups, and ensure used physical capacity is at, or below 95%.
- Large disk groups may tolerate an occupancy alarm that is greater than 95%. An example would be a disk group containing 100 disks.
- It is also important to remember that occupancy is only a warning and not a guarantee, and should be monitored closely. As this space is managed closer to its limit, monitoring the free space closely becomes more critical. Under these circumstances the impact of leveling is minimal upon the array, except in the case of a multiple disk failure or when multiple drives are removed or added simultaneously from an already existing disk group that has been written to its near full capacity. Note that in this case if there are no logical disks in the existing LDAD, there will not be any leveling. Finally, leveling is a background task and is designed to minimize impact to the customer's workload.


hp
invent