



Universal Storage Platform V Architecture and Performance Guidelines

A White Paper

By

Alan Benway

Senior Performance Consultant
Performance Measurements Group

*Hitachi Data Systems
Technical Operations
Santa Clara CA 5050*

January 7, 2008 (Update 9a)

Copyright 2008 Hitachi Data Systems Corporation, ALL RIGHTS RESERVED

**Hitachi Data Systems Confidential
For Internal Use and Distribution to Authorized Parties Only**

Table of Contents

<i>I. Introduction</i>	2
<i>II. Overview of Changes</i>	3
Software	3
Hardware	4
<i>III. Dynamic Provisioning</i>	5
Traditional Host-based Volume Management	6
Traditional USP Storage-based Volume Management	7
Dynamic Provisioning Volume Management	9
HDP Pools	10
V-VOL Groups and DP Volumes	12
Pool Page Details	15
Pool Expansion	18
Miscellaneous HDP Details	19
HDP and Program Products Compatibility	20
USP V Volume Flexibility Example	20
<i>IV. USP V Architecture</i>	21
Processor Upgrade	22
Overview of Features	22
Logic Box Layout	23
Front-End-Directors	24
FED FC-16 port Feature	24
FED FC-8 port, ESCON, FICON Features	25
Back-end Directors	25
BED Feature	26
USP V: HDU and BED Associations by Frame	26
USP: HDU and BED Associations by Frame	27
HDU Switched Loop Details	29
Disk Details	30
<i>V. USP V FED Details</i>	34
MP Distributed I/O	34
Open Fibre 8-Port Feature	35
Open Fibre 16-Port Feature	36
ESCON 8-port Feature	37
FICON 8-port Feature	38
<i>VI. USP V BED Disk Controller Details</i>	39
Back-End Director Details	39
Back-End RAID level Organization	40

<i>VII. USP V Product Examples</i>	<i>42</i>
Small Configuration.....	42
Midsize Configuration	43
Large Configuration	44
<i>VIII. Special Performance Tests</i>	<i>45</i>
Database Test Comparison	45
DP-VOL Performance	45
<i>IX. RAID10 (2D+2D) Back-end Scalability Tests</i>	<i>46</i>
Test Methodology	46
Random I/O Transactions	47
Sequential I/O Throughput	47
<i>X. RAID5 (3D+1P) Back-end Scalability Tests</i>	<i>51</i>
Test Methodology	51
Random I/O Transactions	52
Sequential I/O Throughput	52
<i>XI. RAID5 (7D+1P) Back-end Scalability Tests.....</i>	<i>56</i>
Test Methodology	56
Random I/O Transactions	57
Sequential I/O Throughput	57
<i>XII. RAID6 (6D+2P) Back-end Scalability Tests</i>	<i>61</i>
Test Methodology	61
Random I/O Transactions	62
Sequential I/O Throughput	62
<i>XIII. RAID10 (2D+2D) AMS500 Virtualization Tests</i>	<i>66</i>
Universal Volume Manager	66
Test Overview	67
Test Methodology	67
Random Workloads	68
Sequential Workloads.....	68
<i>XIV. RAID10 (2D+2D) Single Array Group LUN scalability tests.</i>	<i>70</i>
Test Methodology	70
Random Workloads	70
Sequential Workloads.....	75
<i>XV. RAID5 (3D+1P) Single Array Group LUN scalability tests</i>	<i>77</i>

Test Methodology	77
Random Workloads	77
Sequential Workloads.....	82
XV. RAID10 (2D+2D) HDP Comparison Tests.....	84
Test Descriptions	84
CASE 1	84
CASE 2	85
CASE 3	86
CASE 4	86
Test Methodology	87
Test Results.....	88
Random Workloads Comparisons	88
Sequential Workloads Comparisons	92
XVI. Summary.....	95
APPENDIX 1. USP V (Frames, HDUs, and Parity Groups).....	96
APPENDIX 2. USP (Frames, HDUs, and Parity Groups).	97
APPENDIX 3. Open Systems RAID Mechanisms.....	98
APPENDIX 4. Mainframe 3390x and Open-x RAID Mechanisms.....	101
APPENDIX 5. Concatenated Array Groups.....	105
APPENDIX 6. BED: Loop Maps with Parity Group Names	107
APPENDIX 7. FED: 8-Port FC Maps with Processors	108
APPENDIX 8. FED: 16-Port FC Maps with Processors	109
APPENDIX 9. FED: 8-Port FICON Maps with Processors.....	110
APPENDIX 10. FED: 8-Port ESCON Maps with Processors	111
APPENDIX 11. Pool sizing example	112
APPENDIX 12. Veritas™ Volume Manager details.....	114
APPENDIX 13. Pool details	115
APPENDIX 14. Pool example	116
APPENDIX 15. Disk IOPS Examples	117

CONTRIBUTIONS

Gil Rangel

Director of Performance Measurement Group

Ron-an Lee

Performance Consultant, PMG, HDP Tests

James Byun

Performance Consultant, PMG, RAID tests

Michael Netsvetayev

Performance Consultant, PMG, AMS500 tests

Jean-Francois Masse (HDP Pool sizing example)

Solutions Architect, Solutions Development, GSS

Larry Korbus (HDP details)

Director, Product Management (HDP, UVM, VPM, HTSM)

I. Introduction

This document covers the architectural and software features of the Universal Storage Platform V (**USP V**). These features are covered in detail in the following sections. This document is not intended to cover any aspects of program products, databases, customer specific environments, or new features available by the second general release. Areas not covered by this document include:

- TrueCopy/Shadow Image/Universal Replicator – Disaster Recovery Solutions
- Host - Logical Volume Management – General guidelines
- HDLM – General guidelines
- Universal Volume Management (UVM) – virtualization of external storage for Data Lifecycle Management (DLM)
- Virtual Partition Management (VPM) – General guidelines for workload management
- Oracle – General guidelines for storage configuration
- Microsoft Exchange – General guidelines for storage configuration

This document is intended to familiarize Hitachi Data Systems' sales personnel, technical support staff, customers, and value-added resellers with the new software and architectural features of the USP V. The users that will benefit from this document are those who already possess an in depth knowledge of the TagmaStore USP architecture and performance capabilities.

Throughout this paper the terminology used by Hitachi Data Systems - not the factory - will be used. For instance, Hitachi calls this new system "RAID 600" while HDS has renamed it "USP V". Some other storage terminology is used differently in the documentation or by users in the field as well. Here are some definitions as used in this paper:

- **Array Group** – the set of 1 or 2 Parity Groups used to form a single Array Group. When 2 Parity Groups are used (as RAID10 4D+4D, RAID5 7D+1P, RAID6 6D+2P), there is a specific, fixed, location association between the 2 Parity Groups used. **Note:** Although Storage Navigator and the SVP GUI imply that a RAID10 4D+4D is a concatenation of two 2D+2D Parity Groups, it is not. There is a single RAID stripe across 4 pairs of mirrored disks.
- **BED** – Back-end Director package; the pair of Disk Adapter (DKA) PCBs with 8 pairs of FC loops used to attach 128/192 disks to the storage system.
- **Concatenated Array Groups** – a logical set of 2 or 4 RAID5 7D+1P Array Groups using interleaved RAID stripes. Sometimes called "VDEV Disperse". See Appendix 5 for more details.
- **DP-Vol** – a Dynamic Provisioning (DP) Volume; the Virtual Volume from an HDP Pool. Some documents refer to this as a **V-VOL**. It is a member of a V-VOL Group.
- **FED** – Front-end Director package; the pair of Channel Adapter (CHA) PCBs used to attach hosts to the storage system, providing Open FC, FICON, or ESCON attachment.
- **HDU (Hard Disk Unit)** – the 64 disk container with 32 disks on the front side and 32 more on the back. It is further split into left and right halves, each on separate power domains. There are up to four HDUs per expansion frame, and one in the control frame.

- **LDEV** (Logical Device) – a volume created from the space within a particular Array Group. It is a partition from that Array Group’s raw space.
- **LUN** (Logical Unit Number) – the host-visible identifier assigned by the user to an LDEV when it is made visible to a host port.
- **LUSE** (Logical Unit Size Expansion) – A *concatenation* of 2 to 32 LDEVs (up to a 60TB limit) that are then presented to a host as a single LUN.
- **Parity Group** – the 4-disk upgrade package that goes into a specific set of HDU slots and then takes on a fixed name (such as Parity Group 10-2). This term is often used in the place of *Array Group*, although they aren’t the same thing. Some people refer to this as a “disk cluster” or “disk feature”.
- **V-VOL Group** – the container which is specified to be of a user selectable size between 8GB and 2TB. It is connected to an HDP Pool. One or more DP-Vols are configured within this space, using a specified maximum size. The total of these maximum sizes must be less than the associated V-VOL Group.

II. Overview of Changes

Expanding on the proven and superior USP technology, the USP V offers a new level of Enterprise Storage, capable of meeting the most demanding of workloads while maintaining great flexibility. The USP V offers higher performance, higher reliability, and greater flexibility than any competitive offering in existence today. These are the new products and features that distinguish the USP V from the USP:

- **Software**
 - Hitachi Dynamic Provisioning (**HDP**) volume management feature.
- **Hardware**
 - Enhanced Shared Memory system (up to 32GB and 256 paths @ 150MB/s).
 - Faster 800MHz RISC processors on FEDs (Channel Adapters) and BEDs (Disk Adapters).
 - Faster BED 4Gbit/s back-end disk loops.
 - Switched Loop back-end.
 - Half-sized PCBs used now (except for Shared Memory PCBs), allowing for more flexible configuration choices.
 - Support for internal 750GB 7200 RPM SATA HDDs

Software

The USP V software includes **Hitachi Dynamic Provisioning**, a major new Open Systems volume management feature that will allow storage managers and system administrators to more efficiently plan and allocate storage to users or applications. This new feature provides for two new capabilities: *thin provisioning* and *enhanced volume performance*. HDP provides for the creation of one or more HDP Pools of physical space (multiple LDEVs from multiple Array Groups of the same disk types and RAID level), and for the establishment of DP volumes (virtual volumes) that are connected to individual HDP Pools.

Thin provisioning comes from the creation of DP volumes of a user-specified *logical* size without any corresponding *physical* space. Actual physical space (as 42MB Pool pages) is only

assigned to a DP volume by the HDP software from the connected HDP Pool as that volume's logical space is written to over time. A volume does not have any Pool pages assigned to it when it is first created. Technically, it never does – the pages are “loaned out” from its connected Pool to that DP volume until the volume is reformatted and deleted. At that point all of that volume's assigned pages are returned to their original slots (by releasing the pointers) in the Pool's Free Page List.

The **volume performance** feature is an automatic result from the manner in which the individual HDP Pools are created. A Pool is created using 1-256 LDEVs (Pool Volumes) that provide the physical space, and the Pool's 42MB allocation pages are assigned on demand to any of the DP volumes connected to that Pool. Each individual 42MB Pool page is consecutively laid down on a whole number of RAID stripes from one Pool Volume. Other pages assigned over time to that DP volume will randomly originate from the next free Page from other Pool Volumes in that Pool.

As an example, assume that there are 24 LDEVs from 12 RAID10 (2D+2D) Array Groups assigned to an HDP Pool. All 48 disks will contribute their IOPS and throughput power to all of the DP volumes assigned to that Pool. If more random read IOPS horsepower was desired for that Pool, then it could have been created with 32 LDEVs from 16 RAID5 (7D+1P) Array Groups, thus providing 128 disks of IOPS power to that Pool.

As up to 256 such LDEVs may be assigned to a single Pool, this could represent a considerable amount of I/O power under (possibly) just a few DP volumes. This type of aggregation of disks was only possible previously by the use of expensive and somewhat complex host-based volume managers (such as VERITAS VxVM) on the servers. On the USP (and the USP V), the only alternative would be to build Striped Array Groups (“Concatenated” Array Groups) using two or four RAID5 7D+1P Array Groups. This would provide either 16 or 32 disks under the volumes (VDEVs) created there. There is also the LUSE option, but that is merely a simple concatenation of LDEVs, a capacity (not performance) configuration.

Hardware

While there are many physically apparent changes to the USP V chassis and PCB cards from the previous USP model, there are also a number of not-so-evident internal configuration changes that an SE must be aware of when laying out a system. The USP and the USP V are both a collection of frames, HDUs, PCB cards in a Logic Box, and disks (see Figure 1). The frames include the control frame (DKC) and the disk expansion frames (DKUs). Disks are added to the 64-disk HDU containers (up to 18) in sets of four (Parity Group). Parity Groups are installed (following a certain upgrade order) into specific HDU slots on either the left or right half (such as HDU-1R) and front and rear of an HDU box to work with a particular backend controller.

Logic boards are installed in the front and rear slots in the Logic Box in the Control Frame. The logic board types (for both USP and USP V) include these features (or options, each a pair of PCBs):

- **CSWs** – Cache Switch
- **CMAs** – Cache Memory

- **SMA**s – Shared Memory
- **FED**s – Front-end Directors (or Channel Adapters)
- **BED**s – Back-end Directors (or Disk Adapters)

The USP V's new **half-sized PCBs** (now with upper and lower slots in the Logic Box, described later) allow for a less costly, more incremental expansion of a system. For example, there were typically 4-6 FED features installed in a USP600, and they could be a mixture of Open Fibre, ESCON, FICON, and iSCSI. However, this gave you a large number of ports of a single type that you may not need, with a substantial reduction of other port types that you may need to maximize. With the new half-sized cards, you can have 8 FED features (or up to 14 FEDs at the expense of 6 BED features), using any mixture of the interface types as before. Now, as there are half as many ports per board, smaller numbers of lesser used port types may be installed. Features are still installed as pairs of PCB cards just as with the USP.

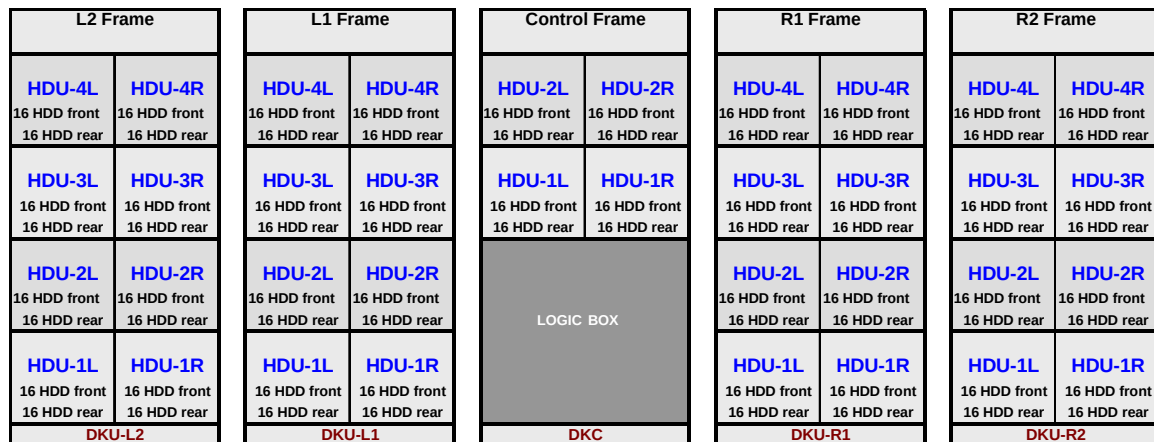


Figure 1. USP and USP V Frames and HDUs.

Other physical USP V changes include:

- The associations between some BEDs and HDUs have changed
- The associations between FEDs, BEDs, and the CSWs are different.
- The locations of 50% of the Parity Group names have shifted around.
- It takes two BED features (4 PCBs) to support the 8-disk Array Groups (just as was the case on the Lightning models). Hence, BEDs must be installed as pairs of features.

A discussion of each of the hardware features (with a comparison to the USP) will follow in Sections IV-VII below.

III. Dynamic Provisioning

This section will begin with a review of the techniques typically used to manage storage volumes. Some of these methods are host-based, while others use existing HDS Enterprise storage system capabilities such as *LUSE* or *Concatenated Array Groups*. The end of this section

will examine the overall details of the new Hitachi Dynamic Provisioning volume management feature and explain how it works.

Traditional Host-based Volume Management

There are two methods typically used to organize storage space on a server today. One method would be the direct use of physical volumes as devices for use either as raw space or as file systems (local or clustered). These are fixed size volumes with a fixed number of disks, and as such each has a certain inherent physical random IOPS or sequential throughput (MB/s) capacity. A system administrator must manage the aggregate server workloads against them. This is shown in Figure 2. As workloads exceed the volume's available space or its IOPS capacity, the user contents must be manually moved onto a larger or faster (more spindles) volume if possible.

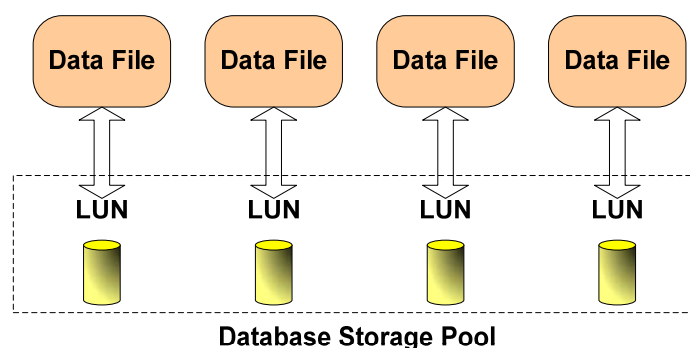


Figure 2. Use of individual volumes on a host.

The alternative method is to use a host-based **Logical Volume Manager (LVM)** when the planned workloads require either more space or IOPS capacity than the individual physical volumes can provide. The LVM can create a single logical (virtual) volume from two or more independent physical volumes. This is shown in Figure 3.

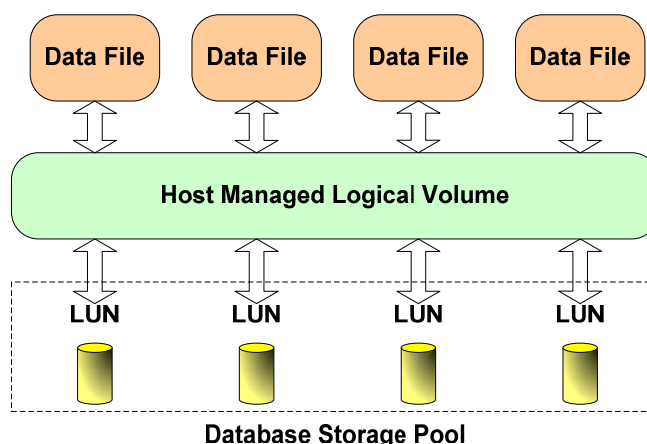


Figure 3. Use of LUNs in host managed logical volumes.

When such a logical volume runs out of space or IOPS capacity, one could replace it with one that was created with even more physical volumes and then copy all of the user data. In some cases it is best to add a second logical volume and then manually relocate just part of the existing

data in order to redistribute the workload across two such volumes. These two logical volumes would probably be mapped to the server using separate host paths as well. However, this manual intervention could become a costly and tedious exercise as workloads grow over time or as new ones are added to the mix. The problem becomes more complex as new servers are added to the scenario.

Traditional USP Storage-based Volume Management

In order to create volumes on an HDS enterprise storage system, an administrator will first create Array Groups from one or two of the 4-disk sets (Parity Groups) on a USP product. These Array Groups could be any mix of five types of RAID levels, to include: RAID10 (2D+2D, 4D+4D), RAID5 (3D+1P, 7D+1P) and RAID6 (6D+2P). The next step would be to create volumes (LDEVs) of the desired size from these individual Array Groups using Open-V (or Open-X, but this isn't recommended) emulation on Open Systems, and 3390-X emulation on a Mainframe. These "Standard" LDEVs will then be individually mapped to one or more host ports as a LUN (an LDEV on a Mainframe). This standard method is shown below in Figure 4.

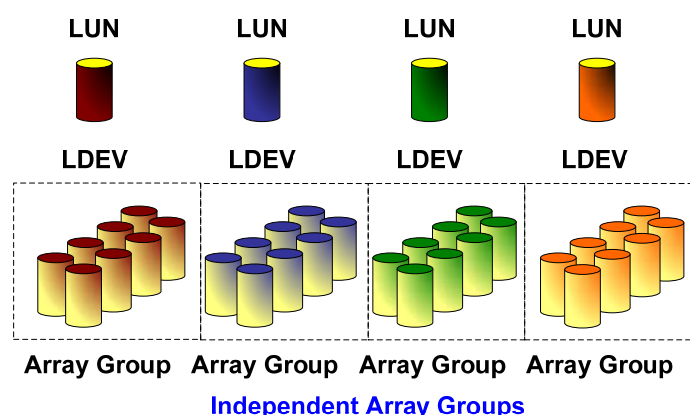


Figure 4. Standard volumes.

In order to change the size of a volume (LDEV) already in use, one must first create a new one (if possible) that was larger than the old one, and then move the contents of the current volume onto the new one. The new volume would then be remapped on the server to take the mount point of the old one, which is then retired.

On a Lightning product, the volume sizes were fixed at a small number of Open-X choices. If you needed something bigger than the largest emulation size (Open-L, at 33GB), the storage system-based solution was to use the **LUSE (Logical Unit Size Expansion)** LDEV concatenation feature (Figure 5). Note that as LUSE is a concatenation of 2-32 LDEVs (up to a 60TB maximum size), it normally operates with just the IOPS power of those disks under a single LDEV (either 4 or 8 disks). On the other hand, the host-based RAID0 striped volume will use the power of all of the disks of the logical volume all of the time.

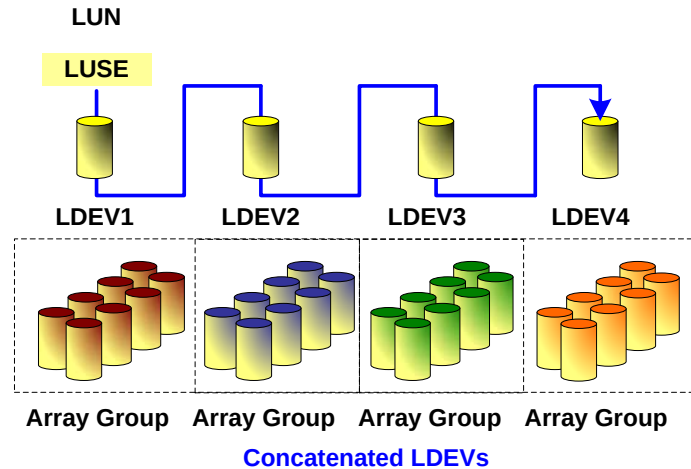


Figure 5. LUSE volume.

On a USP product, with its Open-V emulation, a single LDEV could be as large as the entire Array Group (up to 2TB). However a single Array Group might not have enough space or sufficient IOPS power (disks) to support all of the capacity or IOPS power needed by the workloads. You could build a concatenated LUSE volume as with the Lightning, but the use of the USP's large volume facility known as **Concatenated Array Groups** is preferred. This feature allows you to create a logical volume that is striped across 2 or 4 RAID5 7D+1P Array Groups. The individual RAID stripes from those member Array Groups are interleaved rather than simply concatenated as the name implies. Thus, one could have 2x or 4x amounts of total capacity and IOPS per volume. Note that a host-based LVM uses block striping (usually 64KB) whereas a Concatenated Array Group uses RAID stripe interleaving (of 3,584MB, using 256KB + 256KB RAID chunks × 7 data disks). See Appendix 3 for a detailed discussion of RAID levels, stripe sizes (chunks), and stripe widths. The Hitachi method of using two 256KB chunks from the same disk before switching to the next disk in a RAID stripe is explained there. Also, see Appendix 5 for a more detailed discussion of *Concatenated Array Groups*.

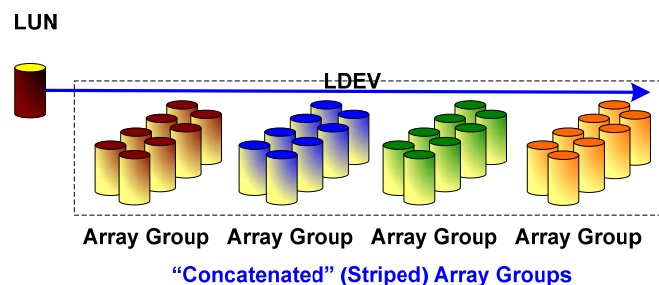


Figure 6. Concatenated Array Group striped volume.

All of these techniques can be awkward and time consuming to configure. As such, there is a lot of potential for error in a busy IT shop. Now there is a new and simpler alternative to the older methods. The **Hitachi Dynamic Provisioning** feature (**HDP**) allows for new volume management techniques that provide a greater degree of freedom of storage management for administrators.

Dynamic Provisioning Volume Management

The new HDP feature allows for the creation of one or more HDP Pools of physical space (multiple LDEVs from multiple Array Groups of any RAID level) to support the creation of DP volumes. There are three distinct and unique capabilities that become available to the user with this feature. These are:

Thin provisioning volumes – a feature where physical space is only allocated to a DP volume as it is used over time. The DP volume may be specified to be any size when created, and this is the size the host sees when scanning the port for this physical volume. The DP volume is allocated physical space in 42MB pages from a Pool as the host writes to this volume. As an example, this is a good method to manage lots of small user accounts and avoid dedicating fixed pools of space to each account up front.

High performance volumes – the ability to create Pools with the RAID level desired and the number of disks needed to support the target IOPS rates. With the use of the Concatenated Array Group feature, one could create volumes that spanned either 16 or 32 disks, but they had to be RAID5. Now, a wide variety of choices exists for establishing the physical space under DP volumes. One might create a Log Pool of 16 LDEVs taken from four 2D+2D RAID10 Array Groups (four LDEVs each) and use this Pool for DP-VOLs assigned to database logs or temp space. One might also then create four Database Pools, each with 32 LDEVs from eight RAID5 7D+1P Array Groups (four LDEVs each), and use these for the database files that contain data. The Log Pool was configured to accept high levels of large block sequential write workloads, whereas the Database Pools are configured to support high levels of random small block (IOPS) workloads.

Management simplification – though related to and embedded within both of the above topics, this requires some special mention as a feature in and of itself. Because of this new ability to create large and/or high performance volumes, a considerable amount of often complicated host-based storage management can be avoided. Rather than use a Logical Volume Manager on each server to organize volume space (using 2 or more LDEVs as previously described), in many cases one can now rely on just the HDP feature for this function.

One example of simplification would be the Log Pool example from above. Another example would be the ability to now manage users and space quite easily in a Microsoft Exchange Server environment. Since only four Storage Pools are available in Exchange (i.e. four volumes), it is necessary to carefully balance the user types (regular or Blackberry) among these four volumes because of their significant workload differences and the IOPS limitations of a standard 4-disk or 8-disk Array Group. Now, in addition to the use of one Standard LDEV of RAID10 for the logs, there could be four DP-VOLs created from one or two Pools for these Storage Groups. A storage administrator is in now control of designing the desired physical IOPS rates available per HDP Pool. As such, Exchange users may now be added to any Storage Group in a simple round-robin method without being concerned with what type of user they may be and how many IOPS they may consume.

In order to use Dynamic Provisioning volumes, there are a few extra steps to take when creating them. One still configures various Array Groups to the desired RAID level, and then creates one or more Open-V volumes (LDEVs) on each of them. The first step in setting up an HDP environment would be to create one or more **HDP Pools** that are each collection of some of these internal LDEVs (**HDP Pool Volumes**). The second step is to create one or more **HDP Volumes (DP-VOLs)** that are connected to a single Pool. The DP-VOL is then mapped to a host port as a LUN and used just like any Standard LDEV.

Figure 7 below is an overview of a Pool environment used for discussion purposes in the following sections. See Appendix 12 and 13 for additional views.

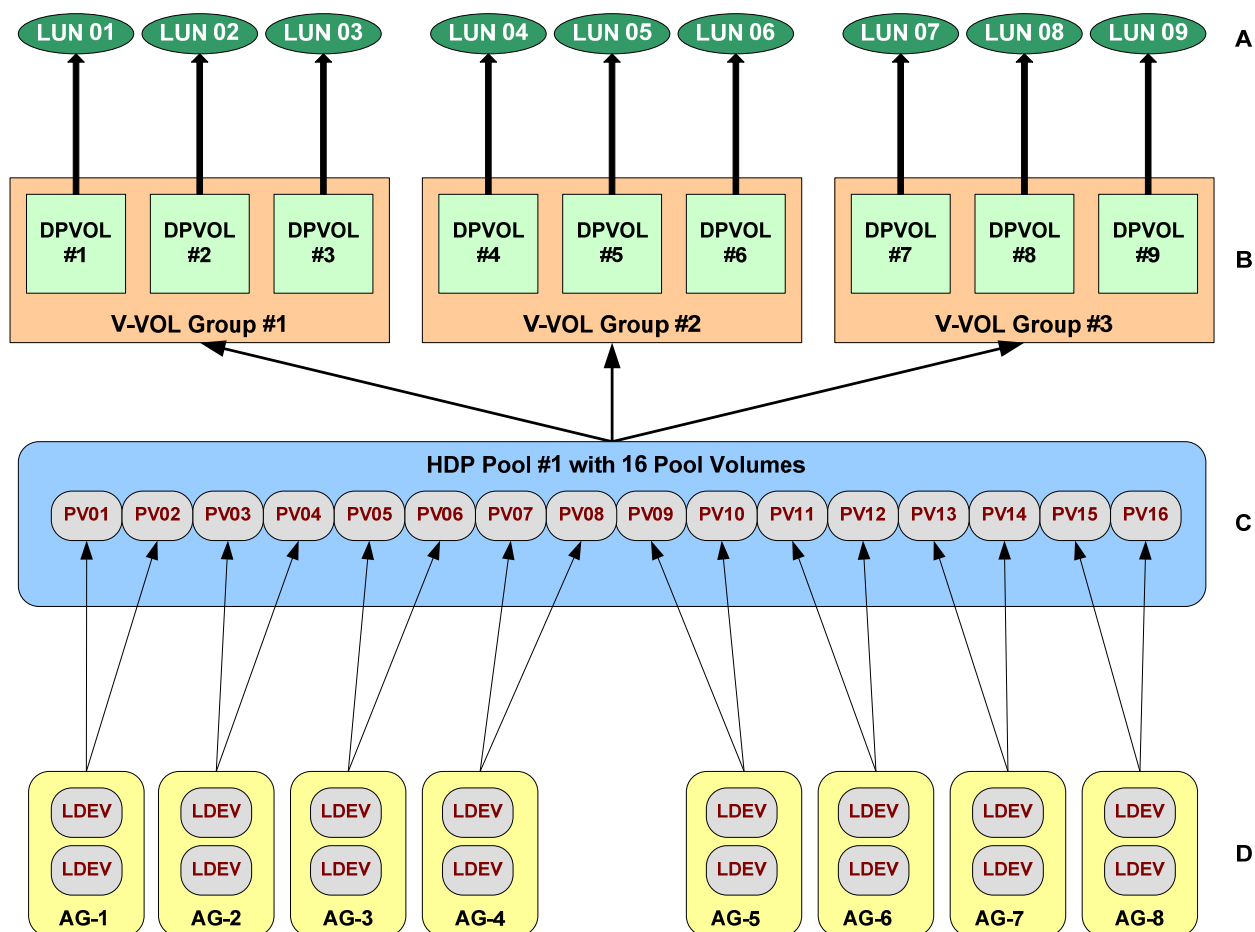


Figure 7. An HDP Pool environment.

HDP Pools

Each HDP Pool (up to 32 Pools per USP V system) provides the physical space for those DP-VOLs which are connected to that Pool. Each HDP Pool is given a name which is used when connecting newly defined DP-VOLs. The LDEVs assigned to a Pool must be of type OPEN-V. The size of each LDEV must be at least 8GB, but less than 2TB. The size used for LDEVs should be a multiple of 42MB, as Pool space is allocated and managed in that size. Once assigned to a Pool, these LDEVs are called **Pool Volumes**. A Pool's size must fall between 8GB

(1 Pool Volume, minimum size) and 286TB in size, although the aggregate size of all Pools (and all non-HDP volumes) on a USP V must be less than 286TB.

The Pool Volume sizes are important in that they control the striping rate in the Free Page List (at the “Optimize” step in Pool setup). If all Pool Volumes are the same size, and if they come from the same type of Array Groups (i.e. 2+2 or 7+1), then the Pool will present an even distribution of I/O across the AGs in that Pool to all of its connected DP-VOLs. The actual size (if uniform for all of the LDEVs) does not matter except that it will determine how soon the Pool runs out of initial physical space. You must scale your usage expectations against how big the Pool needs be when it first starts out. Additional Pool Volumes may be added to a Pool when its space begins to run low (explained in *Pool Expansion* below), with an overall limit of 256 LDEVs per Pool.

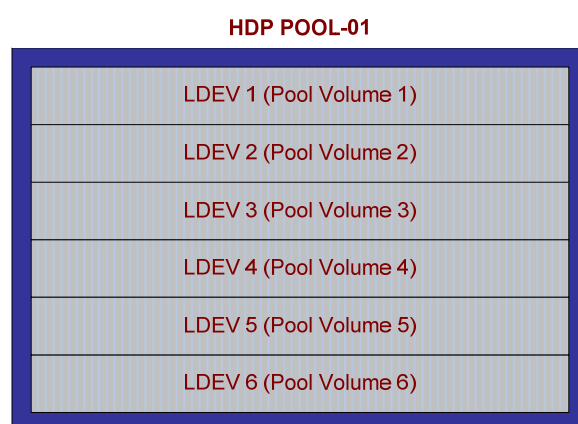


Figure 8. An HDP Pool with six Pool Volumes.

You should use the same disk type and RAID level for every Pool Volume within a Pool. Additionally, all of the LDEVs from an Array Group should be assigned to the same Pool. It is recommended that there be 32 or more disks per Pool in order to have an adequate disk IOPS pool. This could be four 8-disk Array Groups (such as 4D+4D, with four LDEVs each) or eight 4-disk Array Groups (such as 3D+1P, with two LDEVs each). There ought to be 16 LDEVs (or more) per Pool so that there will be a sufficient degree of “stride” (the Pool’s interleaving of small sets of pages per LDEV) by the initialization of the Free Page List (all of this described below). If using a smaller number of Array Groups in a Pool, when adding their LDEVs to the Pool one needs to manually alternate the selection among the LDEVs from each Array Group. This will disperse the Free Page Table entries more evenly across the individual Array Groups in that Pool.

There are various restrictions on the LDEVs used for Pools. These restrictions include:

- There can be a maximum of 256 Pool Volumes assigned to a single HDP Pool.
- Pool Volumes (LDEVs) must come from internal disks, not from external storage. They must all come from one USP V (can’t create Pools across systems).
- Pool Volumes must be of the OPEN-V emulation type.
- A Pool Volume is an LDEV, but it is a normal LDEV created from (or allocated from) an internal RAID Group, so its CU:LDEV address is assigned before the

volume is defined as an HDP pool volume; in the future HDP will support a Pool made up of external LDEVs (or E-LDEVs).

- Once an LDEV becomes a Pool Volume (assigned to a given Pool), there are a wide variety of restrictions that are invoked in terms of what you can and cannot do with it, particularly in terms of replication functions.
- You cannot remove a Pool Volume from an HDP Pool without deleting the entire Pool (first formatting and disconnecting all of its DP-VOLs volumes) from the USP V system.
- An LDEV may not be uninstalled from the system if it is a Pool Volume.
- Each LDEV must be RAID formatted (scrubbed with zeros) before being assigned to a Pool (you can't format a Pool).
- A LUSE volume can not be used as a Pool Volume.
- If used, all Array Groups that are members of a Pool must be assigned to the same cache partition (CLPR). Hence, all associated DP-VOLs will also be assigned to the same VPM cache partition. However, the Array Groups for other Pools may be assigned to different cache partitions.

V-VOL Groups and DP Volumes

There are both **V-VOL Groups** (virtual containers) and the **DP-VOLs** (virtual volumes, also known as **V-VOLs**). The DP-VOLs are seen as real volumes by the hosts. There may be up to 4,096 DP-VOLs created per Pool. DP-VOLs share the USP V system-wide addressability limit of 65,536 volumes with external volumes, Shadow Image volumes, Universal Replicator volumes, and Pool Volumes. The size of each DP-VOL may be between 46MB and 2TB in size.

Once a Pool has been created, one or more V-VOL Groups are then configured. A V-VOL Group is a virtual substitution for an Array Group. When a V-VOL Group is created and connected to an HDP Pool (a V-VOL Group can not span Pools), a user selectable Group size between 46MB and 2TB is specified. The next step is to create one or more virtual DP volumes within the V-VOL Group container. The sum of the sizes of these DP volumes created must be less than or equal to the size of that V-VOL Group. Each DP-VOL is then mapped to a host port as a LUN (just like a standard LDEV). The host will then detect the LUN's size as the virtual size **specified** for that DP-VOL in the V-VOL Group. The sizes specified for the V-VOL Group and each of its DP-VOLs are permanent. Figure 8 shows four DP volumes, two V-VOL Group containers, the associated Pool with 1152GB of capacity, and the four Array Group LDEVs assigned as Pool Volumes (288GB each) to that Pool.

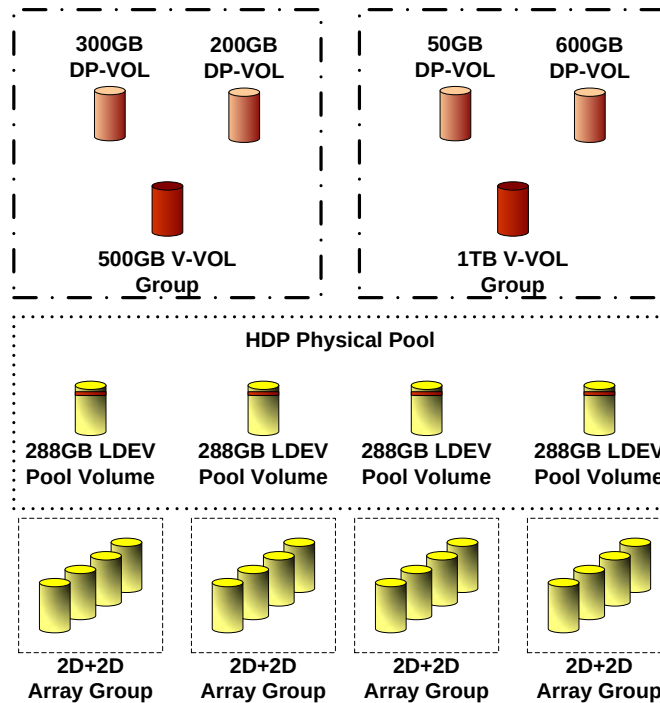


Figure 9. HDP: Association of a DP-VOL to Array Groups.

The **physical space** assigned as needed from the Pool to a V-VOL Group container is evenly distributed across the Pool Volumes by the mechanism of the Pool's Free Page List as described in the sub-section *Pool Pages* below. As a server writes to a DP-VOL, and new pages are required, these are assigned to the V-VOL Group for use by that DP-VOL. Note that this assignment is from the next available page in the Free Page List, not from just anywhere in the Free Page List. It is random in the sense that you do not know from which Pool Volume the next page for an individual DP-VOL will be mapped.

However, the server sees each DP-VOL as having the complete **SCSI Logical Block Address (LBA)** range of 512 byte physical blocks that would match the logical size that was specified when that DP-VOL was created. This is what thin provisioning is all about – space allocated as needed. A DP-VOL is assigned 42MB pages of physical space from the Pool over time as the server writes to blocks in that volume.

Once assigned to a V-VOL Group, a page is mapped into the associated **DP volume Page Table**. This is a table of pointers matching up the host's virtual LBA range and physical 512 byte blocks with that DP volume's assigned pages. The Page Table presents a contiguous space of virtual and assigned 42MB segments to that volume starting at beginning of that DP-VOL's address space (byte 0). A DP-VOL whose size is not a whole number of 42MB pages can acquire a page that will be only partially used at the highest address range for that DP-VOL. The remainder of that page is unused (but this is a very small amount of space).

If a DP-VOL was specified to be a 100GB volume, that would represent 209,715,200 such SCSI blocks sitting on 2,439 42MB Pool pages. Each assigned Pool page will contain a specific (predetermined) range of consecutive 86,016 512 byte blocks. This means that one page in a DP-

VOL will contain LBA-0 to LBA-86015; some other page will contain LBA-86016 to LBA-172031, and so forth. Whether a physical page has been assigned is dependent on whether an LBA within that fixed range has been written to by the host. For example, if the server wrote 8KB file system blocks to a new DP-VOL at LBA locations 0, 86016, 172032, 258048, and 344064 then that DP-VOL would have five new Pool pages assigned. If the server then wrote to LBA locations 16, 86032, 172048, 258064, and 344080 (or such) then those same five pages would be used.

The number of Pool pages assigned to a DP-VOL by file system formatting will depend upon the operating system, the file system, and the size of the volume. Once the server formats a volume, the Pool will have assigned some number of physical pages to it. The amount of the physical space initially required by this formatting operation is determined by the amount of metadata the file system initially writes to that volume, along with the distribution (stride) of these blocks of metadata across the entire volume. Hence, the number of pages initially assigned to volume by the formatting operation will vary.

If, for example, a DP-VOL is formatted by a file system, and the file system's block size is 8KB, there will be 5,376 file system blocks per page (each 8KB block consuming 16 LBA blocks). The first time one of these blocks from a DP-VOL virtual page is written to, the following sequence occurs:

- The next available free Pool page will be assigned from the Pool's Free Page List.
- That page will be mapped into that DP-VOL's Page Table and associated with a certain LBA range.
- The new block(s) will be written to that page.

Usage Mechanisms

If a host **Read** is posted to an area of the DP-VOL that does not yet have physical space (no blocks have been written there yet), the HDP Dynamic Mapping Table will direct the read to a special "Null Page". New Pool pages to cover that region will not be allocated from the Free Page List since it was a Read.

If a **Write** is posted to an area of the DP-VOL which does not yet have physical space, the DP-VOL requests a new page and assigns it to that LBA range in the Page Table.

If a **Write** is posted to an area of the DP-VOL which does not yet have physical space, and the Pool has completely run out of space, the host will get a "Write Protect" error. During this "out of space" condition, all **Read** operations specifying an LBA region of a DP-VOL without an assigned page will now return an "Illegal Request" error rather than the "Null Page". These error conditions will be cleared once additional physical Pool Volume space has been added to that Pool or some DP-VOLs are formatted and disconnected from the Pool, thereby freeing up all of their previously allocated pages.

There will normally be several DP-VOLs connected to a single HDP Pool. As each DP-VOL begins receiving writes from the servers, the number of assigned Pool pages to that DP-VOL will grow. As this group of DP-VOLs consumes Pool Pages over time, the Pool's **Free Space**

Threshold can be reached and the storage administrator will receive an alert telling him to add additional Pool Volumes to that Pool. The amount of physical space a DP-VOL can acquire over time will be limited by the DP-VOL size (up to 2TB) when it was defined. The sum of all the virtual sizes of those DP-VOLs connected to a Pool, rounded up to an integral number of pages, will be the maximum physical Pool size required over time.

DP-VOL Restrictions

There are various restrictions regarding DP-VOLs. These include:

- A DP-VOL may only be connected to one Pool.
- A DP-VOL must be formatted before being deleted. This clears stale data from the Pages before they are released back to the Page Free List. Pages are only released from a DP-VOL when the association to the HDP Pool is deleted. Since these Pages were merely mapped from the Pool to the DP-VOL's local Page Table, these *N* freed Pages "return" to the same slot they originated from, since they didn't actually go anywhere. They are then marked as free in that Pool's Free Page List and they will be reassigned to DP-VOLs by the next *N* requests for Pages.
- There can be up to 4,096 DP-VOLs assigned to an HDP Pool.
- A DP-VOL can range from 46MB to 2TB in size.
- In terms of usage, a DP-VOL is pretty much like a standard LDEV with a few restrictions.
 1. A DP-VOL has a normal CU:LDEV number but this is in a special type of Array Group.
 2. Using the usual HDS tools you can: use DRU to protect a DP-VOL, and create a ShadowImage pair from it.
 3. These are the things you can't do with a DP-VOL: use cache residency, create a TrueCopy, HUR or COW pair from it, or use volume migration on it.
 4. A DP-VOL also differs from a normal LDEV in that you can't resize or sub-divide it using VLVI or glue them together with LUSE.

Pool Page Details

The physical Pool space is assigned over time as needed to the DP-VOLs that are **connected** to that Pool through their V-VOL Group. Space is assigned from the Pool in 42MB allocation units called **Pool pages**. All Pool pages are managed by the HDP software in a **Dynamic Mapping Table** (DMT) in Shared Memory. All metadata for the HDP Pools and their DP-VOLs is maintained in the DMT, which resides in the 12GB-14GB address space of Shared Memory (Shared Memory Module #4). One type of element in the DMT is a **Free Page List** that is maintained for each Pool for all of their 42MB pages.

The **Pool's Free Page List** is a table of pointers to all of the pages located on its set of Pool Volumes as well as some status information. When a DP-VOL needs a new physical page, it is assigned the next available page from this Free Page List. The structure of the Free Page List consists of a matrix of 128 columns by *N* rows of 42MB pages. The number of rows is

determined by how many total pages the Pool contains divided by 128. The concept of *thin provisioned capacity* involves the page capacity actually assigned to a DP-VOL – which grows over time - and not the reported size of the volume from the host’s perspective.

Each **DP-VOL** has a table for keeping track of pointers to its assigned Pool pages. This is a list of all assigned Pool pages and where to find them on the Pool Volumes. Each Page Table is also a lookup table of assigned physical pages to the server’s virtual address space (the LBA range of all the 512 byte blocks in that volume). Note that once a Pool page is assigned to a DP-VOL, it remains there regardless of the state of the data written to blocks on that page. If all data that happens to reside on that page is deleted by the host file system, leaving the page “empty”, it remains under the ownership of that DP-VOL. As described previously, each page – once assigned to the volume – becomes a piece of the physical space under a specific LBA range for the server. Once allocated, the page assignment is permanent until that DP-VOL is reformatted and deleted. This reformatting deletes all data on all of that DP-VOL’s pages so that they are clean when returned to the Pool for reuse. Recall that all LDEVs were first formatted (scrubbed with zeros) before being added to a Pool, and as such were in a “clean” state.

There are some **housekeeping pages** from each Pool that are not available for assignment to DP-VOLs. There is a small 84MB system region (2 pages) in a Pool for each Pool Volume. There is also a small 4GB (98 page) region in every HDP Pool that is reserved by the system to use for a disk mirror copy of the DMT in Shared Memory. For USP V release G01, with the setting of a flag in the SVP, the DMT is copied to the SVP’s internal disk on a system power down (PS-OFF). For G02, the DMT is planned to be mirrored to this reserved 4GB region in the first HDP Pool currently defined in the system.

There is a simple **structure** for Pool pages. An individual 42MB page is laid down on a whole number of consecutive RAID stripes from a single Pool Volume. The Table below illustrates the relationship among different RAID levels, their various RAID stripe sizes, and the assignment of consecutive stripes to a single page.

RAID type	RAID level	Data disks	RAID stripe size (KB)	Number of RAID stripes per HDP page
RAID10	2+2	2	1024	42
RAID10	4+4	4	2048	21
RAID5	3+1	3	1536	28
RAID5	7+1	7	3584	12
RAID6	6+2	6	3072	14

Table 1. Number of RAID stripes per 42MB Pool Page.

Take, for example, the case of an HDP Pool that is composed of several RAID5 7D+1P Pool Volumes (LDEVs). Each LDEV has a formatted RAID stripe width of 3,584KB. This comes from two 256KB consecutive chunks per disk, with (effectively) seven data disks in the stripe (256+256 x 7). [See Appendix 3 for the various RAID level details.] Therefore, twelve such RAID stripes from a single 7D+1P LDEV will be used to create a single 42MB Pool Page. In the case of a Pool consisting of RAID10 2D+2D Array Groups, each LDEV will provide 42 consecutive stripes for each 42MB Pool page.

Pool Page Organization Details

Consider Pool-01 with six Pool Volumes. When Pool-01 was first created, it was a single concatenated space of 42MB pages starting with LDEV01 then working down through LDEV02 to LDEV06. This initial organization is shown in Figure 9.

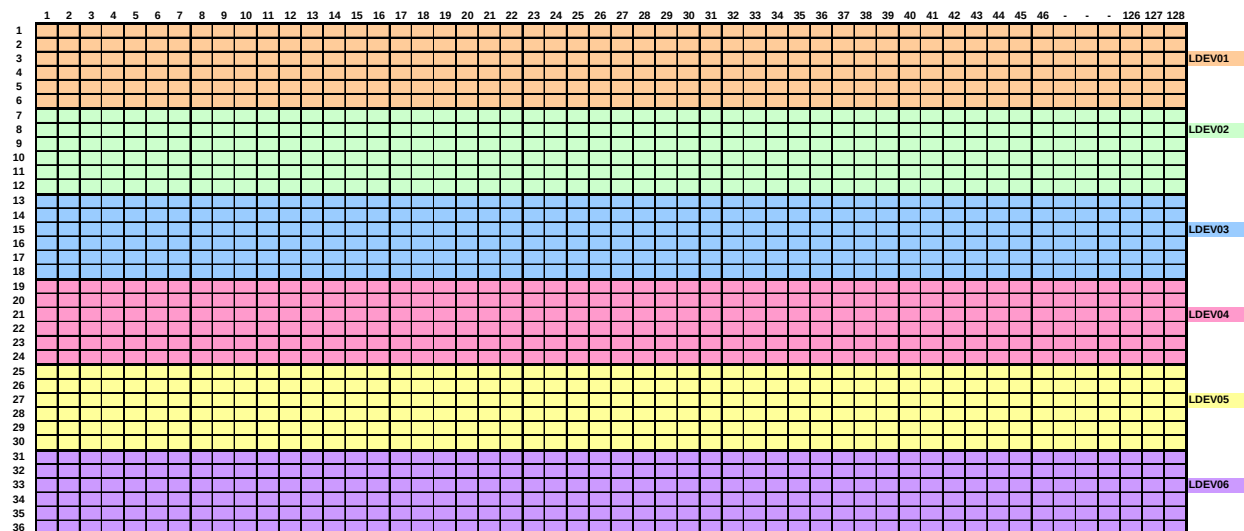


Figure 10. Free Pool pages from 6 LDEVs at Pool creation.

When Pool-01 was "Optimized" in the Storage Navigator GUI (a very important but optional step), this linear space was mapped into *striped* sets of 42MB pages across the entire Pool. This mapping is how the Free Page List is created. Depending on how many Pool Volumes there are in a Pool and how big they are, there will be some number of contiguous pages **per LDEV** inserted into the Free Page List before switching the selection to the next LDEV in the Pool. This distribution (referred to as a stride in this paper) ends up generally looking like the map for Pool-01 as shown in Figure 10. One can visualize this stride as a RAID0 stripe of "super-chunks" across the set of LDEVs, where each super-chunk is a set of N consecutive 42MB pages per LDEV (the stride). The stride value is uniform for the Pool if all LDEVs are the same size. While this is not required, it does affect uniformity of performance as discussed below.

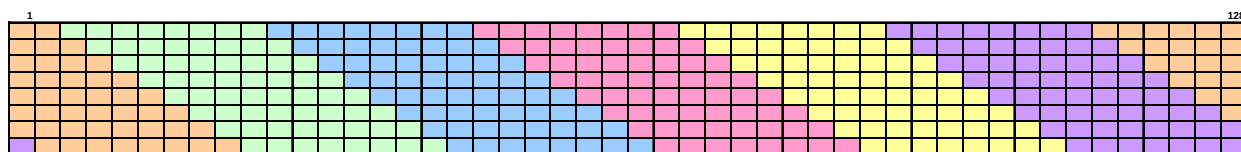


Figure 11. Optimized Free Page List for Pool-01's 6 Pool Volumes.

For a new Pool, if all of the LDEVs are the same size, the stride is essentially determined by dividing 128 (columns) by the number of LDEVs in that Pool. The specific stride calculation is per LDEV and uses the formula $128 * (LDEV \text{ pages} / Pool \text{ Pages})$. Notice that when the LDEVs are of different sizes they will each get a different stride factor. So, if there are 16 equal sized LDEVs in a Pool, the stride will be 8 consecutive 42MB pages per LDEV. The "Optimize" algorithm will create an offset value for the first stride so that a "slanted" distribution will occur in the table. This introduces a randomizing element to the assignment of pages by LDEV to the

DP-VOLs. Otherwise this map would appear to be vertical colored stripes of pages (by LDEV) rather than the dispersal of the slanted stripes.

Returning to the Pool-01 example (Figures 9 and 10), assume that each of the six LDEVs contains 3,000 42MB pages. Thus, the Pool contains 18,000 pages. In this example, the “stride” of consecutive pages per LDEV would be $128 \div 6 \text{ (LDEVs)} = 21 \text{ pages (whole number only)}$ per LDEV. It follows that there will be $18000 \div 128 = 141 \text{ rows}$ in the table (of academic interest only). In this example, the Free Page List will begin with fewer than 21 pages from LDEV01 (the offset factor), then the mapping mechanism will select 21 pages from LDEV02, followed by 21 pages from LDEV03 and so forth, until that row of 128 pages is complete. This pattern continues on the next row by finishing the incomplete 21 page sequence from the previous row, then continuing the 21 page mapping sequence from the next LDEV. The result is a “slanted” striping effect across the Pool Volumes. The number of initial pages mapped at the very beginning (automatically determined by the HDP software) determines the degree of slant.

If the Pool contained LDEVs of various sizes, the selection process to build the Free Page List for that Pool would be different. Assume that the six LDEVs in Pool-01 had sizes of 1000, 2000, 3000, 4000, 5000, and 6000 pages for a total of 21,000 pages (hopefully an unlikely scenario). Using the mapping formula above, they would individually have strides of 6, 12, 18, 24, 30, and 36 pages. If each LDEV came from a different RAID10 2D+2D Array Group, you can see how this imbalance would skew the I/O onto some of the LDEVs. If they were all the same size, the I/O load would be even across the LDEVs.

Figure 11 illustrates a Pool that was initially created with 12 LDEVs, which causes the number of consecutive pages (the stride) from each LDEV to be smaller. Here, with 12 LDEVs of 3,000 pages each (providing 36,000 Pool Pages), the stride would be 10 pages per LDEV. There would be 282 rows in the table. In this Figure, many of the pages are shown in gray to represent pages that have already been assigned over time to the DP-VOLs connected to that Pool.

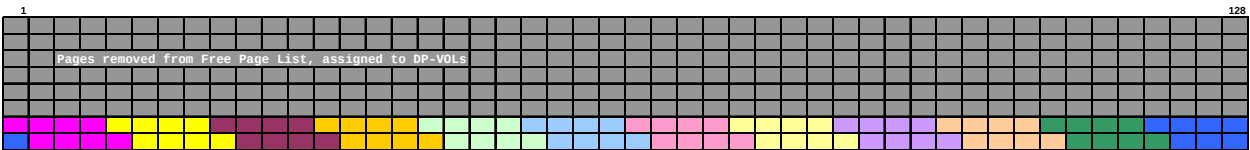


Figure 12. Optimized Free Page List for a 12 LDEV Pool with pages assigned to DP-VOLs.

Pool Expansion

As Pages in the Pool are assigned to DP-VOLs and the number of free Pages drops below a threshold of the initial Pool capacity, more LDEVs will have to be added to that Pool. Once new LDEV(s) are added to the Pool and the Optimize button in Storage Navigator is used, the Free Page List will once again be optimized. The existing Free Page List will be extended (more rows) using pages available in the additional Pool Volumes. The number of consecutive pages from the new Pool volumes will be calculated as described above. Therefore, when new Pool Volumes are added once the remaining Free Pages are low, the HDS recommendation is to add another set of LDEVs that match the initial set (same RAID level, same disk RPM). This will ensure that the performance remains on par with what was previously experienced for that Pool.

For example, a Pool that started with six LDEVs will have a certain performance due to its Pages being distributed by the Free Page List across its Pool Volumes. Assuming that six RAID10 2D+2D Array Groups were used for these six LDEVs, the Pool would have started out with 24 disks of IOPS power (6 times 2D+2D LDEVs). Each individual page would have the IOPS power of the underlying LDEV (such as 2D+2D), but on average the set of pages mapped to a DP-VOL would include pages from all Pool Volumes, thus seeing the 24-disk IOPS power.

Suppose that the initial set of free pages (from the six-LDEV Pool) become depleted (below 20% of the Pool) and two new LDEVs are then added to the Pool (each from different RAID10 2D+2D Array Groups). Once the last 20% of original pages are assigned from Pool's Free Page List, the extended section of the Free Page List will begin assigning new pages from a Pool which now only has an overall IOPS power of 8 disks. There will be a noticeable drop in IOPS rates from a 24 disk Pool as IO to new pages is clustered onto the new 8 disks added to the pool.

The following may be a way to describe this concept in a 10 second "techie spot":

A new Pool is created from a set of LDEVs taken from like Array Groups. The more Array Groups in use, the higher the IOPS performance that is available to that Pool. This space is then "Optimized", where the Free Pool Page List (using 42MB pages) is created in Shared Memory. This list is built in small groups of pages from across the LDEVs, each LDEV contributing its share of free pages to each run of 128 free pages. These pages are then assigned on a first-come first-serve basis as needed to DP-VOLs connected to that Pool.

When new LDEVs are added to the Pool once the space becomes depleted, the same process is followed. A complementary number of LDEVs from enough Array Groups should be added to maintain the IOPS performance level of the Pool.

Miscellaneous HDP Details

The configuration of HDP Pools and Volumes is done via Storage Navigator. This is also how the physical space usage per Pool is monitored. There are two Pool Free Space thresholds: a **user specified threshold** between 5%-95% page utilization and a **system threshold** at **80%** utilization. Both of these thresholds are used as the triggers for the low Pool space notifications.

If more than 2TB of space is required for a single host volume, then multiple DP-VOLs may be used in the ordinary fashion by a logical volume manager on the server to create a single large striped Logical Volume. Note that if these DP-VOLs come from different Pools, then the active disk count will increase as well as the amount of space. If the DP-VOLs are all from the same Pool, then only the usable capacity will increase.

The use of HDP may cause a brief 10-20% increase in port processor (FED MP) utilization for individual reads or writes to the DP-VOL area when physical pages are already allocated from the Pool for that address space. When a Write to a DP-VOL causes a physical Page allocation from the Pool, there may be a brief 80% utilization increase (overhead factor) on the FED MP for that one operation.

HDP and Program Products Compatibility

A DP-VOL (DP-VOL) cannot be used with the TrueCopy, TrueCopy Asynchronous, Universal Replicator, Copy-on-Write Snapshot, Flash Access, LUSE or Volume Migration products. DP-VOLs do work with Shadow Image, LUN Security, CVS, Server Priority Manager, Data Retention, and Virtual Partition Manager.

USP V Volume Flexibility Example

If you have a customer using Exchange 2003, they are limited to the use of four “Storage Groups” (LUNs). A host volume manager cannot be used with Exchange. Due to the way that the underlying Jet database operates, one must use RAID10 volumes to avoid the insert response time issues that a well known characteristic of Jet. Here are some alternative methods on a USP V to provide these four Storage Group LUNs to an Exchange system:

STANDARD VOLUMES: If using RAID10 4+4, this method would provide 32 disks when using four standard LDEVs.

LUSE VOLUMES: If using four LUSE volumes – where each of them has four LDEVs from four 4+4 Array Groups (thus 16 Array Groups), this would provide 128 disks overall. However, you will normally only see the IO power of one LDEV per LUSE volume (8 disks each, or 32 overall) due to the simple concatenation nature of LUSE (fill one LDEV, then start on the next one).

STRIPED (CONCATENATED) ARRAY GROUPS: If using this method with four VDEV volumes, where each was based on four R5 7+1 Array Groups, there would be 128 disks overall. However, due to the nature of Jet, one would not want to use a RAID5 solution. Note that this 'native' interleaved (not concatenated) stripe solution is not available from other vendors.

HDP #1: If using this method, there could be 24 LDEVs from 6 R10 4+4 Array Groups put into the Pool (48 disks in all), and then create four DP-VOLs connected to this Pool. Each DP-VOL would be used as an Exchange Storage Group over a dedicated 4Gbit path. This very large "native" IOPS solution is not available from any other vendor.

HDP #2: If using this method, there could be 32 LDEVs from 32 R10 4+4 Array Groups put into two Pools (256 disks in all), with four DP-VOLs (two per Pool) being created. Each DP-VOL would be used as an Exchange Storage Group. This huge "native" IOPS solution is not available from any other vendor. However, note that the use of only four 4Gbit host ports from a single Exchange server would not be able to effectively use this huge reservoir of random IOPS capacity.

IV. USP V Architecture

At the core of the USP V is the third generation, non-blocking, switched architecture. The Universal Star Network is a fully fault tolerant, high performance, non-blocking, switched architecture. There are no specific models as with the USP products. There is a single base model onto which upgrade features are applied to meet the needs of the customer.

The **data Cache** system has the same path speeds (1064 MB/s) and counts (64 paths), with a peak bandwidth of 68 GB/s. The **Shared Memory** system has been significantly upgraded over the USP version, with 256 (up from 192) paths operating at 150MB/s (up from 83MB/s), with a peak bandwidth of 38.4 GB/s (up from 15.9 GB/s). On the USP, the FED PCBs had 8 Shared Memory paths, and the BEDs had 16. Now, all PCBs have 16 Shared Memory paths.

The Cache memory only contains user data blocks, whereas the Shared Memory system holds all of the metadata about the internal Array Groups, LDEVs, external LDEVs, and runtime tables for various software products. There can be up to 256GB of Cache and 32GB of Shared Memory.

Table 2 lists some of the features and their limits as supported by the USP and the USP V. Some of the USP V's HDP characteristics are repeated here as well. Note that the maximum volume size excludes LUSE volumes, which may be up to 60TB.

Limits	USP	USP-V
data cache (GB)	256	256
cache bandwidth	68 GB/s	68 GB/s
shared memory (GB)	24	32
shared memory bandwidth	15.9 GB/s	38.4 GB/s
FC disks	1152	1152
Logical Volumes	64k	64k
Maximum volume size	2TB	2TB
Queue Depth per port	1024	2048
Queue Depth per LUN	32	32
HDP Pools	-	32
LDEVs per Pool	-	256
HDP Volumes	-	65k
HDP Volumes per Pool	-	4k
HDP Volume size range	-	46MB - 2TB

Table 2. Summary of features.

For reference, Figure 13 below depicts the fully optioned USP V architecture. This is composed of 64 x 1064MB/sec Data paths, representing 68GB/sec of data bandwidth, and 256 x 150MB/sec Control or Metadata paths, representing 38.4GB/sec of control bandwidth.

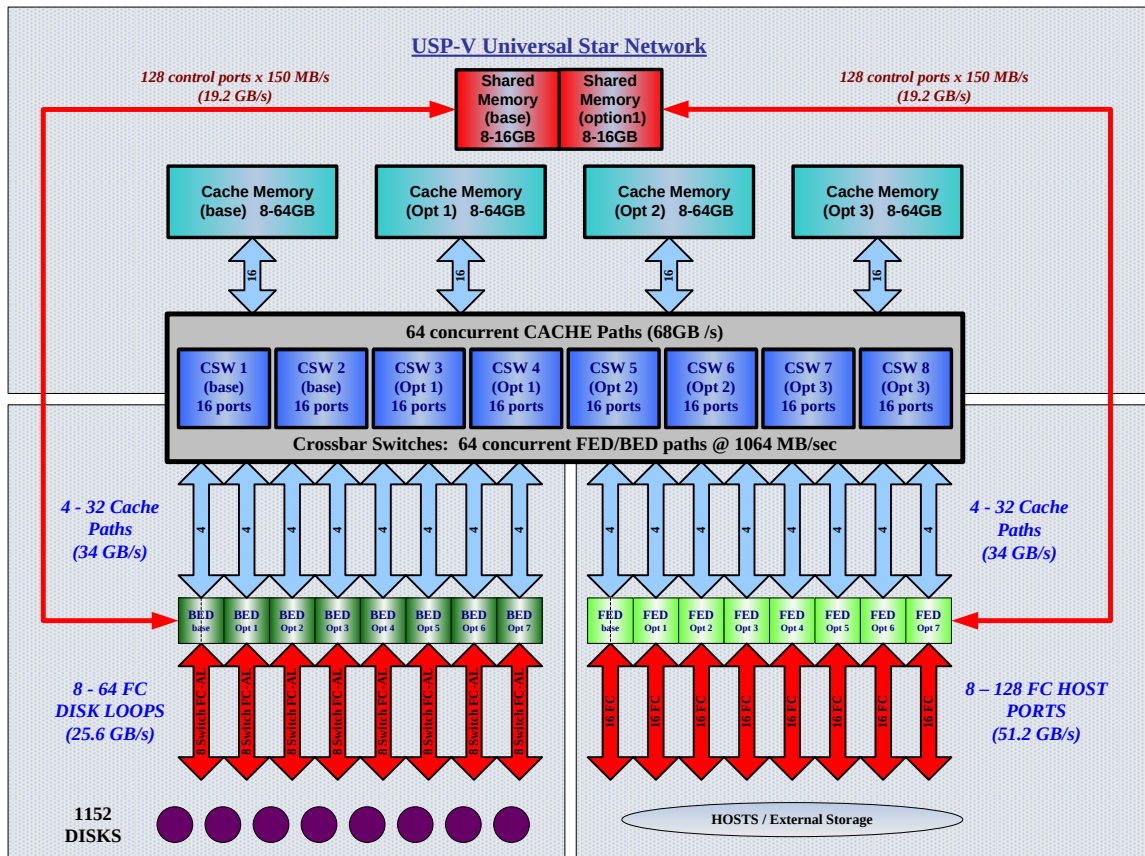


Figure 13. USP V system components and bandwidths overview.

Processor Upgrade

The processor used on the FEDs and BEDs has been improved and the speed has been doubled. The quantities of processors represented in Table 3 are per feature values. Each feature is defined as a pair of PCBs where each board is located on a separate power boundary. As there are twice as many of the FED/BED features in the USP V, the overall processor count for FEDs and BEDs is the same as with a USP.

Feature	USP	USP-V
FED-ESCON	8 x 400 MHz	4 x 800 MHz
FED-FICON	16 x 400 MHz	8 x 800 MHz
FED-FC	16 x 400 MHz	8 x 800 MHz
BED	16 x 400 MHz	8 x 800 MHz

Table 3. USP V Processor Enhancements.

Overview of Features

The table below shows an overview of the available features for both the USP and the USP V. Except for the Shared Memory PCBs, all of the other ones are now half sized, so there are twice as many of them in the USP V.

Features (PCB pairs)	USP	USP-V
FEDs	4	8
BEDs	4	8
Cache	2	4
Shared Memory	2	2
Cache Switches	2	4

Table 4. Comparison of features, USP and USP V.

Logic Box Layout

The Logic Box chassis in the main DKC frame is where all of the different types of PCBs for the features are installed in specific slots (see Figure 14). This layout is very different from the previous USP model (Figure 15). All of the PCBs (except for the SMA Shared Memory cards) are now half sized, and there are upper and lower slots in addition to the front and back layout. The associations among BEDs, FEDs and the CSWs (cache switches) are also different. The factory names of the PCB types and their option numbers are shown as well. Note that FED upgrade features may consume up to six pairs of unused BED slots.

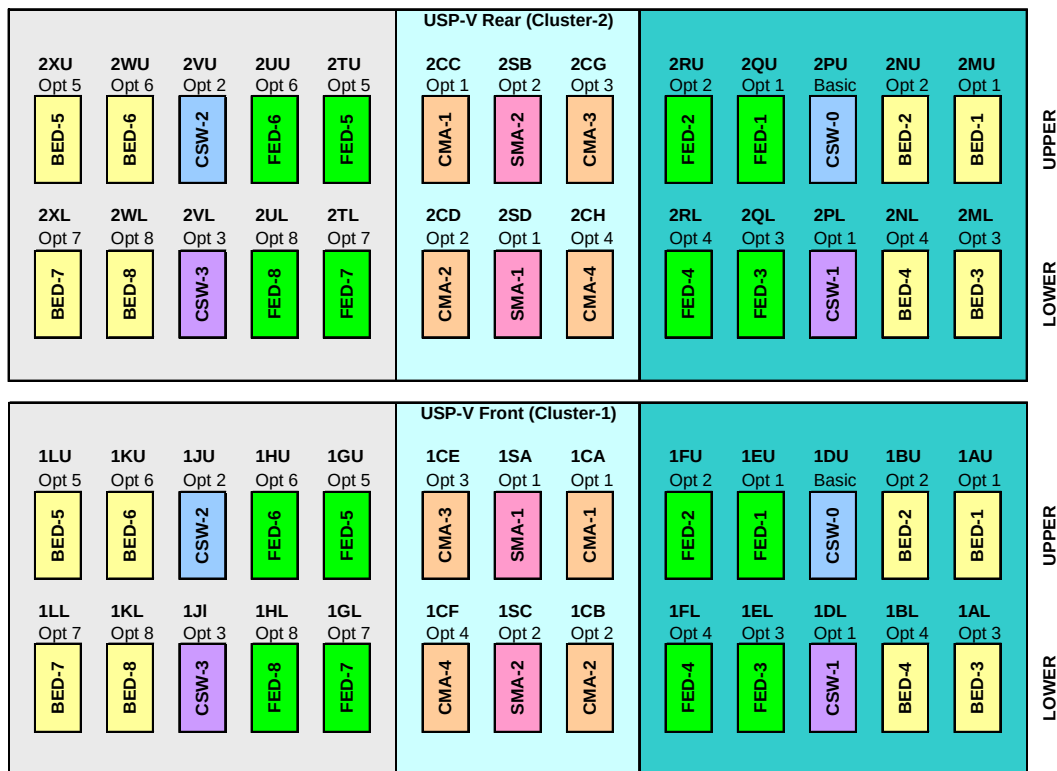


Figure 14. USP V Logic Box slot layouts.

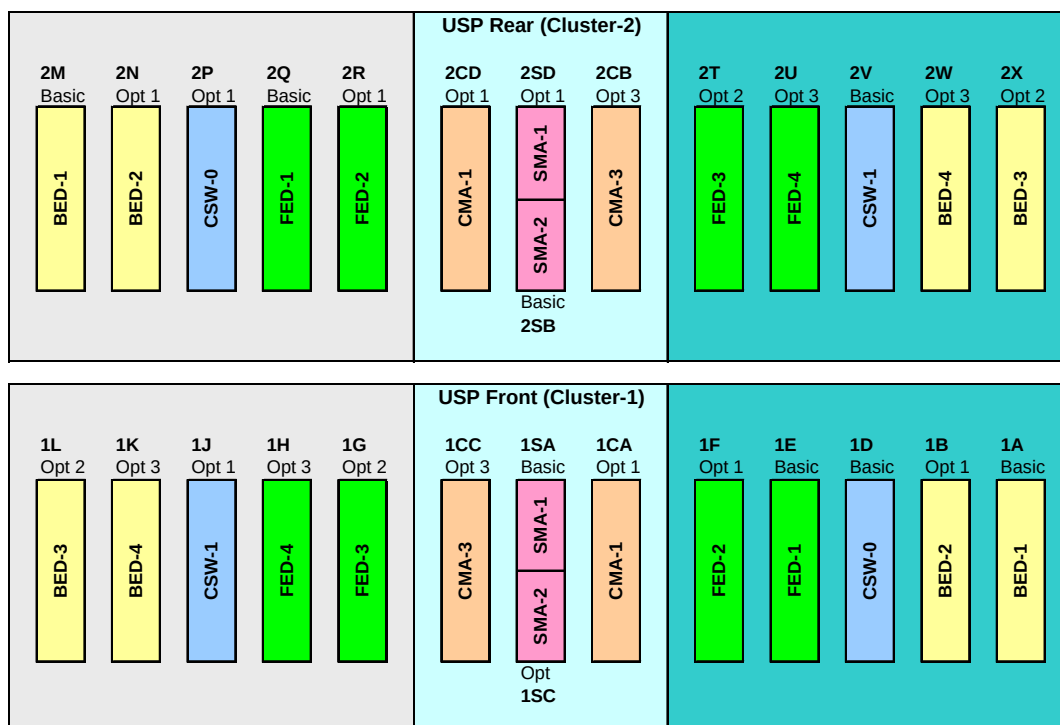


Figure 15. USP Logic Box slot layouts.

Front-End-Directors

The FEDs contain the host ports that interface with the servers or (using Open Fibre) attach to external storage. All I/O, whether it is Reads or Writes, internal or external storage, must pass through the Cache and Shared Memory systems. The various types of interface options can be supported simultaneously by mixing FED features. The specific number of ports for each interface type is shown below.

FED Features	FED Packages		Ports / Package	
	USP	USP-V	USP	USP-V
Open Fibre	4	8	16,32	8,16
ESCON	4	8	16	8
FICON	4	8	8,16	8

Table 5. Comparison of FED features.

FED FC-16 port Feature

Table 6 shows the USP V options for the 16-port Fibre Channel features. The front-end port counts, the FED PCB names, and the associated CSWs are also indicated. This port count can be increased up to 224 (using up to 14 FED features) if the number of BED features is minimal (two) in order to allow for more FED features (by stealing 12 BED “slots”).

FC- 16 port Feature					
	Pair	FED Boards		Ports	CSW
Feature 1	FED-1	FED00	FED08	16	0
Feature 2	FED-2	FED02	FED0A	32	0
Feature 3	FED-3	FED01	FED09	48	1
Feature 4	FED-4	FED03	FED0B	64	1
Feature 5	FED-5	FED04	FED0C	80	2
Feature 6	FED-6	FED06	FED0E	96	2
Feature 7	FED-7	FED05	FED0D	112	3
Feature 8	FED-8	FED07	FED0F	128	3
Feature 9	FED-9	(BED3)		144	1
Feature 10	FED-10	(BED4)		160	1
Feature 11	FED-11	(BED5)		176	2
Feature 12	FED-12	(BED6)		192	2
Feature 13	FED-13	(BED7)		208	3
Feature 14	FED-14	(BED8)		224	3

Table 6. USP V CHA FC-16 features.

FED FC-8 port, ESCON, FICON Features

Table 7 shows the USP V options for the 8-port Fibre Channel, 8-port FICON, and 8-port ESCON features. The front-end port counts, the FED (CHA) PCB names, and the associated CSWs are also indicated.

FC- 8 port, FICON, ESCON Features					
	Pair	FED Boards		Ports	CSW
Feature 1	FED-1	FED00	FED08	8	0
Feature 2	FED-2	FED02	FED0A	16	0
Feature 3	FED-3	FED01	FED09	24	1
Feature 4	FED-4	FED03	FED0B	32	1
Feature 5	FED-5	FED04	FED0C	40	2
Feature 6	FED-6	FED06	FED0E	48	2
Feature 7	FED-7	FED05	FED0D	56	3
Feature 8	FED-8	FED07	FED0F	64	3

Table 7. USP V 8-port CHA features.

Back-end Directors

The BEDs control the Fibre Channel Loops that interface with the internal disks (but not the external storage). Every I/O operation to the disks passes through the Cache and Shared Memory subsystems. Table 8 lists the BED features, loop counts and disk capacities by USP and USP V systems. Note that BEDs must be installed as pairs of Options (four PCBs) due to provide 8 pairs of loops to support the 8-disk Array Groups (RAID10 4D+4D, RAID5 7D+1P, RAID6 6D+2P).

BED Feature (2 PCBs)	Back-end Loop Pairs		Max Disks	
	USP	USP-V	USP	USP-V
1 BED	8	-	384	-
2 BED	16	8	640	256
3 BED	24	-	896	-
4 BED	32	16	1152	640
5 BED	-	-	-	-
6 BED	-	24	-	896
7 BED	-	-	-	-
8 BED	-	32	-	1152

Table 8. Comparison of BED features.

BED Feature

Table 9 shows the USP V options for the BED feature. The total back-end loop counts, the names of the BED PCBs, and the associated CSWs are also indicated.

BED Feature					
Pair		BED Boards		Loop Pairs	CSW
Basic	BED-1	BED10	BED18	-	0
	BED-2	BED12	BED1A	8	0
Feature 2	BED-3	BED11	BED19	-	1
	BED-4	BED13	BED1B	16	1
Feature 3	BED-5	BED14	BED1C	-	2
	BED-6	BED16	BED1E	24	2
Feature 4	BED-7	BED15	BED1D	-	3
	BED-8	BED17	BED1F	32	3

Table 9. USP V BED features.

USP V: HDU and BED Associations by Frame

Figures 16 and 17 illustrate the layouts of the 64-disk containers (HDU), the Frames, and the associated BED ownerships of the USP V. The names of the ranges of Parity Groups are also shown. Figures 18 and 19 show the USP details for comparison. There are two views presented: a regular frontal view and a view of the back as seen from the back.

For example, looking at Figure-16, the bottom of Frame DKU-R1 shows the front of the HDU whose disks are controlled by the eight 4Gbit loops on BED-1 (pg3, yellow) and BED-2 (pg4, orange). The HDU is split in half by power domains, where the 32-disks on the left half (16 on the front and 16 on the back of the Frame) are attached to BED-2 and the 32 on the right half go to BED-1. Note the diagonal offsets of the HDU halves from front to rear of the two associated 16-disk groups by HDU power domains.

The yellow section contains sixteen 4-disk “Parity Groups” whose names (used in Array Group configurations) are 3-1 through 3-16 (or pg3 as indicated). The official name of these 4-disk

groups (2 HDD on the front, 2 on the back of the HDU) is “Parity Group” even though they could be assigned RAID level 10 (no “parity involved”). See Appendix-1 for maps of the USP V Frames showing the names of the Parity Groups and where they are located. See Appendix-2 for the same maps for the USP Frames. Take note of how the locations of the Parity Groups have been shifted between the USP and the USP V models. This can have important considerations when configuring a system if this is not understood.

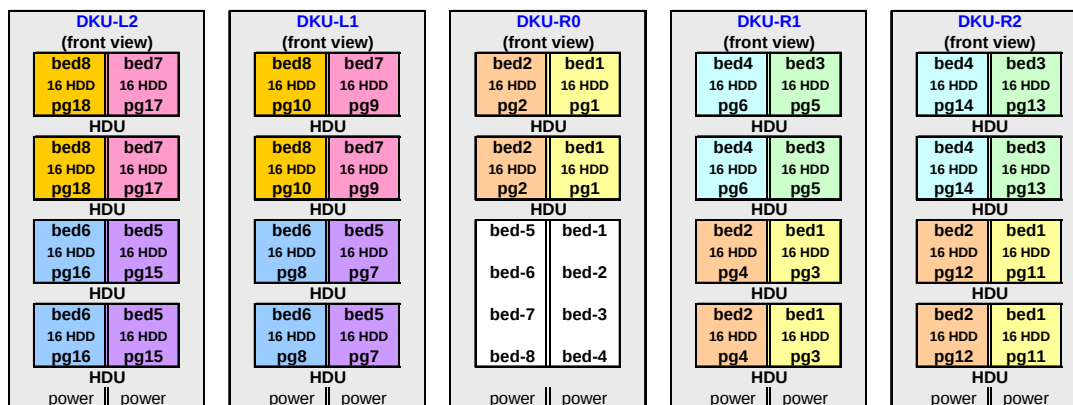


Figure 16. Front view of the USP V Frames, HDUs, with BED ownership.

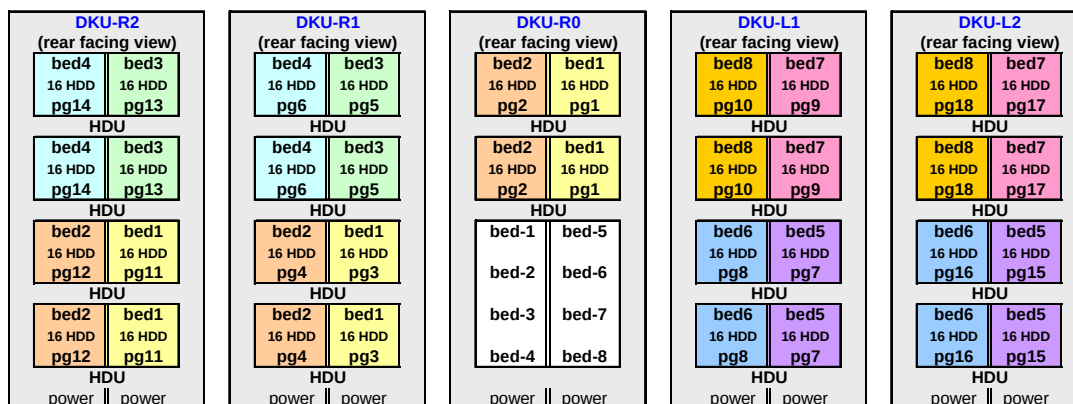


Figure 17. Rear view of the USP V Frames, HDUs, with BED ownership.

USP: HDU and BED Associations by Frame

For comparison purposes, the next three Figures illustrate the USP’s layouts of the 64-disk containers (**HDUs**), the Frames, and the associated BED ownerships. The names of the ranges of Parity Groups are also shown. There are two views presented: a regular frontal view and a view of the back as seen from the back.

For example, looking at Figure-18, the bottom of Frame DKU-R1 shows the front of the HDU whose disks are controlled by the eight 2Gbit loops on BED-1 (pg3, yellow) and BED-2 (pg5, orange). The HDU is split in half by power domains, where the 32-disks on the left half (16 on the front and 16 on the back of the Frame) are attached to BED-2 and the 32 on the right half go

to BED-1. Note the diagonal offsets of the HDU halves from front to rear of the two associated 16-disk groups by HDU power domains.

The yellow section contains sixteen 4-disk “Parity Groups” whose names (used in Array Group configuration) are 3-1 through 3-16 (or pg3 as indicated). The official name of these 4-disk groups (2 HDD on the front, 2 on the back of the HDU) is “Parity Group” even though they could be assigned RAID level 10 (no “parity involved”). See Appendix-2 for maps of the USP Frames showing the names of the Parity Groups.

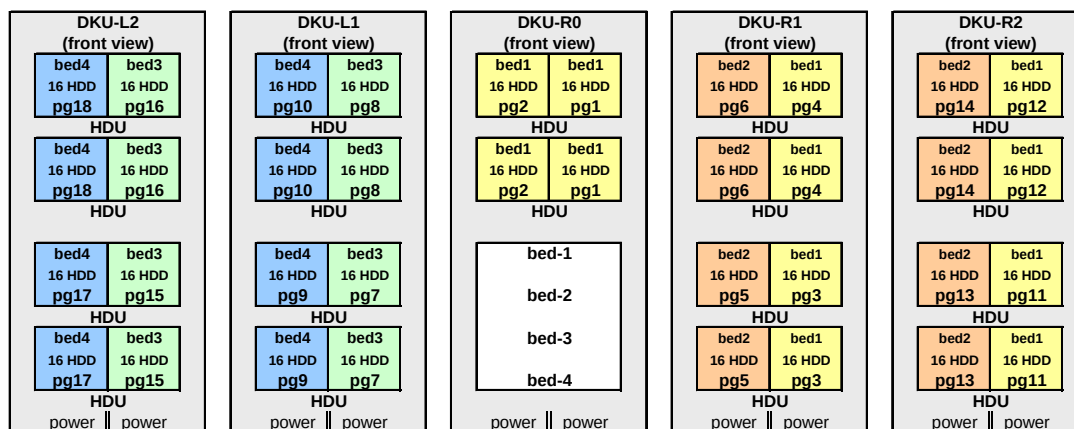


Figure 18. Front view of the USP Frames, HDUs, with BED ownership.

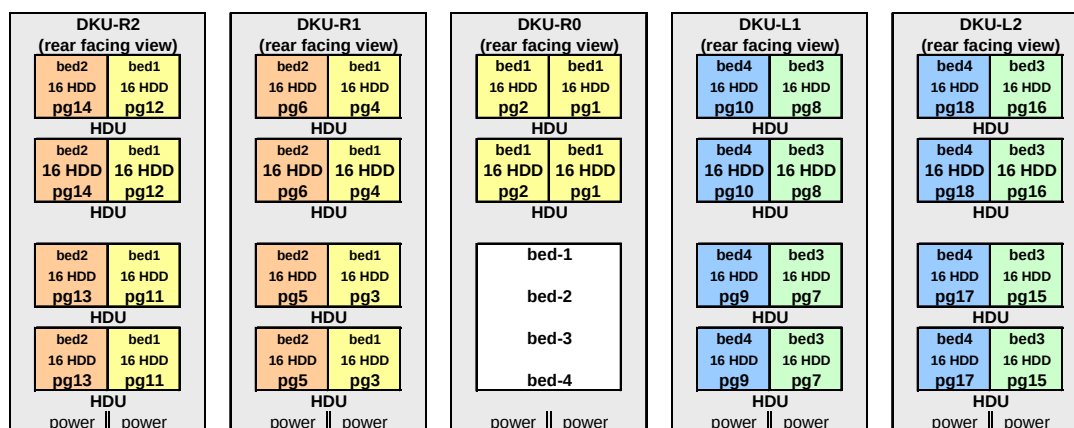


Figure 19. Rear view of the USP Frames, HDUs, with BED ownership.

Figure 20 below shows the details of a single frame, with its four HDUs (showing front and rear elements as viewed through the front). Figure 21 is a zoomed look at the front of just one HDU and the details of the switches and disks.

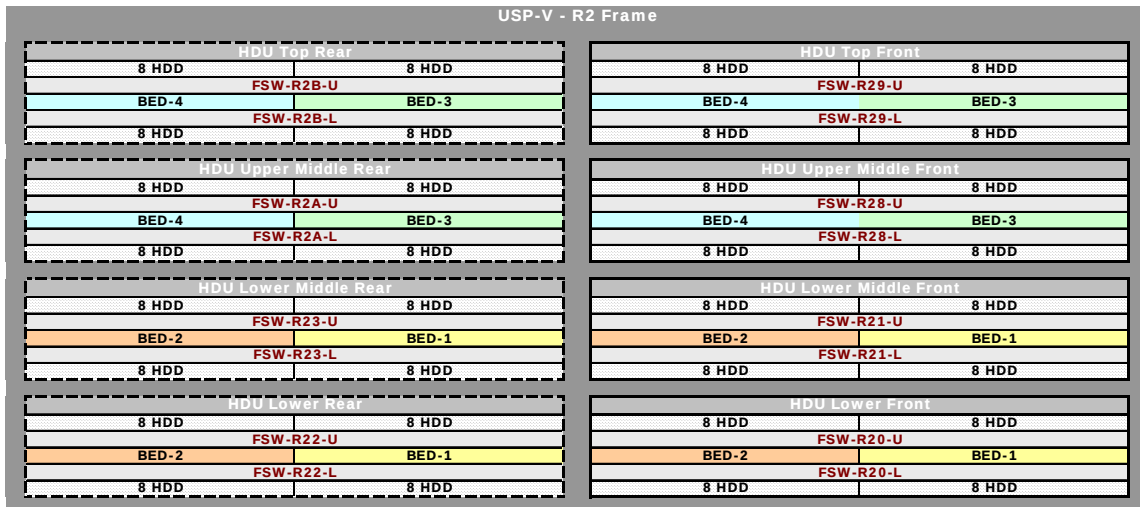


Figure 20. Details of a frame, with the HDUs and FSWs.

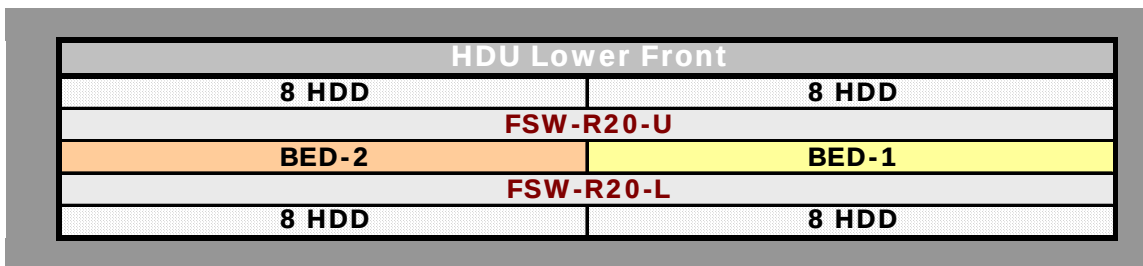


Figure 21. Closer look at one HDU and its two switches.

HDU Switched Loop Details

On the previous USP model, all of the disks attached to a BED PCB were on eight shared Loops. A transfer to the last disk on a loop (could be 32 or 48 disks per loop) would have to pass through bypass logic on all of the disks in front of it. The USP V introduces a switched back end, where the drives are still on Arbitrated Loops (all FC disks use FC-AL, not FC-SW fabrics), but the switch logic bypasses all targets on that loop except for the one disk being addressed. This reduces the propagation time by some microseconds (not noticeable), but the primary effect of this change is improved accessibility to individual disks. Each HDU has four such FSW switches, two on the front side (32 disks) and two more on the back (32 disks). Each pair of switches is either an “upper” or a “lower” switch as seen in Figure 23. Figure 22 is a “zoomed” view showing just the left half of the image in Figure 23.

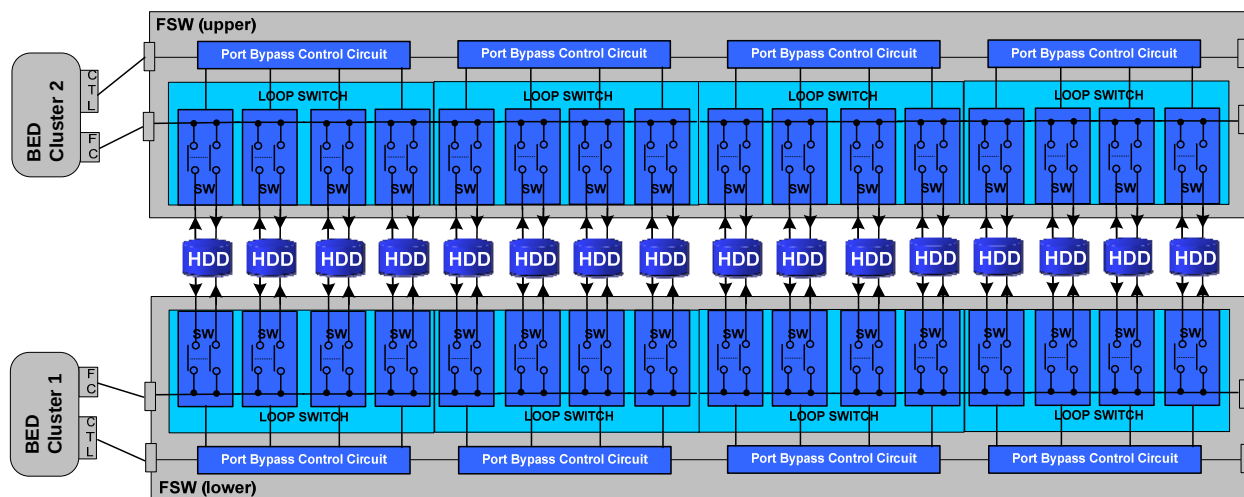


Figure 22. Closer view of the left half of the upper and lower switches on an HDU.

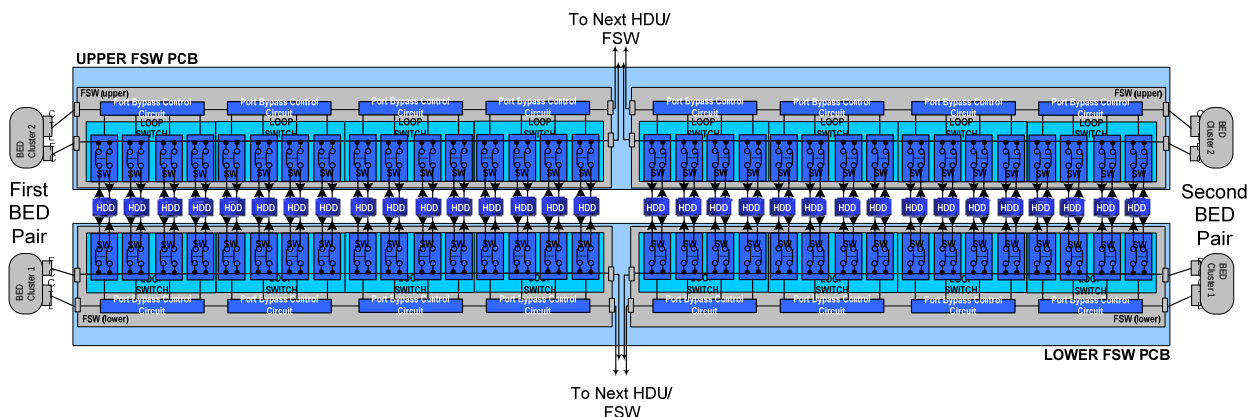


Figure 23. Full view of the two switches on the front or rear half of an HDU.

Disk Details

The USP V supports a mixture (in upgrade units of four disks) of up to 1152 FC and/or SATA disks. Table 10 illustrates the advertised size (base-10), the actual (computer base-2) raw size, and the typical usable (formatted) size of each type of disk. Also shown is the “rule of thumb” average maximum random IOPS rate for each type of disk when using most of the surface. The number and size of the LDEVs created per Array Group determine how much of its disks’ surfaces are in active use.

As more of the disk surface is allocated to LDEVs for an Array Group, the farther the disk heads must seek. This creates higher average seek times. For example (using a 15k RPM disk), when only using the outer 25% of the surface the average read seek time can be around 1.8ms (providing about 267 IOPS), whereas the 100% surface average seek time will be about 3.8ms (about 182 IOPS). For a SATA disk at 7200 RPM, these values are about 4.3ms (119 IOPS) and 8.5ms (79 IOPS). All write IOPS rates for all disks are about 5% lower due to the higher seek precision required. See Appendix 15 for more details on theoretical estimates of disk IOPS rates.

The USP V disk options include:

- **72GB FC, 15k RPM (4Gbit)**
- **146GB FC, 15k RPM (4Gbit)**
- **300GB FC, 10k (2Gbit), 15k RPM (4Gbit)**
- **750GB SATA, 7200 RPM (3Gbit, with native command queuing)**

Note that the use of the 2Gbit 300GB10k disks will slow down the backend loop performance for all disks to the 2Gbit rate on each pair of BED features where these disks are installed.

HDD Type	Advertised Size (GB)	Actual Size (GB)	Typical Formatted Size (GB)	Typical Physical ROT IOPS
750 GB SATA 7.2k RPM	750	738.62	701.7	80
300 GB FC 15k, 10k RPM	300	288.2	273.8	180, 130
146 GB FC 15k RPM	146	143.76	136.6	180
72 GB FC 15k RPM	72	71.5	67.9	180

Table 10. Table of HDD sizes and Rule of Thumb IOPS rates

The type and quantity of disks and the RAID levels chosen for the USP V will vary according to analysis of the user workloads, performance and capacity requirements. The following sections provide information to be considered when selecting which disks to install. Table 11 illustrates usable capacities by RAID and disk types for various disk capacities.

RAID Levels	Array Groups	256 HDDs		Array Groups	512 HDDs		Array Groups	1152 HDDs	
		Usable Space %	Usable Capacity GB		Usable Space %	Usable Capacity GB		Usable Space %	Usable Capacity GB
SATA									
RAID6 6D+2P	32	75%	134,724	64	75%	269,449	144	75%	606,259
FC 300GB									
RAID5 3D+1P	64	75%	52,568	128	75%	105,135	288	75%	236,555
RAID5 7D+1P	32	87.5%	61,329	64	87.5%	122,658	144	87.5%	275,980
RAID10 2D+2D	64	50%	35,045	128	50%	70,090	288	50%	157,703
RAID10 4D+4D	32	50%	35,045	64	50%	70,090	144	50%	157,703
RAID6 6D+2P	32	75%	52,568	64	75%	105,135	144	75%	236,555
FC 146GB									
RAID5 3D+1P	64	75%	26,222	128	75%	52,444	288	75%	117,998
RAID5 7D+1P	32	87.5%	30,592	64	87.5%	61,184	144	87.5%	137,665
RAID10 2D+2D	64	50%	17,481	128	50%	34,962	288	50%	78,665
RAID10 4D+4D	32	50%	17,481	64	50%	34,962	144	50%	78,665
RAID6 6D+2P	32	75%	26,222	64	75%	52,444	144	75%	117,998
FC 72GB									
RAID5 3D+1P	64	75%	13,042	128	75%	26,083	288	75%	58,687
RAID5 7D+1P	32	87.5%	15,215	64	87.5%	30,430	144	87.5%	68,468
RAID10 2D+2D	64	50%	8,694	128	50%	17,389	288	50%	39,125
RAID10 4D+4D	32	50%	8,694	64	50%	17,389	144	50%	39,125

Table 11. Disk types and usable capacities by RAID level

SATA DISKS

LDEVs that use SATA disks must be configured as OPEN-V emulation; therefore these LDEVs are not usable for mainframe volumes. If using the 4-disk RAID levels, there may only be one LDEV configured per Array Group. When using the 8-disk RAID levels, there may be one or two LDEVs configured per Array Group.

SATA disks are constructed with a far lower level of technology sophistication as compared to enterprise-grade FC disks. In general, the differences between the SATA and FC disks include:

- the amount of extreme technology and precision involved with the materials used for FC disks (disk platter materials and magnetic layer sandwiching, head technology, and head fly height control)
- the degree of error detection and correction (FC disks are about double that of SATA) and processing power (Hitachi FC disks are dual processor)
- the expected lifespan with the duty cycles experienced in enterprise class storage
- the overall performance available

When making performance estimates for SATA Array Groups that use RAID6 6D+2P, one may use as a rule of thumb that reads/writes will be 50%/75% lower than with similar RAID6 6D+2P 146GB 15k RPM Array Groups.

The back-end disk loops on the USP V are operated at 4Gbit/s, with all disks connected to these loops via switches. All SATA drives operate at either 1.5Gbit/s or 3Gbit/s depending on the model. SATA drives also only have one interface connection, whereas FC disks are dual ported (they may only talk to one loop at a time, but both ports are active). The SATA disk canisters fit into the same Disk Chassis slots as do FC disks because each SATA disk canister has a built-in controller module. One side of the controller presents two 4Gbit FC paths (mimicking a standard FC disk), while the inner side operates the 3Gbit SATA-II protocol. There is buffering in this controller that provides the speed level matching between the different protocol rates. [Note: This buffering keeps the BED loops from slowing down to a rate below 4Gbit as happens with the 2Gbit 300GB 10k FC disks.] The SATA canister controllers also perform a read-after-write check on every block written to their disk. After a block is written to the disk, it is read from the surface and compared to the buffer in the canister controller. While this reduces the write rate by 50%, it does guarantee that the data is accurately written to disk.

Table 12 shows the actual physical disk IOPS required for read or write operations as a function of disks type and RAID level. Although RAID6 does protect against a double disk failure, the write costs are high (especially for SATA disks), so this must be considered when designing a storage solution. Note that in the case where full stripe writes occur, the write penalty across the Array Group will be lower since the old parity block from each disk is not read.

	HDD IOPS per Host Read	HDD IOPS per Host Write
FC		
RAID10	1	2
RAID5	1	4
RAID6	1	6
SATA		
RAID10	1	4
RAID5	1	6
RAID6	1	8

Table 12. Physical disk IOPS by RAID level and type

Other considerations on the use of SATA versus FC when using RAID6 include:

- the no-load format times are about 18X longer than with FC disks
- the no-load corrective copy times are about 4.4X longer than with FC disks
- the no-load disk copy time is about 10X longer than with FC disks
- there should be 1 spare SATA disk for every 16 disks in use

Hitachi's recommendation on suitable uses for SATA volumes is to use them for Nearline storage with a low frequency of access but online availability, such as:

- Email archives, backups, 2nd and 3rd copies
- Historical data with limited access (medical records, etc.)
- Archived documents for regulatory purposes
- Sound or video files
- Tape replacement

The decision to use up valuable internal disk slots with SATA disks rather than the much higher performing FC disks should be carefully considered. In many cases, the use of SATA disks within virtualized storage on a midrange product might make better sense. However, there will be individual cases where the use of some internal slots for SATA drives will solve a specific customer business solution in a straight forward manner. It is not expected that a USP V would normally be configured with a high percentage of SATA disks.

V. USP V FED Details

The Front-end Directors manage the access and scheduling of requests to and from the host servers. The FED processors (**CHPs**, more commonly called **MPs**) also manage the Shared Memory and Cache areas, as well as execute the optional software features such as Dynamic Provisioning, Universal Replicator, Volume Migrator, and TrueCopy. The Data Adapter (**DTA**) is a special ASIC for communication between the Open Fibre Front-End-Director and the Cache Switches. The Microprocessor Adapter (**MPA**) is an ASIC for communication between the FED and the Shared Memory Adapters.

MP Distributed I/O

The Open Fibre FEDs (not FICON or ESCON) on the USP V have a new feature known as “MP Distributed I/O”. Some of the work pertaining to an individual I/O on a single port may be processed by another MP on the same PCB rather than the MP that normally controls that port. [See Appendix 8 for the port-MP mappings.] A heavily loaded MP can hand off some tasks pertaining to data, metadata, software (SI, TC, HUR), and HDP Pool administrative tasks to another MP that is less busy.

The Distributed I/O feature is limited to the four MPs on an individual PCB board. Some of this technique was actually begun in the USP with code levels V08 and higher, but most of it is unique to the USP V with its much higher powered MPs and faster Shared Memory system. The degree to which Distributed I/O is available is dependent on the pattern of I/O requests from the host as well as the internal status of work already in the system, back-end processing of pending requests, and so forth.

Distributed I/O on an individual MP begins when it reaches a 50% busy level, and is managed on a round-robin basis with the other available MPs on the same board, taking into account their current busy state. A table is kept in Shared Memory to assist in this workload dispatching. An MP that receives an I/O command over a host path and is currently more than 50% busy will look for another MP on that board to assist with the operation. This offload doesn't represent load balancing among host ports as such, just some degree of parallel processing for an individual command based on available cycles among the MPs on a board.

If an MP is doing processing for other MPs and then receives its own host I/O command, this workload could be distributed among the other MPs on the board. As the processors all get busy, the degree of sharing rapidly drops off. MPs that currently have no host I/O of their own can work up to a 100% busy rate on Distributed I/O loads. In the presence of host I/O on a port associated with an MP, the sum of that MP's host I/O + distributed I/O must be less than 50%. Once host I/O on an MP exceeds the 50% busy rate, there is no further acceptance of Distributed I/O by that MP.

When FC FED ports are configured for external storage, the associated pair of “owning” MPs no longer participates in the Distributed I/O mode. At least two MPs (1 and 2, or 3 and 4) per PCB are involved in virtualization of pairs of ports. If all ports on a PCB are used for external attachment, then all of those MPs are operated in initiator mode, and there is no Distributed I/O function on that PCB.

Open Fibre 8-Port Feature

The 8-port Open Fibre feature consists of two PCBs, each with four 4Gbit/sec Open Fibre ports. Each PCB has four 800MHz Channel Host Interface Processors (CHP) and two Tachyon-DX4 dual ported chips. The Shared Memory (MPA) port count is now eight paths per PCB (double the USP PCB). Figure 24 is a high level diagram of the Open Fibre 8-Port PCBs. Figure 25 is a diagram of the USP 16-port Open Fibre FED.

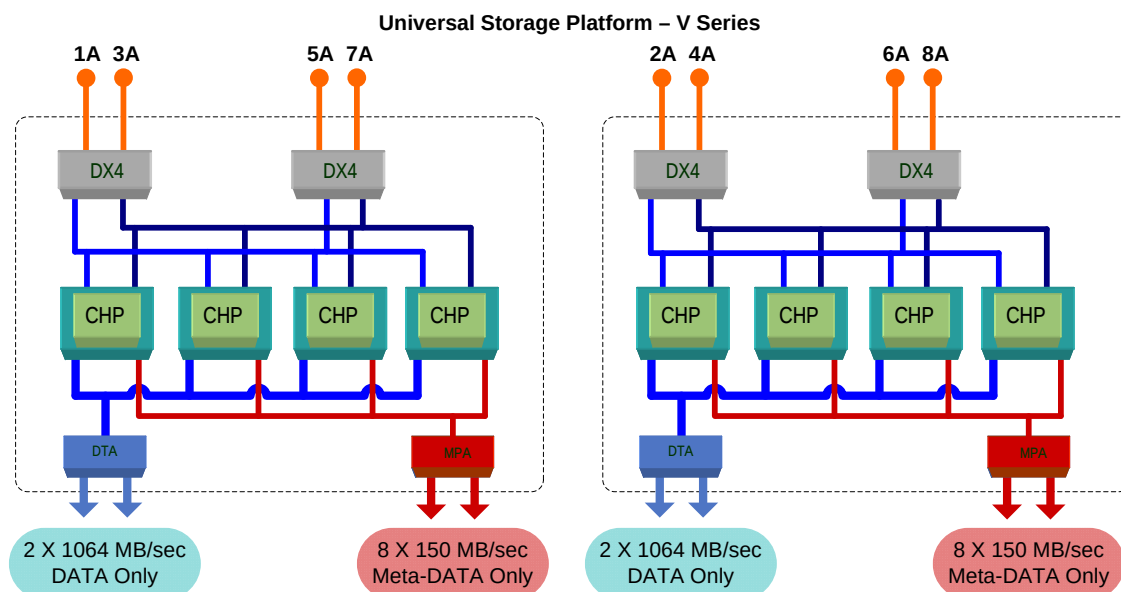


Figure 24. USP V 8-port FC feature (showing both PCBs).

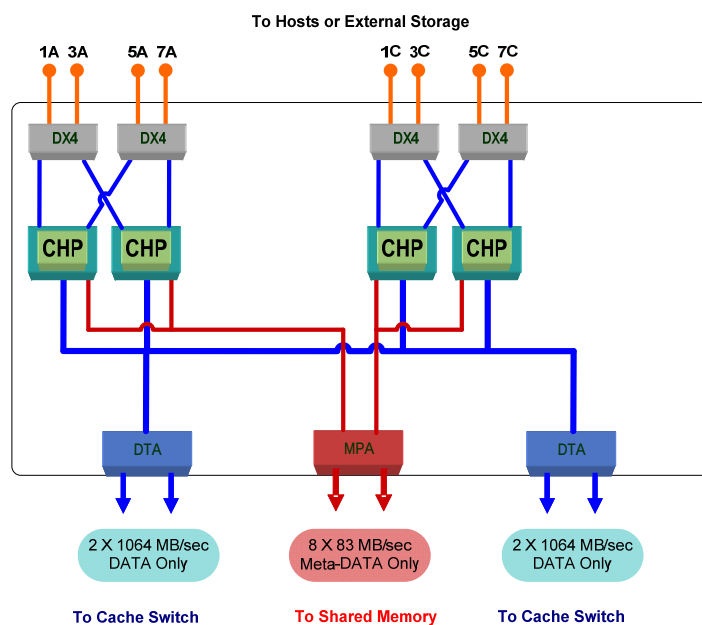
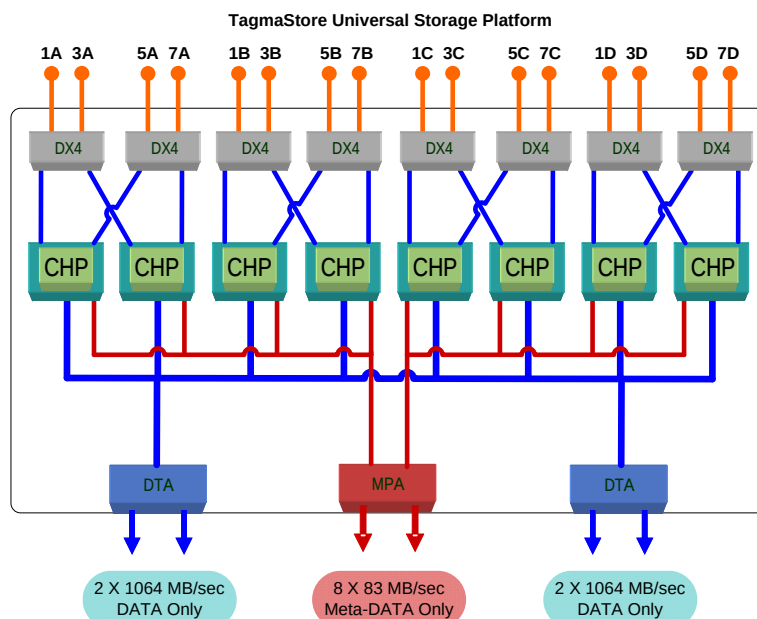
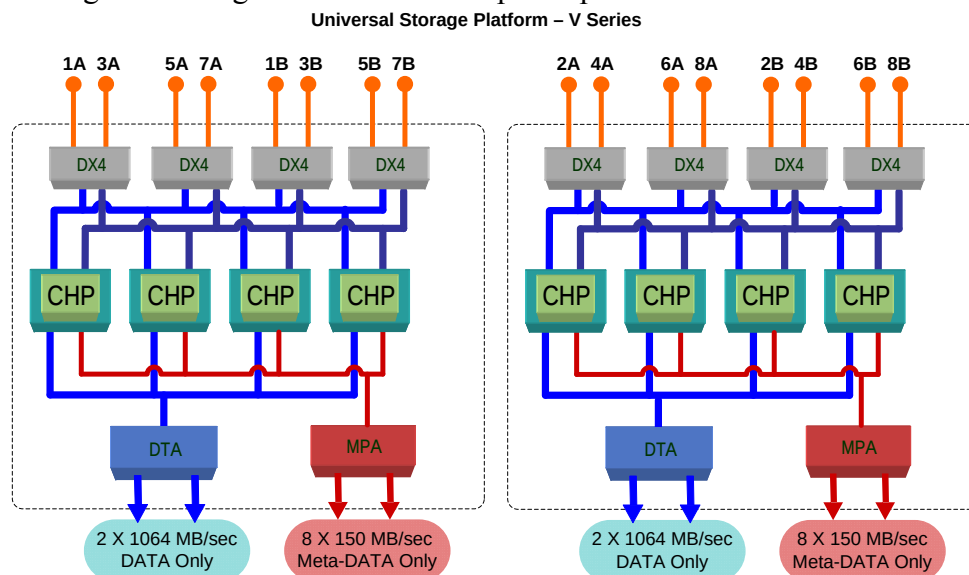


Figure 25. USP 16-port FC feature (showing one PCB).

Open Fibre 16-Port Feature

The 16-port Open Fibre feature consists of two PCBs, each with eight 4Gbit/sec Open Fibre ports. Each PCB has four 800MHz Channel Host Interface Processors (CHP) and four Tachyon-DX4 dual ported chips. There are eight Shared Memory (MPA) paths per PCB (double the USP PCB). Figure 26 is a high level diagram of the Open Fibre 16-Port PCBs. Figure 27, for reference, is a high level diagram of the USP 32-port Open Fibre FED.



ESCON 8-port Feature

The 8-port ESCON feature consists of two PCBs, each with four 17MB/s ESCON Fibre ports. Each PCB has two 800MHz Channel Host Interface Processors (CHP) and one ESA0 interface. There are two Cache paths and eight Shared Memory (MPA) paths per PCB. Figure 28 is a high level diagram of the ESCON Channel Adapter PCBs. Figure 29, for reference, is a high level diagram of the USP ESCON FED PCB.

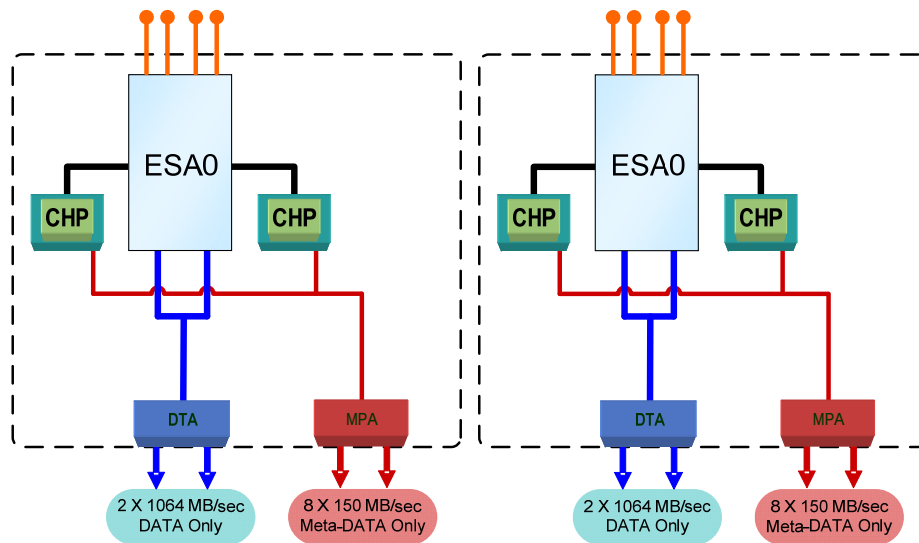


Figure 28. USP V 8-port ESCON feature (both PCBs shown).

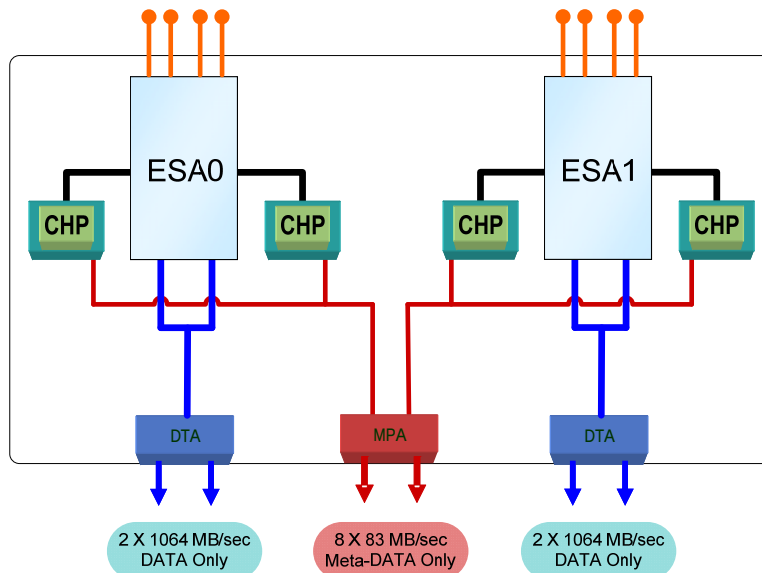


Figure 29. USP ESCON feature (one PCB shown).

FICON 8-port Feature

The 8-port FICON feature consists of two PCBs, each with four 4Gbit/sec FICON ports. Each PCB has four 800MHz Channel Host Interface Processors (CHP) and two HTP interface chips. There are two cache Switch paths and eight Shared Memory (MPA) paths per PCB. Figure 30 is a high level diagram of the 8-Port FICON PCBs. Figure 31, for reference, is a high level diagram of the USP 16-port FICON FED.

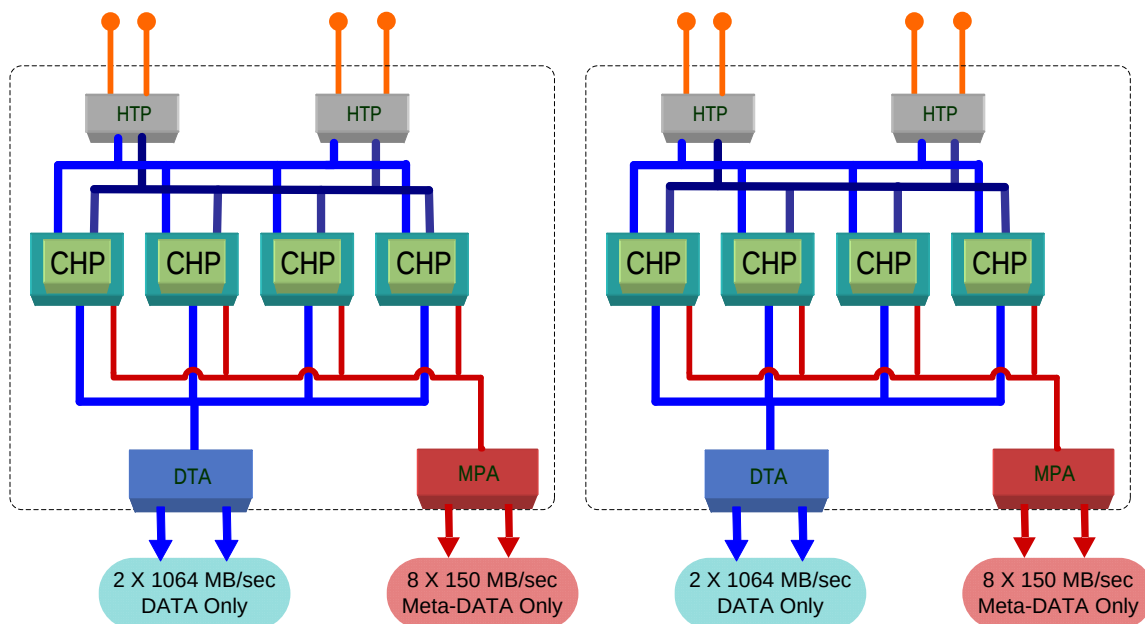


Figure 30. USP V 8-port FICON feature (both PCBs shown).

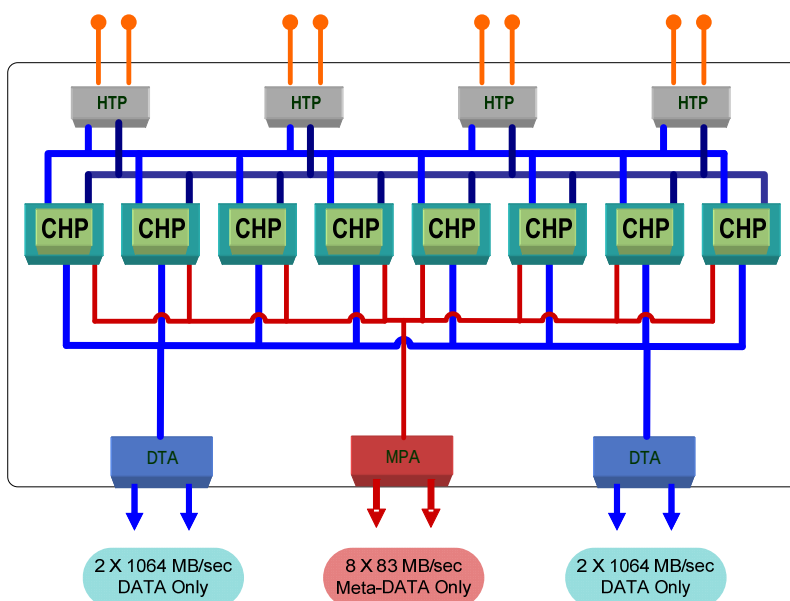


Figure 31. USP 16-port FICON feature (one PCB shown).

VI. USP V BED Disk Controller Details

The Back-end Directors manage access and scheduling of requests to and from the physical disks. The BEDs also monitor utilization of the loops, array groups, processors, and status of the PCBs in a pair. The USP V BED feature (2 PCBs) has four pairs of 4Gbit/sec loops supporting up to 128 disks. Each PCB has four 800MHz DKP processors and four DRR RAID processors. The USP BED pair has eight pairs of 2Gbit/sec Fibre Loops supporting up to 256 disks. There are up to eight BED features (16 PCBs) in a USP V, while a USP has four BED features (8 PCBs), both having the same number of backend loop pairs. Table 13 lists the loop and disk capacities by feature for the USP and the USP V.

BED Feature (2 PCBs)	Back-end Loop Pairs		Max Disks	
	USP	USP-V	USP	USP-V
1 BED	8	-	384	-
2 BED	16	8	640	256
3 BED	24	-	896	-
4 BED	32	16	1152	640
5 BED	-	-	-	-
6 BED	-	24	-	896
7 BED	-	-	-	-
8 BED	-	32	-	1152

Table 13. Comparison table of BEDs and disks, USP and USP V.

Back-End Director Details

Figure 32 is a high level diagram of the pair of PCBs for the USP V BED feature. Figure 33 is a high level diagram of one of the PCBs for the USP BED feature.

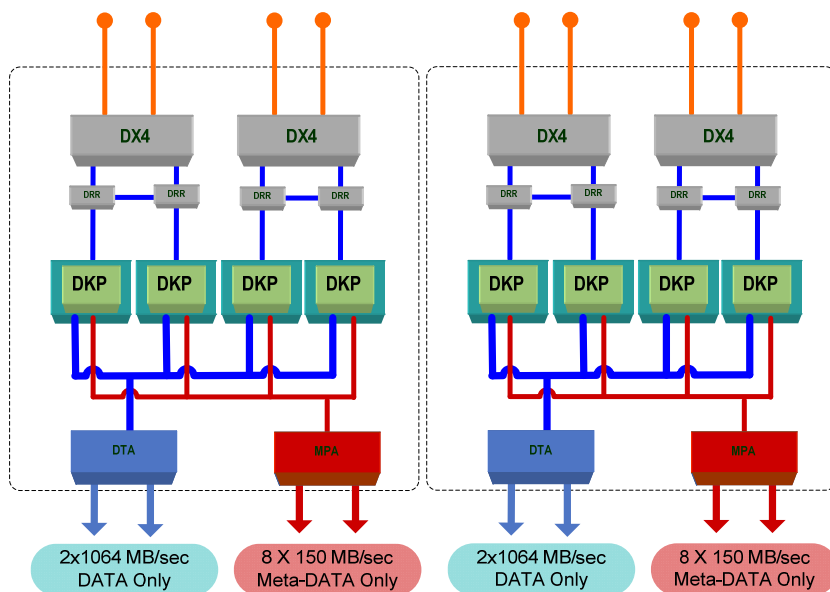


Figure 32. USP V BED feature (both PCBs shown).

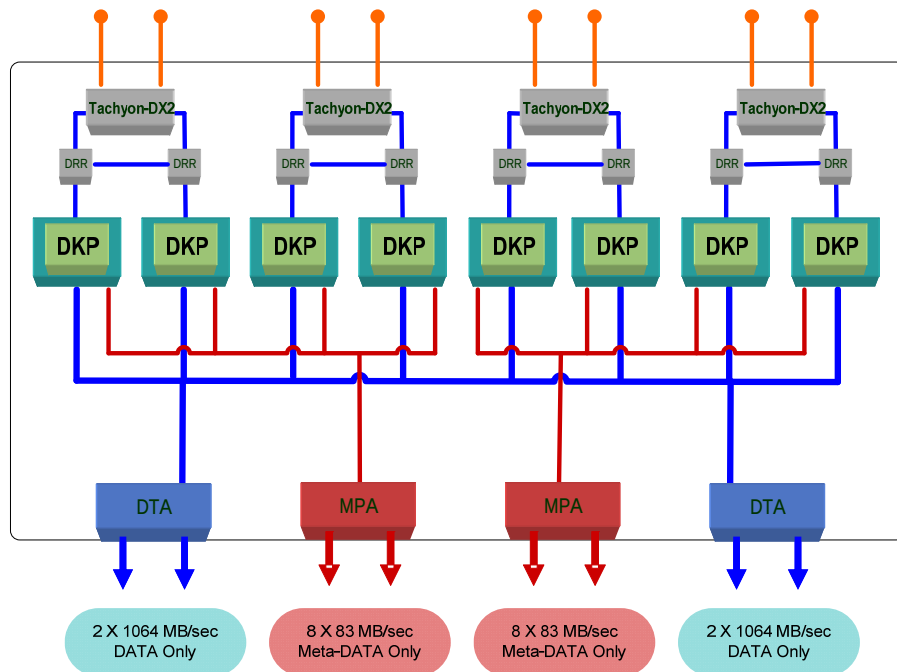


Figure 33. USP BED (one PCB shown).

Back-End RAID level Organization

Figure 34 is a view of the USP V's BED-to-disk organization. This is also how the Lightning 9900V was organized. These two BED features (4 PCBs) can support 256 disks, with the exception being BED1 and BED2, which also control the 128 disks in the DKC control frame (hence 384 disks overall). It takes two BED options to support all of the Array Group types due to the smaller 4-loop PCBs. Because of this, pairs of BEDs (4 PCBs) must always be installed.

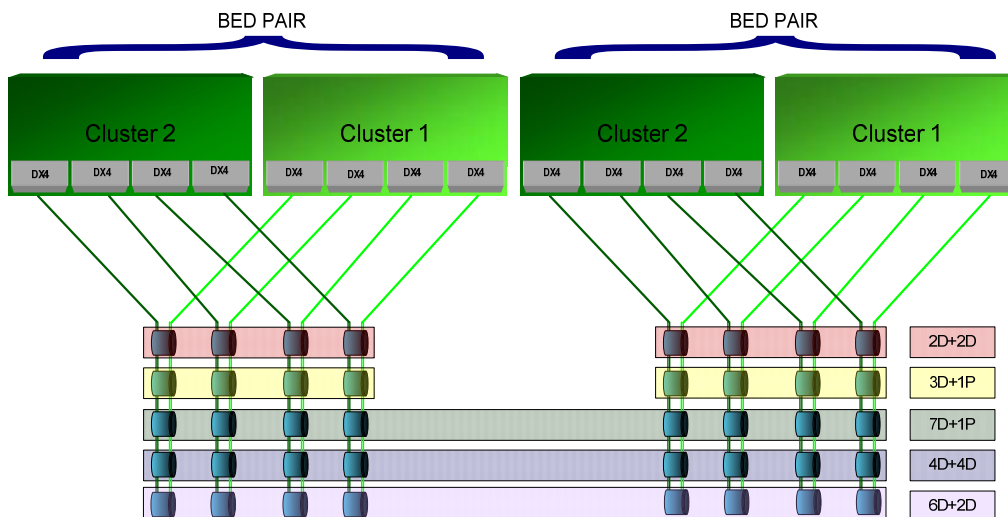


Figure 34. USP V Back-end disk layout. (2 BED features)

Figure 35 is the high level diagram of the USP's BED-to-disk organization. Each BED feature (2 PCBs) can support 256 disks, with the exception of BED1, which also controls the 128 disks located in the DKC frame (hence 384 disks overall). All Array Group types were configured behind one BED option.

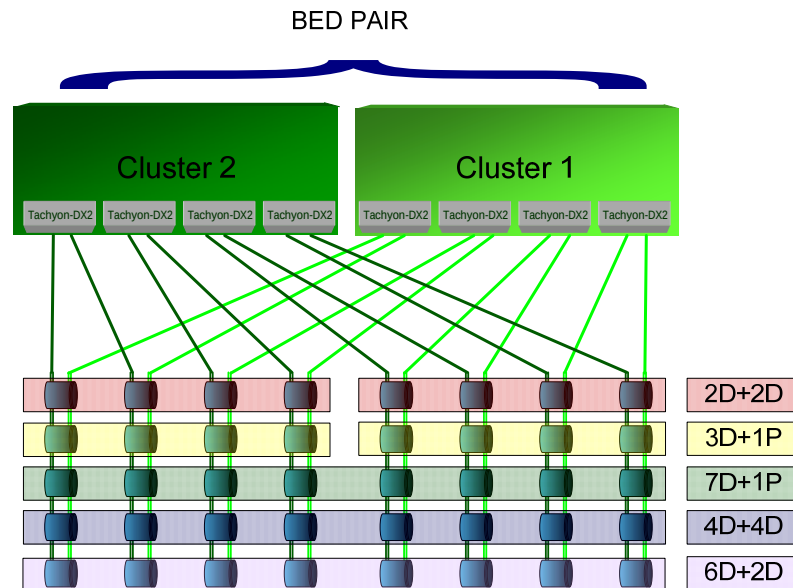


Figure 35. USP Array Group organization by BED.

VII. USP V Product Examples

The USP V is not being offered in specific models as is the USP, but as basic model that is a starting point for configurations that meet a variety of customer requirements.

Small Configuration

The USP V depicted in Figure 36 is comprised of 16 x 1064MB/sec Data paths, representing 17GB/sec of data bandwidth, and 64 x 150MB/sec Control or Metadata paths, representing 9.6GB/sec of control bandwidth.

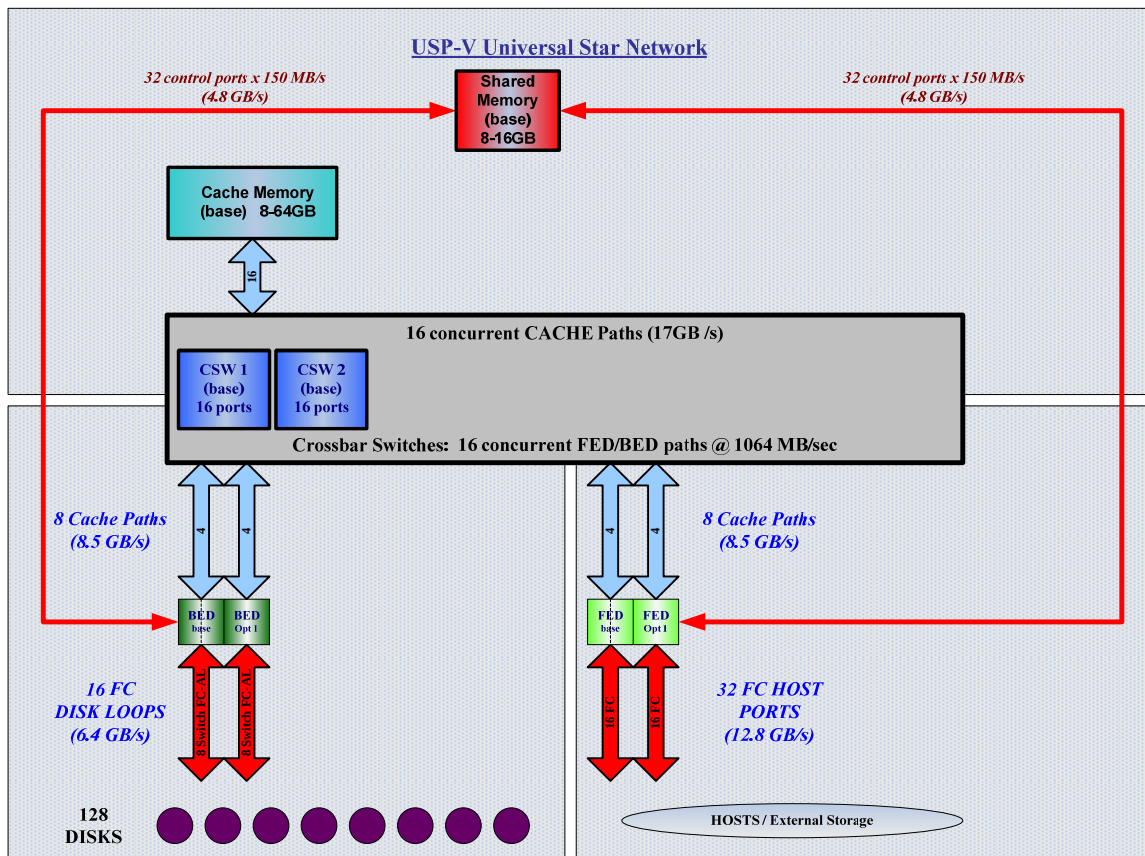


Figure 36. USP V, small configuration.

This system is shown with:

- Up to 128 physical disks in the controller frame.
- Two Back-end-Director features (BED-1, BED-2; 4 PCBs), consisting of 16 4Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 6.4GB/sec.
- One Cache Card feature (CMA-1; 2 PCBs) supporting up to 64GB of cache.
- One Cache Switch feature (CSW-0; 2 PCBs) with a total internal data bandwidth of 17GB/sec.
- One Shared Memory feature (SMA-1; 2 PCBs) with a total metadata bandwidth of 9.6GB/sec.

- Two Front-End-Director features (FED-1, FED-2; 4 PCBs) supporting either 16 – ESCON, 16 4Gbit/sec FICON, 16 4Gbit/sec Fibre Channel, or 32 4Gbit/sec Fibre Channel ports.

Midsize Configuration

The USP V depicted in Figure 37 is comprised of 32 x 1064MB/sec Data paths, representing 34GB/sec of data bandwidth, and 128 x 150MB/sec Control or Metadata paths, representing 19.2GB/sec of control bandwidth.

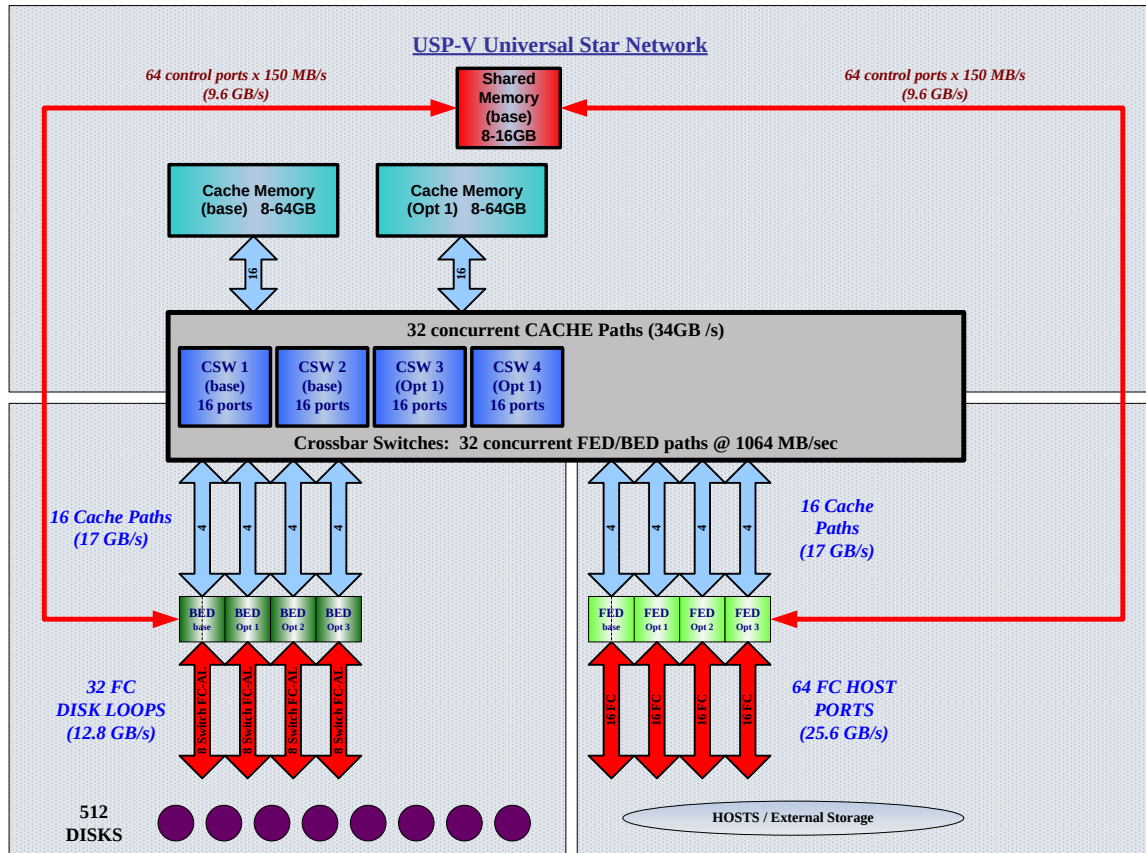


Figure 37. USP V, enhanced configuration.

This system is shown with:

- Up to 512 physical disks, 256 in the R1 frame and 256 in the R2 expansion frame.
- Four Back-end-Director features (BED-1, BED-2, BED-3, BED-4; 8 PCBs), consisting of 32 4Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 12.8GB/sec.
- Two Cache Card features (CMA-1, CMA-2; 4 PCBs) supporting up 128GB of cache.
- Two Cache Switch features (CSW-0, CSW-1; 4 PCBs) with a total internal data bandwidth of 34GB/sec.
- One Shared Memory feature (SMA-1; 2 PCBs) with a total metadata bandwidth of 19.2GB/sec.

- Four Front-End-Director features (FED-1, FED-2, FED-3, FED-4; 8 PCBs) supporting either 32 ESCON, 32 4Gbit/sec FICON, 32 4Gbit/sec Fibre Channel, or 64 4Gbit/sec Fibre Channel ports (more if fewer BED features are installed).

Large Configuration

The USP V depicted in Figure 38 is comprised of 64 x 1064MB/sec Data paths, representing 68GB/sec of data bandwidth, and 256 x 150MB/sec Control or Metadata paths, representing 38.4GB/sec of control bandwidth.

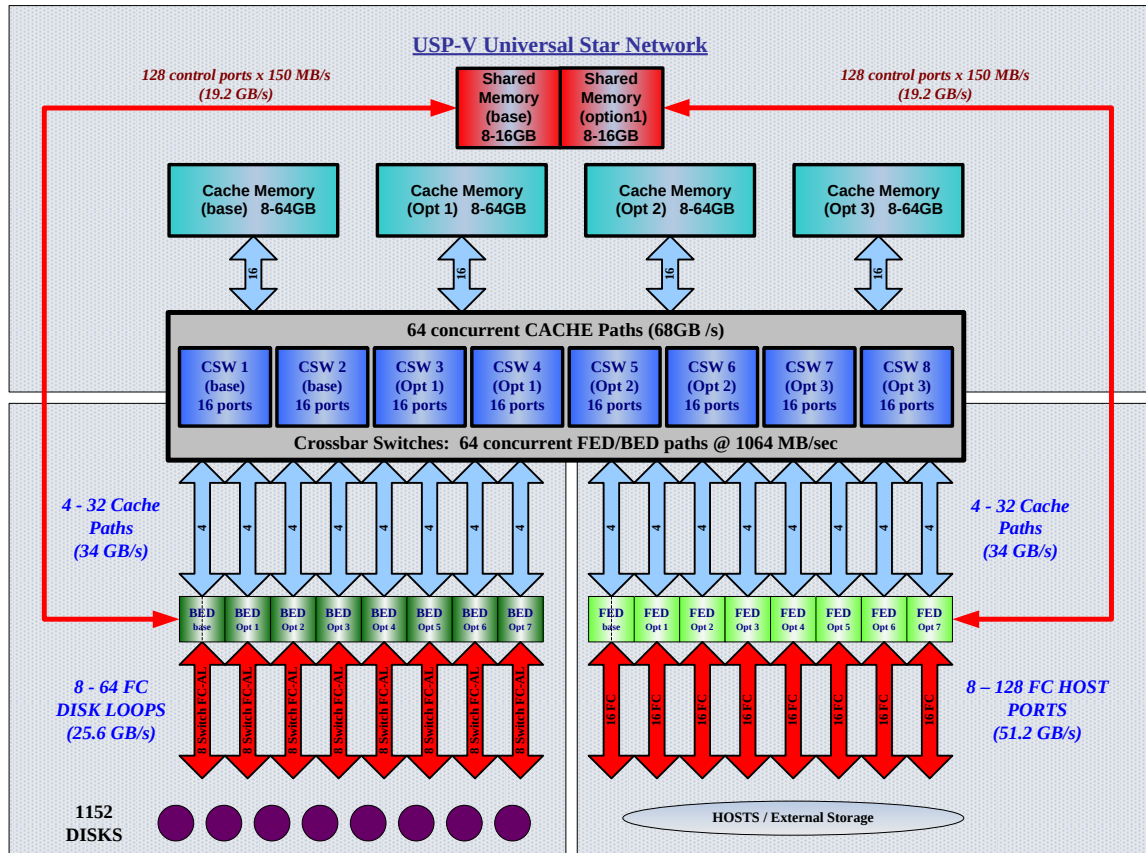


Figure 38. USP V, maximum configuration.

This system is shown with:

- Up to 1152 physical disks, 128 in the controller frame, 256 in each R1, R2, L1, and L2 expansion frames.
- Eight Back-end-Director features (16 PCBs), consisting of 64 4Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 25.6GB/sec.
- Four Cache Card features (8 PCBs) supporting up to 256GB of cache.
- Four Cache Switch features (8 PCBs) with a total internal data bandwidth of 68GB/sec.
- Two Shared Memory features (4 PCBs) with a total metadata bandwidth of 38.4GB/sec.
- Eight Front-End-Director features (16 PCBs) supporting either 64 ESCON, 64 4Gbit/sec FICON, 64 4Gbit/sec Fibre Channel, or 128 4Gbit/sec Fibre Channel ports (up to 224 in 14 features if only 2 BED features are installed).

VIII. Special Performance Tests

A variety of initial performance tests have been conducted using the USP V. Full sets of tests and papers will not begin until availability of GA microcode. This section shows two specialized tests that have been conducted on the USP V. The first results show how the USP V fares on a standardized mixed workload database test. The second tests show how the performance of the new DP-VOLs compares to the use of Standard LDEV volumes.

Database Test Comparison

The Figure below shows a comparison of a standard type of mixed database workload against a USP1100 and a fully loaded USP V. The USP V and IBM DS8300 Turbo results are from published SPC1 tests. This workload is based on 90% random (8KB), 10% sequential (256KB), 60% read and 40% write, with a 20% cache hit ratio.

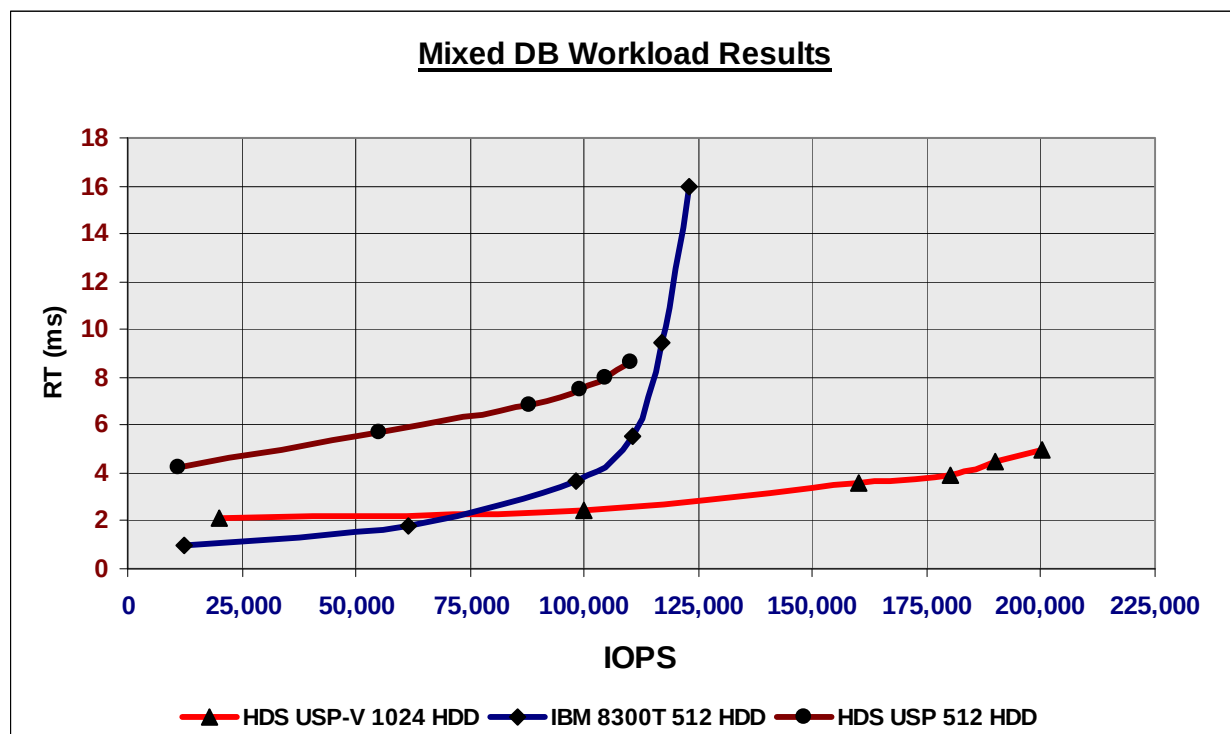


Figure 39. Simulated database workload performance comparison.

DP-VOL Performance

The Figure below shows the results of tests that compare the standard independent use of LDEVs, a host-based LVM stripe across the 8 LDEVs, and the use of HDP on 8 LDEVs. There were 32 146gb15k HDDs used, configured as 8 2+2 Array Groups. There was a single 4Gbit host port used for all tests. Notice how HDP outperformed the host LVM (IBM AIX's LVM) by a small amount in both IOPS and Response Times. The "8 LDEV aggregate" column shows the maximum results possible for this test. This test used 8 individual LUNs driven by concurrent workloads by the test tool.

8kb Random workload, 75% Read, uncached								
Workload Threads	Single LDEV		8 LDEV aggregate		LVM Volume		HDP Volume	
	IOPS	RT(ms)	IOPS	RT(ms)	IOPS	RT(ms)	IOPS	RT(ms)
8	620	12.9	1,513	5.3	1,339	6.0	1,366	5.9
16	736	21.7	2,689	5.9	2,308	6.9	2,370	6.7
32	898	35.6	3,926	8.1	3,377	9.5	3,453	9.3
64	1,075	59.5	4,959	12.9	4,381	14.6	4,437	14.4
128	1,158	110.5	5,945	21.5	5,370	23.8	5,427	23.6
256	1,167	219.0	7,103	36.0	6,452	39.7	6,640	38.5

Figure 40. Single host port HDP tests.

IX. RAID10 (2D+2D) Back-end Scalability Tests

The purpose of this testing was to conduct a back-end disk (cache miss) scalability analysis using the Hitachi Data Systems Universal Storage Platform V enterprise storage system. The quantity of Front-end Directors (FEDs) and Back-end Directors (BEDs) used were varied to scale the system size. The first set of tests evaluated how a **small** USP V (2 FED, 2 BED, and 256 HDDs) could perform. The second set of tests measured a **midsized** USP V (4 FEDs, 4 BEDs, and 512 HDDs). The final set of tests measured a **large** USP V (8 FEDs, 8 BEDs, and 1024 HDDs). A fairly powerful IBM p595 server was used for these tests.

There was a single test system that had 256GB of cache and 22 GB of Shared Memory. The disks used were 146GB 15K RPM. The front-end and back-end paths are all 4Gbit FC. The test configurations used for these three sets of tests included:

- **USP V small**
 - o 2 FED – 16 unshared FC host ports
 - o 2 BED – 256 HDDs (16 disk loops, 64 Array Groups, 64 LUNs)
- **USP V midsize**
 - o 4 FEDs – 32 unshared FC host ports
 - o 4 BEDs – 512 HDDs (32 disk loops, 128 Array Groups, 128 LUNs)
- **USP V large**
 - o 8 FEDs – 64 unshared FC host ports
 - o 8 BEDs – 1024 HDDs (64 disk loops, 256 Array Groups, 256 LUNs)

Test Methodology

The test methodology was to first configure 256 Array Groups in the USP V using RAID10 2D+2D, and to create one 33GB LDEV per RAID Group. These LDEVs were then mapped to FED ports. For each USP V “model”, only those LDEVs that were behind a certain set of BEDs were accessed for those tests. A specific pair of FEDs was associated with each pair of BEDs, so that certain host ports always saw certain LUNs. Hence, there were always four LUNS per port. The “small” system used the first 16 host ports and 64 LUNs. The “midsize” system used those plus the next group of 16 ports and 64 LUNs. The “large” system used all four sets of ports and LUNs. Each test run for a system began with one LUN and scaled up to the maximum for that

system. There were 8 concurrent VDBench threads per LUN for Random workloads, and just 1 thread per LUN for Sequential workloads.

These are the tests that were used:

- **Random I/O, Transactions per second and Average Response Time**, at 2, 4, 8, and 16KB block sizes with 8 Threads (100% - 75% - 50% Read, and 100% Write)
- **Sequential I/O, Throughput**, at 64, 128, 256, and 512KB block sizes with 1 Thread (100% - 75% - 50% Read, and 100% Write)

Random I/O Transactions

The following table shows the peak Random workload results for the RAID10 2D+2D scalability tests. The number of LUNs used per test run was steadily increased from 1 to the maximum (64, 128, 256), and the peak results achieved during those steps are reported in this table.

Random I/O	USP-V	256-1024 HDD	2 FED 2 BED		4 FED 4 BED		8 FED 8 BED	
cache miss	2+2	146GB 15k	16 host paths		32 host paths		64 host paths	
standard LDEVs		1-256 LUNs	16 disk loops		32 disk loops		64 disk loops	
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	peak average	49,540	10.3	95,503	10.7	176,566	11.6
100	4k		49,455	10.3	95,032	10.8	176,240	11.6
100	8k		48,900	10.5	94,359	10.8	174,232	11.8
100	16k		48,015	10.7	92,785	11.0	171,903	11.9
			49,540	10.7	95,503	11.0	176,566	11.9
			48,977	10.5	94,419	10.8	174,735	11.7
75	2k	peak average	46,562	11.0	91,441	11.2	178,680	11.5
75	4k		44,250	11.6	91,285	11.2	177,825	11.5
75	8k		40,195	12.7	87,977	11.7	175,973	11.6
75	16k		39,893	12.8	77,412	13.2	165,549	12.4
			46,562	12.8	91,441	13.2	178,680	12.4
			42,725	12.0	87,029	11.8	174,507	11.7
50	2k	peak average	31,024	16.5	56,134	18.2	104,715	19.5
50	4k		28,051	18.2	55,210	18.5	111,200	18.5
50	8k		27,968	18.3	54,969	18.6	105,621	19.4
50	16k		27,668	18.5	54,166	18.9	102,847	19.9
			31,024	18.5	56,134	18.9	111,200	19.9
			28,678	17.9	55,120	18.6	106,096	19.4
0	2k	peak average	18,810	27.2	36,479	28.1	71,061	28.8
0	4k		18,437	27.8	35,990	28.5	71,992	28.5
0	8k		18,298	28.0	36,218	28.3	72,608	28.2
0	16k		18,365	27.9	36,158	28.3	73,334	27.9
			18,810	28.0	36,479	28.5	73,334	28.8
			18,478	27.7	36,212	28.3	72,249	28.4

Table 14. Summary of all peak Random results.

Sequential I/O Throughput

The following table shows the maximum Throughput results by block size for the three system types for all four workloads. The number of LUNs used per test run was steadily increased from 1 to the maximum (64, 128, 256), and the peak results achieved during those steps are reported in this table.

Sequential I/O cache miss standard LDEVs	USP-V 2+2	256-1024 HDD 146GB 15k 1-256 LUNs	2 FED 2 BED	4 FED 4 BED	8 FED 8 BED
Read %	Block	Thread	MB/sec	MB/sec	MB/sec
100	64k	1	3,768	5,606	10,143
100	128k	1	3,972	5,403	11,484
100	256k	1	3,705	5,893	10,841
100	512k	1	3,048	4,958	9,434
		peak	3,972	5,893	11,484
		average	3,623	5,465	10,475
75	64k	1	2,239	4,045	7,667
75	128k	1	2,643	4,451	8,298
75	256k	1	2,618	4,667	8,651
75	512k	1	2,590	4,550	8,225
		peak	2,643	4,667	8,651
		average	2,522	4,428	8,210
50	64k	1	1,636	2,985	5,695
50	128k	1	1,875	3,311	6,346
50	256k	1	2,107	3,550	6,626
50	512k	1	2,047	3,562	6,528
		peak	2,107	3,562	6,626
		average	1,916	3,352	6,299
0	64k	1	1,660	2,965	5,827
0	128k	1	1,705	2,938	5,858
0	256k	1	1,668	2,986	5,841
0	512k	1	1,706	2,964	5,788
		peak	1,706	2,986	5,858
		average	1,685	2,963	5,828

Table 15. Summary of all peak Sequential results.

These results show that the USP V model can deliver far more performance than a USP when using Random workloads, and somewhat higher performance than a USP when using Sequential workloads. The USP is already able to considerably surpass all competitive storage systems in sequential throughput, typically achieving 3-4 times higher performance. The “small, midsize, and large” USP V configurations would be compared to a USP100, USP600, and USP1100. The two tables below show the USP V scalability ratios by configuration.

Random I/O cache miss standard LDEVs Read %	USP-V linearity 2+2 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD TPS ratio	512 HDD TPS ratio	1024 HDD TPS ratio
100	64k	1	1	1.93	3.56
100	128k	1	1	1.92	3.56
100	256k	1	1	1.93	3.56
100	512k	1	1	1.93	3.58
75	64k	1	1	1.96	3.84
75	128k	1	1	2.06	4.02
75	256k	1	1	2.19	4.38
75	512k	1	1	1.94	4.15
50	64k	1	1	1.81	3.38
50	128k	1	1	1.97	3.96
50	256k	1	1	1.97	3.78
50	512k	1	1	1.96	3.72
0	64k	1	1	1.94	3.78
0	128k	1	1	1.95	3.90
0	256k	1	1	1.98	3.97
0	512k	1	1	1.97	3.99

Table 16. Random workload scalability ratios by USP-V model.

Sequential I/O cache miss standard LDEVs Read %	USP-V linearity 2+2 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD MB/sec	512 HDD MB/sec	1024 HDD MB/sec
100	64k	1	1	1.49	2.69
100	128k	1	1	1.36	2.89
100	256k	1	1	1.59	2.93
100	512k	1	1	1.63	3.10
75	64k	1	1	1.81	1.90
75	128k	1	1	1.68	1.86
75	256k	1	1	1.78	1.85
75	512k	1	1	1.76	1.81
50	64k	1	1	1.82	1.91
50	128k	1	1	1.77	1.92
50	256k	1	1	1.68	1.87
50	512k	1	1	1.74	1.83
0	64k	1	1	1.79	1.97
0	128k	1	1	1.72	1.99
0	256k	1	1	1.79	1.96
0	512k	1	1	1.74	1.95

Table 17. Sequential workload scalability ratios by USP-V model.

The next two tables compare the percent differences of Random and Sequential results for the USP V to the USP for the same tests using the same configurations and tests. These differences are calculated by using the standard $[A-B]/B$ method, where A is the USP V result and B is the USP result. A positive TPS result means the USP V is showing higher performance than the USP for that workload. A negative RT percentage means the USP V has a lower response time than the USP for that workload.

Random I/O cache miss standard LDEVs Read %	product comparison 2+2 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD		512 HDD		1024 HDD	
			ratios		ratios		ratios	
			TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	8	85%	-46%	77%	-43%	80%	-44%
100	4k	8	84%	-46%	76%	-43%	80%	-44%
100	8k	8	82%	-45%	75%	-43%	78%	-44%
100	16k	8	79%	-44%	72%	-42%	75%	-43%
75	2k	8	121%	-55%	129%	-56%	139%	-58%
75	4k	8	110%	-52%	130%	-57%	138%	-58%
75	8k	8	91%	-48%	121%	-55%	136%	-58%
75	16k	8	89%	-47%	94%	-49%	122%	-55%
50	2k	8	102%	-50%	85%	-46%	82%	-45%
50	4k	8	82%	-45%	82%	-45%	93%	-48%
50	8k	8	82%	10%	81%	-45%	84%	-45%
50	16k	8	80%	11%	79%	-44%	79%	-44%
0	2k	8	76%	14%	75%	14%	79%	12%
0	4k	8	73%	16%	72%	16%	81%	-45%
0	8k	8	71%	17%	73%	16%	83%	9%
0	16k	8	70%	17%	73%	16%	86%	-46%

Table 18. Percent delta, Random workloads, USP-V to USP.

Sequential I/O cache miss standard LDEVs Read %	product comparison 2+2 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD	512 HDD	1024 HDD
			MB/sec	MB/sec	MB/sec
			ratio	ratio	ratio
100	64k	1	43%	-7%	8%
100	128k	1	41%	-10%	23%
100	256k	1	28%	4%	15%
100	512k	1	3%	-15%	0%
50	64k	1	40%	40%	48%
50	128k	1	37%	29%	37%
50	256k	1	48%	33%	41%
50	512k	1	44%	34%	38%
0	64k	1	28%	26%	44%
0	128k	1	32%	25%	45%
0	256k	1	28%	27%	44%
0	512k	1	32%	26%	43%

Table 19. Percent delta, Sequential workloads, USP-V to USP.

X. RAID5 (3D+1P) Back-end Scalability Tests

The purpose of this testing was to conduct a back-end disk (cache miss) scalability analysis using the Hitachi Data Systems Universal Storage Platform V enterprise storage system. The quantity of Front-end Directors (FEDs) and Back-end Directors (BEDs) used were varied to scale the system size. The first set of tests evaluated how a **small** USP V (2 FED, 2 BED, and 256 HDDs) could perform. The second set of tests measured a **midsized** USP V (4 FEDs, 4 BEDs, and 512 HDDs). The final set of tests measured a **large** USP V (8 FEDs, 8 BEDs, and 1024 HDDs). A fairly powerful IBM p595 server was used for these tests.

There was a single test system that had 256GB of cache and 22 GB of Shared Memory. The disks used were 146GB 15K RPM. The front-end and back-end paths are all 4Gbit FC. The test configurations used for these three sets of tests included:

- **USP V small**
 - o 2 FED – 16 unshared FC host ports
 - o 2 BED – 256 HDDs (8 pairs of disk loops, 64 Array Groups, 64 LUNs)
- **USP V midsize**
 - o 4 FEDs – 32 unshared FC host ports
 - o 4 BEDs – 512 HDDs (16 pairs of disk loops, 128 Array Groups, 128 LUNs)
- **USP V large**
 - o 8 FEDs – 64 unshared FC host ports
 - o 8 BEDs – 1024 HDDs (32 pairs of disk loops, 256 Array Groups, 256 LUNs)

Test Methodology

The test methodology was to first configure 256 Array Groups in the USP V using RAID5 3D+1P, and to create one 33GB LDEV per RAID Group. These LDEVs were then mapped to FED ports. For each USP V “model”, only those LDEVs that were behind a certain set of BEDs were accessed for those tests. A specific pair of FEDs was associated with each pair of BEDs, so that certain host ports always saw certain LUNs. Hence, there were always four LUNS per port. The “small” system used the first 16 host ports and 64 LUNs. The “midsize” system used those plus the next group of 16 ports and 64 LUNs. The “large” system used all four sets of ports and LUNs. Each test run for a system began with one LUN and scaled up to the maximum for that system. There were 8 concurrent VDBench threads per LUN for Random workloads, and just 1 thread per LUN for Sequential workloads.

These are the tests that were used:

- **Random I/O, Transactions per second and Average Response Time**, at 2, 4, 8, and 16KB block sizes with 8 Threads (100% - 75% - 50% Read, and 100% Write)
- **Sequential I/O, Throughput**, at 64, 128, 256, and 512KB block sizes with 1 Thread (100% - 75% - 50% Read, and 100% Write)

Random I/O Transactions

The following table shows the peak Random workload results for the RAID5 3D+1P scalability tests. The number of LUNs used per test run was steadily increased from 1 to the maximum (64, 128, 256), and the peak results achieved during those steps are reported in this table.

Random I/O	USP-V	256-1024 HDD	2 FED 2 BED		4 FED 4 BED		8 FED 8 BED	
cache miss	3+1	146GB 15k	16 host paths		32 host paths		32 host paths	
standard LDEVs		1-256 LUNs	16 disk loops		32 disk loops		32 disk loops	
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	peak average	45,152	11.3	88,224	11.6	164,919	12.4
100	4k		45,002	11.4	87,632	11.7	164,625	12.4
100	8k		44,571	11.5	87,031	11.8	163,182	12.5
100	16k		43,686	11.7	85,254	12.0	160,415	12.8
			45,152	11.7	88,224	12.0	164,919	12.8
			44,603	11.5	87,035	11.8	163,285	12.5
75	2k	peak average	29,572	17.5	55,598	18.4	107,543	19.1
75	4k		28,180	18.2	55,639	18.4	107,851	19.0
75	8k		28,137	18.2	55,365	18.5	107,357	19.1
75	16k		27,927	18.3	54,974	18.6	106,521	19.2
			29,572	18.3	55,639	18.6	107,851	19.2
			28,454	18.1	55,394	18.5	107,318	19.1
50	2k	peak average	17,741	28.9	34,942	29.3	67,801	30.2
50	4k		17,730	28.9	34,823	29.4	67,841	30.2
50	8k		17,710	28.9	34,812	29.4	67,584	30.3
50	16k		17,566	29.1	34,610	29.6	67,233	30.5
			17,741	29.1	34,942	29.6	67,841	30.5
			17,687	28.9	34,797	29.4	67,615	30.3
0	2k	peak average	10,477	48.9	20,517	49.9	39,677	51.6
0	4k		10,448	49.0	20,510	49.9	39,680	51.6
0	8k		10,482	48.9	20,456	50.1	39,631	51.7
0	16k		10,387	49.3	20,429	50.1	39,480	51.9
			10,482	49.3	20,517	50.1	39,680	51.9
			10,449	49.0	20,478	50.0	39,617	51.7

Table 20. Summary of all peak Random results.

Sequential I/O Throughput

The following table shows the maximum Throughput results by block size for the three system types for all four workloads. The number of LUNs used per test run was steadily increased from 1 to the maximum (64, 128, 256), and the peak results achieved during those steps are reported in this table.

Sequential I/O cache miss standard LDEVs	USP-V 3+1	256-1024 HDD 146GB 15k 1-256 LUNs	2 FED 2 BED	4 FED 4 BED	8 FED 8 BED
Read %	Block	Thread	MB/sec	MB/sec	MB/sec
100	64k	1	3,682	5,611	10,270
100	128k	1	3,961	5,412	11,423
100	256k	1	3,695	5,687	10,651
100	512k	1	3,109	4,927	9,395
		peak	3,961	5,687	11,423
		average	3,612	5,409	10,435
75	64k	1	1,227	2,044	3,778
75	128k	1	1,404	2,367	4,365
75	256k	1	1,629	3,139	5,818
75	512k	1	1,849	3,169	5,913
		peak	1,849	3,169	5,913
		average	1,527	2,680	4,968
50	64k	1	817	1,446	2,642
50	128k	1	1,061	1,756	3,251
50	256k	1	1,406	2,435	4,555
50	512k	1	1,407	2,451	4,538
		peak	1,407	2,451	4,555
		average	1,173	2,022	3,746
0	64k	1	2,006	2,935	4,609
0	128k	1	1,999	2,966	4,613
0	256k	1	1,985	2,962	4,612
0	512k	1	1,941	2,945	4,592
		peak	2,006	2,966	4,613
		average	1,983	2,952	4,607

Table 21. Summary of all peak Sequential results.

These results show that the USP V model can deliver far more performance than a USP when using Random workloads, and somewhat higher performance than a USP when using Sequential workloads. The USP is already able to considerably surpass all competitive storage systems in sequential throughput, typically achieving 3-4 times higher performance. The “small, midsize, and large” USP V configurations would be compared to a USP100, USP600, and USP1100. The two tables below show the USP V scalability ratios by configuration.

Random I/O cache miss standard LDEVs Read %	USP-V linearity 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD TPS ratio	512 HDD TPS ratio	1024 HDD TPS ratio
100	64k	1	1	1.95	3.65
100	128k	1	1	1.95	3.66
100	256k	1	1	1.95	3.66
100	512k	1	1	1.95	3.67
75	64k	1	1	1.88	3.64
75	128k	1	1	1.97	3.83
75	256k	1	1	1.97	3.82
75	512k	1	1	1.97	3.81
50	64k	1	1	1.97	3.82
50	128k	1	1	1.96	3.83
50	256k	1	1	1.97	3.82
50	512k	1	1	1.97	3.83
0	64k	1	1	1.96	3.79
0	128k	1	1	1.96	3.80
0	256k	1	1	1.95	3.78
0	512k	1	1	1.97	3.80

Table 22. Random workload scalability ratios by USP-V model.

Sequential I/O cache miss standard LDEVs Read %	USP-V linearity 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD MB/sec ratio	512 HDD MB/sec ratio	1024 HDD MB/sec ratio
100	64k	1	1	1.52	2.79
100	128k	1	1	1.37	2.88
100	256k	1	1	1.54	2.88
100	512k	1	1	1.58	3.02
75	64k	1	1	1.67	1.85
75	128k	1	1	1.69	1.84
75	256k	1	1	1.93	1.85
75	512k	1	1	1.71	1.87
50	64k	1	1	1.77	1.83
50	128k	1	1	1.66	1.85
50	256k	1	1	1.73	1.87
50	512k	1	1	1.74	1.85
0	64k	1	1	1.46	1.57
0	128k	1	1	1.48	1.56
0	256k	1	1	1.49	1.56
0	512k	1	1	1.52	1.56

Table 23. Sequential workload scalability ratios by USP-V model.

The next two tables compare the percent differences of Random and Sequential results for the USP V to the USP for the same tests using the same configurations and tests. These differences are calculated by using the standard $[A-B]/B$ method, where A is the USP V result and B is the USP result. A positive TPS result means the USP V is showing higher performance than the USP

for that workload. A negative RT percentage means the USP V has a lower response time than the USP for that workload.

Random I/O cache miss standard LDEVs Read %	product comparison 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD		512 HDD		1024 HDD	
			ratios		ratios		ratios	
			TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	8	92%	-49%	67%	-41%	67%	-40%
100	4k	8	94%	-51%	66%	-40%	67%	-40%
100	8k	8	93%	-50%	65%	-40%	65%	-39%
100	16k	8	89%	-49%	61%	-38%	63%	-39%
75	2k	8	119%	-10%	126%	-56%	129%	-56%
75	4k	8	108%	-7%	127%	-56%	128%	-57%
75	8k	8	109%	-7%	126%	-56%	128%	-56%
75	16k	8	107%	-6%	124%	-56%	127%	-56%
50	2k	8	116%	84%	120%	-54%	124%	-55%
50	4k	8	117%	84%	120%	-54%	124%	-56%
50	8k	8	116%	85%	119%	-54%	123%	-55%
50	16k	8	114%	86%	118%	-54%	122%	-55%
0	2k	8	112%	88%	116%	84%	125%	-56%
0	4k	8	111%	89%	117%	84%	126%	-56%
0	8k	8	112%	89%	116%	85%	126%	-56%
0	16k	8	111%	90%	116%	84%	125%	-56%

Table 24. Percent delta, Random workloads, USP-V to USP.

Sequential I/O cache miss standard LDEVs Read %	product comparison 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD	512 HDD	1024 HDD
			MB/sec	MB/sec	MB/sec
			ratio	ratio	ratio
100	64k	1	25%	2%	9%
100	128k	1	37%	-9%	21%
100	256k	1	26%	1%	12%
100	512k	1	5%	-15%	0%
50	64k	1	-4%	16%	41%
50	128k	1	7%	9%	38%
50	256k	1	22%	27%	56%
50	512k	1	13%	23%	51%
0	64k	1	18%	21%	33%
0	128k	1	19%	24%	33%
0	256k	1	21%	24%	33%
0	512k	1	20%	23%	33%

Table 25. Percent delta, Sequential workloads, USP-V to USP.

XI. RAID5 (7D+1P) Back-end Scalability Tests

The purpose of this testing was to conduct a back-end disk (cache miss) scalability analysis using the Hitachi Data Systems Universal Storage Platform V enterprise storage system. The quantity of Front-end Directors (FEDs) and Back-end Directors (BEDs) used were varied to scale the system size. The first set of tests evaluated how a **small** USP V (2 FED, 2 BED, and 256 HDDs) could perform. The second set of tests measured a **midsized** USP V (4 FEDs, 4 BEDs, and 512 HDDs). The final set of tests measured a **large** USP V (8 FEDs, 8 BEDs, and 1024 HDDs). A fairly powerful IBM p595 server was used for these tests.

There was a single test system that had 256GB of cache and 22 GB of Shared Memory. The disks used were 146GB 15K RPM. The front-end and back-end paths are all 4Gbit FC. The test configurations used for these three sets of tests included:

- **USP V small**
 - o 2 FED – 16 unshared FC host ports
 - o 2 BED – 256 HDDs (8 pairs of disk loops, 32 Array Groups, 32 LUNs)
- **USP V midsize**
 - o 4 FEDs – 32 unshared FC host ports
 - o 4 BEDs – 512 HDDs (16 pairs of disk loops, 64 Array Groups, 64 LUNs)
- **USP V large**
 - o 8 FEDs – 64 unshared FC host ports
 - o 8 BEDs – 1024 HDDs (32 pairs of disk loops, 128 Array Groups, 128 LUNs)

Test Methodology

The test methodology was to first configure 128 Array Groups in the USP V using RAID5 7D+1P, and to create one 33GB LDEV per RAID Group. These LDEVs were then mapped to FED ports. For each USP V “model”, only those LDEVs that were behind a certain set of BEDs were accessed for those tests. A specific pair of FEDs was associated with each pair of BEDs, so that certain host ports always saw certain LUNs. Hence, there were always four LUNS per port. The “small” system used the first 16 host ports and 32 LUNs. The “midsize” system used 32 ports and 64 LUNs. The “large” system used all 128 ports and 128 LUNs. Each test run for a system began with one LUN and scaled up to the maximum for that system. There were 8 concurrent VDBench threads per LUN for Random workloads, and just 1 thread per LUN for Sequential workloads.

These are the tests that were used:

- **Random I/O, Transactions per second and Average Response Time**, at 2, 4, 8, and 16KB block sizes with 8 Threads (100% - 75% - 50% Read, and 100% Write)
- **Sequential I/O, Throughput**, at 64, 128, 256, and 512KB block sizes with 1 Thread (100% - 75% - 50% Read, and 100% Write)

Random I/O Transactions

The following table and charts show the peak Random workload results for the RAID5 7D+1P scalability tests. The number of LUNs used per test run was steadily increased from 1 to the maximum (32, 64, 128), and the peak results achieved during those steps are reported in this table.

Random I/O	USP-V	256-1024 HDD	2 FED 2 BED		4 FED 4 BED		8 FED 8 BED	
cache miss	7+1	146GB 15k	16 host paths		32 host paths		32 host paths	
standard LDEVs		1-128 LUNs	16 disk loops		32 disk loops		32 disk loops	
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	16	44,080	11.6	85,999	11.9	161,533	12.7
100	4k	16	43,956	11.6	85,766	11.9	161,100	12.7
100	8k	16	43,493	11.8	84,766	12.1	159,624	12.8
100	16k	16	42,571	12.0	83,069	12.3	156,941	13.0
peak			44,080	12.0	85,999	12.3	161,533	13.0
average			43,525	11.8	84,900	12.1	159,800	12.8
75	2k	16	28,343	18.3	53,662	19.1	103,947	19.7
75	4k	16	27,269	18.8	53,748	19.0	104,207	19.6
75	8k	16	27,235	18.8	53,573	19.1	103,795	19.7
75	16k	16	27,071	18.9	53,280	19.2	103,320	19.8
peak			28,343	18.9	53,748	19.2	104,207	19.8
average			27,479	18.7	53,566	19.1	103,817	19.7
50	2k	16	16,993	30.1	33,575	30.5	65,479	31.3
50	4k	16	17,091	29.9	33,686	30.4	65,467	31.3
50	8k	16	17,042	30.1	33,613	30.5	65,327	31.3
50	16k	16	16,962	30.2	33,388	30.7	65,075	31.5
peak			17,091	30.2	33,686	30.7	65,479	31.5
average			17,022	30.1	33,566	30.5	65,337	31.3
0	2k	16	10,135	50.5	19,969	51.3	38,544	53.1
0	4k	16	10,243	50.0	19,972	51.3	38,593	53.1
0	8k	16	10,176	50.3	19,965	51.3	38,547	53.1
0	16k	16	10,179	50.3	19,872	51.5	38,465	53.2
peak			10,243	50.5	19,972	51.5	38,593	53.2
average			10,183	50.3	19,944	51.4	38,537	53.1

Table 26. Summary of all peak random results.

Sequential I/O Throughput

The following table and charts show the maximum Throughput results by block size for the three system types for all four workloads. The number of LUNs used per test run was steadily increased from 1 to the maximum (32, 64, 128), and the peak results achieved during those steps are reported in this table.

Sequential I/O cache miss standard LDEVs	USP-V 7+1	256-1024 HDD 146GB 15k 1-128 LUNs	2 FED 2 BED MB/sec	4 FED 4 BED MB/sec	8 FED 8 BED MB/sec
Read %	Block	Thread			
100	64k	2	3,718	5,305	9,413
100	128k	2	3,976	5,523	10,328
100	256k	2	3,751	4,898	9,046
100	512k	2	2,942	5,219	9,346
		peak	3,976	5,523	10,328
		average	3,597	5,236	9,533
75	64k	2	1,279	2,042	3,801
75	128k	2	1,364	2,375	4,412
75	256k	2	1,706	3,114	6,074
75	512k	2	1,930	3,342	6,212
		peak	1,930	3,342	6,212
		average	1,570	2,719	5,125
50	64k	2	883	1,542	2,880
50	128k	2	1,087	1,869	3,474
50	256k	2	1,555	2,685	4,959
50	512k	2	1,649	2,727	4,954
		peak	1,649	2,727	4,959
		average	1,294	2,206	4,067
0	64k	2	1,856	2,889	5,306
0	128k	2	1,810	2,860	5,302
0	256k	2	2,304	3,198	5,289
0	512k	2	2,339	3,220	5,299
		peak	2,339	3,220	5,306
		average	2,077	3,042	5,299

Table 27. Summary of all peak Sequential results.

These results show that the USP V models, when using RAID5 7D+1P Random workloads, scale quite nicely. However, there are currently some issues with the Test Microcode with scaling in the case of Sequential workloads, particularly when there is a significant Write component. The Sequential values don't match the Random scaling factors. The two tables below report the scalability ratios by configuration.

Random I/O cache miss standard LDEVs Read %	USP-V linearity 7+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD TPS ratio	512 HDD TPS ratio	1024 HDD TPS ratio
100	64k	1	1	1.95	3.66
100	128k	1	1	1.95	3.67
100	256k	1	1	1.95	3.67
100	512k	1	1	1.95	3.69
75	64k	1	1	1.89	3.67
75	128k	1	1	1.97	3.82
75	256k	1	1	1.97	3.81
75	512k	1	1	1.97	3.82
50	64k	1	1	1.98	3.85
50	128k	1	1	1.97	3.83
50	256k	1	1	1.97	3.83
50	512k	1	1	1.97	3.84
0	64k	1	1	1.97	3.80
0	128k	1	1	1.95	3.77
0	256k	1	1	1.96	3.79
0	512k	1	1	1.95	3.78

Table 28. Random workload 7+1 scalability ratios by USP-V model.

Sequential I/O cache miss standard LDEVs Read %	USP-V linearity 7+1 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	256 HDD MB/sec ratio	512 HDD MB/sec ratio	1024 HDD MB/sec ratio
100	64k	1	1	1.43	2.53
100	128k	1	1	1.39	2.60
100	256k	1	1	1.31	2.41
100	512k	1	1	1.77	3.18
75	64k	1	1	1.60	1.86
75	128k	1	1	1.74	1.86
75	256k	1	1	1.82	1.95
75	512k	1	1	1.73	1.86
50	64k	1	1	1.74	1.87
50	128k	1	1	1.72	1.86
50	256k	1	1	1.73	1.85
50	512k	1	1	1.65	1.82
0	64k	1	1	1.56	1.84
0	128k	1	1	1.58	1.85
0	256k	1	1	1.39	1.65
0	512k	1	1	1.38	1.65

Table 29. Sequential workload scalability ratios by USP-V model.

The next two tables compare the percent differences of results for the USP V when using RAID5 7D+1P versus RAID5 3D+1P and equivalent configurations. These are calculated by using the standard [A-B]/B method, where A is the 7D+1P result and B is the 3D+1P result. As you can see, the 7D+1P Random results are uniformly slower by just 2-4%. In the case of Sequential, the 7D+1P results are generally faster, with some random exceptions.

Random I/O cache miss standard LDEVs Read %	RAID5 comparison 7+1 : 3+1 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	256 HDD		512 HDD		1024 HDD	
			ratios		ratios		ratios	
			TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	8	-2%	2%	-3%	3%	-2%	2%
100	4k	8	-2%	2%	-2%	2%	-2%	2%
100	8k	8	-2%	2%	-3%	3%	-2%	2%
100	16k	8	-3%	3%	-3%	3%	-2%	2%
75	2k	8	-4%	4%	-3%	4%	-3%	3%
75	4k	8	-3%	3%	-3%	3%	-3%	3%
75	8k	8	-3%	3%	-3%	3%	-3%	3%
75	16k	8	-3%	3%	-3%	3%	-3%	3%
50	2k	8	-4%	4%	-4%	4%	-3%	3%
50	4k	8	-4%	4%	-3%	3%	-4%	4%
50	8k	8	-4%	4%	-3%	4%	-3%	3%
50	16k	8	-3%	4%	-4%	4%	-3%	3%
0	2k	8	-3%	3%	-3%	3%	-3%	3%
0	4k	8	-2%	2%	-3%	3%	-3%	3%
0	8k	8	-3%	3%	-2%	2%	-3%	3%
0	16k	8	-2%	2%	-3%	3%	-3%	3%

Table 30. Percent delta, Random workloads, USP-V, 7D+1P to 3D+1P.

Sequential I/O cache miss standard LDEVs Read %	RAID5 comparison 7+1 : 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD	512 HDD	1024 HDD
			MB/sec	MB/sec	MB/sec
			ratio	ratio	ratio
100	64k	1	1%	-5%	-8%
100	128k	1	0%	2%	-10%
100	256k	1	2%	-14%	-15%
100	512k	1	-5%	6%	-1%
75	64k	1	4%	0%	1%
75	128k	1	-3%	0%	1%
75	256k	1	5%	-1%	4%
75	512k	1	4%	5%	5%
50	64k	1	8%	7%	9%
50	128k	1	2%	6%	7%
50	256k	1	11%	10%	9%
50	512k	1	17%	11%	9%
0	64k	1	-7%	-2%	15%
0	128k	1	-9%	-4%	15%
0	256k	1	16%	8%	15%
0	512k	1	21%	9%	15%

Table 31. Percent delta, Sequential workloads, USP-V, 7D+1P to 3D+1P.

XII. RAID6 (6D+2P) Back-end Scalability Tests

The purpose of this testing was to conduct a back-end disk (cache miss) scalability analysis using the Hitachi Data Systems Universal Storage Platform V enterprise storage system. The quantity of Front-end Directors (FEDs) and Back-end Directors (BEDs) used were varied to scale the system size. The first set of tests evaluated how a **small** USP V (2 FED, 2 BED, and 256 HDDs) could perform. The second set of tests measured a **midsized** USP V (4 FEDs, 4 BEDs, and 512 HDDs). The final set of tests measured a **large** USP V (8 FEDs, 8 BEDs, and 1024 HDDs). A fairly powerful IBM p595 server was used for these tests.

There was a single test system that had 256GB of cache and 22 GB of Shared Memory. The disks used were 146GB 15K RPM. The front-end and back-end paths are all 4Gbit FC. The test configurations used for these three sets of tests included:

- **USP V small**
 - o 2 FED – 16 unshared FC host ports
 - o 2 BED – 256 HDDs (8 pairs of disk loops, 32 Array Groups, 32 LUNs)
- **USP V midsize**
 - o 4 FEDs – 32 unshared FC host ports
 - o 4 BEDs – 512 HDDs (16 pairs of disk loops, 64 Array Groups, 64 LUNs)
- **USP V large**
 - o 8 FEDs – 64 unshared FC host ports
 - o 8 BEDs – 1024 HDDs (32 pairs of disk loops, 128 Array Groups, 128 LUNs)

Test Methodology

The test methodology was to first configure 128 Array Groups in the USP V using RAID6 6D+2P, and to create one 33GB LDEV per RAID Group. These LDEVs were then mapped to FED ports. For each USP V “model”, only those LDEVs that were behind a certain set of BEDs were accessed for those tests. A specific pair of FEDs was associated with each pair of BEDs, so that certain host ports always saw certain LUNs. Hence, there were always four LUNS per port. The “small” system used 16 host ports and 32 LUNs. The “midsize” system used 32 ports and 64 LUNs. The “large” system used all 128 ports and 128 LUNs. Each test run for a system began with one LUN and scaled up to the maximum for that system. There were 8 concurrent VDBench threads per LUN for Random workloads, and just 1 thread per LUN for Sequential workloads.

These are the tests that were used:

- **Random I/O, Transactions per second and Average Response Time**, at 2, 4, 8, and 16KB block sizes with 8 Threads (100% - 75% - 50% Read, and 100% Write)
- **Sequential I/O, Throughput**, at 64, 128, 256, and 512KB block sizes with 1 Thread (100% - 75% - 50% Read, and 100% Write)

Random I/O Transactions

The following table and charts show the peak Random workload results for the RAID6 6D+2P scalability tests. The number of LUNs used per test run was steadily increased from 1 to the maximum (32, 64, 128), and the peak results achieved during those steps are reported in this table.

Random I/O	USP-V	256-1024 HDD	2 FED 2 BED		4 FED 4 BED		8 FED 8 BED	
cache miss	6+2	146GB 15k	16 host paths		32 host paths		32 host paths	
standard LDEVs		1-128 LUNs	16 disk loops		32 disk loops		32 disk loops	
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	16	44,092	11.6	85,983	11.9	161,614	12.7
100	4k	16	44,024	11.6	85,813	11.9	161,279	12.7
100	8k	16	43,559	11.8	84,873	12.1	159,760	12.8
100	16k	16	42,695	12.0	83,283	12.3	156,980	13.0
peak			44,092	12.0	85,983	12.3	161,614	13.0
average			43,593	11.7	84,988	12.0	159,908	12.8
75	2k	16	17,493	29.3	36,470	28.1	70,996	28.9
75	4k	16	17,888	28.7	37,160	27.9	70,759	29.0
75	8k	16	17,811	28.8	35,609	29.0	70,483	29.1
75	16k	16	17,790	28.8	36,145	28.6	70,122	29.2
peak			17,888	29.3	37,160	29.0	70,996	29.2
average			17,745	28.9	36,346	28.4	70,590	29.0
50	2k	16	10,451	49.0	20,366	50.3	40,859	50.2
50	4k	16	10,386	49.3	20,143	50.8	41,050	50.0
50	8k	16	10,353	49.5	20,042	51.1	40,760	50.3
50	16k	16	10,362	49.4	19,922	51.4	39,305	52.1
peak			10,451	49.5	20,366	51.4	41,050	52.1
average			10,388	49.3	20,118	50.9	40,494	50.6
0	2k	16	5,877	88.1	11,401	89.9	22,246	92.1
0	4k	16	6,234	82.4	11,601	88.3	23,076	88.8
0	8k	16	5,961	86.0	11,438	89.6	22,247	92.1
0	16k	16	5,980	85.7	11,440	89.6	21,891	93.7
peak			6,234	88.1	11,601	89.9	23,076	93.7
average			6,013	85.6	11,470	89.3	22,365	91.7

Table 32. Summary of all 6D+2P peak random results.

Sequential I/O Throughput

The following table and charts show the maximum Throughput results by block size for the three system types for all four workloads. The number of LUNs used per test run was steadily increased from 1 to the maximum (32, 64, 128), and the peak results achieved during those steps are reported in this table.

Sequential I/O cache miss standard LDEVs Read %	USP-V 6+2 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	2 FED 2 BED MB/sec	4 FED 4 BED MB/sec	8 FED 8 BED MB/sec
100	64k	2	3,713	5,337	9,435
100	128k	2	3,971	5,519	10,328
100	256k	2	3,718	4,766	9,058
100	512k	2	2,906	5,253	9,518
		peak	3,971	5,519	10,328
		average	3,577	5,218	9,585
75	64k	2	1,114	1,801	2,839
75	128k	2	1,271	2,086	3,339
75	256k	2	1,551	2,711	5,026
75	512k	2	1,850	2,953	5,092
		peak	1,850	2,953	5,092
		average	1,446	2,388	4,074
50	64k	2	854	1,396	2,369
50	128k	2	1,044	1,704	2,795
50	256k	2	1,397	2,415	4,103
50	512k	2	1,581	2,494	4,028
		peak	1,581	2,494	4,103
		average	1,219	2,003	3,324
0	64k	2	1,574	2,509	4,577
0	128k	2	1,529	2,469	4,588
0	256k	2	1,997	2,787	4,579
0	512k	2	2,007	2,808	4,584
		peak	2,007	2,808	4,588
		average	1,777	2,643	4,582

Table 33. Summary of all peak Sequential results.

These results show that the USP V models, when using RAID6 6d+2 Random workloads, scale extremely well. However, there are currently some issues with the Test Microcode with scaling in the case of Sequential workloads, particularly when there is a significant Write component. The Sequential values don't match the Random scaling factors. The two tables below report the scalability ratios by configuration.

Random I/O cache miss standard LDEVs Read %	USP-V linearity 6+2 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	256 HDD TPS ratio	512 HDD TPS ratio	1024 HDD TPS ratio
100	64k	1	1	1.95	3.67
100	128k	1	1	1.95	3.66
100	256k	1	1	1.95	3.67
100	512k	1	1	1.95	3.68
75	64k	1	1	2.08	4.06
75	128k	1	1	2.08	3.96
75	256k	1	1	2.00	3.96
75	512k	1	1	2.03	3.94
50	64k	1	1	1.95	3.91
50	128k	1	1	1.94	3.95
50	256k	1	1	1.94	3.94
50	512k	1	1	1.92	3.79
0	64k	1	1	1.94	3.79
0	128k	1	1	1.86	3.70
0	256k	1	1	1.92	3.73
0	512k	1	1	1.91	3.66

Table 34. Random workload 6+2 scalability ratios by USP-V model.

Sequential I/O cache miss standard LDEVs Read %	USP-V linearity 6+2 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	256 HDD MB/sec ratio	512 HDD MB/sec ratio	1024 HDD MB/sec ratio
100	64k	1	1	1.44	2.54
100	128k	1	1	1.39	2.60
100	256k	1	1	1.28	2.44
100	512k	1	1	1.81	3.28
75	64k	1	1	1.62	1.58
75	128k	1	1	1.64	1.60
75	256k	1	1	1.75	1.85
75	512k	1	1	1.60	1.72
50	64k	1	1	1.63	1.70
50	128k	1	1	1.63	1.64
50	256k	1	1	1.73	1.70
50	512k	1	1	1.58	1.61
0	64k	1	1	1.59	1.82
0	128k	1	1	1.62	1.86
0	256k	1	1	1.40	1.64
0	512k	1	1	1.40	1.63

Table 35. Sequential workload 6+2 scalability ratios by USP-V model.

The next two tables compare the percent differences of results for the USP V when using RAID6 6D+2P versus RAID5 7D+1P and equivalent configurations. These are calculated by using the standard [A-B]/B method, where A is the 6D+2P result and B is the 7D+1P result. As you can see, the 6D+2P Random results are uniformly and considerably slower (31-43% for TPS) except for the 100% Random Read workloads. In the case of Sequential, the 6D+2P results are only

somewhat slower than 7D+1P overall, except for the 100% Read workloads. This is the usual behavior of RAID6 and Sequential workloads.

Random I/O cache miss standard LDEVs Read %	RAID comparison 6+2 : 7+1 Block	256-1024 HDD 146GB 15k 1-128 LUNs Thread	256 HDD		512 HDD		1024 HDD	
			ratios		ratios		ratios	
			TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100	2k	8	0%	0%	0%	0%	0%	0%
100	4k	8	0%	0%	0%	0%	0%	0%
100	8k	8	0%	0%	0%	0%	0%	0%
100	16k	8	0%	0%	0%	0%	0%	0%
75	2k	8	-38%	61%	-32%	48%	-32%	46%
75	4k	8	-34%	53%	-31%	47%	-32%	47%
75	8k	8	-35%	53%	-34%	52%	-32%	47%
75	16k	8	-34%	52%	-32%	49%	-32%	47%
50	2k	8	-38%	63%	-39%	65%	-38%	60%
50	4k	8	-39%	65%	-40%	67%	-37%	60%
50	8k	8	-39%	65%	-40%	68%	-38%	61%
50	16k	8	-39%	64%	-40%	68%	-40%	66%
0	2k	8	-42%	74%	-43%	75%	-42%	73%
0	4k	8	-39%	65%	-42%	72%	-40%	67%
0	8k	8	-41%	71%	-43%	75%	-42%	73%
0	16k	8	-41%	70%	-42%	74%	-43%	76%

Table 36. Percent delta, Random workloads, USP-V, 6D+2P to 7D+1P.

Sequential I/O cache miss standard LDEVs Read %	RAID comparison 6+2 : 7+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD	512 HDD	1024 HDD
			MB/sec	MB/sec	MB/sec
			ratio	ratio	ratio
100	64k	1	0%	1%	0%
100	128k	1	0%	0%	0%
100	256k	1	-1%	-3%	0%
100	512k	1	-1%	1%	2%
75	64k	1	-13%	-12%	-25%
75	128k	1	-7%	-12%	-24%
75	256k	1	-9%	-13%	-17%
75	512k	1	-4%	-12%	-18%
50	64k	1	-3%	-9%	-18%
50	128k	1	-4%	-9%	-20%
50	256k	1	-10%	-10%	-17%
50	512k	1	-4%	-9%	-19%
0	64k	1	-15%	-13%	-14%
0	128k	1	-16%	-14%	-13%
0	256k	1	-13%	-13%	-13%
0	512k	1	-14%	-13%	-13%

Table 37. Percent delta, Sequential workloads, USP-V, 6D+2P to 7D+1P.

XIII. RAID10 (2D+2D) AMS500 Virtualization Tests

Universal Volume Manager

The purpose of the USP V's **UVM** virtualization feature is to provide for the attachment of external storage (from Hitachi or a variety of other storage vendors) to a USP V and to manage these external volumes as if they originated within the USP V. Those volumes appear to hosts attached to the USP V as logical devices originating from within the USP V. This is the basis of the Hitachi solution for *storage virtualization*. There is also a USP V software feature available for the logical partitioning of USP cache, disks, and host ports (**Virtual Partition Manager**) as well as a feature to non-disruptively manage data migration between local volumes and external volumes (**Tiered Storage Manager**).

There are two UVM cache management modes available when mapping the external storage LDEVs through to USP V host ports. These are ***Cached Mode Enabled*** and ***Cached Mode Disabled*** settings which are selectable on an individual LDEV basis. Note that these configuration settings are independent of those operating within the external storage system. The external systems will continue to operate as if they were still direct or SAN attached to servers. Also note that read or write operations from/to external storage will always pass through the USP V's cache system. The difference between these cache modes is in how the data blocks are managed.

Cache Mode Disabled

For random reads, the use of the cache management MRU/LRU (Most Recently Used/Least Recently Used) queues are disabled in "Cache Mode Disabled" so data blocks do not remain in cache for further reuse. In the case of random writes, data blocks are written to the USP cache and then out to external storage, just as with internal disks, but the write operation completion signal is not reported to the host until the data has been written to the external storage device's cache. For sequential reads, the prefetch function is not enabled.

Cache Mode Enabled

For random reads, the use of the cache management MRU/LRU queues and sequential prefetch are enabled. Enabling Cache Mode can benefit externally attached subsystems of older storage technologies that do not contain sophisticated cache routines. In the case of random writes, data blocks are written to the USP cache, a write operation completion signal is returned to the host, and then the data is written to external storage cache (seen as a disk volume by the USP) just as with internal disks.

For sequential read or write patterns, I/O's are never subject to the MRU/LRU system. These are handled in bursts, with a number of cache slots being allocated for the duration of the operation and placed back on the free queue when complete. For sequential reads, the prefetch function will read ahead from the storage device. Note that if the external storage system supports a similar sequential prefetch function then a similar and independent read ahead operation will likely occur within the external system.

Test Overview

The purpose of this testing was to conduct a port scalability performance analysis of external storage using an Adaptable Modular Storage 500 (AMS500) virtualized behind a Universal Storage Platform V (USP V). Various combinations of USP V 4Gbit ports and associated processors (MPs) were used between the USP V and the AMS500. Pairs of ports on the USP V share a common processor (MP). The single MP limits the simultaneous small block random workload performance on these two ports (perhaps to 55% each). Sequential workloads can show perhaps 70% of maximum single port rates across the shared MP.

There were four fixed paths from the USP V to the test server. Both “Cache Mode Enabled” and “Cache Mode Disabled” LDEV options (see below) were tested using the following port combinations:

- **1 Port – 1 MP**
- **2 Ports – 2 MPs**
- **4 Ports – 4 MPs**

The AMS500’s ports and LUNs were remapped on the USP V for each set of tests. There were 128 146GB 10K rpm disks used in the AMS for these tests. The RAID level used was RAID-10 (2D+2D). Both Random and Sequential workloads were used, with varying block sizes for each. These scalability tests used 1, 2, 4, 8, 16, and 32 RAID Groups on the AMS. There was one 72GB LUN created per RAID Group. These LUNs were mapped to USP V server ports using the three “port-MP” combinations shown above. The two sets of workloads were specified to achieve a 100% and a 0% cache hit ratio (100% seek per I/O).

Test Methodology

The USP V used a single 8-port Open Fibre 4Gbit FED PCB. The paths between the USP V and the server varied from 1-4 4Gbit paths. The AMS500 LUNs were remapped to the USP V ports for each test combination. Matching 1-4 4Gbit paths were used between the USP V external ports and the AMS500 for these tests. The external LUNs were tested with the Cache Mode enabled as well as disabled. The USP V-to-AMS500 FED ports were set up to use the combinations shown in Table 38, where the notation means X-ports with Y-processors (MPs).

Port-MP combinations	AMS RAID Groups	AMS LUNs Per Port	AMS controller ports	USP external ports	USP Host Ports (fixed)
1p-1mp	4, 8, 16, 32	1-32	0A	2C	1A
2p-2mp	4, 8, 16, 32	1-16	0A, 1A	2C, 4C	1A, 3A
4p-4mp	4, 8, 16, 32	1-8	0A, 1A, 1B, 1B	2C, 4C, 2D, 4D	1A, 3A, 1B,

Table 38. Port and LUN mappings used.

Each test run for a system began with one LUN and scaled up to all 32 LUNs (as 1, 2, 4, 8, 16, and 32). There were 8 concurrent VDBench threads per LUN for Random workloads, and just 1 thread per LUN for Sequential workloads.

These are the tests that were used:

- **Random I/O, Transactions per second and Average Response Time**, using an 8KB block size with 8 Threads (100% - 75% - 50% Read, and 100% Write)

- **Sequential I/O, Throughput**, using a 256KB block size with 1 Thread (100% - 75% - 50% Read, and 100% Write)

It is important to understand that these test results were all measured in the absence of any internal or additional external workloads. Thus, all of the resources of the USP V were dedicated to the processing of I/O to the external LUNs. These results are not to be expected in the presence of significant competing workloads. With regard to the “cache hit” tests, most Open Systems will see less than a 20% cache hit ratio with most workloads. Therefore, these results are only indicative of how hard the hardware can process I/O if a 100% cache hit ratio were encountered.

Random Workloads

The following two tables (Cache Miss, Cache Hit) present the peak IOPS values (and the Response Times) achieved for the Random workload tests. The cache miss tests were designed to have a 0% cache hit ratio (back-end disk test), but there will always be some residual of caching, even if it is only about 2%. For all of these tests, the peak rates achieved are included in these tables. The results for “Cache Mode = Enabled” (CM on) and “Cache Mode = Disabled” (CM off) are both shown. This is the USP V Universal Volume Manager’s setting (per LUN) for how individual external LUNs are managed.

Random I/O 2+2 USP-V external storage cache miss			AMS500 8GB 146gb10k				1p-1mp CM off 1p-1mp CM on				2p-2mp CM off 2p-2mp CM on				4p-4mp CM off 4p-4mp CM on			
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100%	8k	8	11,395	21.8	11,396	21.8	17,084	14.5	17,381	14.3	17,375	14.7	17,385	14.7	17,385	14.7	17,385	14.7
75%	8k	8	8,320	29.8	8,823	14.7	14,663	16.9	13,934	17.9	14,792	17.3	14,258	18.3	14,258	18.3	14,258	18.3
50%	8k	8	6,660	19.2	5,435	47.4	12,318	20.1	10,064	25.0	12,935	19.8	13,050	19.6	13,050	19.6	13,050	19.6
0%	8k	8	4,625	27.7	426	37.7	9,055	27.4	1,076	31.1	9,090	28.2	9,067	28.3	9,067	28.3	9,067	28.3

Table 39. Random workload results, cache miss.

Random I/O 2+2 USP-V external storage cache hit			AMS500 8GB 146gb10k				1p-1mp CM off 1p-1mp CM on				2p-2mp CM off 2p-2mp CM on				4p-4mp CM off 4p-4mp CM on			
Read %	Block	Thread	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100%	8k	8	32,174	0.5	32,192	0.2	52,937	2.4	52,838	2.4	104,401	2.4	104,217	2.4	104,401	2.4	104,217	2.4
75%	8k	8	15,648	16.4	34,304	1.9	16,921	15.1	56,461	4.5	27,456	9.3	111,002	2.3	27,456	9.3	111,002	2.3
50%	8k	8	11,072	23.1	36,790	1.7	11,377	22.5	57,784	4.4	16,628	15.4	111,873	2.3	16,628	15.4	111,873	2.3
0%	8k	8	8,624	29.7	31,985	1.0	8,475	30.2	54,514	2.3	9,426	27.2	107,214	2.4	9,426	27.2	107,214	2.4

Table 40. Random workloads results, cache hit.

From these results you can see that it would be advisable to use Cache Mode Enabled for Open Systems workloads with their likely 20% cache hit ratios in order to control the Response Times and achieve the highest TPS rates.

Sequential Workloads

The following Tables and Figures show the peak performance results by test type for the Sequential workloads. The first set is for a back-end (cache miss), while the second if for a front-end test (cache hit).

Sequential I/O 2+2	256kb	AMS500 8GB	1p-1mp CM off	1p-1mp CM on	2p-2mp CM off	2p-2mp CM on	4p-4mp CM off	4p-4mp CM on
USP-V external storage cache miss		146gb10k						
Read %	Block	Thread	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
100%	256kb	8	306	370	515	520	607	641
50%	256kb	8	172	171	341	336	343	335
0%	256kb	8	167	197	328	409	328	373

Table 41. Sequential workload results, cache miss.

Sequential I/O 2+2	256kb	AMS500 8GB	1p-1mp CM off	1p-1mp CM on	2p-2mp CM off	2p-2mp CM on	4p-4mp CM off	4p-4mp CM on
USP-V external storage cache hit		146gb10k						
Read %	Block	Thread	MB/s	MB/s	MB/s	MB/s	MB/s	MB/s
100%	256kb	8	385	385	537	514	1,020	1,037
50%	256kb	8	316	443	528	531	621	1,063
0%	256kb	8	166	336	327	523	327	1,008

Table 42. Sequential workload results, cache hit.

From these results you can see that it would be advisable to use Cache Mode Enabled for Open Systems workloads with their likely 20% cache hit ratios in order to achieve the highest throughput rates. There was little difference due to Cache Mode on the “cache miss” workloads, but often a substantial improvement for “cache hit” workloads with significant write content. These tests displayed some anomalies with the Test Code running on the USP V, and will be rerun once GS code is available.

XIV. RAID10 (2D+2D) Single Array Group LUN scalability tests.

The purpose of this testing was to conduct a workload and LUN scalability performance analysis of a single 4-disk RAID10 2D+2D Array Group. There was a single 4Gbit FC path from the USP V to the test server. There were (4) 146GB 15K rpm disks used for these tests. Both Random and Sequential workloads were used, with an 8KB block size for Random and 256KB block size for Sequential.

Test Methodology

Each test run began with one LUN and scaled up to 8 LUNs (as 1, 2, 4, 8 LUNs). Additionally, each of these tests was scaled from 4-256 concurrent VDBench threads per LUN (Random), or 1-64 threads per LUN (Sequential). The two sets of workloads were specified to achieve a 0% cache hit ratio (100% seek per I/O).

These are the workloads that were used:

- **Random I/O: Transactions per second and Average Response Time**, using an 8KB block size, and 100% - 75% - 50% Read, and 100% Write.
- **Sequential I/O: Throughput**, using a 256KB block size, and 100% - 75% - 50% Read, and 100% Write.

Random %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	8KB Block
100 %	0 %	0 %	0 %	4 - 256	✓
	50 %			4 - 256	✓
	75 %			4 - 256	✓
	100 %			4 - 256	✓

Table 43. Random workloads overview.

Sequential %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	256K Block
100 %	0 %	0 %	0 %	1 - 64	✓
	50 %			1 - 64	✓
	75 %			1 - 64	✓
	100 %			1 - 64	✓

Table 44. Sequential workloads overview.

Random Workloads

The following charts show the results of the Random testing. One chart shows the TPS, the other shows the Response Times. As can be seen in the charts, the highest performance almost always came from a single LUN per Array Group, with the single LUN test often showing a large performance lead over the others

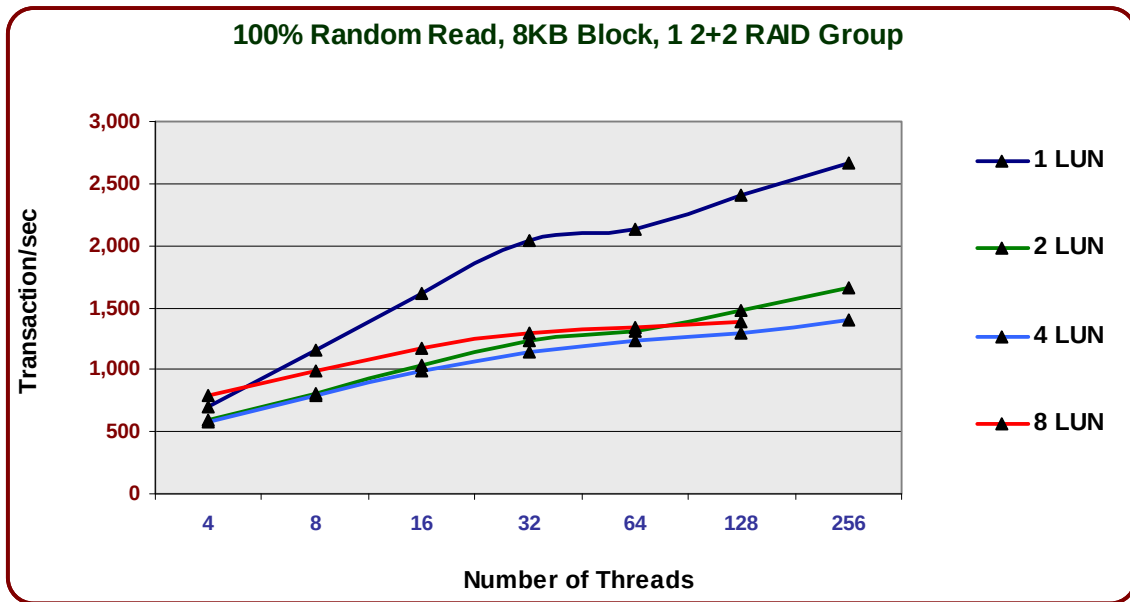


Figure 41. 100% Random Read TPS.

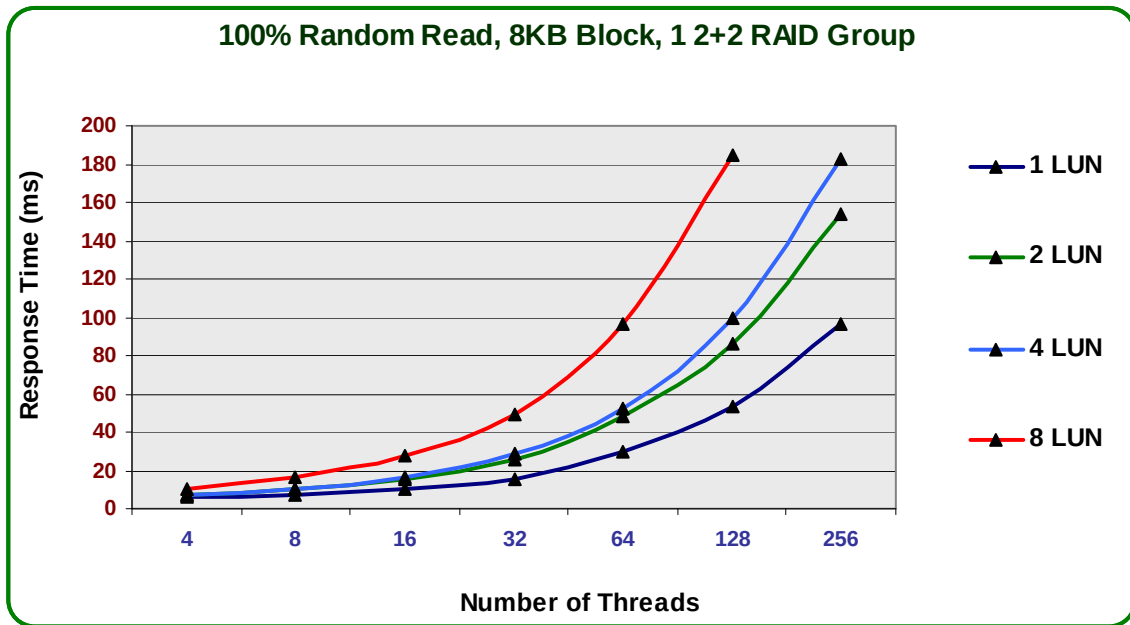


Figure 42. 100% Random Read RT.

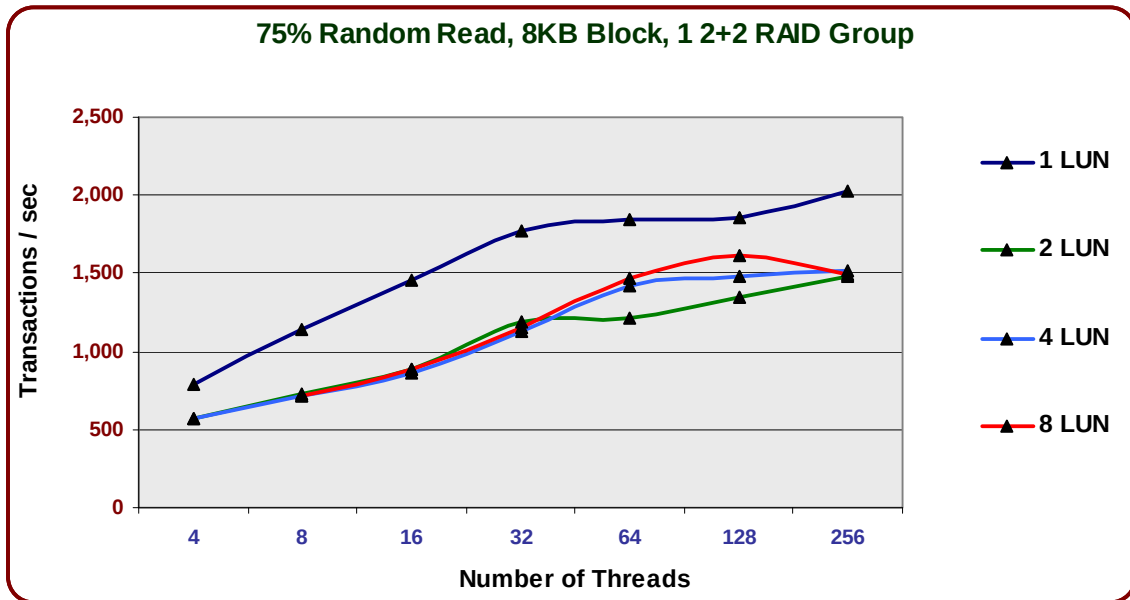


Figure 43. 75% Random Read TPS.

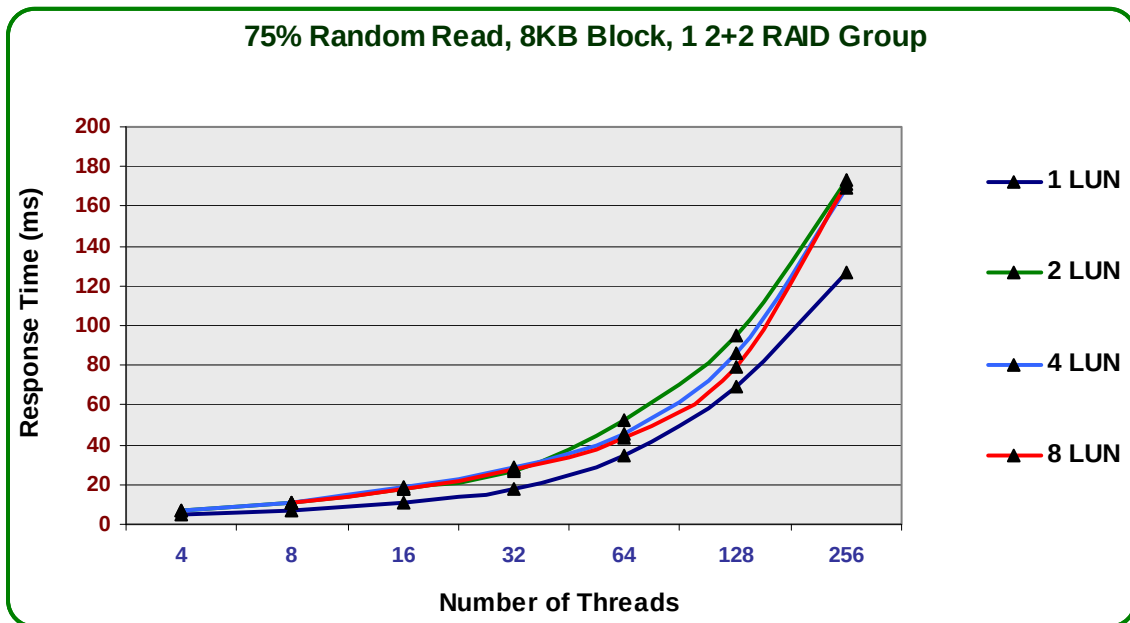


Figure 44. 75% Random Read RT.

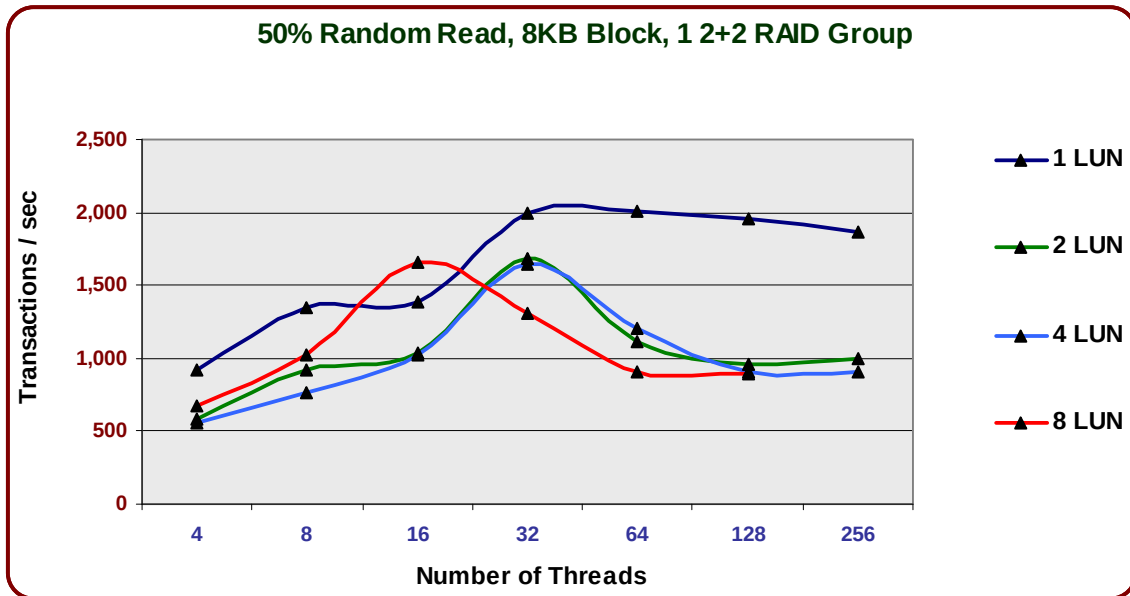


Figure 45. 50% Random Read TPS.

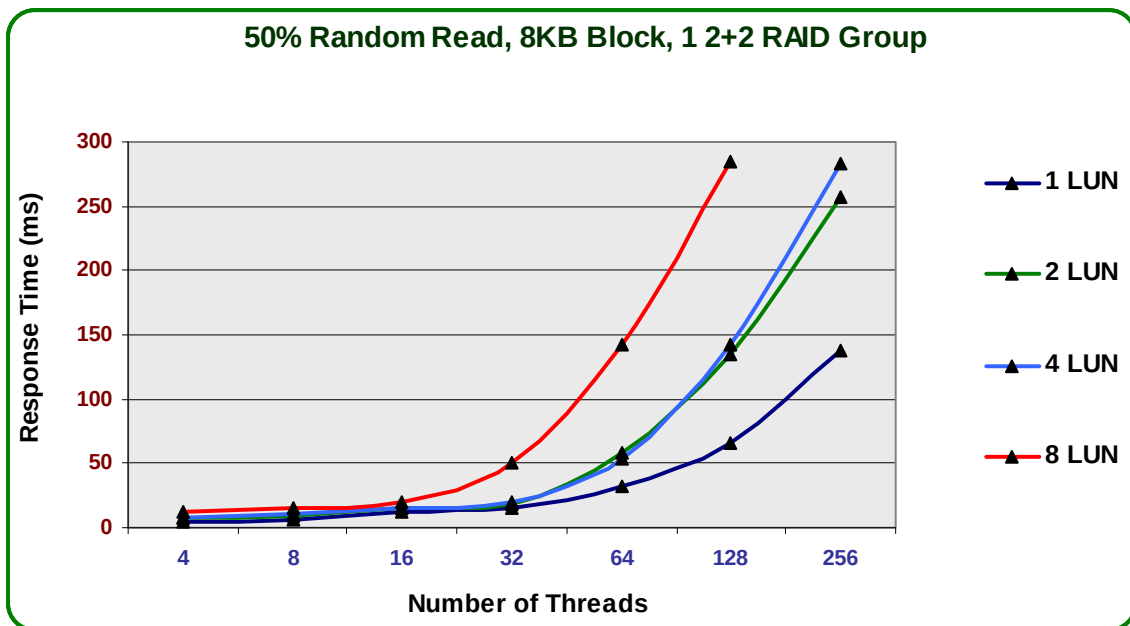


Figure 46. 50% Random Read RT.

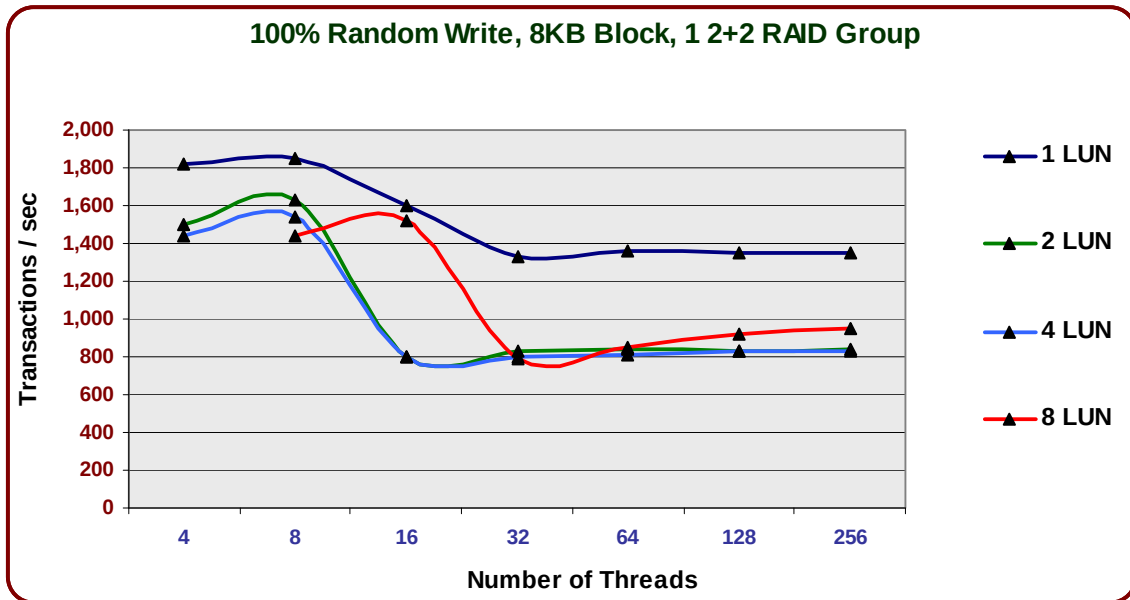


Figure 47. 100% Random Write TPS.

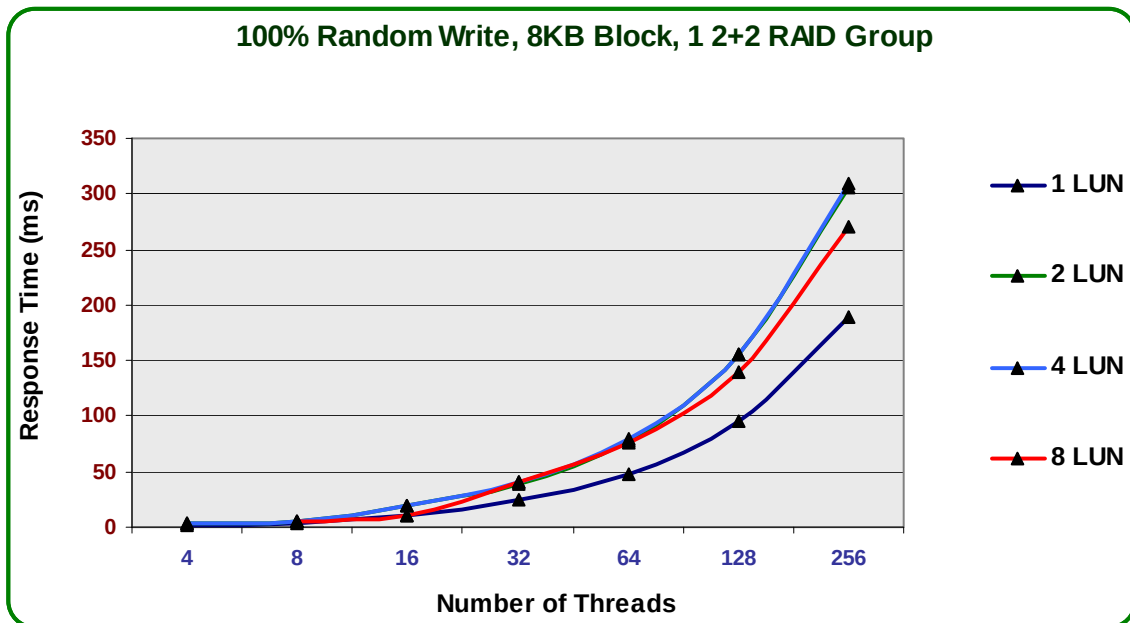


Figure 48. 100% Random Write RT.

Sequential Workloads

The following charts show the Throughput results (MB/s) of the Sequential testing. As can be seen in the charts, the highest performance always came from a single LUN per Array Group, with the single LUN test showing a large performance lead over the others

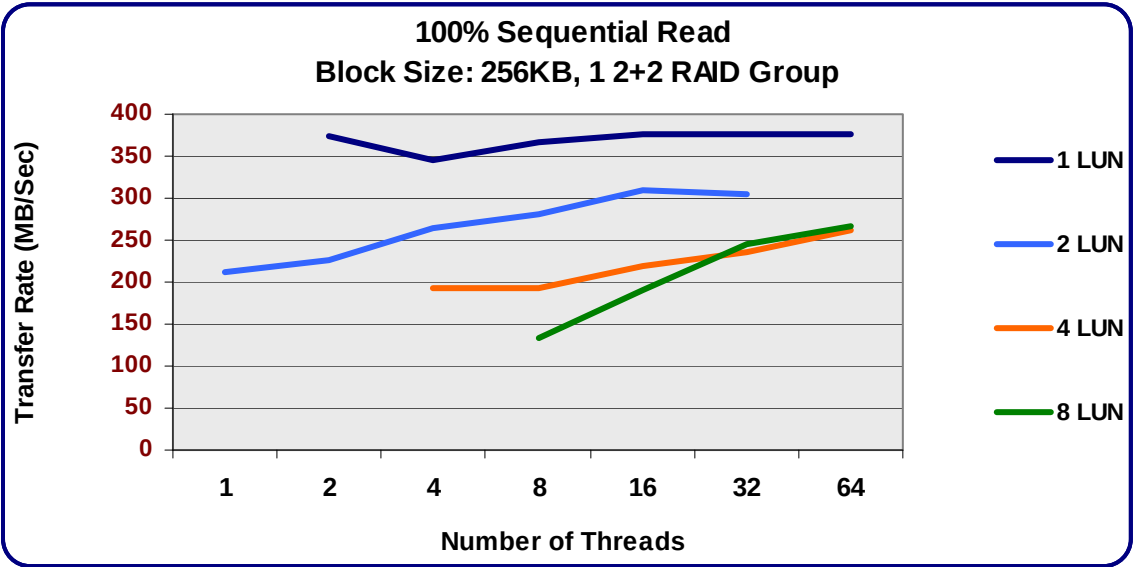


Figure 49. 100% Sequential Read.

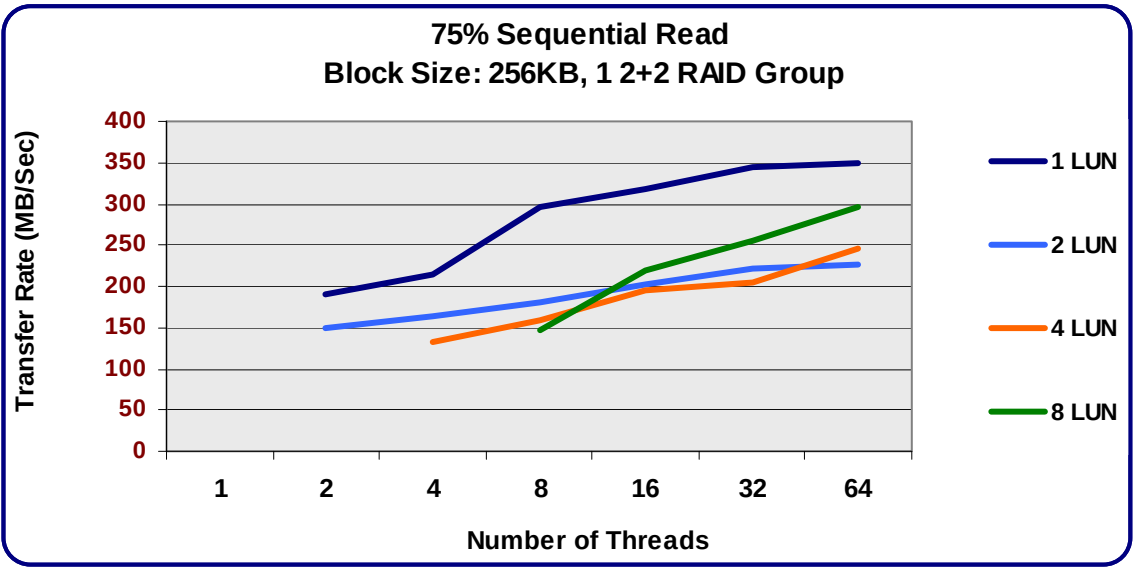


Figure 50. 75% Sequential Read.

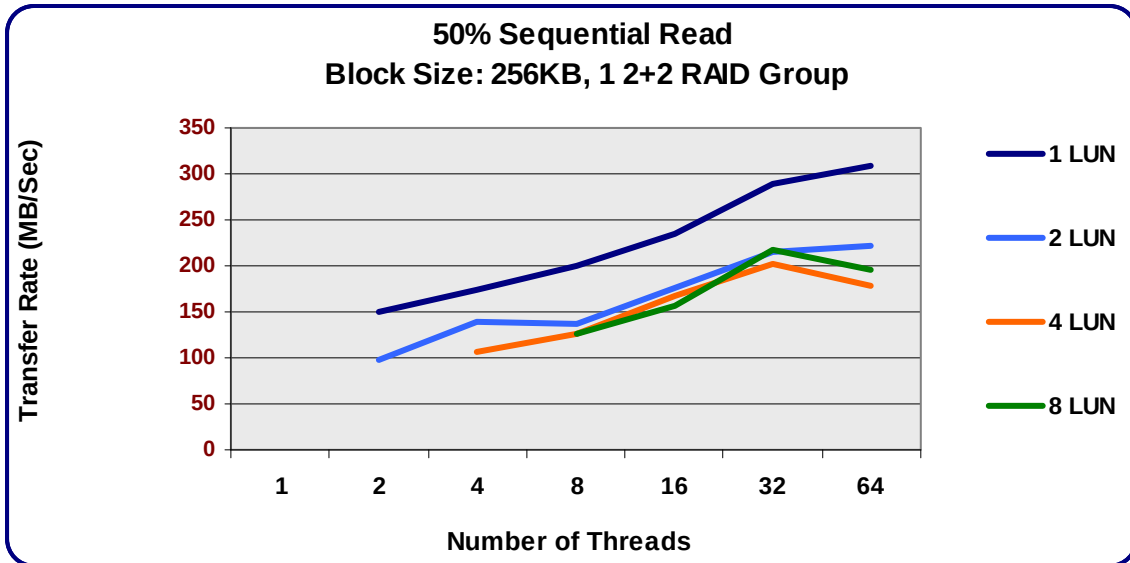


Figure 51. 50% Sequential Read.

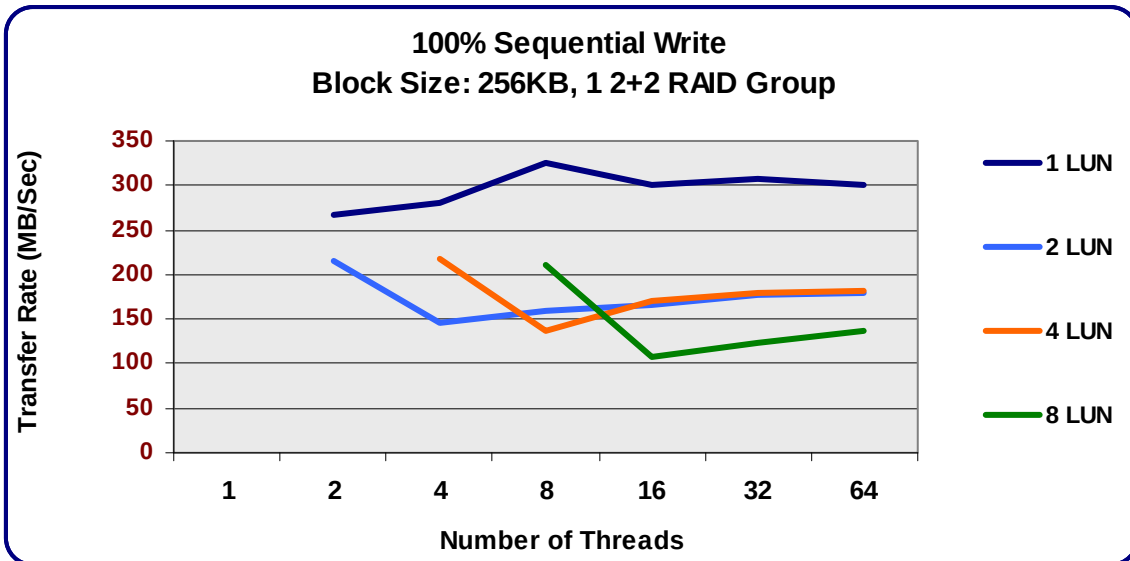


Figure 52. 100% Sequential Write.

XV. RAID5 (3D+1P) Single Array Group LUN scalability tests

The purpose of this testing was to conduct a workload and LUN scalability performance analysis of a single RAID5 4-disk 3d+1d Array Group. There was a single 4Gbit FC path from the USP V to the test server. There were (4) 146GB 15K rpm disks used for these tests. Both Random and Sequential workloads were used, with an 8KB block size for Random and 256KB block size for Sequential.

Test Methodology

Each test run began with one LUN and scaled up to 8 LUNs (as 1, 2, 4, 8 LUNs). Additionally, each of these tests was scaled from 4-256 concurrent VDBench threads per LUN (Random), or 1-64 threads per LUN (Sequential). The two sets of workloads were specified to achieve a 0% cache hit ratio (100% seek per I/O).

These are the workloads that were used:

- **Random I/O: Transactions per second and Average Response Time**, using an 8KB block size, and 100% - 75% - 50% Read, and 100% Write.
- **Sequential I/O: Throughput**, using a 256KB block size, and 100% - 75% - 50% Read, and 100% Write.

Random %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	8KB Block
100 %	0 %	0 %	0 %	4 - 256	✓
	50 %			4 - 256	✓
	75 %			4 - 256	✓
	100 %			4 - 256	✓

Table 45. Random workloads overview.

Sequential %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	256K Block
100 %	0 %	0 %	0 %	1 - 64	✓
	50 %			1 - 64	✓
	75 %			1 - 64	✓
	100 %			1 - 64	✓

Table 46. Sequential workloads overview.

Random Workloads

The following charts show the results of the Random testing. One chart shows the TPS, the other shows the Response Times. As can be seen in the charts, the highest performance almost always came from a single LUN per Array Group, with the single LUN test often showing a large performance lead over the others.

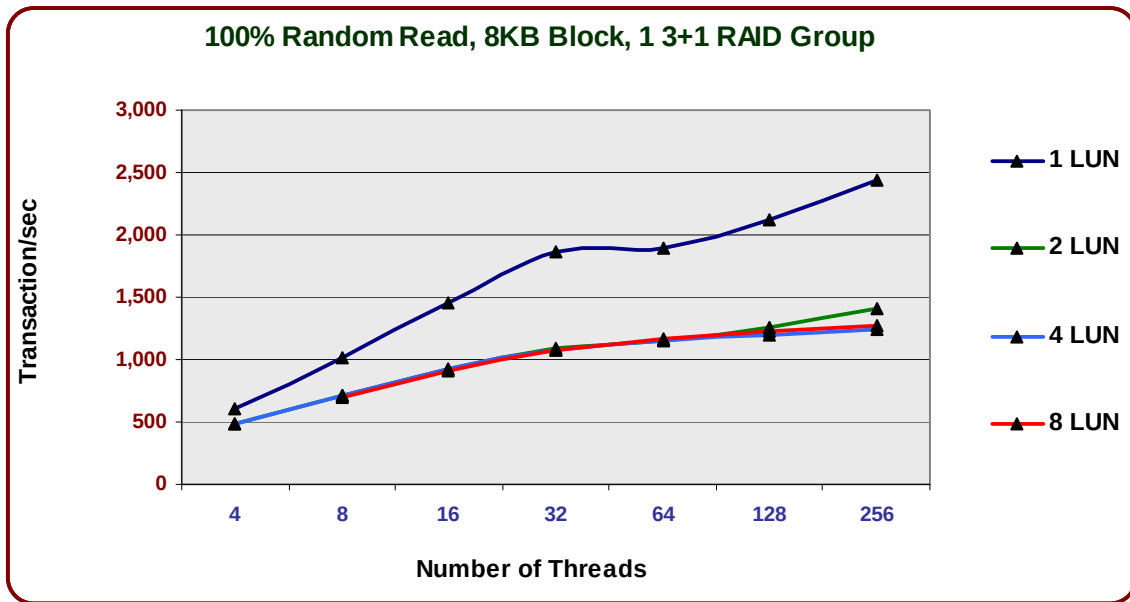


Figure 53. 100% Random Read TPS.

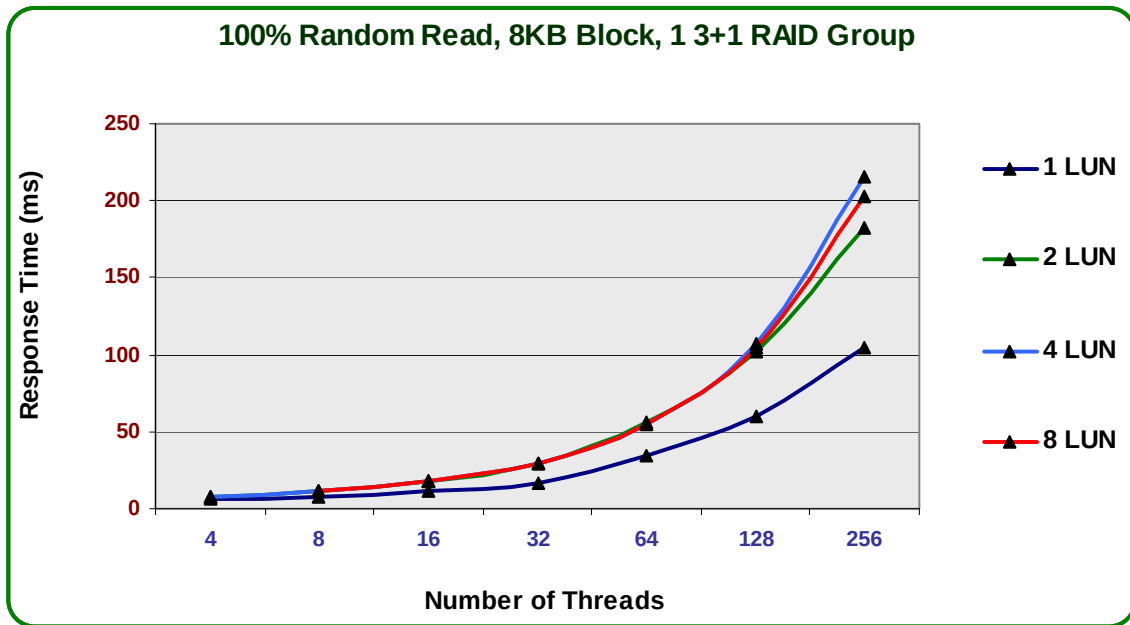


Figure 54. 100% Random Read RT.

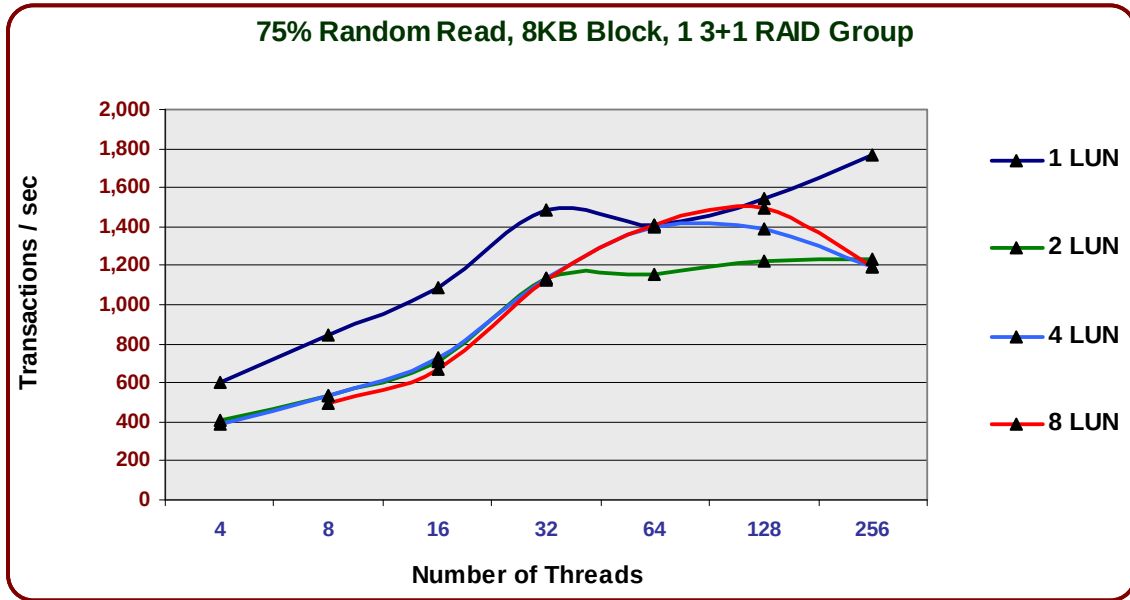


Figure 55. 75% Random Read TPS.

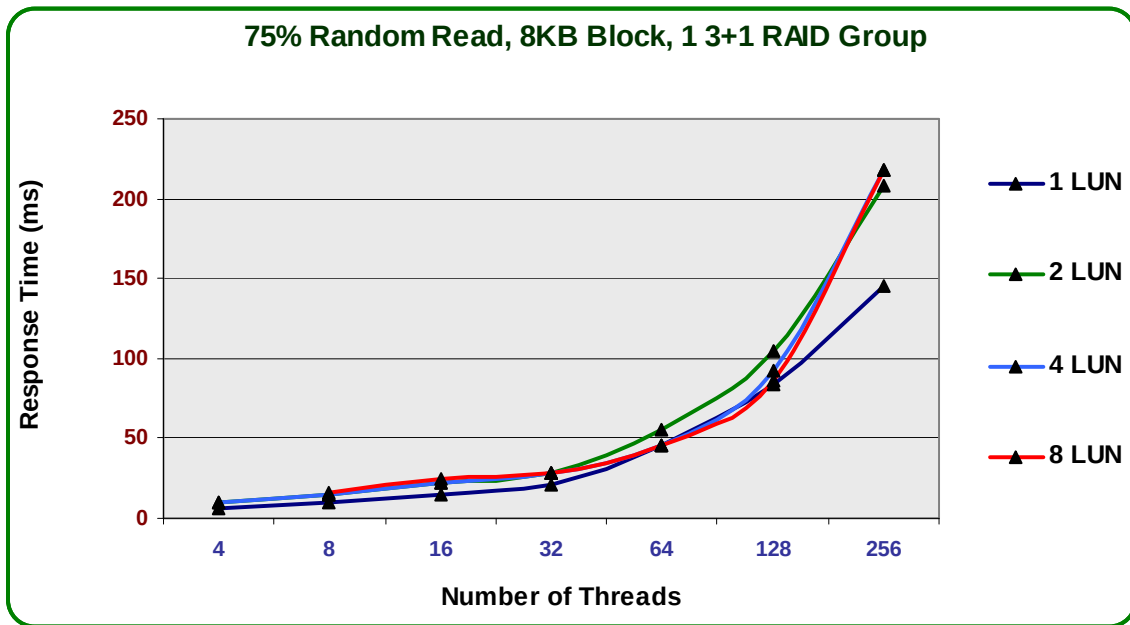


Figure 56. 75% Random Read RT.

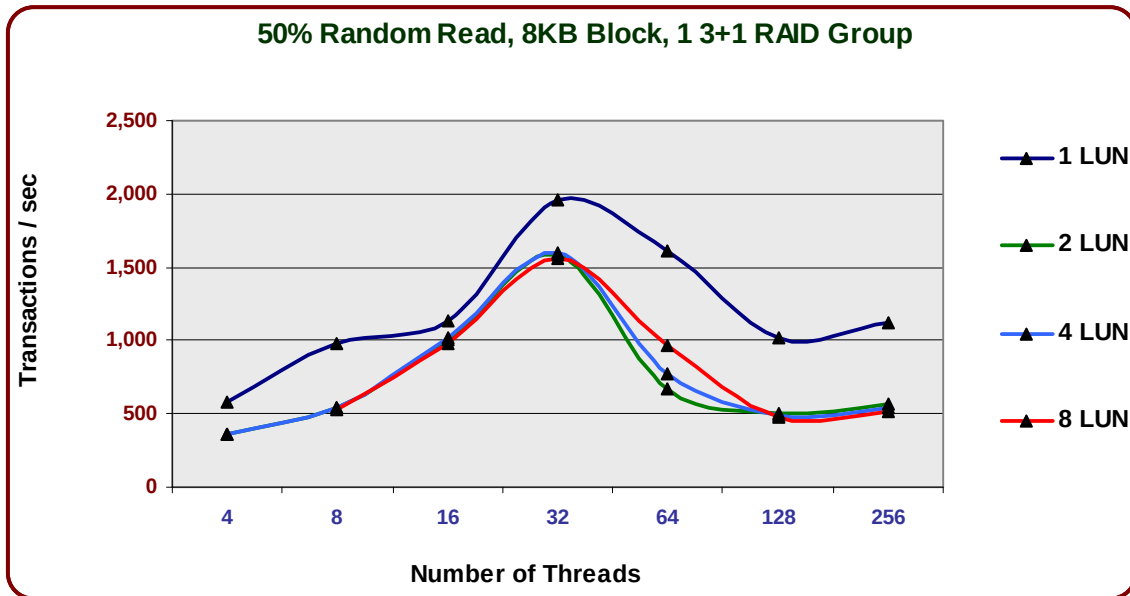


Figure 57. 50% Random Read TPS.

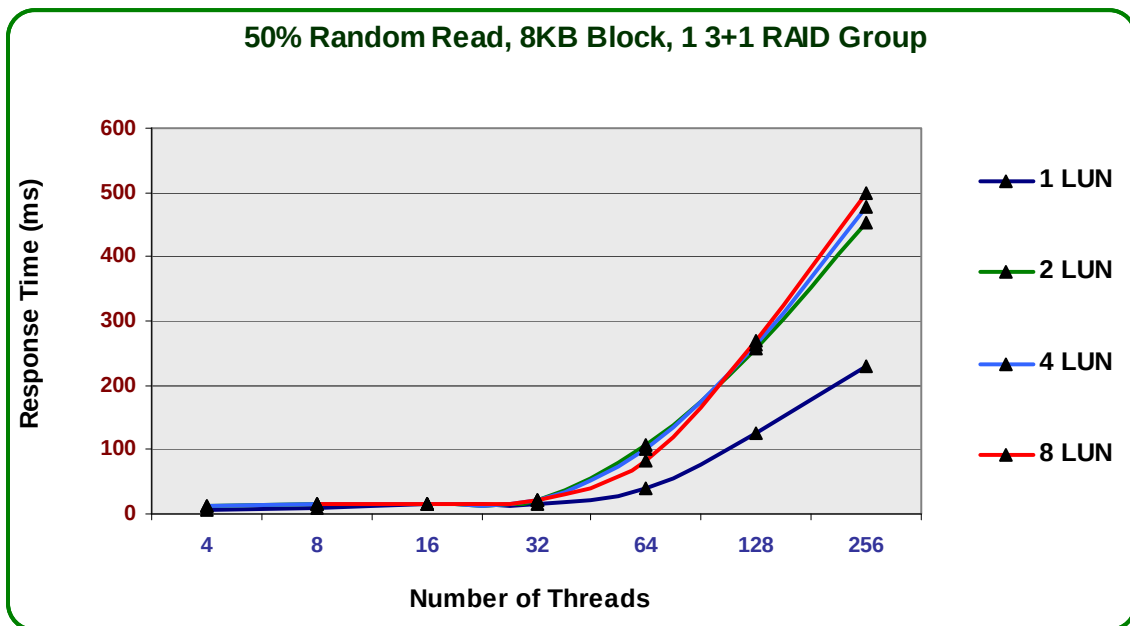


Figure 58. 50% Random Read RT.

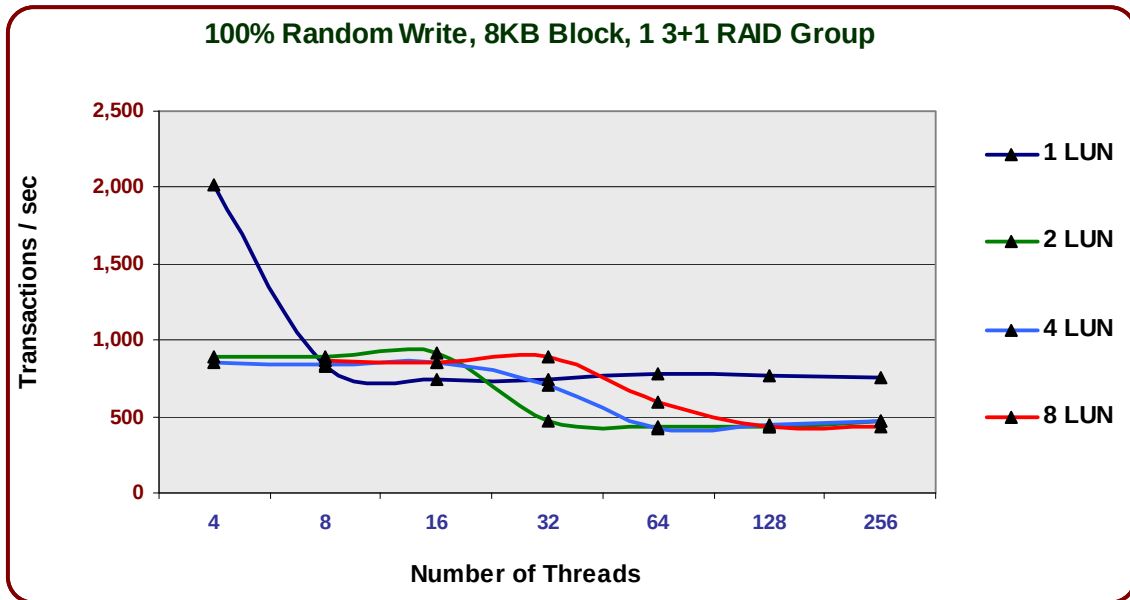


Figure 59. 100% Random Write TPS.

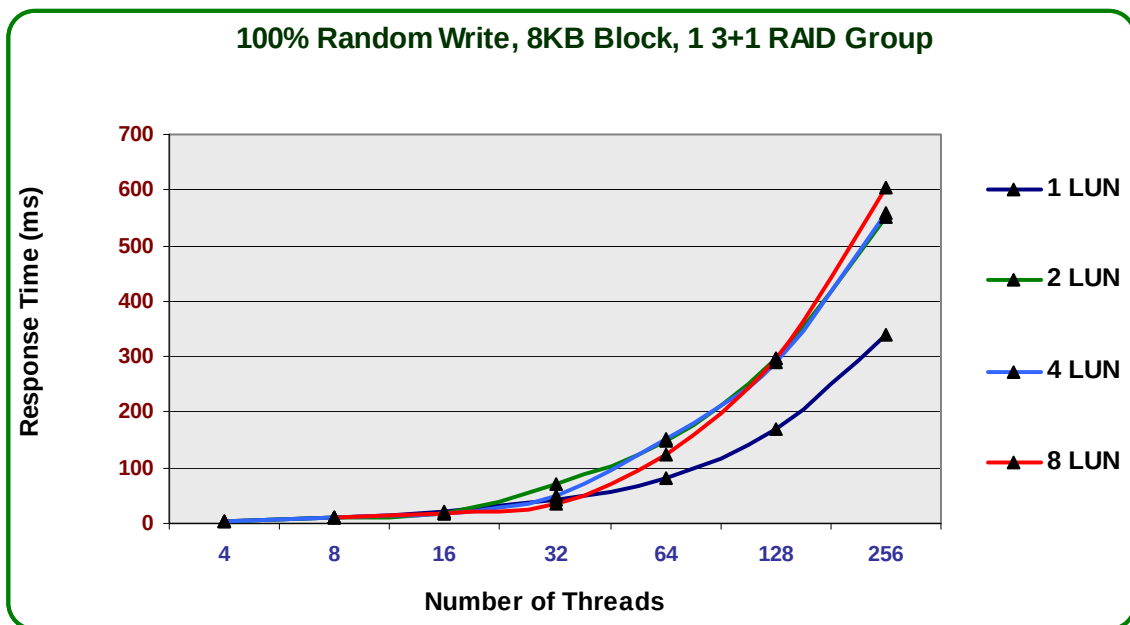


Figure 60. 100% Random Write RT.

Sequential Workloads

The following charts show the Throughput results (MB/s) of the Sequential testing. As can be seen in the charts, the highest performance always came from a single LUN per Array Group, with the single LUN test showing a large performance lead over the others.

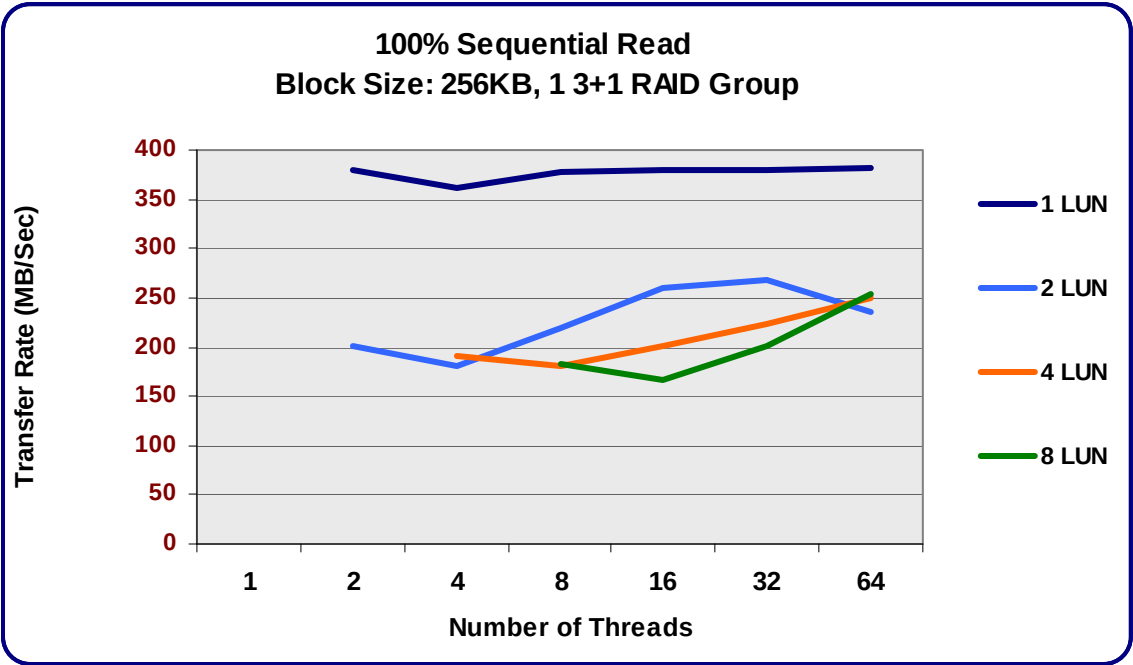


Figure 61. 100% Sequential Read.

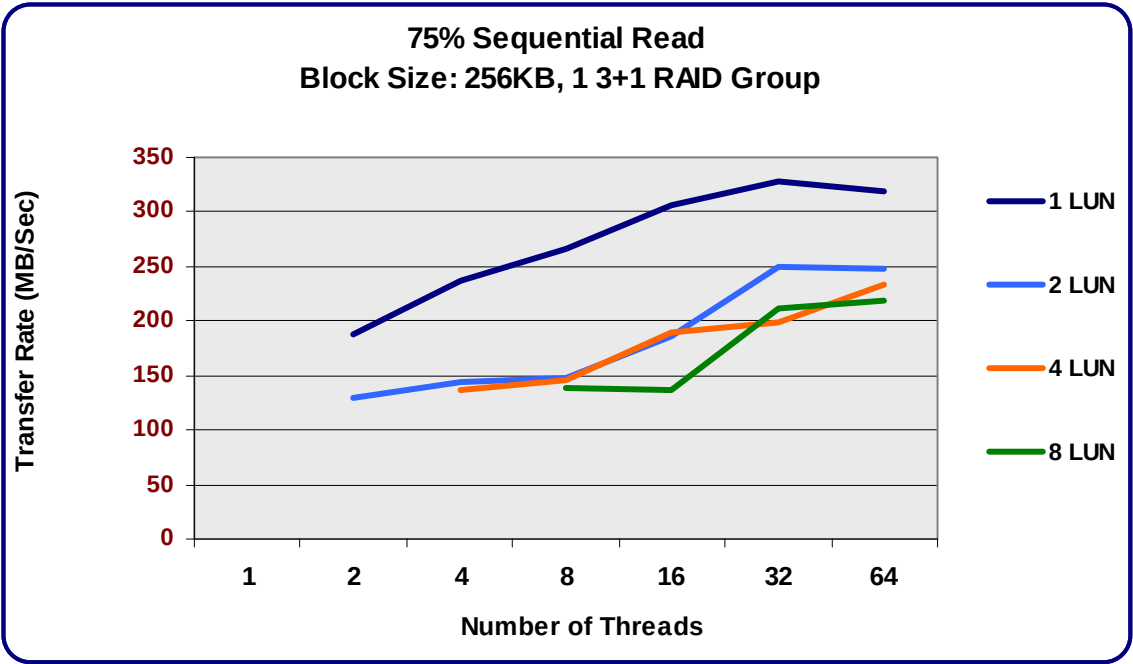


Figure 62. 75% Sequential Read.

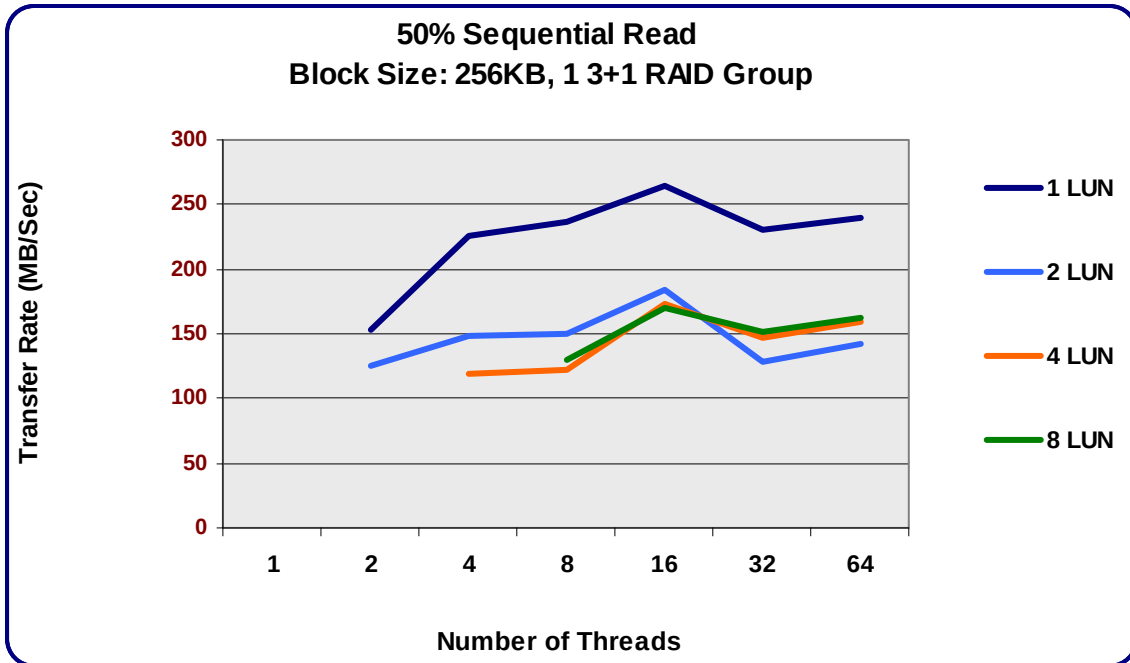


Figure 63. 50% Sequential Read.

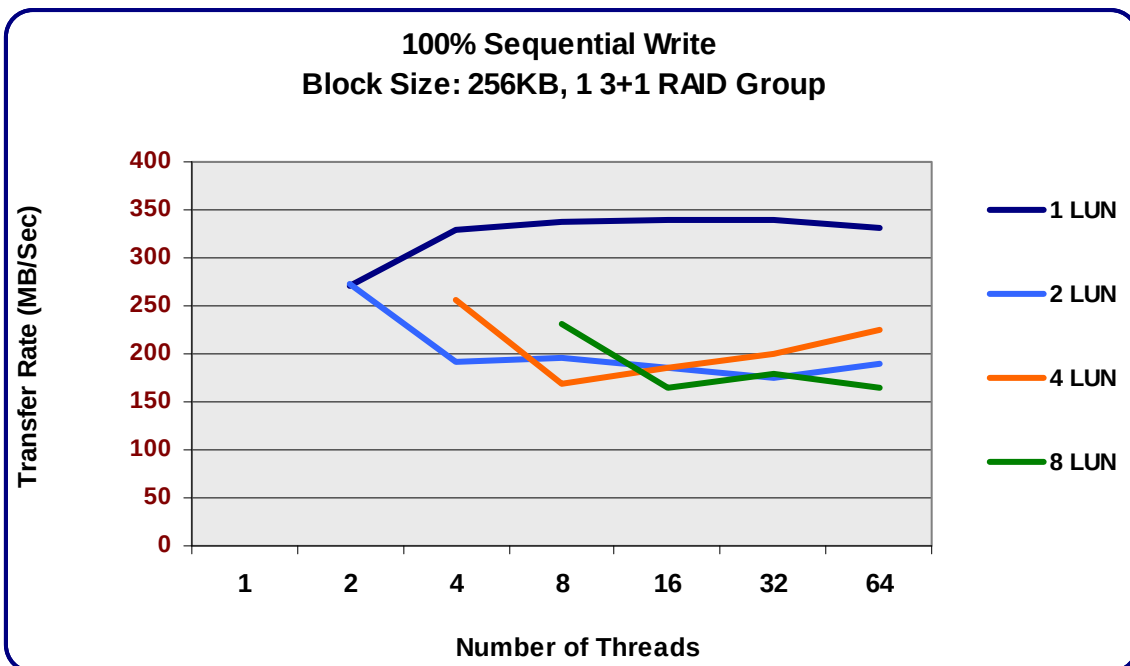


Figure 64. 100% Sequential Write.

XV. RAID10 (2D+2D) HDP Comparison Tests

The purpose of this testing was to compare the results of various random and sequential workloads on a USP V storage system using various storage configurations on the server. These were “cache miss” workloads designed to measure the back-end disk performance. The USP V was configured with 16, 32, or 64 disks as RAID10 (2D+2D) Array Groups using 146GB 15k RPM disks. Four 4Gbit host paths were used for the 16-disk tests, while 8 paths were used for both the 32-disk and 64-disk tests. None of the FED ports used shared a FED MP (processor). The server’s queue depth was set to 16 on each FC port.

These were the four classes of USP V volumes tested:

1. **Standard volumes**
2. **Host LVM managed volumes**
3. **HDP volume-1**
4. **HDP volume-2**

The configurations used were designed to keep the four test environments as matched as possible in order to evaluate the effects of the method rather than any hardware differences. The same numbers of Array Groups, host paths, and LDEVs were used. Case #4 compared to case #3 defined more DP volumes and Pool Volumes from the same number of Array Groups. In general, Standard volumes should show a performance advantage for these tests as they have the least amount of associated overhead. Note that for Cases 1 and 3, the workload generator will balance the workload threads across the individual volumes, but this is not the case for the single host volume of Case 2 (LVM). Also note that because the test workloads were constant across the host volumes (unlike a normal environment with constantly shifting hot spots), the inherent load distribution advantage of HDP was not the usual distinguishing positive factor for those tests. The fundamental intent of these tests was to compare the efficiencies of the four techniques.

The performance results presented below are intended to help answer general questions about the kind of performance trade-offs to expect on the USP V system when using these four methods. The results are presented in the tables and charts below. Note that other combinations of RAID levels, the quantity of Pool volumes, the number of V-VOL Groups per Pool, and the number of DP volumes per V-VOL Group could yield very different results.

Test Descriptions

CASE 1

For the first case, “**Standard volumes**”, the server volumes (4, 8 or 16 LUNs) were individually addressed as raw devices (no file systems) and accessed in parallel by the VDbench workloads. Each of these LUNs was created as a single “Open-V” LDEV on a dedicated Array Group that used 100% of the available space. See the configuration overview in Table 47 below. The use of Standard volumes is probably the most common technique for using storage on servers, especially for database workloads. The problem with this method is that it doesn’t scale well. While a dozen or so host volumes is easily managed, this simple technique is very burdensome with many dozens of volumes. This leads many users to use the alternate method of case #2.

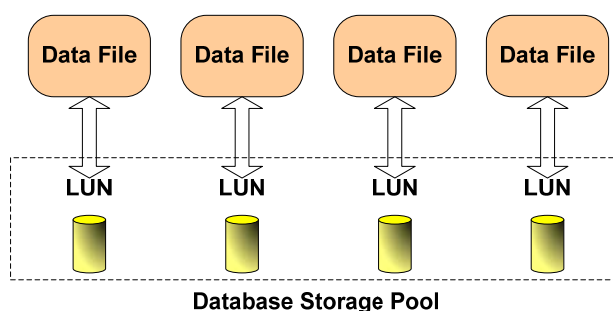


Figure 65. Use of four Standard volumes on a host

Standard and LVM Volume Configurations					
# of RAID10 2+2 Array Groups	# of LDEVs per Array Group	# of LDEVs in test	LDEV size	% of AG used per LDEV	4Gb host paths
4	1	4	1070GB	100 %	4
8	1	8	1070GB	100 %	8
16	1	16	1070GB	100 %	8

Table 47. Configuration summary for Standard volume and LVM volume testing

CASE 2

For the second case, “**Host LVM managed volumes**”, the same 4, 8, or 16 LUNs were used by the native Logical Volume Manager on the host to create a single RAID0 striped volume. This volume was used as a raw device (no file system) by the server for the same workloads as used for Case 1. Refer back to Table 47 for the configuration overview and to Figure 66 for an illustration of a generic LVM environment. The use of an LVM on the server is a very typical method for organizing space on many servers, where there is a need for more capacity or raw disk power than could be provided by the single Array Group behind each Standard volume. One advantage to this method is the extra level of caching of data blocks provided by the LVM buffer system. The principal drawbacks of using LVMs are that they are more complex to configure and manage, and can be fairly costly when advanced products such as Veritas VxVM is deployed.

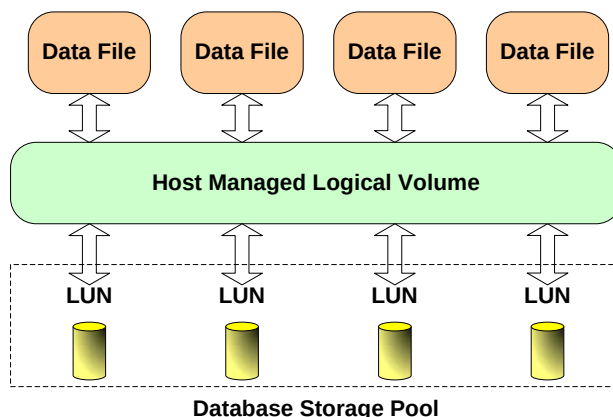


Figure 66. Use of a host-managed logical volume with four LUNs

CASE 3

For the third case, “**HDP volume-1**”, one HDP Pool was created using 4, 8 or 16 LDEVs as Pool Volumes. This test was designed to examine how an equally configured HDP environment would compare to Cases 1 and 2. There were 1, 2, and 4 V-VOL Groups created as 1TB containers and all were connected to the one HDP Pool. Next, there were four 240GB DP volumes created per V-VOL Group. See Table 48 for the configuration overview and Figure 67 for an illustration of a generic HDP environment. Each DP volume was used as a raw device (no file system) by the server for the same workloads as used for Case 1. The discarded ramp-up phase of each test sequence pre-populated all of the physical pages for these DP volumes, so the incremental overhead of assigning physical pages is not present in these tests.

HDP Pool	Pool Volumes (LDEVs)	V-VOL Groups	V-VOL size	DP-Vols (host LUNs)	DP-Vol size	DP-Vols per 4Gb path	4Gb host paths
1	4	1	1070GB	4	240GB	1	4
1	8	2	1070GB	8	240GB	1	8
1	16	4	1070GB	16	240GB	2	8

Table 48. HDP volume-1 configurations

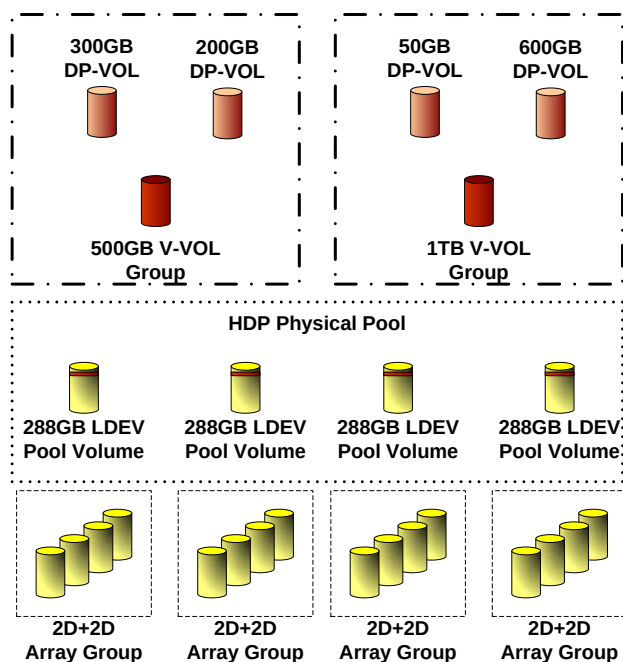


Figure 67. Components of the HDP system

CASE 4

The fourth case, “**HDP volume-2**”, was similar to the third case, except that each Array Group was reconfigured to provide 4, 8, or 16 equal sized LDEVs (rather than have a single large LDEV per Array Group). This was to test the effect on Pool performance when using many LDEVs per Array Group, which may increase the degree of striping of Pool pages across the Pool Volumes. The question being examined for this case was to determine if there was any performance advantage in doing so. All of these LDEVs were assigned as Pool Volumes to the

same HDP Pool. There were 1, 2, and 4 V-VOL Groups created as 1TB containers and all were connected to this Pool. There were four 240GB DP volumes created per V-VOL Group. See Table 49 for a configuration overview. Each DP volume was used as a raw device (no file system) by the server for the same workloads as used for Case 1. Again, the ramp-up phase of the tests pre-populated all of the physical pages for the DP volumes.

HDP Pool	Pool Volumes (LDEVs)	V-VOL Groups	V-VOL size	DP-Vols (host LUNs)	DP-Vol size	DP-Vols per 4Gb path	4Gb host paths
1	16	1	1070GB	4	240GB	1	4
1	64	2	1070GB	8	240GB	1	8
1	256	4	1070GB	16	240GB	2	8

Table 49. HDP volume-2 configurations

Test Methodology

Either Random or Sequential workloads were used, where the workload block size was 8KB for Random and 256KB for Sequential. These workloads used 100% Read, 75% Read, 50% Read, 25% Read, or 100% Write patterns. All tests were run with the workload generator's "cache miss" option set in order to measure the back-end of the USP V and not the front-end. Typically, there is about a 2% cache hit rate even with this option enabled. This type of test is generally called a "back-end test" as it tests the back-end controllers and disks. In all test cases, 70% of the space presented by the volumes on the server was used for the test's data patterns (i.e. the tests generally had long seeks).

Each test run began with a few ramp-up runs (all discarded), and then the measured target workloads were run, each with a greater number of parallel threads. The 16-disk tests started with 4 threads, the 32-disk tests started with 8 threads, and the 64-disk tests started with 16 threads. All tests (random and sequential) scaled up to 512 threads, with the thread counts doubling for each run.

Workload	Read %	Cache Read Hit %	Cache Write Hit %	Max Threads per test	8KB Block
Random	100, 75, 50, 25, 0	0%	0%	512	√

Table 50. Random workloads overview

Workload	Read %	Cache Read Hit %	Cache Write Hit %	Max Threads per test	256KB Block
Sequential	100, 75, 50, 25, 0	0%	0%	512	√

Table 51. Sequential workloads overview

Table 52 below lists the LDEV configuration by test case and the number of Array Groups in use. The scaling of workload threads used for each configuration is also shown.

Test Type	2+2 Array Groups	<u>TOTAL LDEVs USED</u>				Workload Thread Scaling
		Standard volumes	LVM volume	HDP volume-1	HDP volume-2	
16-disk	4	4	4	4	16	4, 8, 16, 32, 64, 128, 256, 512
32-disk	8	8	8	8	64	8, 16, 32, 64, 128, 256, 512
64-disk	16	16	16	16	256	16, 32, 64, 128, 256, 512

Table 52. Summary of LDEV usage by configuration

Test Results

Due to the large number of results to consider, the following tables and charts only present the maximum values achieved for a particular thread count. The same number of threads is used for all samples in each summary table to keep the results comparable. It was normally the case that the same thread count produced the performance plateau of each test case. Again, all tests were designed to have a 0% cache hit ratio to better expose performance characteristics of the backend hardware.

For the Write tests using DP volumes, their HDP Pool pages had already been pre-allocated by ramp-up tests, so there was no additional overhead from the Pool Page assignment process.

Overall, there were no performance advantages when using HDP volume-2 rather than the simpler HDP volume-1 configuration. There is no benefit in using a more complex HDP configuration in order to improve on a Pool's performance.

Random Workloads Comparisons

The following Tables and charts show the peak results of the Random testing for the three different disk counts. In each section, one chart shows the Transactions per Second while the other shows the Response Times.

16-DISK TESTS

As can be seen in the table and charts, there was little difference between the HDP volume-1 (1 LDEV per Array Group) and HDP volume-2 (4 LDEVs per Array Group) configurations. The expected advantages when using the simpler Standard volumes were not seen in the '25% Read' and '100% Write' workloads. For the other three workloads, the HDP volumes had the same performance as the LVM volume. The LVM volume had a slight advantage at the 25% Read (75% Write) workload. This might be due to the small number of reads getting some boost from additional cache hits in the LVM's buffer pool.

Note that while the response time for the Standard volumes for the 100% read test stands out from the other 100% read tests, the achieved IOPS rate was also much higher. This test was

rerun with the same result, so it appears that some internal limit was reached with this particular test. This limit was probably due to the small number of disks (16) since it didn't reappear for the 32-disk and 64-disk tests.

16 HDD, 2+2		Random		USP-V		Cache Miss		Best results < 30ms (for an entire row)			
8KB block		Threads used for this data	Standard Volumes		LVM volume		HDP volume-1		HDP volume-2		
Workload			IOPS	RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	
100% Read	64	3,810	29.8	3,440	18.6	3,397	18.8	3,392	18.9		
75% Read	64	3,347	19.1	3,021	21.2	3,027	21.2	3,024	21.2		
50% Read	64	3,807	16.8	3,244	19.7	3,308	19.4	3,309	19.3		
25% Read	64	2,669	24.0	3,175	21.5	2,910	23.1	2,903	23.3		
100% Write	64	2,568	24.9	2,588	24.7	2,471	25.9	2,494	25.7		

Table 53. Summary of peak results, 16-disk tests

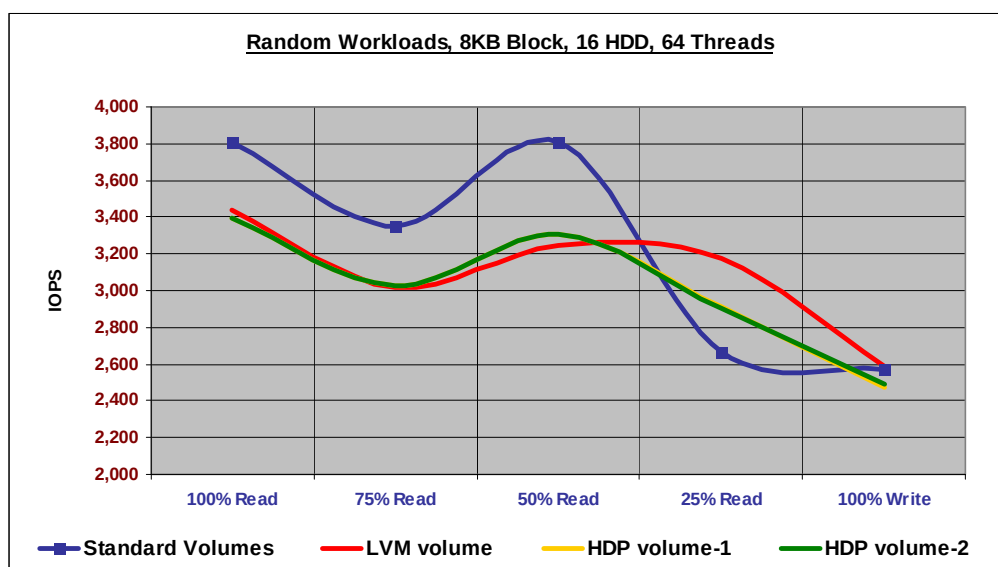


Figure 68. Random workloads TPS, 16-disk tests

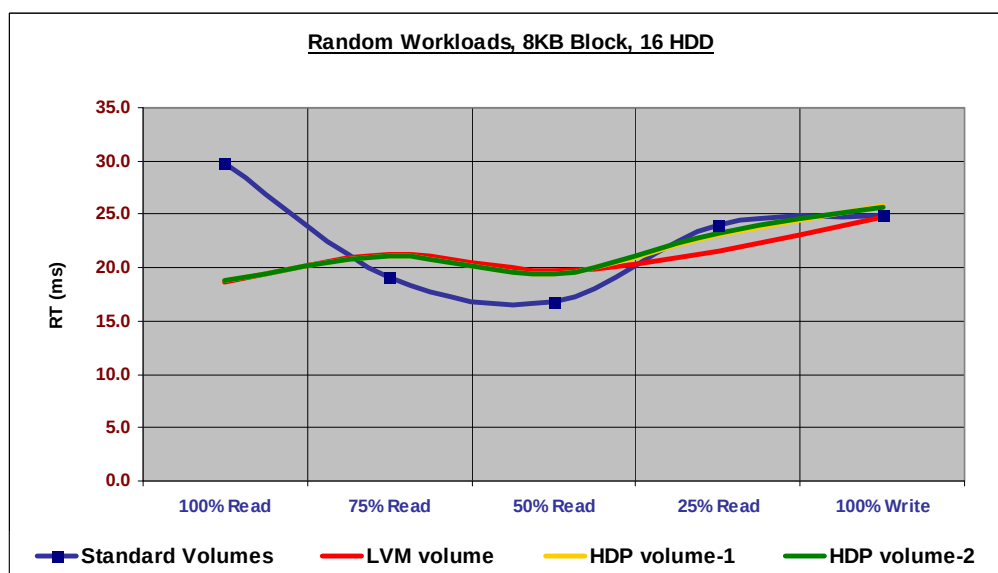


Figure 69. Random workloads RT, 16-disk tests

32-DISK TESTS

As can be seen in the table and charts, there was little difference between the HDP volume-1 (1 LDEV per Array Group) and HDP volume-2 (8 LDEVs per Array Group) configurations. Note how the performance of the LVM volume tracked the HDP volumes. Again, the Standard volumes lost their advantage for the last two workloads, with all volume types delivering about the same rates and response times.

32 HDD, 2+2		Random		USP-V		Cache Miss		Best results < 30ms (for an entire row)			
8KB block		Threads used for this data		Standard Volumes		LVM volume		HDP volume-1		HDP volume-2	
Workload	IOPS			RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	
100% Read	128	7,669	16.7	6,919	18.5	6,710	19.1	6,685	19.1		
75% Read	128	6,631	19.3	6,215	20.6	6,063	21.1	6,051	21.2		
50% Read	128	7,354	17.4	6,335	20.2	6,440	19.9	6,436	19.9		
25% Read	128	5,111	25.0	5,213	24.6	4,992	25.6	4,995	25.6		
100% Write	128	5,051	25.3	4,997	25.6	4,835	26.5	4,848	26.4		

Table 54. Summary of peak results, 32-disk tests

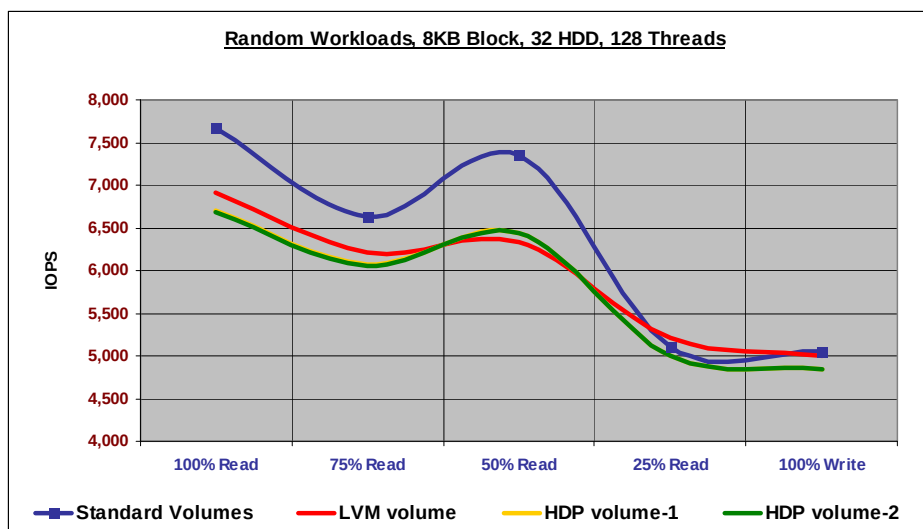


Figure 70. Random workloads TPS, 32-disk tests

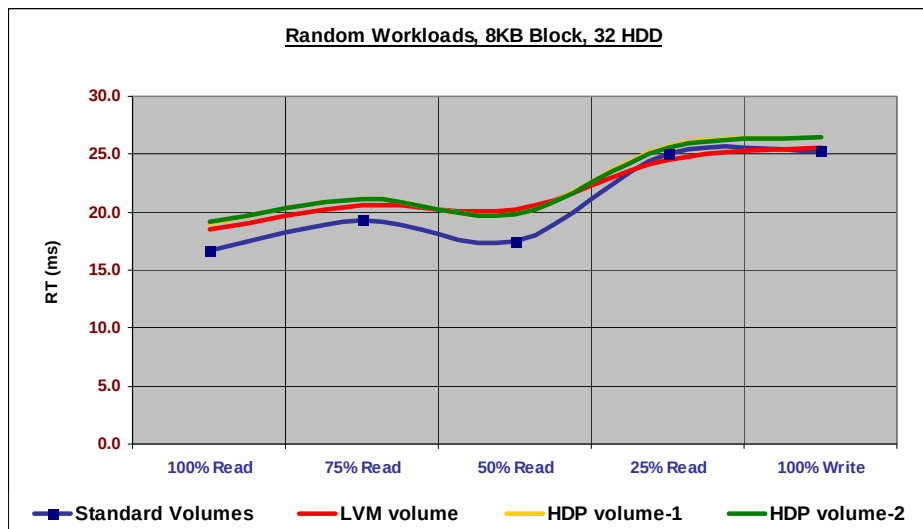


Figure 71. Random workloads RT, 32-disk tests

64-DISK TESTS

As can be seen in the table and charts, there was little difference between the HDP volume-1 (1 LDEV per Array Group) and HDP volume-2 (16 LDEVs per Array Group) configurations. The shape of the curves and their relationships for the 64-disk results are very similar to those for the 32-disk tests. The levels of IOPS performance are much higher than with 32 disks while maintaining similar response times.

64 HDD, 2+2		Random		USP-V		Cache Miss		Best results < 30ms (for an entire row)			
8KB block		Threads used for this data	Standard Volumes		LVM volume		HDP volume-1		HDP volume-2		
Workload			IOPS	RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	IOPS	RT (ms)	
100% Read		256	15,215	16.8	13,451	19.0	13,265	19.3	13,324	19.2	
75% Read		256	13,287	19.3	11,696	21.9	11,840	21.6	11,864	21.6	
50% Read		256	12,252	21.0	11,677	21.9	12,543	20.4	12,509	20.5	
25% Read		256	10,013	25.6	10,024	25.5	9,806	26.1	9,814	26.1	
100% Write		256	9,755	26.3	9,850	26.0	9,630	26.6	9,633	26.6	

Table 55. Summary of peak results, 64-disk tests

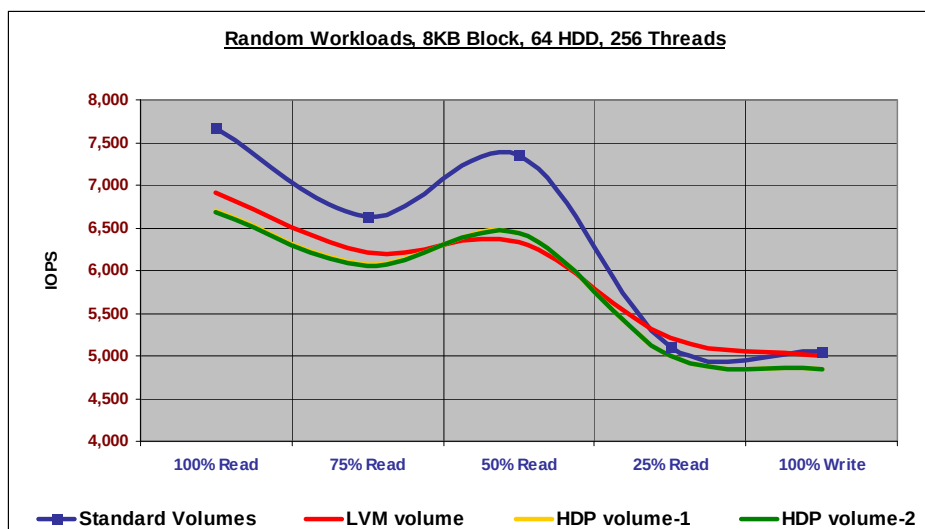


Figure 72. Random workload TPS, 64-disk tests

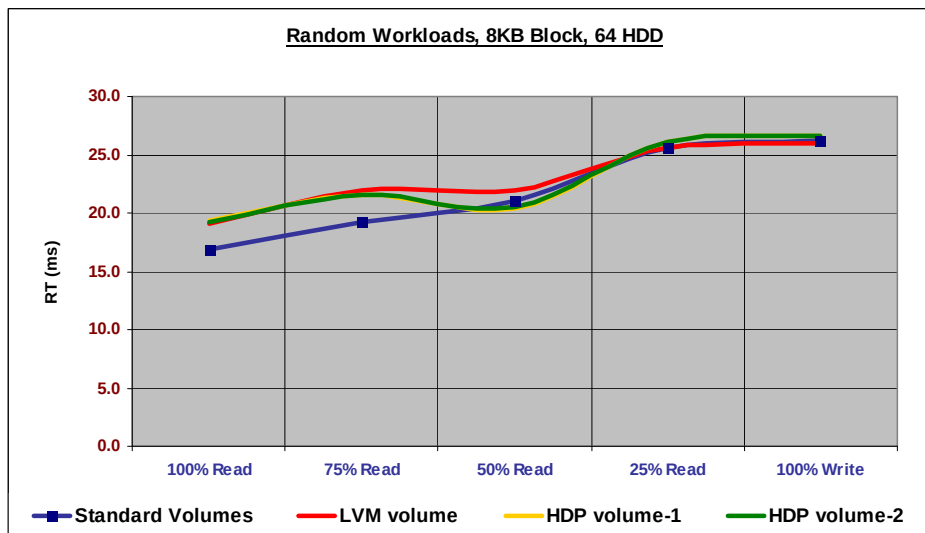


Figure 73. Random workload RT, 64-disk tests

Sequential Workloads Comparisons

The following summary tables and charts show the peak Throughput results (MB/s) for the Sequential workloads.

16-DISK TESTS

As can be seen in the table and chart, there was little difference between the HDP volume-1 (1 LDEV per AG) and HDP volume-2 (4 LDEVs per AG) configurations. The 100% Read workload was the only one from all tests conducted where the LVM volume could outperform the Standard volumes. Notice how both of the HDP tests slightly outperformed the use of a host LVM volume for mixed read-write workloads.

16 HDD, 2+2	Sequential		USP V		Cache Miss
Workload	Threads used for this data	Standard volumes MB/sec	LVM volume MB/sec	HDP volume-1 MB/sec	HDP volume-2 MB/sec
100% Read	512	856	935	715	750
75% Read	64	674	582	621	635
50% Read	64	518	463	544	540
25% Read	64	475	377	430	427
100% Write	64	573	430	387	387

Table 56. Summary of peak sequential results, 16-disk

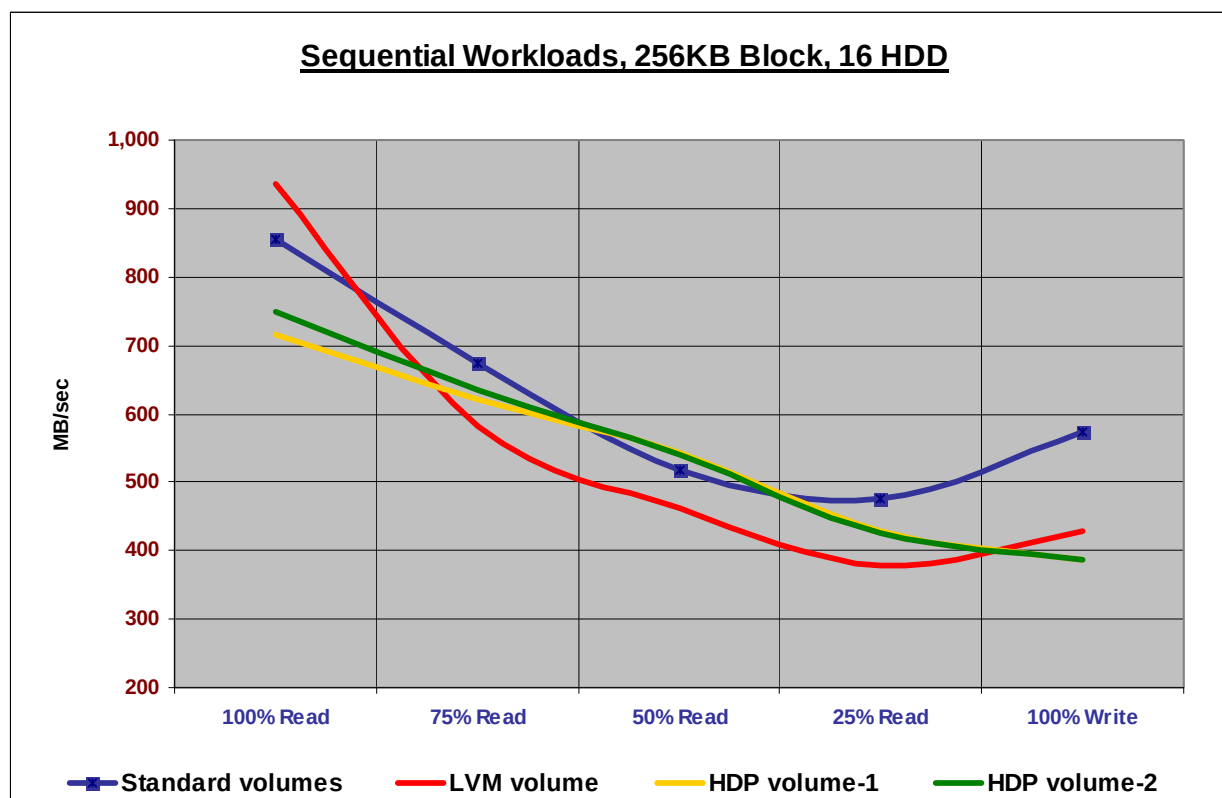


Figure 74. Sequential workloads, MB/sec, 16-disk tests

32-DISK TESTS

As can be seen in the table and chart, once again there was little difference between the HDP volume-1 (1 LDEV per AG) and HDP volume-2 (8 LDEVs per AG) configurations. The HDP results tests fell in line with the Standard volume results for the 50% and 25% Read workloads. Interestingly, the performance of the LVM volume fell well short of the HDP volumes for the 50% and 25% Read workloads.

32 HDD, 2+2	Sequential		USP V		Cache Miss
Workload	Threads used for this data	Standard volumes MB/sec	LVM volume MB/sec	HDP volume-1 MB/sec	HDP volume-2 MB/sec
100% Read	512	1,595	1,417	1104	1035
75% Read	128	1,195	1,035	965	950
50% Read	128	947	694	922	923
25% Read	128	901	605	845	853
100% Write	128	996	827	773	774

Table 57. Summary of peak sequential results, 32-disk.

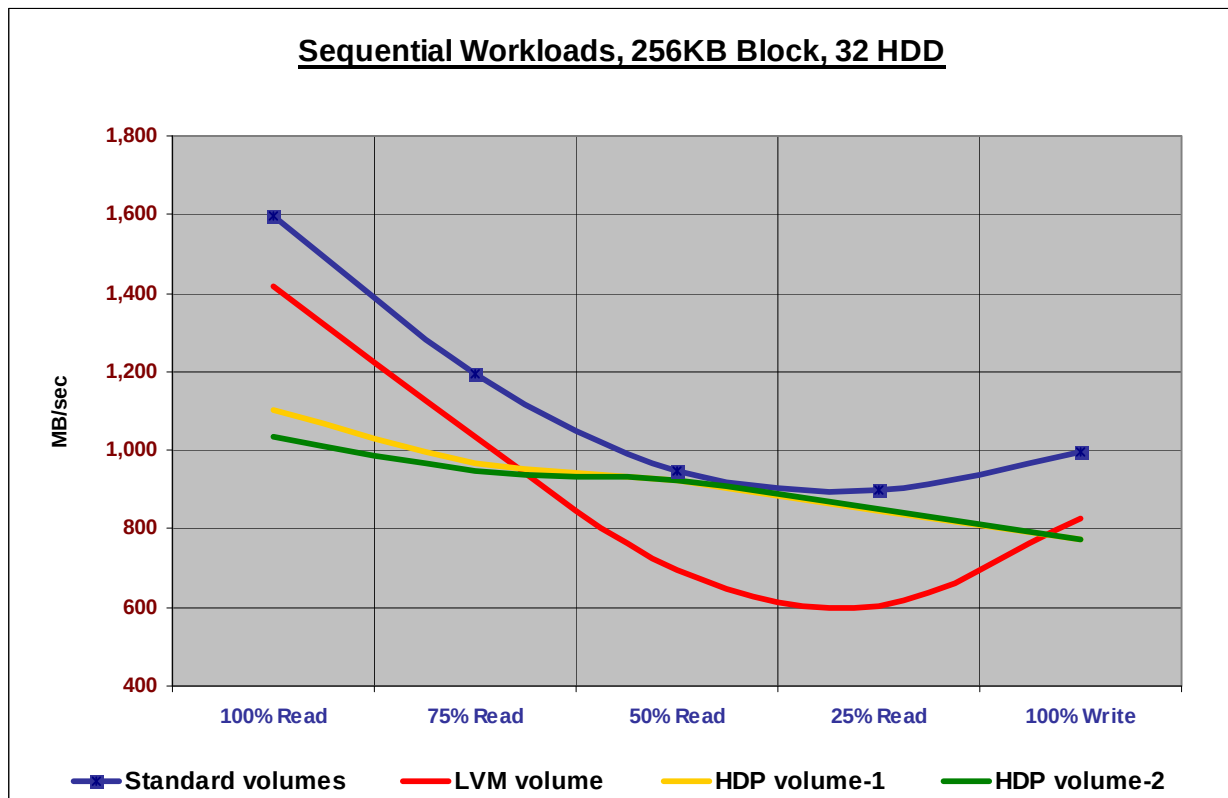


Figure 75. Sequential workloads, MB/sec, 32-disk tests.

64-DISK TESTS

As can be seen in the table and chart, there was some divergence for the 100% Read and 75% Read workloads for the HDP volume-1 (1 LDEV per AG) and HDP volume-2 (16 LDEVs per AG) configurations. The Standard LDEV tests again fell in line with the HDP results for the higher Write component tests. The HDP volumes easily outperformed the LVM volume except for the 75% Read tests. The LVM results were very poor for the 50%, 25% and 100% Write workloads.

64 HDD, 2+2	Sequential USP V				
		Standard volumes	LVM volume	HDP volume-1	HDP volume-2
Workload	Threads used for this data	MB/sec	MB/sec	MB/sec	MB/sec
100% Read	512	1,748	1,484	1,696	1,610
75% Read	256	1,908	1,698	1,463	1,389
50% Read	256	1,730	1,256	1,623	1,578
25% Read	256	1,649	1,140	1,618	1,633
100% Write	256	1,543	1,348	1,562	1,530

Table 58. Summary of peak sequential results, 64-disk.

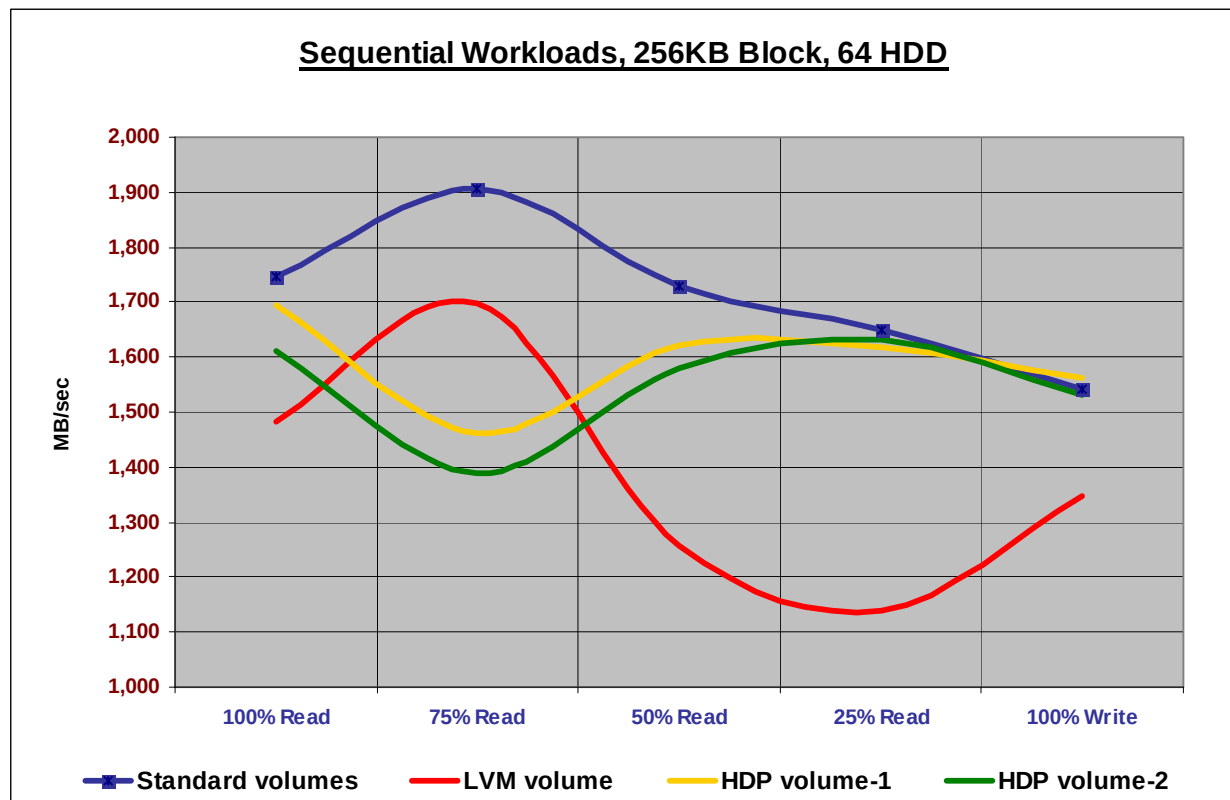


Figure 76. Sequential workloads, MB/sec, 64-disk tests.

XVI. Summary

The Universal Storage Platform V offers great flexibility and it should perform superbly in any environment. At the hardware level, it offers considerably more IOPS power than does the equivalent USP line of products. Even though the USP already delivered best-of-class Sequential performance (about 3x higher than any competition with 9GB/s Read, 4GB/s write), the USP V is outperforming the USP on initial tests. Newer versions of microcode should open up that advantage even more.

It is important for users to understand the two configuration methods now available with the new Dynamic Provisioning feature. The HDP feature can be used for configuring Thin Provisioning and for the creation of High Performance volumes. While only certain applications will benefit from the Thin Provisioning aspect, most can take advantage of the high performance per volume feature.

It is vital, however, that Hitachi Data Systems' sales personnel, technical support staff, value-added resellers, and others who are responsible delivery of solutions, invest the time required to design the best possible solution to meet each customer's unique requirements, whether it be capacity, reliability, performance or cost. To assist in delivering the highest quality solution, Hitachi Data Systems' Global Solution Services should be engaged.

APPENDIX 1. USP V (Frames, HDUs, and Parity Groups).

USP-V - L2 Frame (256 HDD)																				
DKA 8								H D U	DKA 7											
18-16	18-14	18-12	18-10	18-8	18-6	18-4	18-2		17-16	17-14	17-12	17-10	17-8	17-6	17-4	17-2				
12:0F	12:0D	12:0B	12:09	12:07	12:05	12:03	12:01		11:0F	11:0D	11:0B	11:09	11:07	11:05	11:03	11:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
18-15	18-13	18-11	18-9	18-7	18-5	18-3	18-1	H D U	17-15	17-13	17-11	17-9	17-7	17-5	17-3	17-1				
12:0E	12:0C	12:0A	12:08	12:06	12:04	12:02	12:00		11:0E	11:0C	11:0A	11:08	11:06	11:04	11:02	11:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
DKA 6								H D U	DKA 5											
16-16	16-14	16-12	16-10	16-8	16-6	16-4	16-2		15-16	15-14	15-12	15-10	15-8	15-6	15-4	15-2				
10:0F	10:0D	10:0B	10:09	10:07	10:05	10:03	10:01		0F:0F	0F:0D	0F:0B	0F:09	0F:07	0F:05	0F:03	0F:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
16-15	16-13	16-11	16-9	16-7	16-5	16-3	16-1	H D U	15-15	15-13	15-11	15-9	15-7	15-5	15-3	15-1				
10:0E	10:0C	10:0A	10:08	10:06	10:04	10:02	10:00		0F:0E	0F:0C	0F:0A	0F:08	0F:06	0F:04	0F:02	0F:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				

USP-V - L1 Frame (256 HDD)																			
DKA 8								H D U	DKA 7										
10-16	10-14	10-12	10-10	10-8	10-6	10-4	10-2		9-16	9-14	9-12	9-10	9-8	9-6	9-4	9-2			
0A:0F	0A:0D	0A:0B	0A:09	0A:07	0A:05	0A:03	0A:01		09:0F	09:0D	09:0B	09:09	09:07	09:05	09:03	09:01			
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD			
10-15	10-13	10-11	10-9	10-7	10-5	10-3	10-1	H D U	9-15	9-13	9-11	9-9	9-7	9-5	9-3	9-1			
0A:0E	0A:0C	0A:0A	0A:08	0A:06	0A:04	0A:02	0A:00		09:0E	09:0C	09:0A	09:08	09:06	09:04	09:02	09:00			
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD			
DKA 6								H D U	DKA 5										
8-16	8-14	8-12	8-10	8-8	8-6	8-4	8-2		7-16	7-14	7-12	7-10	7-8	7-6	7-4	7-2			
08:0F	08:0D	08:0B	08:09	08:07	08:05	08:03	08:01		07:0F	07:0D	07:0B	07:09	07:07	07:05	07:03	07:01			
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD			
8-15	8-13	8-11	8-9	8-7	8-5	8-3	8-1	H D U	7-15	7-13	7-11	7-9	7-7	7-5	7-3	7-1			
08:0E	08:0C	08:0A	08:08	08:06	08:04	08:02	08:00		07:0E	07:0C	07:0A	07:08	07:06	07:04	07:02	07:00			
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD			

USP-V - Center Frame (128 HDD)																	
DKA 2									DKA 1								
2-16	2-14	2-12	2-10	2-8	2-6	2-4	2-2	H D U	1-16	1-14	1-12	1-10	1-8	1-6	1-4	1-2	
02:0F	02:0D	02:0B	02:09	02:07	02:05	02:03	02:01		01:0F	01:0D	01:0B	01:09	01:07	01:05	01:03	01:01	
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
2-15	2-13	2-11	2-9	2-7	2-5	2-3	2-1	H D U	1-15	1-13	1-11	1-9	1-7	1-5	1-3	1-1	
02:0E	02:0C	02:0A	02:08	02:06	02:02	02:02	02:00		01:0E	01:0C	01:0A	01:08	01:06	01:04	01:02	01:00	
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

RAID 600 - R1 Frame (256 HDD)																				
DKA 4								H D U	DKA 3											
6-16	6-14	6-12	6-10	6-8	6-6	6-4	6-2		5-16	5-14	5-12	5-10	5-8	5-6	5-4	5-2				
06:0F	06:0D	06:0B	06:09	06:07	06:05	06:03	06:01		05:0F	05:0D	05:0B	05:09	05:07	05:05	05:03	05:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
								H D U												
6-15	6-13	6-11	6-9	6-7	6-5	6-3	6-1		5-15	5-13	5-11	5-9	5-7	5-5	5-3	5-1				
06:0E	06:0C	06:0A	06:08	06:06	06:04	06:02	06:00		05:0E	05:0C	05:0A	05:08	05:06	05:04	05:02	05:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
DKA 2								H D U	DKA 1											
4-16	4-14	4-12	4-10	4-8	4-6	4-4	4-2		3-16	3-14	3-12	3-10	3-8	3-6	3-4	3-2				
04:0F	04:0D	04:0B	04:09	04:07	04:05	04:03	04:01		03:0F	03:0D	03:0B	03:09	03:07	03:05	03:03	03:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
								H D U												
4-15	4-13	4-11	4-9	4-7	4-5	4-3	4-1		3-15	3-13	3-11	3-9	3-7	3-5	3-3	3-1				
04:0E	04:0C	04:0A	04:08	04:06	04:04	04:02	04:00		03:0E	03:0C	03:0A	03:08	03:06	03:04	03:02	03:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				

USP-V - R2 Frame (256 HDD)																				
DKA 4								H D U	DKA 3											
14-16	14-14	14-12	14-10	14-8	14-6	14-4	14-2		13-16	13-14	13-12	13-10	13-8	13-6	13-4	13-2				
0E:0F	0E:0D	0E:0B	0E:09	0E:07	0E:05	0E:03	0E:01		0D:0F	0D:0D	0D:0B	0D:09	0D:07	0D:05	0D:03	0D:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
14-15	14-13	14-11	14-9	14-7	14-5	14-3	14-1	H D U	13-15	13-13	13-11	13-9	13-7	13-5	13-3	13-1				
0E:0E	0E:0C	0E:0A	0E:08	0E:06	0E:04	0E:02	0E:00		0D:0E	0D:0C	0D:0A	0D:08	0D:06	0D:04	0D:02	0D:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
DKA 2									DKA 1											
12-16	12-14	12-12	12-10	12-8	12-6	12-4	12-2	H D U	11-16	11-14	11-12	11-10	11-8	11-6	11-4	11-2				
0C:0F	0C:0D	0C:0B	0C:09	0C:07	0C:05	0C:03	0C:01		0B:0F	0B:0D	0B:0B	0B:09	0B:07	0B:05	0B:03	0B:01				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				
12-15	12-13	12-11	12-9	12-7	12-5	12-3	12-1		H D U	11-15	11-13	11-11	11-9	11-7	11-5	11-3	11-1			
0C:0E	0C:0C	0C:0A	0C:08	0C:06	0C:04	0C:02	0C:00	0B:0E		0B:0C	0B:0A	0B:08	0B:06	0B:04	0B:02	0B:00				
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD				

APPENDIX 2. USP (Frames, HDUs, and Parity Groups).

USP - L2 Frame (256 HDD)																							
DKA 4												DKA 3											
18-16	18-14	18-12	18-10	18-8	18-6	18-4	18-2	H	16-16	16-14	16-12	16-10	16-8	16-6	16-4	16-2	D	10-16	10-14	10-12	10-10	10-8	10-6
12:0F	12:0D	12:0B	12:09	12:07	12:05	12:03	12:01	U	10:0F	10:0D	10:0B	10:09	10:07	10:05	10:03	10:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
18-15	18-13	18-11	18-9	18-7	18-5	18-3	18-1	H	16-15	16-13	16-11	16-9	16-7	16-5	16-3	16-1	D	10-15	10-13	10-11	10-9	10-7	10-5
12:0E	12:0C	12:0A	12:08	12:06	12:04	12:02	12:00	U	10:0E	10:0C	10:0A	10:08	10:06	10:04	10:02	10:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
DKA 4												DKA 3											
17-16	17-14	17-12	17-10	17-8	17-6	17-4	17-2	H	15-16	15-14	15-12	15-10	15-8	15-6	15-4	15-2	D	13-16	13-14	13-12	13-10	13-8	13-6
11:0F	11:0D	11:0B	11:09	11:07	11:05	11:03	11:01	U	0F:0F	0F:0D	0F:0B	0F:09	0F:07	0F:05	0F:03	0F:01	U	0E:0F	0E:0D	0E:0B	0E:09	0E:07	0E:05
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
17-15	17-13	17-11	17-9	17-7	17-5	17-3	17-1	H	15-15	15-13	15-11	15-9	15-7	15-5	15-3	15-1	D	13-15	13-13	13-11	13-9	13-7	13-5
11:0E	11:0C	11:0A	11:08	11:06	11:04	11:02	11:00	U	0F:0E	0F:0C	0F:0A	0F:08	0F:06	0F:04	0F:02	0F:00	U	0E:0E	0E:0C	0E:0A	0E:08	0E:06	0E:04
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

USP - L1 Frame (256 HDD)																							
DKA 4												DKA 3											
10-16	10-14	10-12	10-10	10-8	10-6	10-4	10-2	H	8-16	8-14	8-12	8-10	8-8	8-6	8-4	8-2	D	0A:0F	0A:0D	0A:0B	0A:09	0A:07	0A:05
0A:0F	0A:0D	0A:0B	0A:09	0A:07	0A:05	0A:03	0A:01	U	08:0F	08:0D	08:0B	08:09	08:07	08:05	08:03	08:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
10-15	10-13	10-11	10-9	10-7	10-5	10-3	10-1	H	8-15	8-13	8-11	8-9	8-7	8-5	8-3	8-1	D	0A:0E	0A:0C	0A:0A	0A:08	0A:06	0A:04
0A:0E	0A:0C	0A:0A	0A:08	0A:06	0A:04	0A:02	0A:00	U	08:0E	08:0C	08:0A	08:08	08:06	08:04	08:02	08:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
DKA 4												DKA 3											
9-16	9-14	9-12	9-10	9-8	9-6	9-4	9-2	H	7-16	7-14	7-12	7-10	7-8	7-6	7-4	7-2	D	09:0F	09:0D	09:0B	09:09	09:07	09:05
09:0F	09:0D	09:0B	09:09	09:07	09:05	09:03	09:01	U	07:0F	07:0D	07:0B	07:09	07:07	07:05	07:03	07:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
9-15	9-13	9-11	9-9	9-7	9-5	9-3	9-1	H	7-15	7-13	7-11	7-9	7-7	7-5	7-3	7-1	D	09:0E	09:0C	09:0A	09:08	09:06	09:04
09:0E	09:0C	09:0A	09:08	09:06	09:04	09:02	09:00	U	07:0E	07:0C	07:0A	07:08	07:06	07:04	07:02	07:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

USP - Center Frame (128 HDD)																							
DKA 1												DKA 1											
2-16	2-14	2-12	2-10	2-8	2-6	2-4	2-2	H	1-16	1-14	1-12	1-10	1-8	1-6	1-4	1-2	D	02:0F	02:0D	02:0B	02:09	02:07	02:05
02:0F	02:0D	02:0B	02:09	02:07	02:05	02:03	02:01	U	01:0F	01:0D	01:0B	01:09	01:07	01:05	01:03	01:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
2-15	2-13	2-11	2-9	2-7	2-5	2-3	2-1	H	1-15	1-13	1-11	1-9	1-7	1-5	1-3	1-1	D	02:0E	02:0C	02:0A	02:08	02:06	02:04
02:0E	02:0C	02:0A	02:08	02:06	02:04	02:02	02:00	U	01:0E	01:0C	01:0A	01:08	01:06	01:04	01:02	01:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

USP - R1 Frame (256 HDD)																							
DKA 2												DKA 1											
6-16	6-14	6-12	6-10	6-8	6-6	6-4	6-2	H	4-16	4-14	4-12	4-10	4-8	4-6	4-4	4-2	D	06:0F	06:0D	06:0B	06:09	06:07	06:05
06:0F	06:0D	06:0B	06:09	06:07	06:05	06:03	06:01	U	04:0F	04:0D	04:0B	04:09	04:07	04:05	04:03	04:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
6-15	6-13	6-11	6-9	6-7	6-5	6-3	6-1	H	4-15	4-13	4-11	4-9	4-7	4-5	4-3	4-1	D	06:0E	06:0C	06:0A	06:08	06:06	06:04
06:0E	06:0C	06:0A	06:08	06:06	06:04	06:02	06:00	U	04:0E	04:0C	04:0A	04:08	04:06	04:04	04:02	04:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
DKA 2												DKA 1											
5-16	5-14	5-12	5-10	5-8	5-6	5-4	5-2	H	3-16	3-14	3-12	3-10	3-8	3-6	3-4	3-2	D	05:0F	05:0D	05:0B	05:09	05:07	05:05
05:0F	05:0D	05:0B	05:09	05:07	05:05	05:03	05:01	U	03:0F	03:0D	03:0B	03:09	03:07	03:05	03:03	03:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
5-15	5-13	5-11	5-9	5-7	5-5	5-3	5-1	H	3-15	3-13	3-11	3-9	3-7	3-5	3-3	3-1	D	05:0E	05:0C	05:0A	05:08	05:06	05:04
05:0E	05:0C	05:0A	05:08	05:06	05:04	05:02	05:00	U	03:0E	03:0C	03:0A	03:08	03:06	03:04	03:02	03:00	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD		4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

USP - R2 Frame (256 HDD)																									
DKA 2												DKA 1													
14-16	14-14	14-12	14-10	14-8	14-6	14-4	14-2	H	12-16	12-14	12-12	12-10	12-8	12-6	12-4	12-2	D	0E:0F	0E:0D	0E:0B	0E:09	0E:07	0E:05	0E:03	0E:01
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	0C:0F	0C:0D	0C:0B	0C:09	0C:07	0C:05	0C:03	0C:01	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
14-15	14-13	14-11	14-9	14-7	14-5	14-3	14-1	H	12-15	12-13	12-11	12-9	12-7	12-5	12-3	12-1	D	0C:0E	0C:0C	0C:0A	0C:08	0C:06	0C:04	0C:02	0C:00
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
DKA 2												DKA 1													
13-16	13-14	13-12	13-10	13-8	13-6	13-4	13-2	H	11-16	11-14	11-12	11-10	11-8	11-6	11-4	11-2	D	0B:0F	0B:0D	0B:0B	0B:09	0B:07	0B:05	0B:03	0B:01
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD
13-15	13-13	13-11	13-9	13-7	13-5	13-3	13-1	H	11-15	11-13	11-11	11-9	11-7	11-5	11-3	11-1	D	0B:0E	0B:0C	0B:0A	0B:08	0B:06	0B:04	0B:02	0B:00
4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	U	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD	4 HDD

APPENDIX 3. Open Systems RAID Mechanisms

This section describes how the Open-V Open Systems Groups function on both the USP and the USP V. The track size (chunk) used for Open-V is 256KB. The available 9900 Lightning style Open-X emulation is available, with its track (chunk) size of 48KB. One should avoid mixing Open-X and the standard Open-V emulations on the same storage system.

When using generic RAID5, one would normally calculate the stripe width (row size) by multiplying the RAID chunk size (or “stripe size”) by 7 (if 7D+1P). In the case of Open Systems (Open-V volumes), the USP V uses a native chunk size of 256KB. Hence you would expect a row size of 1,792KB. However, the USP V manages RAID5 somewhat differently than any other vendor. For both USP V RAID5+ types, there are two adjacent RAID chunks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the two Figures below. Thus, there are twice as many chunks per row as normal. So the USP V’s RAID5+ 7D+1P stripe width is actually 3,584KB (2 x 7 x 256KB). The USP V’s RAID5+ 3+1p stripe width is actually 1,536KB (2 x 3 x 256KB).

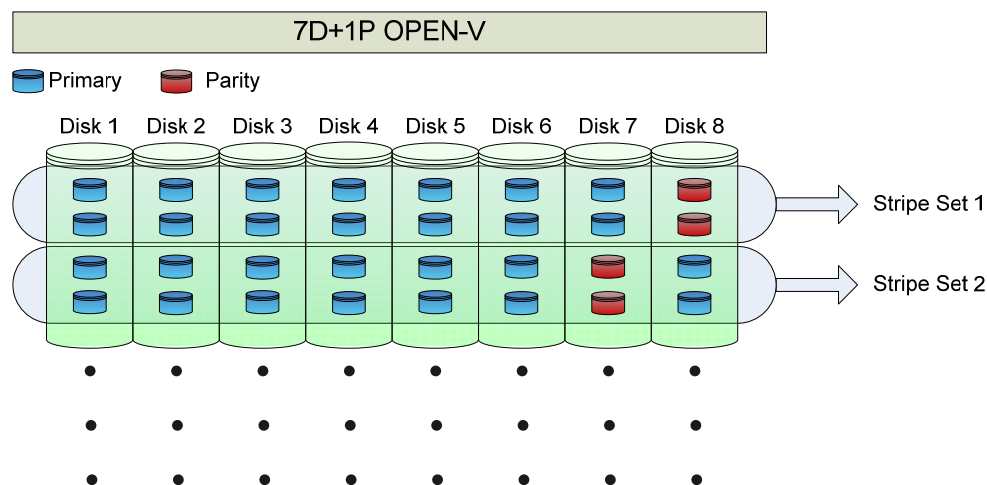


Figure 77. RAID5+ (7D+1P) layout (3,584KB stripe width).

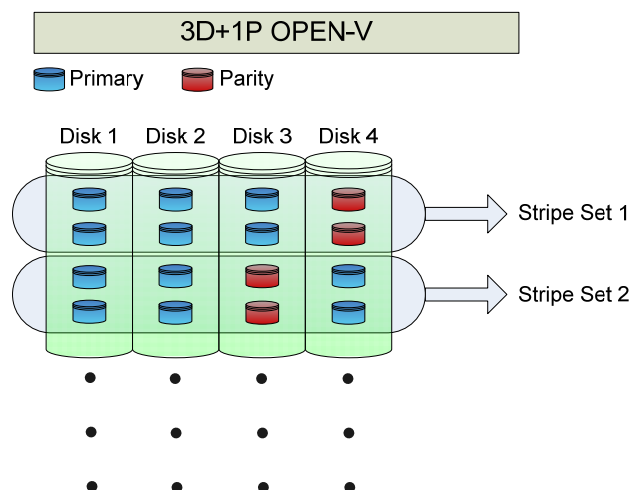


Figure 78. RAID5+ (3D+1P) layout (1,536KB stripe width).

When using generic RAID10, one would normally calculate the stripe width (row size) by multiplying the RAID chunk size (or “stripe size”) by 4 (if 4D+4D). In the case of Open Systems (Open-V volumes), the USP V uses a native chunk size of 256KB. Hence you would expect a row size of 1,024KB.

However, the USP V manages RAID10+ in the same manner as RAID5+, where there are two adjacent RAID chunks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the two Figures below. Thus, there are twice as many chunks per row as normal. So the USP V’s RAID10+ (4D+4D) stripe width is 2,048KB (2 x 4 x 256KB), and for 2D+2D it is 1,024KB (2 x 4 x 256KB). Also, note that the 4D+4D is truly a single 8-disk set of mirrors. The Storage Navigator setup GUI and some out-of-date factory documentation makes it look like there are two 2D+2D LDEVs in a concatenated mode, which is incorrect.

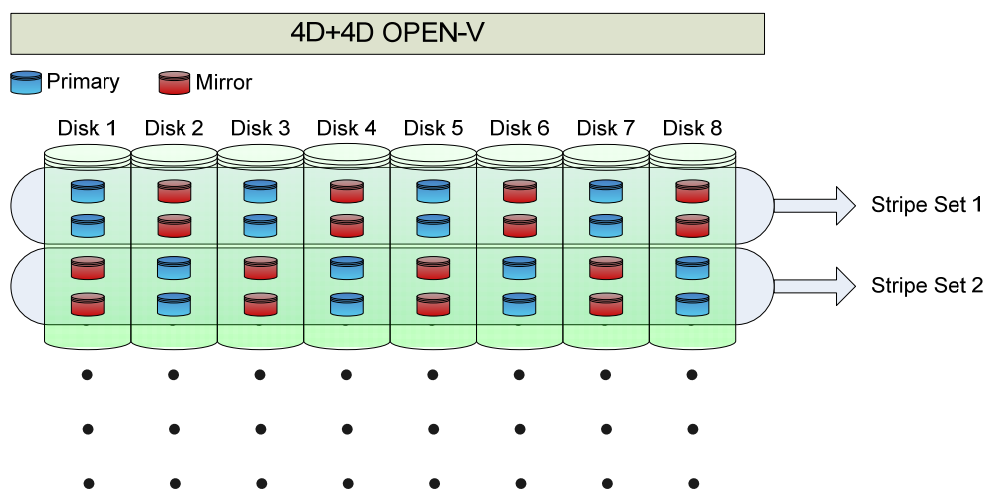


Figure 79. RAID10+ (4D+4D) layout (stripe width 2,048KB).

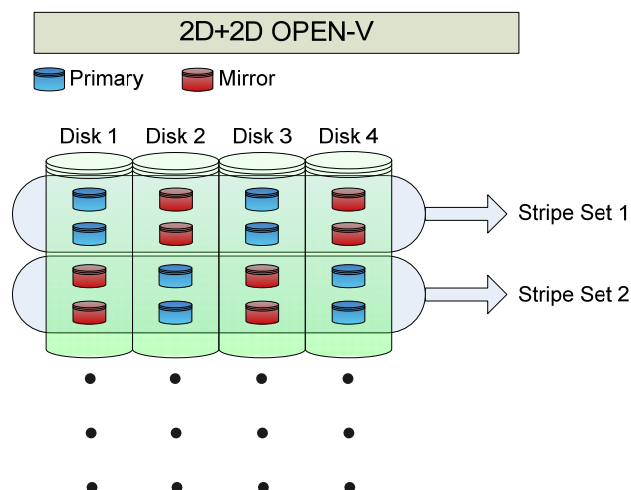


Figure 80. RAID10+ (2D+2D) layout (stripe width 1024KB).

The USP V manages RAID6+ in the same manner as RAID5+, where there are two adjacent RAID chunks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the Figure below. Thus, there are twice as many chunks per row as normal. So the USP V's RAID6 6D+2P stripe width is 3,072KB (12 x 256KB).

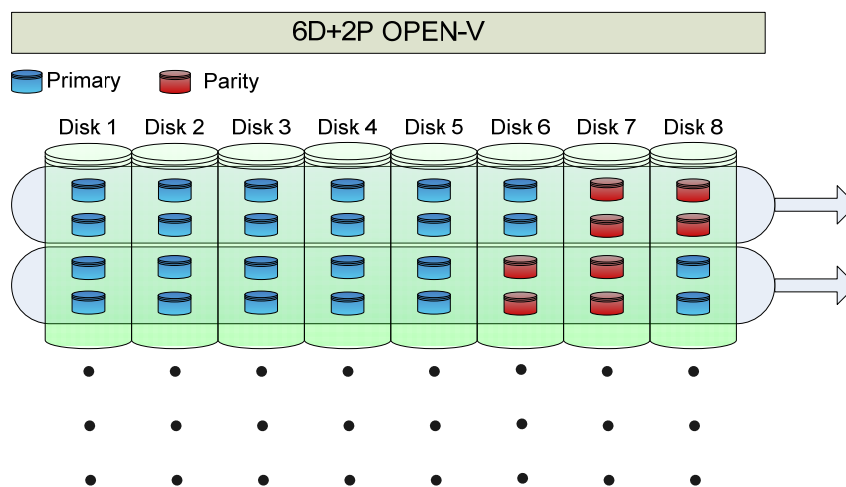


Figure 81. RAID6+ (6D+2P) layout (stripe width 3,072KB).

APPENDIX 4. Mainframe 3390x and Open-x RAID Mechanisms

This section describes how the 3390x (mainframe) and Open-X (9900V Lightning style Open Systems volumes) Array Groups function on both the USP and the USP V. The track size (chunk) used for 3390x is 58KB, and the Open-X track (chunk) size is 48KB. One should avoid mixing Open-X and the standard Open-V on the same storage system.

RAID5+

When using generic RAID5, one would normally calculate the stripe width (row size) by multiplying the RAID chunk size (or “stripe size”) by 7 (if 7D+1P). In the case of mainframe emulation (3390-x volumes), the USP V uses a native track size of 58KB. Hence you would expect a row size of 406KB (7 x 58KB). For the Open-X emulation, the track size is 48KB, with an expected row size of 336KB (7 x 48KB).

However, the USP V manages RAID5 somewhat differently than any other vendor. For both USP V RAID5+ types, there are eight adjacent tracks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the two Figures below. The USP V’s RAID5+ 7D+1P stripe width for 3390-X is 3,248KB (7 x 8 x 58KB), and for Open-X it is 2,688KB (7 x 8 x 48KB). The USP V’s RAID5+ 3+1p stripe width for 3390-X is 1,392KB (3 x 8 x 256KB), and for Open-X it is 1,152KB (3 x 8 x 48KB).

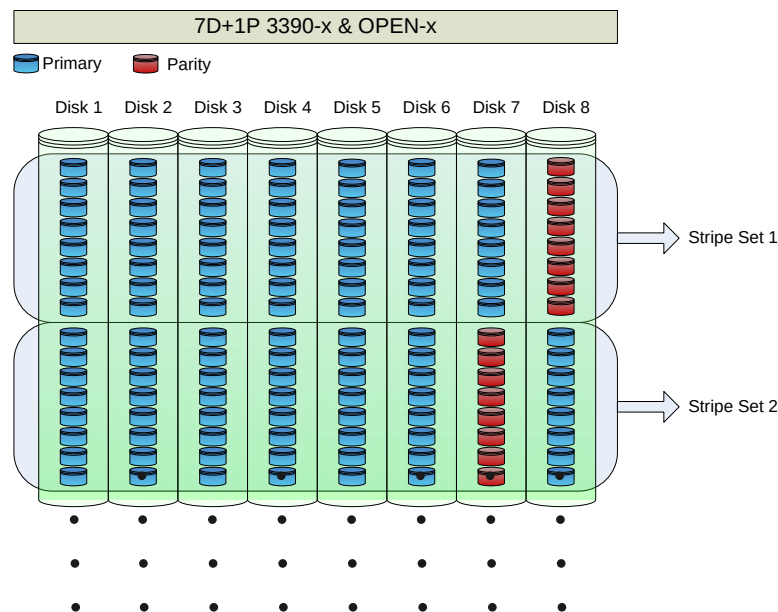


Figure 82. RAID5+ (7D+1P) layout, 3390-X and Open-X.

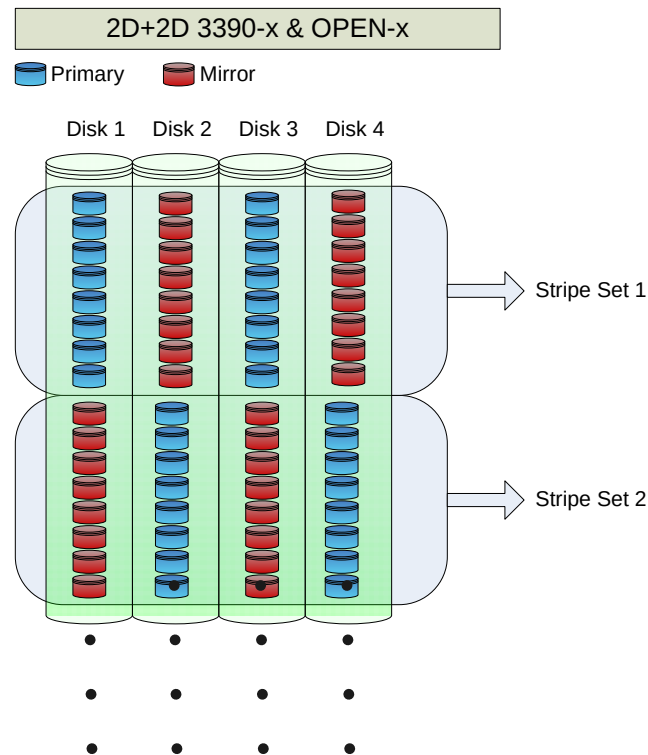


Figure 83. RAID5+ (3D+1P) layout, 3390-X and Open-X.

RAID10+

When using generic RAID10, one would normally calculate the stripe width (row size) by multiplying the RAID chunk size (or “track size”) by 4 (if 4D+4D). In the case of mainframe emulation (3390-x volumes), the USP V uses a native track size of 58KB. Hence you would expect a row size of 232KB (4 x 58KB). For the Open-X emulation, the track size is 48KB, with an expected row size of 192KB (4 x 48KB).

However, the USP V manages RAID10+ in the same manner as RAID5+, where there are eight adjacent tracks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the two Figures below. So the USP V’s RAID10+ (4D+4D) stripe width for 3390-X is 1,856KB (4 x 8 x 58KB), and for 2D+2D it is 928KB (2 x 8 x 58KB). The USP V’s RAID10+ (4D+4D) stripe width for Open-X is 1,536KB (4 x 8 x 48KB), and for 2D+2D it is 768KB (2 x 8 x 48KB).

Also, note that the 4D+4D is truly a single 8-disk set of mirrors. The Storage Navigator setup GUI makes it look like there are two 2D+2D LDEVs in a concatenated mode, which is incorrect.

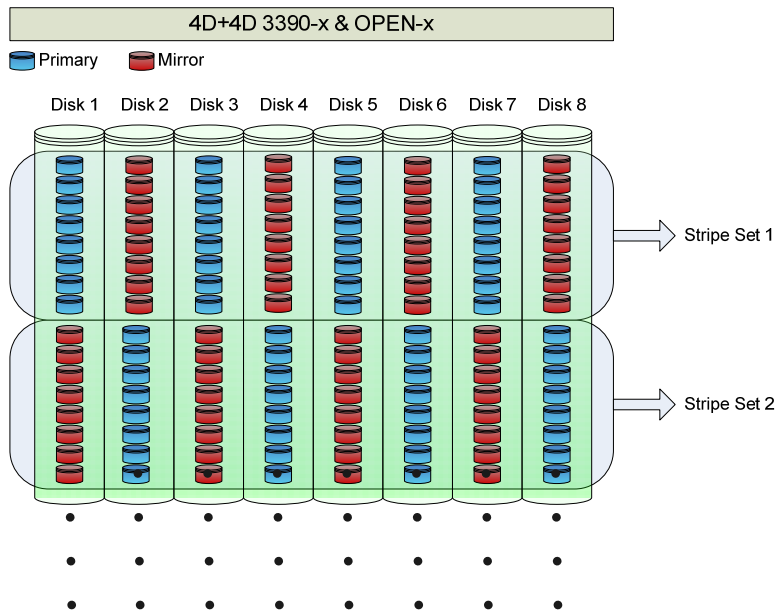


Figure 84. RAID10+ (4D+4D) layout, 3390-X and Open-X.

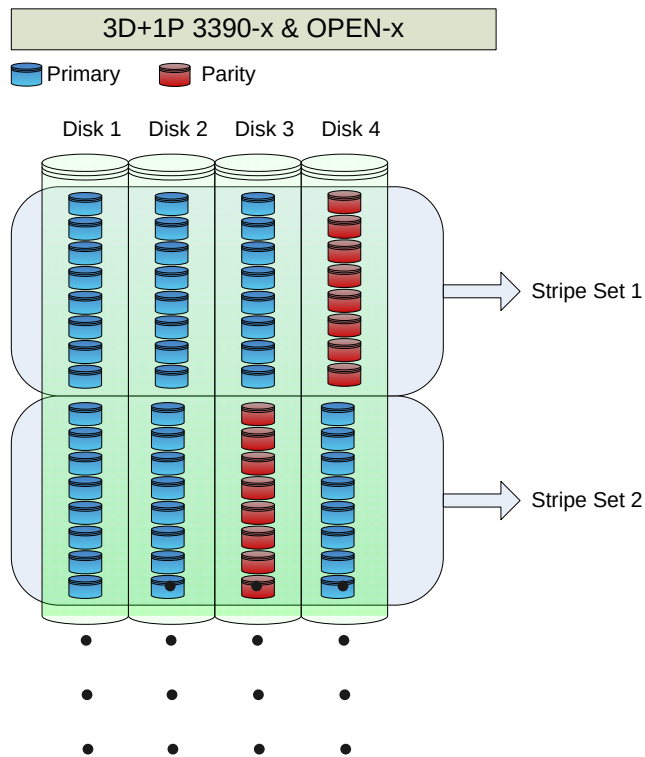


Figure 85. RAID10+ (2D+2D) layout, 3390-X and Open-X.

RAID6+

The USP V manages RAID6+ in the same manner as RAID5+ where there are eight adjacent tracks taken from the *same* disk before switching to the next disk in the row. This is illustrated in the Figure below. Thus, there are twice as many chunks per row as normal. So, for 3390-X, the USP V's RAID6 6D+2P stripe width is actually 2,784KB (6 x 8 x 58KB), and for Open-X it is 2,304KB (6 x 8 x 48KB).

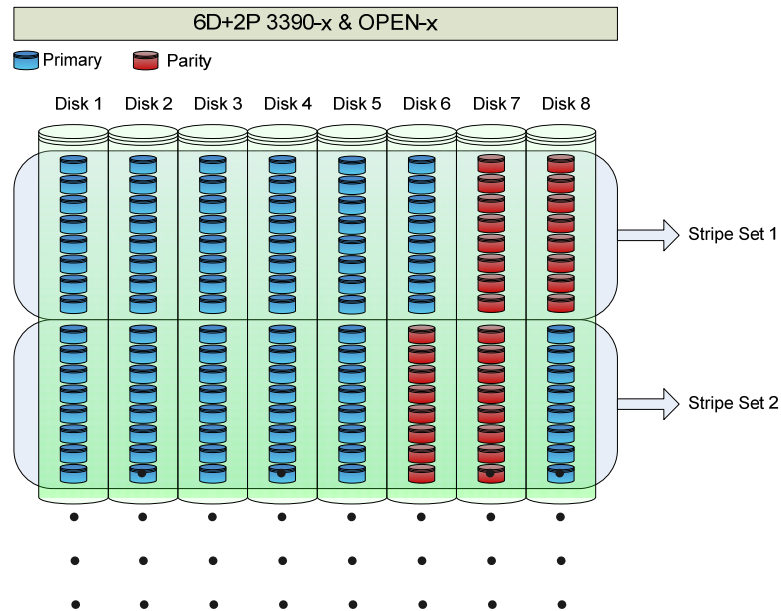


Figure 86. RAID6+ (6D+2P) layout, 3390-X and Open-X.

APPENDIX 5. Concatenated Array Groups

The Concatenated Array Group feature allows you to configure all of the space from either two or four RAID-5 7D+1P Array Groups into an association whereby all LDEVs created on these Array Groups are actually striped across all of the elements. Recall that a slice (or partition) created on a standard Array Group is an **LDEV** (Logical Device), becoming a **LUN** (Logical Unit) once it has been given a name and mapped to a host port. Now there is a new mechanism known as the **VDEV** (Virtual Device). It is the name of a row of LDEVs created across the members of the striped Array Groups. When the VDEV is mapped to a host port, it – not an Array Group’s LDEV - then becomes a LUN.

Looking at the example in the Figure below, four Array Groups are used in a Concatenated Array Group configuration. This shows how the VDEV “slices” are created across each Array Group. The VDEV is a container that encompasses four LDEVs, where each LDEV comes from a different Array Group. The VDEV is not a visible entity in Storage Navigator, but is more of a concept. When you create an LDEV on an Array Group, you specify the overall size that you wish to carve out from the entire Stripe. It appears that you are creating the entire space on the Array Group in question, but it is actually automatically dividing the space request into four equal parts, then creating an LDEV on each Array Group. You cannot see these other LDEVs in Storage Navigator, and the LDEV that you just made appears to have all of the space. Think of this “primary” LDEV as the anchor point for the other three “hidden” LDEVs that are automatically generated and make up the rest of that particular VDEV. [Note: “primary” and “hidden” are my terms for these elements in the absence of any official terms.]

In order to use all of the space available in the Concatenated Array Group, you must create “primary” LDEVs on each Array Group element in turn. You cannot just go through one of the Array Groups and create these “primary” LDEVs to use up all the space from the entire striped set. You may create LDEVs on an Array Group until all of the space available in that Array Group appears to have been consumed. You then move on to the next Array Group in the set and create other LDEVs. You should make a map of what you are doing in order to keep track of the remaining space which can be allocated from each Array Group. The Storage Navigator tool does show you the unused space per Array Group, but it lacks the ability to convey the idea of “primary” or “hidden” LDEV space.

The best way to describe the Concatenated Array Group concept is by example. We will look at a four Array Group scenario that uses 32 300GB FC drives. Look at the Figure below for a map of the VDEVs. The first step is to create the four individual RAID-5 7D+1P Array Groups as per usual. The new step is to then associate these four groups together as a Concatenated Array. From this step, you will then create LDEVs on each Array Group as you normally would, but with the caveat that the assignable space on each Array Group appears differently than you would expect. In the Figure, the bold LDEV names represent the “primaries” and those in italics (on the same row) are the “hidden” elements. Notice that for this example there are three primary LDEVs established on each Array Group. On the left side of the Figure are the VDEV’s name and its approximate usable size.

More specifically, look at the first row, or VDEV-1. This was set up as an 800GB Open-V LUN. In Storage Navigator it appears that all 800GB of the space came from Array Group 1. In actuality, there is 200GB taken from each Array Group, including Array Group 1. So, once you have created the first three LDEVs on Array Group 1 (800GB, 720GB, 454GB), all of the “space” on that Array Group has been consumed, even though in reality there is but 493.5GB of real space used.

In order to use all of the space in Array Group 1 (and the others), you must next create LDEVs in the other three Array Groups in turn. Again, the sum of the space “allocated” in each must be less than the apparent maximum available (about 1,974GB in this case). Note that a “300GB” drive is actually about 282GB. In RAID-5, using 7D+1P, this means that the raw space visible to the host is about $7 \times 282\text{GB}$.

LUNs	VDEVs	ARRAY GROUP 1	ARRAY GROUP 2	ARRAY GROUP 3	ARRAY GROUP 4
LUN-1	VDEV 1 800GB	LDEV-1 "800GB"	(LDEV-1) 200GB	(LDEV-1) 200GB	(LDEV-1) 200GB
LUN-2	VDEV 2 720GB	LDEV-2 "720GB"	(LDEV-2) 180GB	(LDEV-2) 180GB	(LDEV-2) 180GB
LUN-3	VDEV 3 240GB	LDEV-3 "454GB"	(LDEV-3) 113.5GB	(LDEV-3) 113.5GB	(LDEV-3) 113.5GB
LUN-4	VDEV 4 200GB	(LDEV-4) 93.5GB	LDEV-4 "374GB"	(LDEV-4) 93.5GB	(LDEV-4) 93.5GB
LUN-5	VDEV 5 800GB	(LDEV-5) 200GB	LDEV-5 "800GB"	(LDEV-5) 200GB	(LDEV-5) 200GB
LUN-6	VDEV 6 800GB	(LDEV-6) 200GB	LDEV-6 "800GB"	(LDEV-6) 200GB	(LDEV-6) 200GB
LUN-7	VDEV 7 720GB	(LDEV-7) 180GB	(LDEV-7) 180GB	LDEV-7 "720GB"	(LDEV-7) 180GB
LUN-8	VDEV 8 300GB	(LDEV-8) 75GB	(LDEV-8) 75GB	LDEV-8 "300GB"	(LDEV-8) 75GB
LUN-9	VDEV 9 954GB	(LDEV-9) 238.5GB	(LDEV-9) 238.5GB	LDEV-9 "954GB"	(LDEV-9) 238.5GB
LUN-10	VDEV 10 1000GB	(LDEV-10) 250GB	(LDEV-10) 250GB	(LDEV-10) 250GB	LDEV-10 "1000GB"
LUN-11	VDEV 11 100GB	(LDEV-11) 25GB	(LDEV-11) 25GB	(LDEV-11) 25GB	LDEV-11 "100GB"
LUN-12	VDEV 12 874GB	(LDEV-12) 218.5GB	(LDEV-12) 218.5GB	(LDEV-12) 218.5GB	LDEV-12 "874GB"

VDEVs	ARRAY GROUP 1	ARRAY GROUP 2	ARRAY GROUP 3	ARRAY GROUP 4
(32 disks) (28d+4p)	8 FC disks 7d+1p 300GB 10k (282GB raw)	8 FC disks 7d+1p 300GB 10k (282GB raw)	8 FC disks 7d+1p 300GB 10k (282GB raw)	8 FC disks 7d+1p 300GB 10k (282GB raw)
Usable raw space = 7,896GB	Usable space = 1974GB	Usable space = 1974GB	Usable space = 1974GB	Usable space = 1974GB
Stripe Depth = 512KB	Stripe Depth 2 x 256KB	Stripe Depth 2 x 256KB	Stripe Depth 2 x 256KB	Stripe Depth 2 x 256KB
Striped LDEV Width = 14,336KB	Stripe Width = 3,584KB	Stripe Width = 3,584KB	Stripe Width = 3,584KB	Stripe Width = 3,584KB

Figure 87. Concatenated Array Groups Overview.

APPENDIX 6. BED: Loop Maps with Parity Group Names

USP-V (Front View) Cluster-1

BED-6 BED16 Slot 1KU		BED-5 BED14 Slot 1LU		BED-2 BED12 Slot 1BU		BED-1 BED10 Slot 1AU	
Loop	PGrp	Loop	PGrp	Loop	PGrp	Loop	PGrp
0	8, 16	0	7, 15	0	2, 4, 12	0	1, 3, 11
1	8, 16	1	7, 15	1	2, 4, 12	1	1, 3, 11
2	8, 16	2	7, 15	2	2, 4, 12	2	1, 3, 11
3	8, 16	3	7, 15	3	2, 4, 12	3	1, 3, 11

BED-8 BED17 Slot 1KL		BED-7 BED15 Slot 1LL		BED-4 BED13 Slot 1BL		BED-3 BED11 Slot 1AL	
Loop	PGrp	Loop	PGrp	Loop	PGrp	Loop	PGrp
0	10, 18	0	9, 17	0	6, 14	0	5, 13
1	10, 18	1	9, 17	1	6, 14	1	5, 13
2	10, 18	2	9, 17	2	6, 14	2	5, 13
3	10, 18	3	9, 17	3	6, 14	3	5, 13

USP-V (Rear View) Cluster-2

BED-1 BED18 Slot 2MU		BED-2 BED1A Slot 2NU		BED-5 BED1C Slot 2XU		BED-6 BED1E Slot 2WU	
Loop	PGrp	Loop	PGrp	Loop	PGrp	Loop	PGrp
0	1, 3, 11	0	2, 4, 12	0	7, 15	0	8, 16
1	1, 3, 11	1	2, 4, 12	1	7, 15	1	8, 16
2	1, 3, 11	2	2, 4, 12	2	7, 15	2	8, 16
3	1, 3, 11	3	2, 4, 12	3	7, 15	3	8, 16

BED-3 BED19 Slot 2ML		BED-4 BED1B Slot 2NL		BED-7 BED1D Slot 2XL		BED-8 BED1F Slot 2WL	
Loop	PGrp	Loop	PGrp	Loop	PGrp	Loop	PGrp
0	5, 13	0	6, 14	0	9, 17	0	10, 18
1	5, 13	1	6, 14	1	9, 17	1	10, 18
2	5, 13	2	6, 14	2	9, 17	2	10, 18
3	5, 13	3	6, 14	3	9, 17	3	10, 18

APPENDIX 7. FED: 8-Port FC Maps with Processors

USP-V (Front View) OPEN Fibre 8-port Cluster-1

<u>FED-6</u> FED06 Slot 1HU		<u>FED-5</u> FED04 Slot 1GU		<u>FED-2</u> FED02 Slot 1FU		<u>FED-1</u> FED00 Slot 1EU	
Port	MP	Port	MP	Port	MP	Port	MP
1N	mp1	1J	mp1	1E	mp1	1A	mp1
3N	mp2	3J	mp2	3E	mp2	3A	mp2
5N	mp3	5J	mp3	5E	mp3	5A	mp3
7N	mp4	7J	mp4	7E	mp4	7A	mp4

<u>FED-8</u> FED07 Slot 1HL		<u>FED-7</u> FED05 Slot 1GL		<u>FED-4</u> FED03 Slot 1FL		<u>FED-3</u> FED01 Slot 1EL	
Port	MP	Port	MP	Port	MP	Port	MP
1Q	mp1	1L	mp1	1G	mp1	1C	mp1
3Q	mp2	3L	mp2	3G	mp2	3C	mp2
5Q	mp3	5L	mp3	5G	mp3	5C	mp3
7Q	mp4	7L	mp4	7G	mp4	7C	mp4

USP-V (Rear View) OPEN Fibre 8-port Cluster-2

<u>FED-1</u> FED08 Slot 2QU		<u>FED-2</u> FED0A Slot 2RU		<u>FED-5</u> FED0C Slot 2TU		<u>FED-6</u> FED0E Slot 2UU	
Port	MP	Port	MP	Port	MP	Port	MP
2A	mp1	2E	mp1	2J	mp1	2N	mp1
4A	mp2	4E	mp2	4J	mp2	4N	mp2
6A	mp3	6E	mp3	6J	mp3	6N	mp3
8A	mp4	8E	mp4	8J	mp4	8N	mp4

<u>FED-3</u> FED09 Slot 2QL		<u>FED-4</u> FED0B Slot 2RL		<u>FED-7</u> FED0D Slot 2TL		<u>FED-8</u> FED0F Slot 2UL	
Port	MP	Port	MP	Port	MP	Port	MP
2C	mp1	2G	mp1	2L	mp1	2Q	mp1
4C	mp2	4G	mp2	4L	mp2	4Q	mp2
6C	mp3	6G	mp3	6L	mp3	6Q	mp3
8C	mp4	8G	mp4	8L	mp4	8Q	mp4

APPENDIX 8. FED: 16-Port FC Maps with Processors

USP-V (Front View) OPEN Fibre 16-port Cluster-1

FED-6 FED06 Slot 1HU		FED-5 FED04 Slot 1GU		FED-2 FED02 Slot 1FU		FED-1 FED00 Slot 1EU	
Port	MP	Port	MP	Port	MP	Port	MP
1N	mp1	1J	mp1	1E	mp1	1A	mp1
3N	mp2	3J	mp2	3E	mp2	3A	mp2
5N	mp1	5J	mp1	5E	mp1	5A	mp1
7N	mp2	7J	mp2	7E	mp2	7A	mp2
1P	mp3	1K	mp3	1F	mp3	1B	mp3
3P	mp4	3K	mp4	3F	mp4	3B	mp4
5P	mp3	5K	mp3	5F	mp3	5B	mp3
7P	mp4	7K	mp4	7F	mp4	7B	mp4

FED-8 FED07 Slot 1HL		FED-7 FED05 Slot 1GL		FED-4 FED03 Slot 1FL		FED-3 FED01 Slot 1EL	
Port	MP	Port	MP	Port	MP	Port	MP
1Q	mp1	1L	mp1	1G	mp1	1C	mp1
3Q	mp2	3L	mp2	3G	mp2	3C	mp2
5Q	mp1	5L	mp1	5G	mp1	5C	mp1
7Q	mp2	7L	mp2	7G	mp2	7C	mp2
1R	mp3	1M	mp3	1H	mp3	1D	mp3
3R	mp4	3M	mp4	3H	mp4	3D	mp4
5R	mp3	5M	mp3	5H	mp3	5D	mp3
7R	mp4	7M	mp4	7H	mp4	7D	mp4

USP-V (Rear View) OPEN Fibre 16-port Cluster-2

FED-1 FED08 Slot 2QU		FED-2 FED0A Slot 2RU		FED-5 FED0C Slot 2TU		FED-6 FED0E Slot 2UU	
Port	MP	Port	MP	Port	MP	Port	MP
2A	mp1	2E	mp1	2J	mp1	2N	mp1
4A	mp2	4E	mp2	4J	mp2	4N	mp2
6A	mp1	6E	mp1	6J	mp1	6N	mp1
8A	mp2	8E	mp2	8J	mp2	8N	mp2
2B	mp3	2F	mp3	2K	mp3	2P	mp3
4B	mp4	4F	mp4	4K	mp4	4P	mp4
6B	mp3	6F	mp3	6K	mp3	6P	mp3
8B	mp4	8F	mp4	8K	mp4	8P	mp4

FED-3 FED09 Slot 2QL		FED-4 FED0B Slot 2RL		FED-7 FED0D Slot 2TL		FED-8 FED0F Slot 2UL	
Port	MP	Port	MP	Port	MP	Port	MP
2C	mp1	2G	mp1	2L	mp1	2Q	mp1
4C	mp2	4G	mp2	4L	mp2	4Q	mp2
6C	mp1	6G	mp1	6L	mp1	6Q	mp1
8C	mp2	8G	mp2	8L	mp2	8Q	mp2
2D	mp3	2H	mp3	2M	mp3	2R	mp3
4D	mp4	4H	mp4	4M	mp4	4R	mp4
6D	mp3	6H	mp3	6M	mp3	6R	mp3
8D	mp4	8H	mp4	8M	mp4	8R	mp4

APPENDIX 9. FED: 8-Port FICON Maps with Processors

USP-V (Front View) FICON Cluster-1

<u>FED-6</u> FED06 Slot 1HU		<u>FED-5</u> FED04 Slot 1GU		<u>FED-2</u> FED02 Slot 1FU		<u>FED-1</u> FED00 Slot 1EU	
Port	MP	Port	MP	Port	MP	Port	MP
1N	1, 2	1J	1, 2	1E	1, 2	1A	1, 2
3N	1, 2	3J	1, 2	3E	1, 2	3A	1, 2
5N	3, 4	5J	3, 4	5E	3, 4	5A	3, 4
7N	3, 4	7J	3, 4	7E	3, 4	7A	3, 4

<u>FED-8</u> FED07 Slot 1HL		<u>FED-7</u> FED05 Slot 1GL		<u>FED-4</u> FED03 Slot 1FL		<u>FED-3</u> FED01 Slot 1EL	
Port	MP	Port	MP	Port	MP	Port	MP
1Q	1, 2	1L	1, 2	1G	1, 2	1C	1, 2
3Q	1, 2	3L	1, 2	3G	1, 2	3C	1, 2
5Q	3, 4	5L	3, 4	5G	3, 4	5C	3, 4
7Q	3, 4	7L	3, 4	7G	3, 4	7C	3, 4

USP-V (Rear View) FICON Cluster-2

<u>FED-1</u> FED08 Slot 2QU		<u>FED-2</u> FED0A Slot 2RU		<u>FED-5</u> FED0C Slot 2TU		<u>FED-6</u> FED0E Slot 2UU	
Port	MP	Port	MP	Port	MP	Port	MP
2A	1, 2	2E	1, 2	2J	1, 2	2N	1, 2
4A	1, 2	4E	1, 2	4J	1, 2	4N	1, 2
6A	3, 4	6E	3, 4	6J	3, 4	6N	3, 4
8A	3, 4	8E	3, 4	8J	3, 4	8N	3, 4

<u>FED-3</u> FED09 Slot 2QL		<u>FED-4</u> FED0B Slot 2RL		<u>FED-7</u> FED0D Slot 2TL		<u>FED-8</u> FED0F Slot 2UL	
Port	MP	Port	MP	Port	MP	Port	MP
2C	1, 2	2G	1, 2	2L	1, 2	2Q	1, 2
4C	1, 2	4G	1, 2	4L	1, 2	4Q	1, 2
6C	3, 4	6G	3, 4	6L	3, 4	6Q	3, 4
8C	3, 4	8G	3, 4	8L	3, 4	8Q	3, 4

APPENDIX 10. FED: 8-Port ESCON Maps with Processors

USP-V (Front View) ESCON Cluster-1

<u>FED-6</u> FED06 Slot 1HU		<u>FED-5</u> FED04 Slot 1GU		<u>FED-2</u> FED02 Slot 1FU		<u>FED-1</u> FED00 Slot 1EU	
Port	MP	Port	MP	Port	MP	Port	MP
1N	1, 2	1J	1, 2	1E	1, 2	1A	1, 2
3N	1, 2	3J	1, 2	3E	1, 2	3A	1, 2
5N	1, 2	5J	1, 2	5E	1, 2	5A	1, 2
7N	1, 2	7J	1, 2	7E	1, 2	7A	1, 2

<u>FED-8</u> FED07 Slot 1HL		<u>FED-7</u> FED05 Slot 1GL		<u>FED-4</u> FED03 Slot 1FL		<u>FED-3</u> FED01 Slot 1EL	
Port	MP	Port	MP	Port	MP	Port	MP
1P	1, 2	1K	1, 2	1F	1, 2	1B	1, 2
3P	1, 2	3K	1, 2	3F	1, 2	3B	1, 2
5P	1, 2	5K	1, 2	5F	1, 2	5B	1, 2
7P	1, 2	7K	1, 2	7F	1, 2	7B	1, 2

USP-V (Rear View) ESCON Cluster-2

<u>FED-1</u> FED08 Slot 2QU		<u>FED-2</u> FED0A Slot 2RU		<u>FED-5</u> FED0C Slot 2TU		<u>FED-6</u> FED0E Slot 2UU	
Port	MP	Port	MP	Port	MP	Port	MP
2A	1, 2	2E	1, 2	2J	1, 2	2N	1, 2
4A	1, 2	4E	1, 2	4J	1, 2	4N	1, 2
6A	1, 2	6E	1, 2	6J	1, 2	6N	1, 2
8A	1, 2	8E	1, 2	8J	1, 2	8N	1, 2

<u>FED-3</u> FED09 Slot 2QL		<u>FED-4</u> FED0B Slot 2RL		<u>FED-7</u> FED0D Slot 2TL		<u>FED-8</u> FED0F Slot 2UL	
Port	MP	Port	MP	Port	MP	Port	MP
2B	1, 2	2F	1, 2	2K	1, 2	2P	1, 2
4B	1, 2	4F	1, 2	4K	1, 2	4P	1, 2
6B	1, 2	6F	1, 2	6K	1, 2	6P	1, 2
8B	1, 2	8F	1, 2	8K	1, 2	8P	1, 2

APPENDIX 11. Pool sizing example

The size of an HDP Pool, in number of Pages, can be calculated using the following procedure.

1. Number of Pages in a Pool Volume:

$$\text{PVPages} = (\downarrow (\downarrow \text{number of 512 byte blocks in pool-VOL} \div 512 \downarrow) / 168 \downarrow) - 2$$

By dividing the volume's number of 512 byte blocks by 512, then truncating this value to be an integer, you get the number of 256KB units in the volume. When dividing this result by 168, you get the number of 42MB pages in the volume. You then subtract "2" for the Pool volume's system area.

Note: There are 2 Pages per Volume used for system management purposes.

2. Total Number of Usable Pages in Pool:

$$\text{TtlPages} = (\Sigma \text{PVPages}_n) - 98$$

Notes:

- There are 98 Pages (4GB) reserved from each Pool for HDP Shared Memory backup purposes. *The HDP's Shared Memory DMT backup at power down is not done to each Pool but only to the Pool considered to be the "first" Pool. The Backup function is not enabled in G01 support.*

Note: $\downarrow$ means: *truncate the result of the formula between the arrows after the decimal point (round down to nearest integer).*

All of the capacity in the Pool is managed as 42MB Page units. This includes the HDP Shared Memory DMT backup and system management areas. Individual Pool volume sizes should be a multiple of 42MB in order to reduce small amounts of wasted space.

Additional recommendations:

- All space in an Array Group should be allocated to LDEVs (1 or more) which will become Pool volumes. Multiple Array Groups should be defined to a Pool.
- The number of Array Groups should complement the anticipated IO load to the associated DP-VOLs.
- The RAID level should be the same for all Pool Volumes in a Pool. The RAID level should be consistent with the write performance level desired for the DP-VOLs associated with the Pool. The RAID Level does not influence the sizing calculation (above) since the size of a Page was selected by the factory engineers to evenly map to all RAID types.

Pool Size Example (demonstration purposes only, not a good sizing example):

The screenshot shows the Hitachi Storage Navigator interface. The top bar includes the 'Storage Navigator' logo and the 'HITACHI DATA SYSTEM' logo. The menu bar has 'File', 'Go', and 'Help'. The user ID is 'root'. The main window is divided into several sections:

- Pool Information:**
 - Pool ID : 10
 - Status : Normal
 - Total Capacity : 3209.02 GB
 - Used Capacity : 0.00 GB
 - Usage Rate : 0 %
 - Threshold : 70 %
 - Attribute : Dynamic Provisioning
 - CLPR : 01:CLPR1 - HDP test
- Pool-VOL Table:**

LDKC:CU:LDEV	Capacity	Emulation	RAID	PG	Int. VOL Li
00:70:67	50.32 GB	OPEN-V	5(3D+1P)	4-7	
00:70:78	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:79	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7A	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7B	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7C	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7D	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7E	50.32 GB	OPEN-V	5(3D+1P)	4-8	
00:70:7F	50.34 GB	OPEN-V	5(3D+1P)	4-8	

Total Capacity: 3209.02 GB Used LDEVs: 64 Remaining LDEVs: 960
- Free LDEVs:**

LDKC: Select CU: Select

LDKC:CU:LDEV	Capacity	Emulation	RAID	PG	Int. VOL Info	Ext.

Selected LDEVs: 0

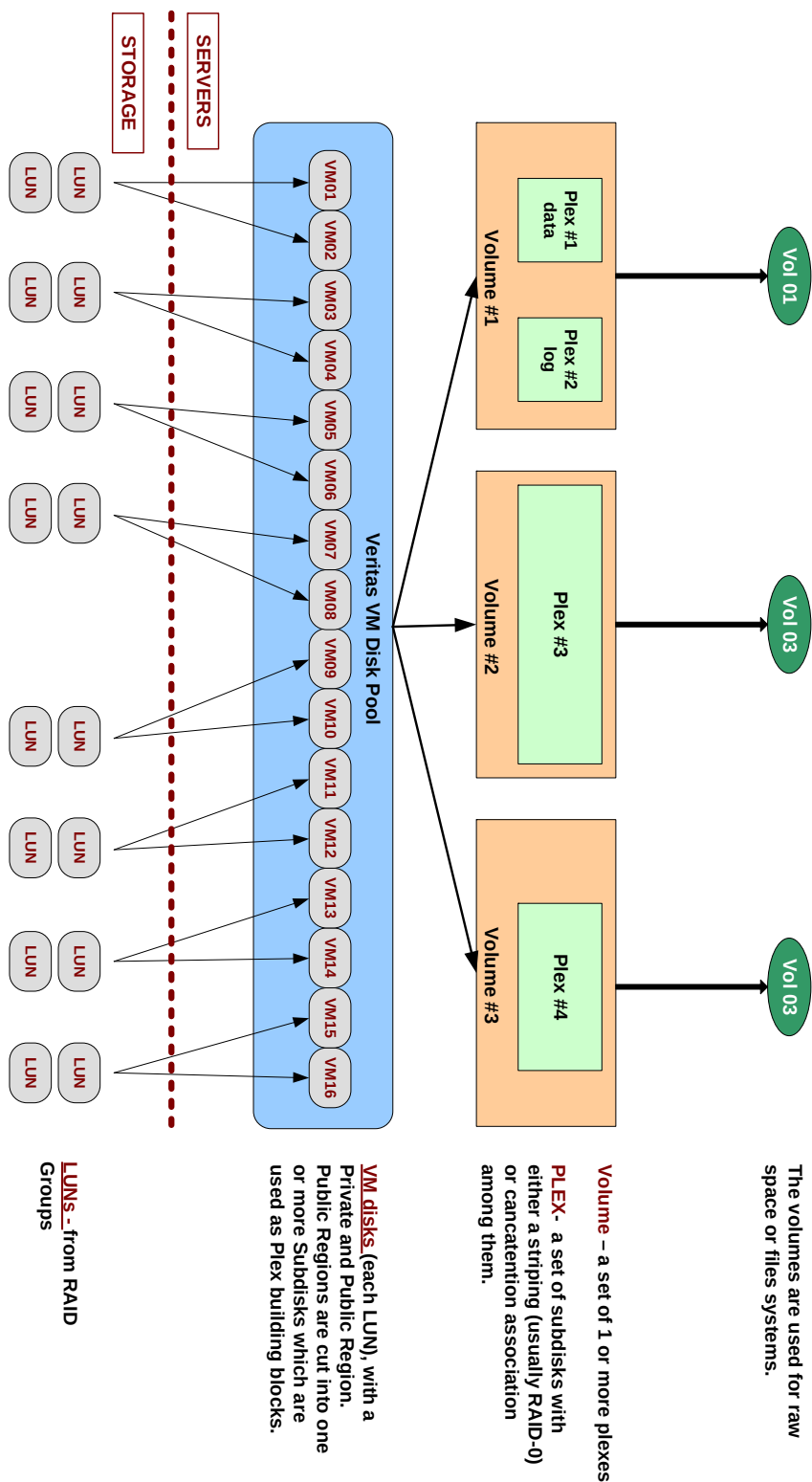
Buttons at the bottom: Initialize, ☐ SIM Complete Request, Apply, Cancel.

The Pool Volume's LDEV sizes add up to 3220.50GB (63 x 50.32GB plus 1 x 50.34GB). However, the total capacity of the Pool is reported as 3209.02GB. The Pool capacity calculation is:

- 63 Pool volumes of 50.32 GB = 51,527.68 MB (Rounded down to a whole number of pages: 51,492 MB = 1226 Pool pages)
- The 63 volumes contain 63 x 1226 pages = 77,238 pages.
- 1 pool volume of 50.34 GB = 51,548.16 MB (Rounded down to a whole number of pages: 51,534 MB = 1227 pages).
- Total number of pages in this pool: 78,465 = (77,238 + 1227).
- HDP DMT backup area: 4,116 MB = 98 pages.
- HDP Management area on each pool volume: 2 pages.
- Data pages: 78,465 – (64 x 2) - 98 = 78,239 pages.
- Final result for the pool size: 3,209,021,484,375 which equals 3209.02 GB. This is shown in the "Total Capacity" under "Pool Information" in the screen shot above.

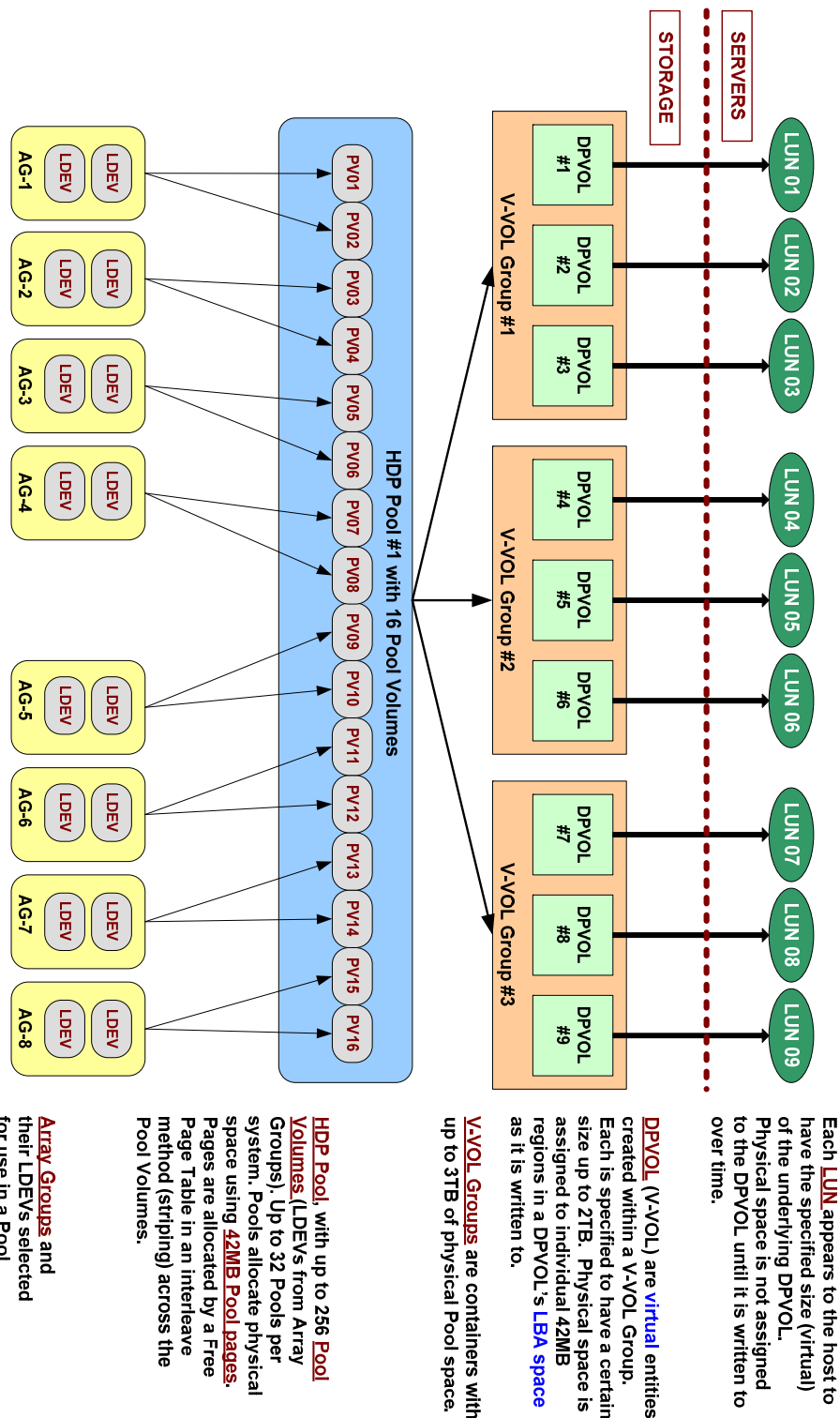
APPENDIX 12. Veritas™ Volume Manager details

This diagram illustrates a typical Veritas host LVM configuration and naming conventions.



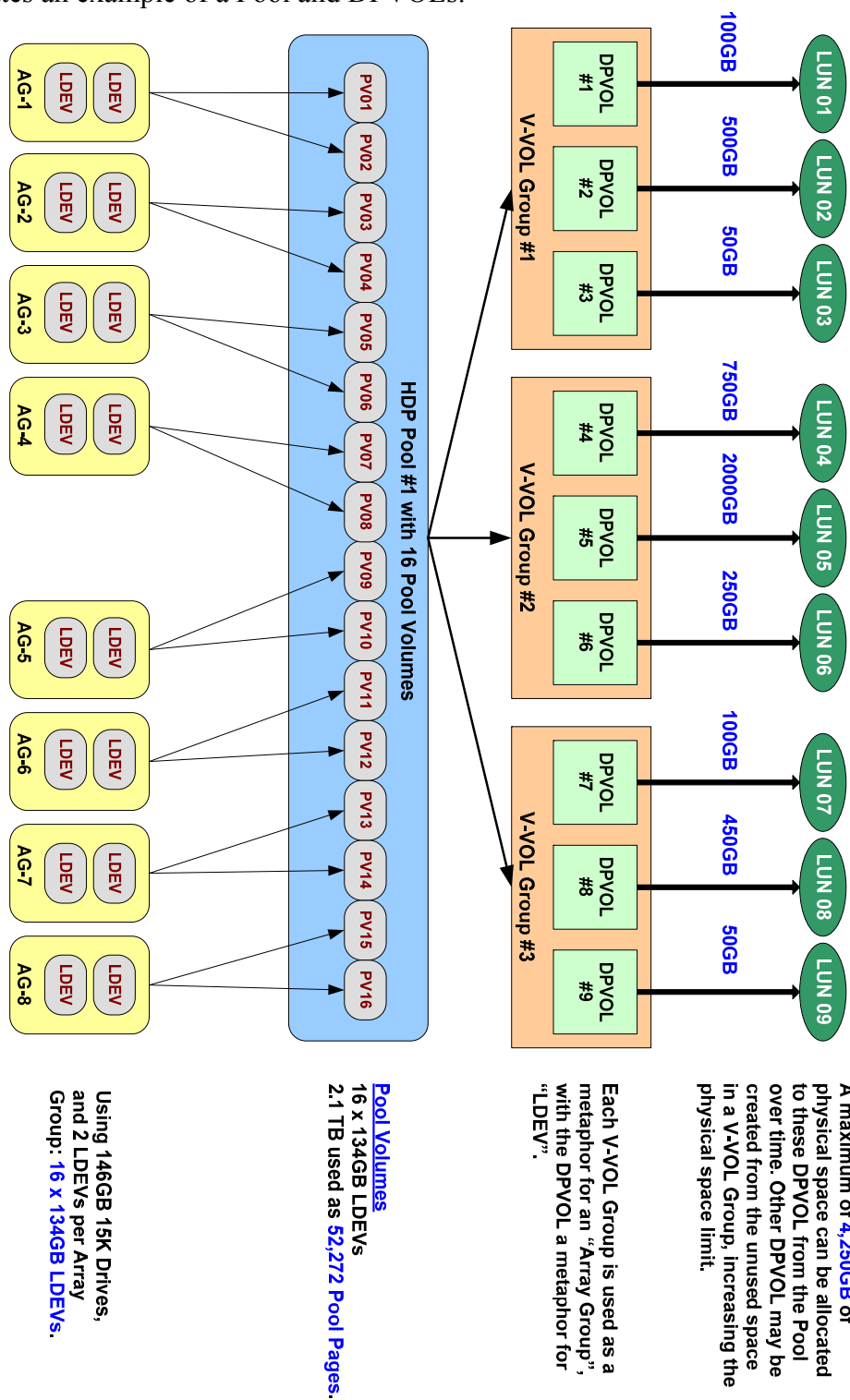
APPENDIX 13. Pool details

This section illustrates the HDP environment with naming conventions. Note the similarities in concepts to the previous Veritas example.



APPENDIX 14. Pool example

This illustrates an example of a Pool and DPVOLs.



APPENDIX 15. Disk IOPS Examples

These are examples of maximum expected physical disk IOPS rates for three types of disks. The “Cylinders” columns show the percentage of surface space in use as disk partitions. As more disk surface is brought into use, the heads must seek farther inwards, which increases the average seek times. Note that the Write table shows reduced maximum IOPS due to the higher track centering precision required for writes.

READS - 100% busy, high RT's		Cylinders			
300 GB Seagate FC 15k rpm		25%	50%	75%	100%
Disk rpm:		15,000	15,000	15,000	15,000
Average Read Seek Time:		1.8	2.5	2.8	3.5
Average rotational delay:		2.0	2.0	2.0	2.0
Average access time:		3.8	4.5	4.8	5.5
Random 8KB IOPS rate:		267	225	208	182
146 GB Hitachi FC 10k rpm		25%	50%	75%	100%
Disk rpm:		10,000	10,000	10,000	10,000
Average Read Seek Time:		2.4	3.3	3.8	4.7
Average rotational delay:		3.0	3.0	3.0	3.0
Average access time:		5.4	6.3	6.8	7.7
Random 8KB IOPS rate:		187	159	148	130
750 GB Seagate SATA 7.2k rpm		25%	50%	75%	100%
Disk rpm:		7,200	7,200	7,200	7,200
Average Read Seek Time:		4.3	6.0	6.8	8.5
Average rotational delay:		4.2	4.2	4.2	4.2
Average access time:		8.4	10.1	11.0	12.7
Random 8KB IOPS rate:		119	99	91	79

Table 59. Calculated maximum disk Read IOPS rates by type and active surface usage

WRITES - 100% busy, high RT's		Cylinders			
300 GB Seagate FC 15k rpm		25%	50%	75%	100%
Disk rpm:		15,000	15,000	15,000	15,000
Average Write Seek Time:		2.0	2.8	3.2	4.0
Average rotational delay:		2.0	2.0	2.0	2.0
Average access time:		4.0	4.8	5.2	6.0
Random 8KB IOPS rate:		250	208	192	167
146 GB Hitachi FC 10k rpm		25%	50%	75%	100%
Disk rpm:		10,000	10,000	10,000	10,000
Average Write Seek Time:		2.6	3.6	4.1	5.1
Average rotational delay:		3.0	3.0	3.0	3.0
Average access time:		5.6	6.6	7.1	8.1
Random 8KB IOPS rate:		180	152	141	123
750 GB Seagate SATA 7.2k rpm		25%	50%	75%	100%
Disk rpm:		7,200	7,200	7,200	7,200
Average Write Seek Time:		5.0	7.0	8.0	10.0
Average rotational delay:		4.2	4.2	4.2	4.2
Average access time:		9.2	11.2	12.2	14.2
Random 8KB IOPS rate:		109	90	82	71

Table 60. Calculated maximum disk Write IOPS rates by type and active surface usage