



HDS USP V Backend Scalability (7d+1p, 2-8 FEDs & BEDs, 256-1024 HDD)

A Performance Brief

Alan Benway
Senior Performance Consultant

James Byun
Performance Consultant

*Performance Measurement Group,
Enterprise Labs, Technical Operations
Santa Clara, CA 95050*

September 2007

Copyright 2007 Hitachi Data Systems Corporation, ALL RIGHTS RESERVED

Table of Contents

1. PURPOSE OF THIS TESTING	2
2. VDBENCH INFORMATION	2
3. WORKLOAD CHARACTERISTICS	3
4. TEST CONFIGURATION OVERVIEW.....	3
TEST INFORMATION	3
HOST CONFIGURATION	3
STORAGE CONFIGURATION	3
VDBENCH CONFIGURATION (RANDOM I/O BACKEND DISK TEST)	4
VDBENCH CONFIGURATION (SEQUENTIAL I/O BACKEND DISK TEST).....	4
TEST METHODOLOGY	5
5. TEST RESULTS SUMMARY	5
RAID5 RANDOM RESULTS SUMMARY	6
RAID5 SEQUENTIAL RESULTS SUMMARY	6
6. TEST RESULT DETAILS.....	7
RANDOM I/O TRANSACTIONS	7
SEQUENTIAL I/O THROUGHPUT	12
7. OBSERVATIONS	15
APPENDIX A. USP V ARCHITECTURAL OVERVIEWS.....	17

1. Purpose of this Testing

The purpose of this testing was to conduct a RAID5 7d+1p back-end disk 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 **mid-sized** 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.

The system was populated with 256GB of Cache, 22GB of Shared Memory, and 1024 146GB 15k RPM disks. The RAID level used for all Array Groups was RAID5 (7d+1p). These small/midsize/large configurations used 32/64/128 Array Groups, 16/32/64 4Gbit disk loops, and 16/32/64 4Gbit FC ports. Both Random and Sequential workloads were used for these tests, with four block sizes for each. The workloads were specified to achieve a 0% cache hit ratio (cache miss, or 100% physical seek per I/O). There was one LUN created per Array Group, and these LUNs were mapped to alternate front-end ports (no shared FED processors). Only standard LDEVs were used for these tests (no Striped LDEVs or HDP Pool Volumes).

Note that while the FED and BED components were scaled (by not using some of them), the Cache and Shared Memory sections remained constant at the full-size USP V levels. It would have made the testing much more complicated to remove hardware and reconfigure the storage system for each level. Hence the “small” and “mid-sized” each benefited somewhat from having more cache and shared memory than would be typical. Since there were two Shared Memory features (four PCBs) installed, then all of the Shared Memory paths on the FEDs and BEDs used for a test came into use. When a configuration has only one Shared Memory feature, then only half of the Shared Memory ports per FED or BED are usable.

These results will help answer questions about the kind of scalability to expect on the USP V system. The performance results are presented in the charts in Section 6 of this report. For convenience, the tests will be labeled as *Small*, *Midsize*, or *Large* even though there are many possible configuration variations for the USP V. This paper includes comparisons to previous USP results that used the same type of tests and configurations.

2. VDbench Information

VDbench was used to generate a variety of I/O workloads against raw volumes (no file systems). Various workload parameters such as I/O rates, file sizes, transfer sizes, thread counts, read/write ratios, read and write cache hit percentages, and random versus sequential (or any combination of the two) were controlled by parameter files. By using raw volumes, the tests bypassed the host file system and its cache, thus more accurately reflecting the I/O performance capabilities of the storage unit.

3. Workload Characteristics

Most applications, from a disk subsystem point of view, can be categorized into one of seven types of application profiles (though there are the extreme cases that are not covered here):

1. Online Transaction Processing – Random Read/Write
2. Decision Support Systems – Sequential or Random Read, depending on access method.
3. General Purpose file systems – Random Read/Write
4. Messaging systems – Random Read/Write
5. Web Server systems – Mixed Random and Sequential with various block sizes.
6. Content Rich Media Streaming and/or Video Streaming – Sequential Read/Write
7. High Performance Computing – Sequential Read/Write

The applications that do not directly fit into one of these seven categories will most likely be a combination of two or more of these categories.

4. Test Configuration Overview

Test Information

Test Period	Report Date	Location	Tester
April 2007	April 2007	Santa Clara CA, USA	James Byun

Host Configuration

Server	Operating System	CPU	Memory	HBA
IBM p595 (Scruggs)	AIX 5.3	24 x 1.65 GHz Power5	96GB RAM	16-64 x 4Gb QLogic HBA

Storage Configuration

Storage	Microcode	Number of Front-end Ports	Cache Sizes (Cache + SM)
USP V	T-code	2-8 FEDs, 16-64 FC ports (4 Gbit)	256GB+22GB

Type of Drive	Size of Drive	Number of Drives	RPM
FC-AL	146GB	256-1024	15,000 RPM

RAID Level	RAID Configuration	Number of Array Groups	Number of LUNs	Size of each LUN
RAID5	7D + 1P	32, 64, 128	32, 64, 128	33GB (OpenV)

VDbench Configuration (Random I/O Backend Disk Test)

Seek %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	2KB Block	4KB Block	8KB Block	16KB Block
100 %	0 %	0 %	0 %	8	✓	✓	✓	✓
	25 %			8	✓	✓	✓	✓
	50 %			8	✓	✓	✓	✓
	75 %			8	✓	✓	✓	✓
	100 %			8	✓	✓	✓	✓

VDbench Configuration (Sequential I/O Backend Disk Test)

Seek %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	64K Block	128K Block	256K Block	512K Block
0 %	0 %	0 %	0 %	1	✓	✓	✓	✓
	25 %			1	✓	✓	✓	✓
	50 %			1	✓	✓	✓	✓
	75 %			1	✓	✓	✓	✓
	100 %			1	✓	✓	✓	✓

The tests were set up to use the alternate ports (unshared processors) from each FED blade, such as 1A, 3A, 1B, 3B, 2A, 4A, 2B, 4B.

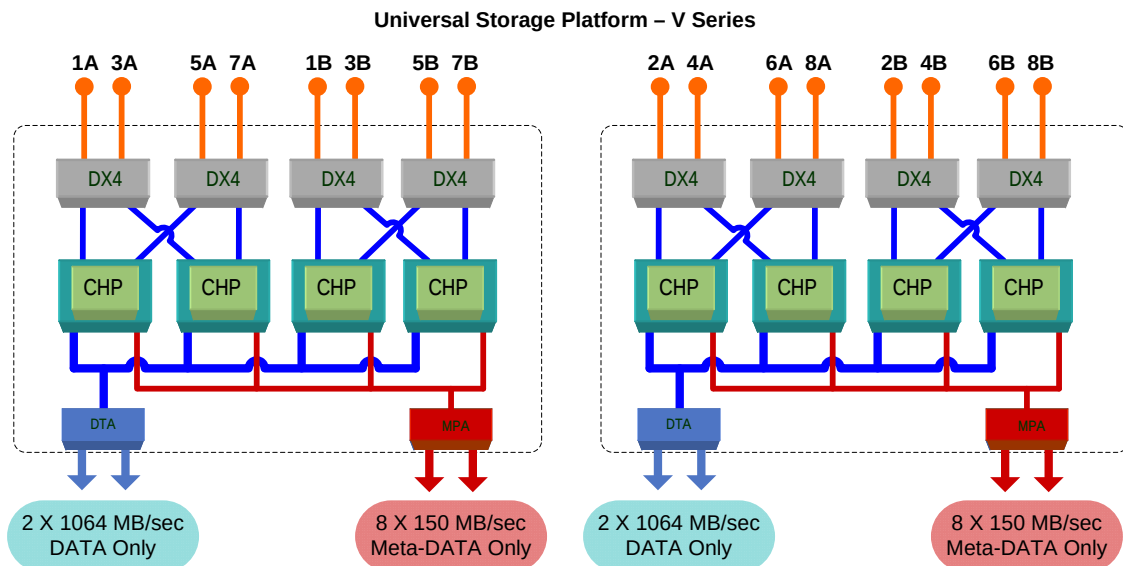


Figure 1. Overview of FED cluster 1 and its port names.

There was a single test system that had 256GB of cache and 22 GB of Shared Memory. The disks used were 146GB 15K RPM, configured using RAID5 7d+1p. The front-end and back-end paths are all 4Gbit. 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, 32 Array Groups, 32 LUNs)
- **USP V midsize**
 - o 4 FEDs – 32 unshared FC host ports
 - o 4 BEDs – 512 HDDs (32 disk loops, 64 Array Groups, 64 LUNs)
- **USP V large**
 - o 8 FEDs – 64 unshared FC host ports
 - o 8 BEDs – 1024 HDDs (64 disk loops, 128 Array Groups, 128 LUNs)

Test Methodology

The test methodology was to first configure all 128 Array Groups using RAID5 7d+1p, and then 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 64 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.

5. Test Results Summary

Due to the large number of results to consider, the following results charts only present the maximum values achieved for a particular test. This greatly reduces the number of data points to consider. The tests were designed to have a 0% cache hit ratio, but that isn’t always possible. There will always be some residual of caching even if it is only about 2%.

All of the Random results are shown as Transactions per Second (TPS). It is important to keep in mind that TPS is what the application will see, whereas disk IOPS is what the storage system will see internally. One usually has no idea of the relationship between the TPS rate and the corresponding physical IOPS rate of the disks.

There is a difference in the physical IOPS due to cache hits or misses. Another factor is the different IOPS cost for Writes by RAID level, where RAID10 has a cost of 2 IOPS, RAID5 is 4 IOPS, and RAID6 is 6 IOPS. The random alignment of a data block on the RAID stripe may cause twice the number of IOPS if adjacent disks are involved when the data block crosses RAID chunk boundaries. Other internal processes are also involved at driving up the physical I/O rates, such as TrueCopy, ShadowImage, or Universal Replicator. Each of these will add a significant number of physical IOPS for a single application write operation.

RAID5 RANDOM RESULTS SUMMARY

The highest levels of Random performance achieved are shown below.

USP-V Random cache miss	2 FED 2 BED		4 FED 4 BED		8 FED 8 BED	
	TPS	RT (ms)	TPS	RT (ms)	TPS	RT (ms)
100% Read	44,080	12	85,999	12.3	161,533	13.0
75% Read	28,343	19	53,748	19.2	104,207	19.8
50% Read	17,091	30	33,686	30.7	65,479	31.5
100% Write	10,243	51	19,972	51.5	38,593	53.2

Table 1. Maximum Random workload results.

RAID5 SEQUENTIAL RESULTS SUMMARY

The highest levels of Sequential performance achieved are shown below.

USP-V Sequential cache miss	2 FED 2 BED	4 FED 4 BED	8 FED 8 BED
	MB/sec	MB/sec	MB/sec
100% Read	3,976	5,523	10,328
75% Read	1,930	3,342	6,212
50% Read	1,649	2,727	4,959
100% Write	2,339	3,220	5,306

Table 2. Maximum Sequential workload results.

6. Test Result Details

The following two sections present the results in this manner:

- **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						
standard LDEVs		1-128 LUNs						
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 3. Summary of all peak random results.

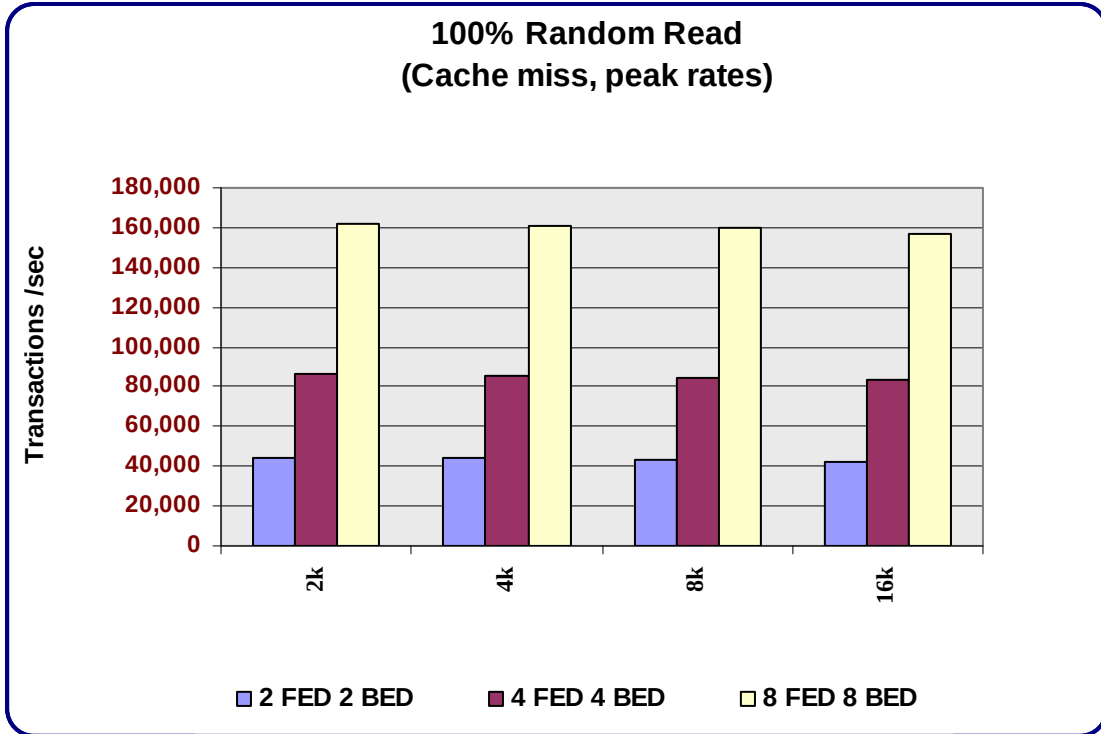


Figure 2. Peak 100% Random Read rates.

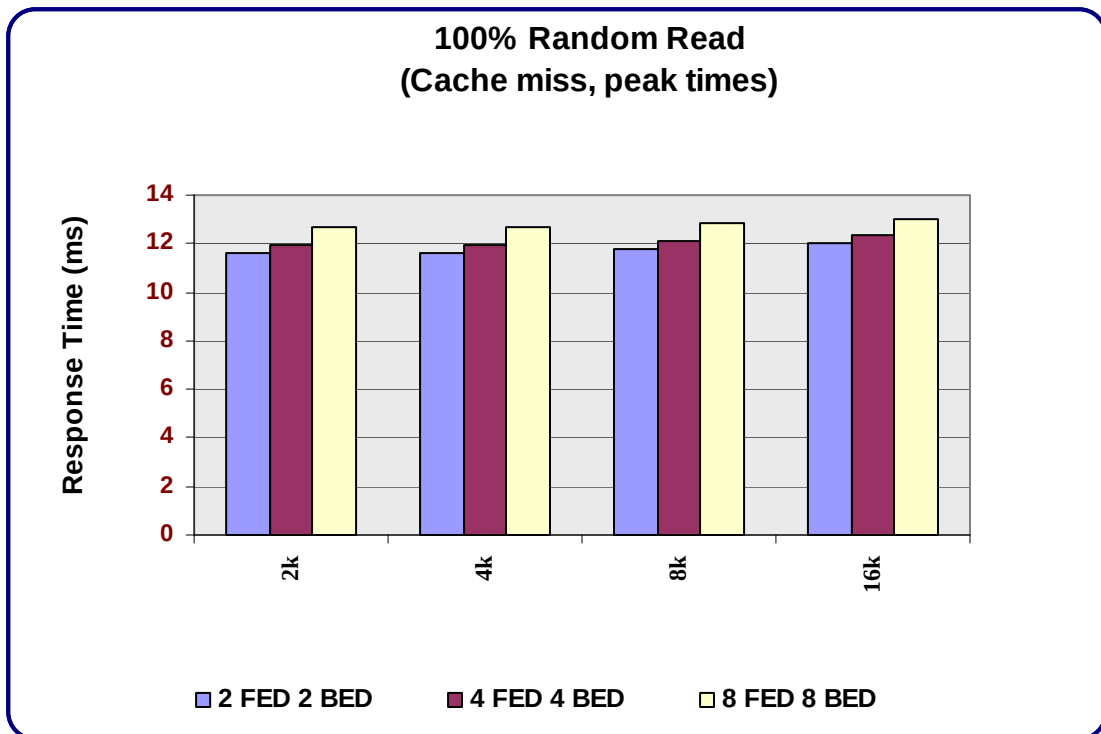


Figure 3. Peak 100% Random Read response times.

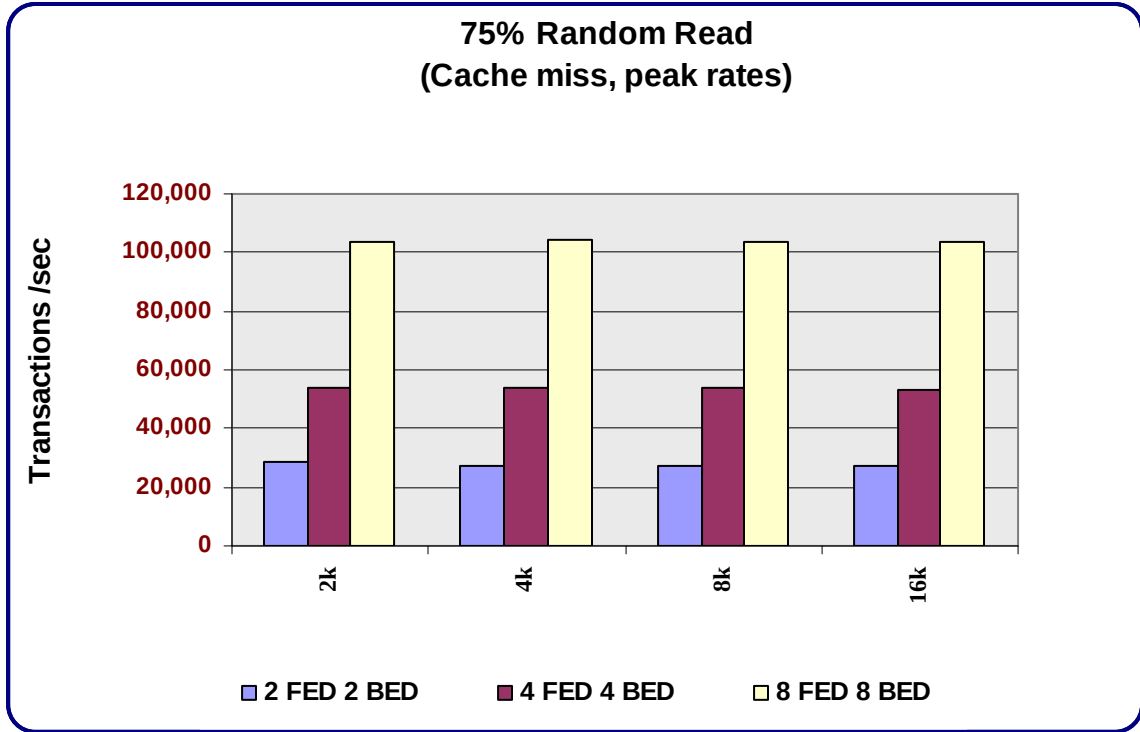


Figure 4. Peak 75% Random Read rates.

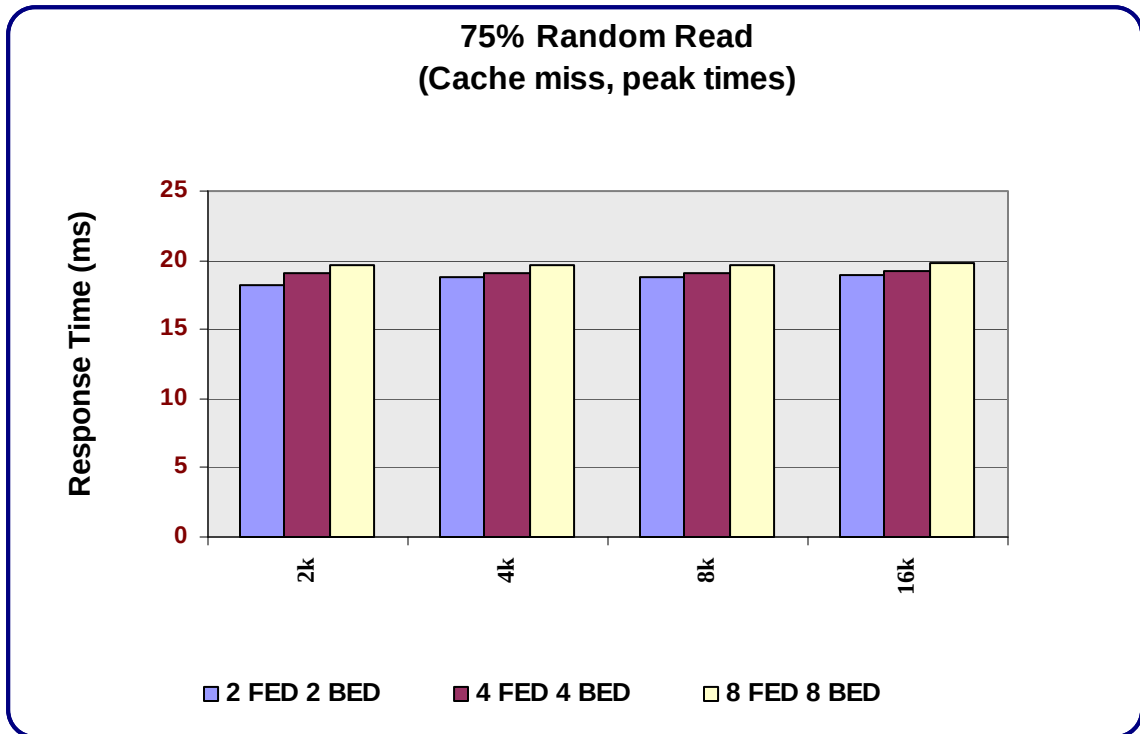


Figure 5. Peak 75% Random Read response times.

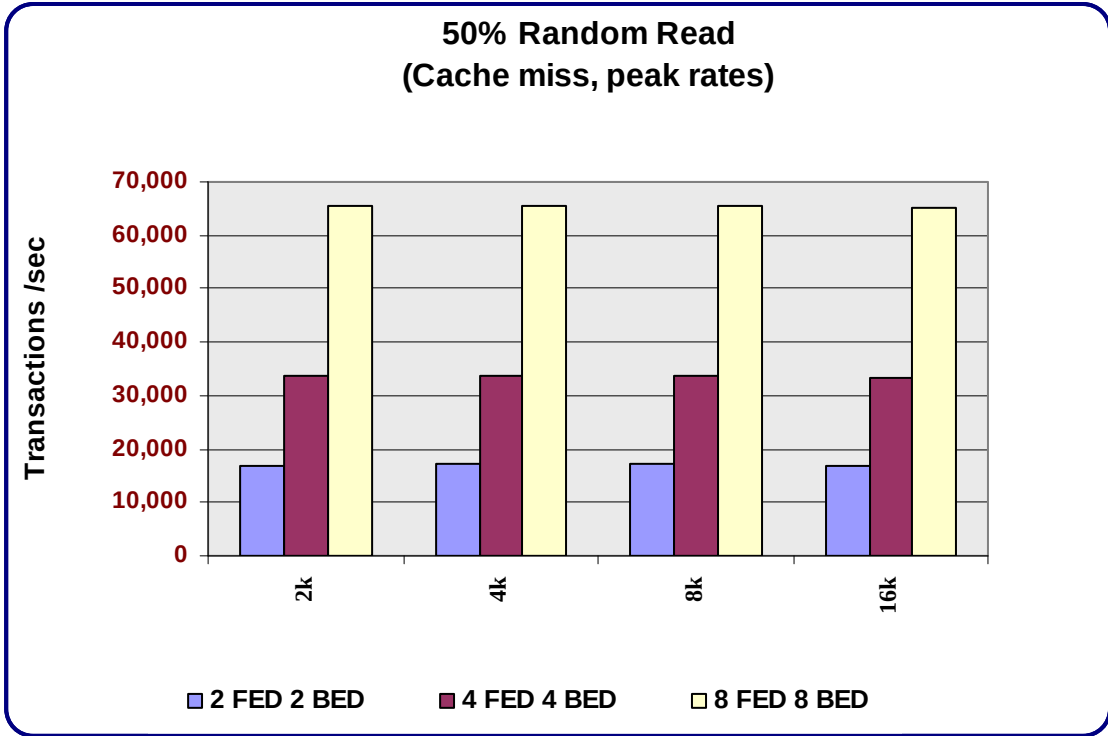


Figure 6. Peak 50% Random Read rates.

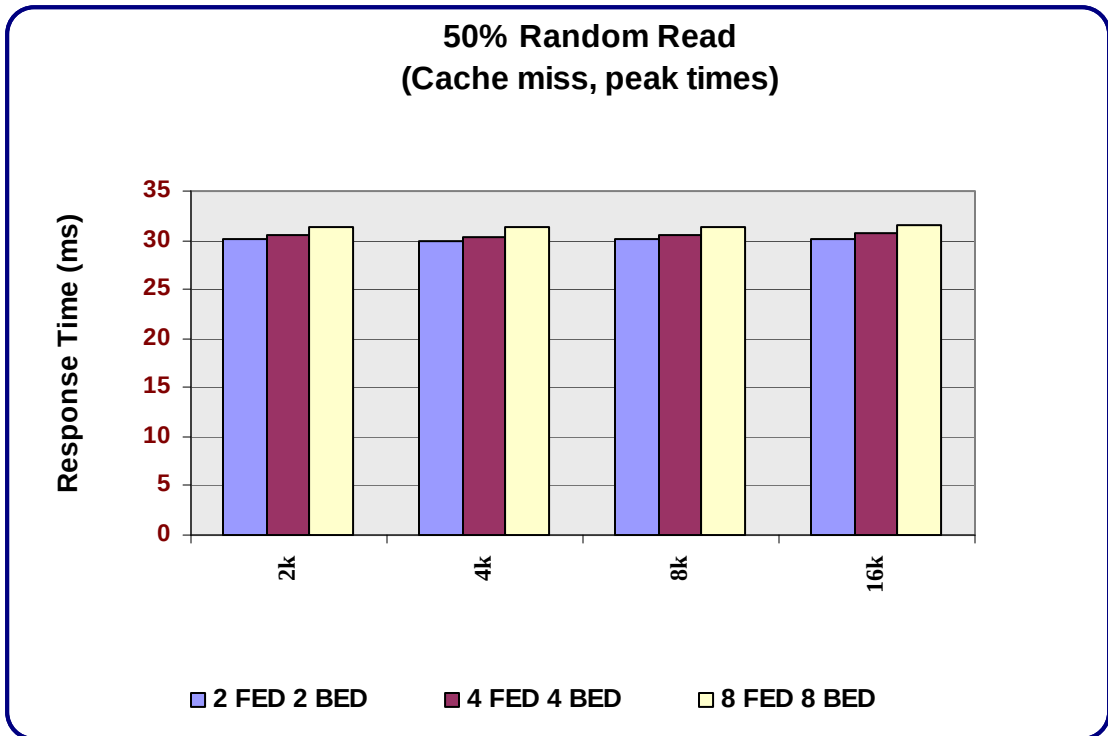


Figure 7. Peak 50% Random Read response times.

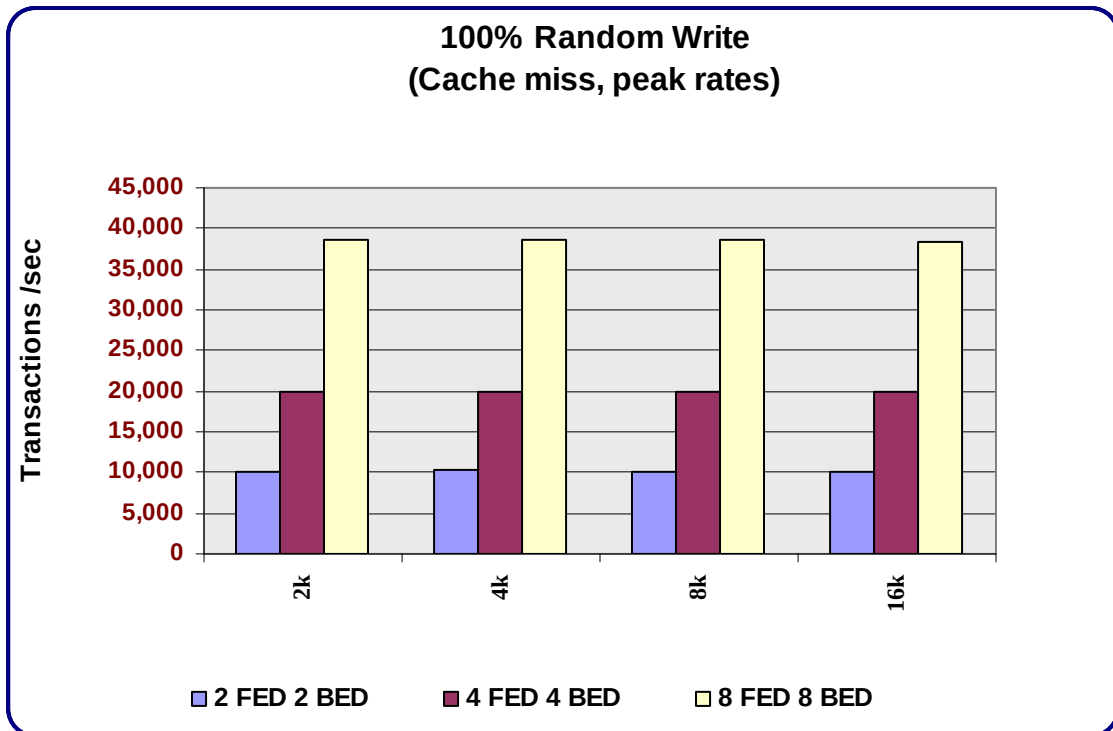


Figure 8. Peak 100% Random Write rates.

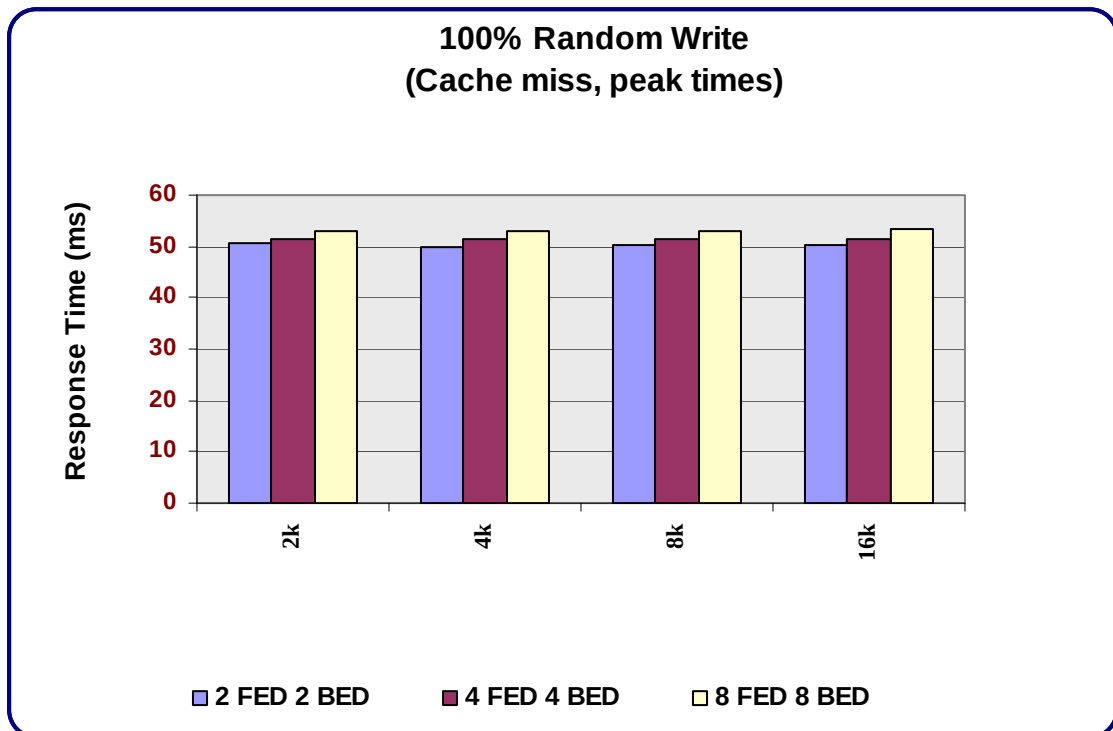


Figure 9. Peak 100% Random Write response times.

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 7+1 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	3,718	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 4. Summary of all peak Sequential results.

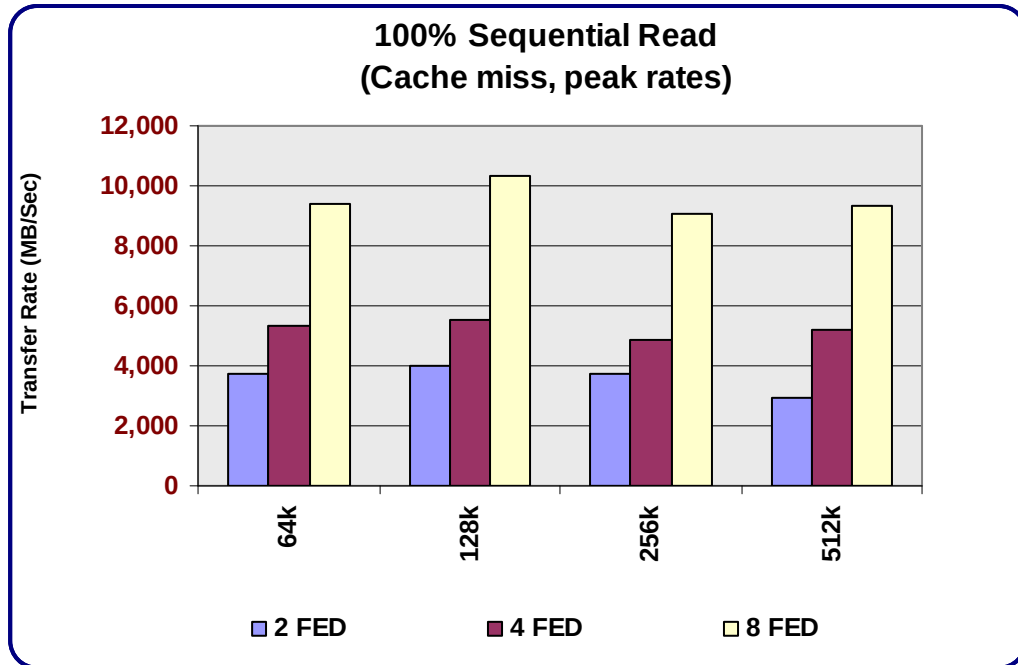


Figure 10. Peak 100% Sequential Read rates.

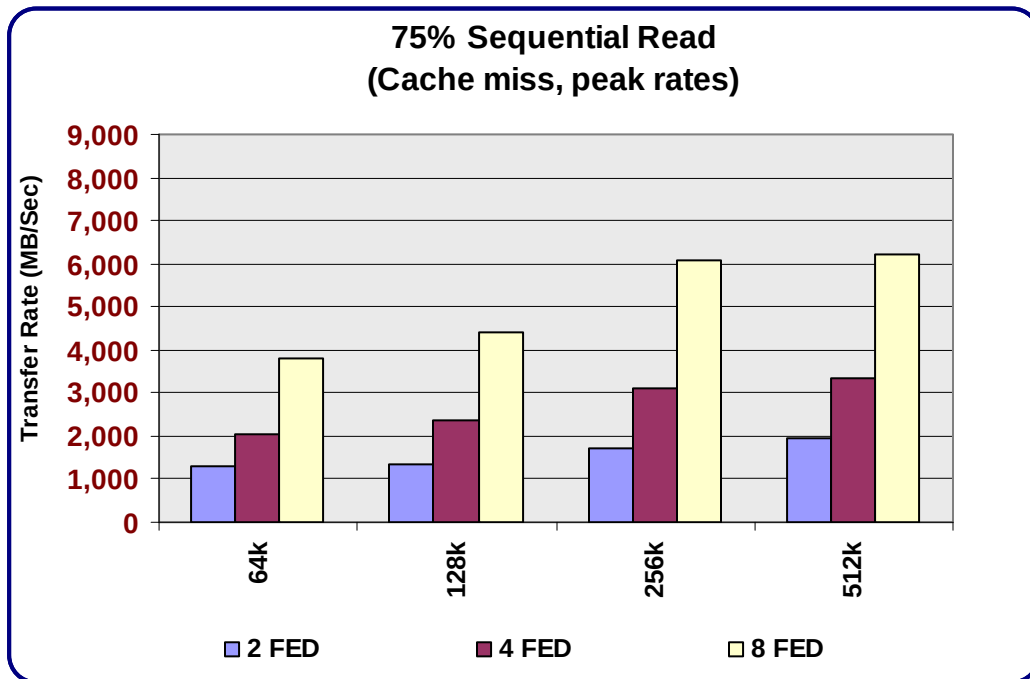


Figure 11. Peak 75% Sequential Read rates.

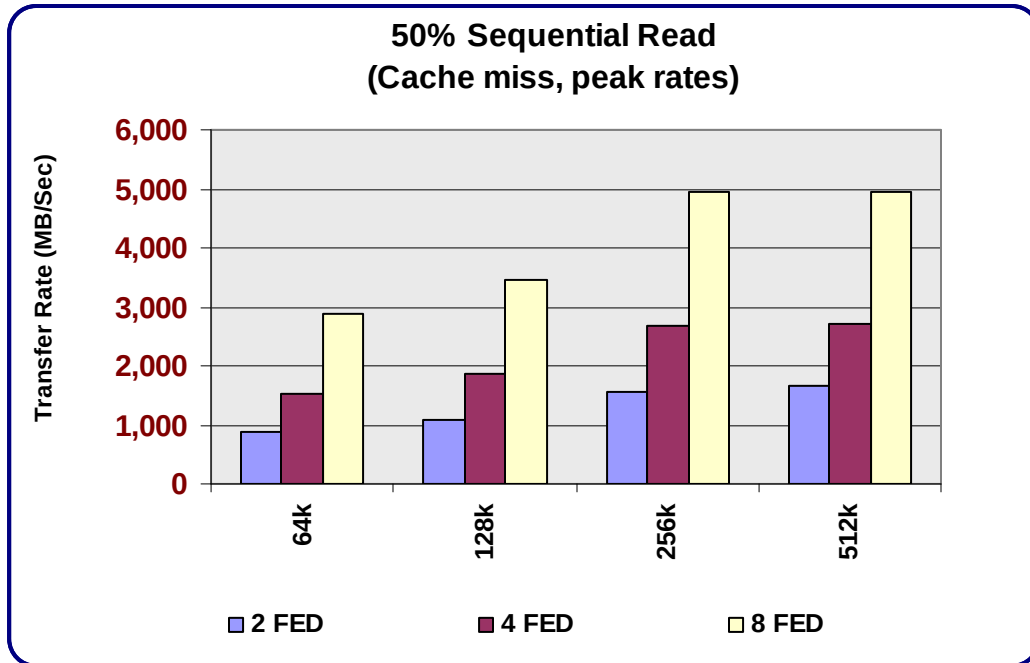


Figure 12. Peak 50% Sequential Read rates.

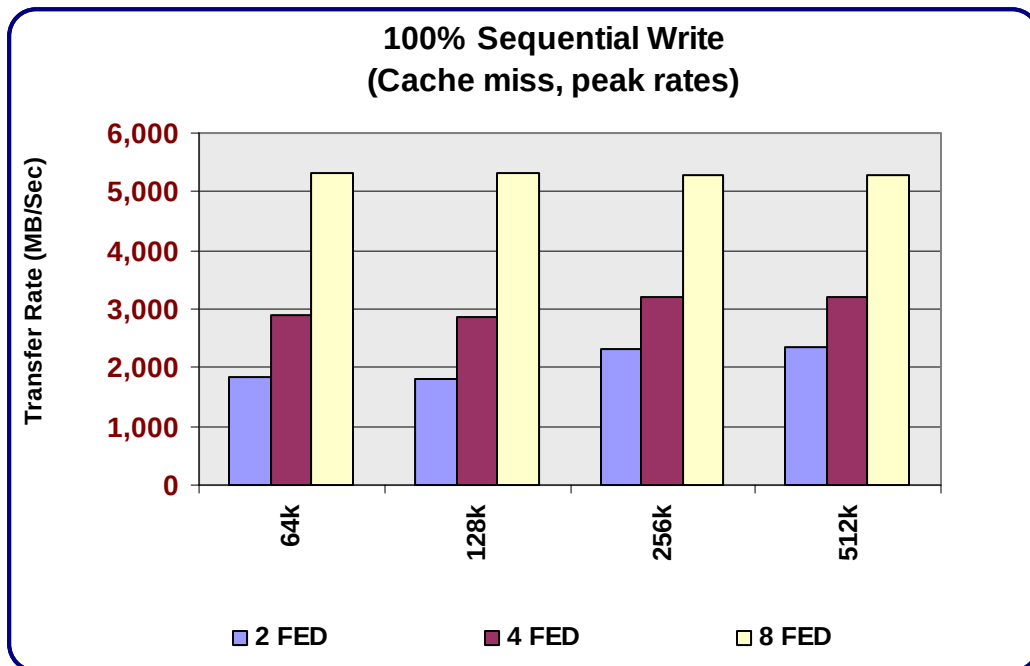


Figure 13. Peak 100% Sequential Write rates.

7. Observations

These results show that the USP V models, when using RAID5 7d+1p Random workloads, scale quite nicely. However, there were some issues with the Test Microcode with scaling in the case of Sequential workloads, particularly when there is a 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 5. 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 6. 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 7. Percent delta, Random workloads, USP-V, 7d+1p to 3d+1p.

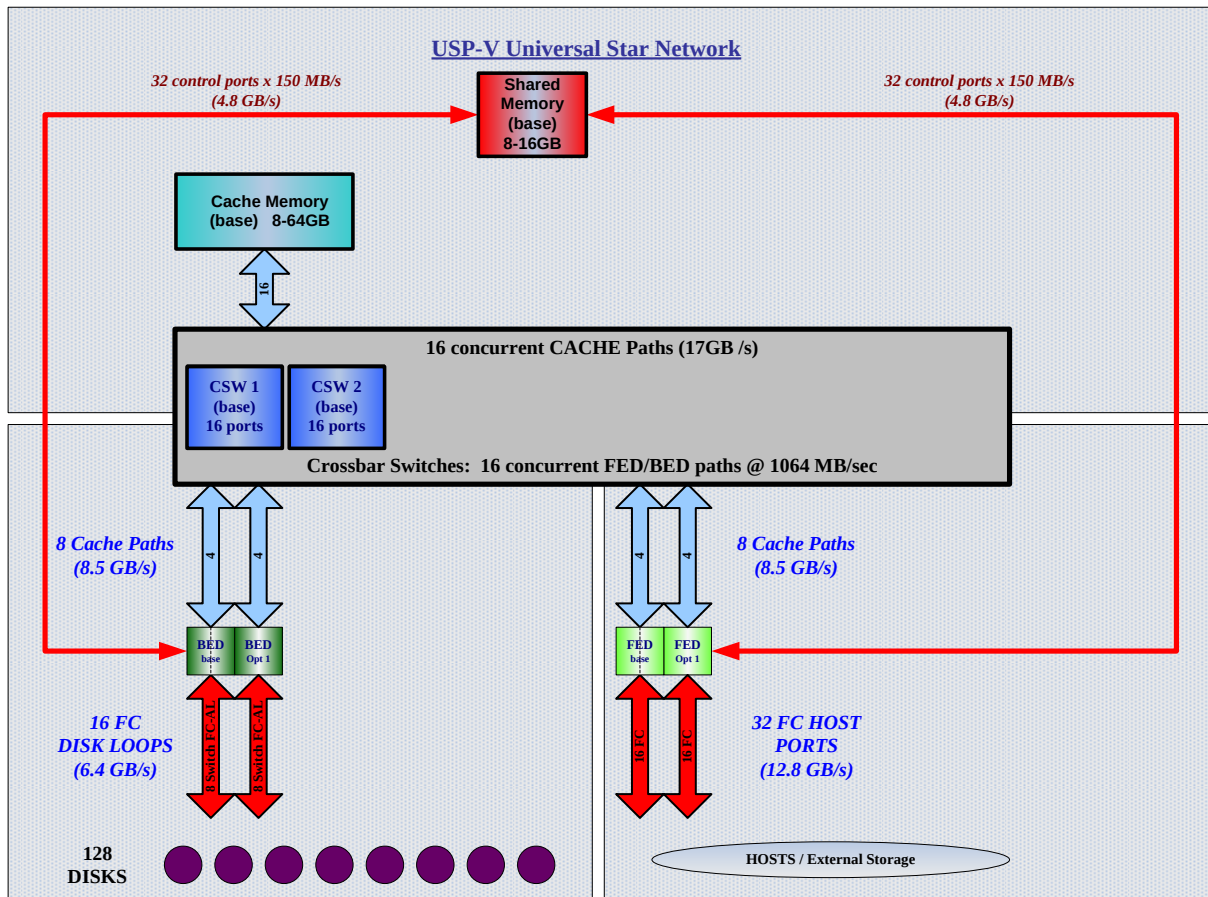
Sequential I/O cache miss standard LDEVs Read %	RAID5 comparison 7+1 : 3+1 Block	256-1024 HDD 146GB 15k 1-256 LUNs Thread	256 HDD	512 HDD	1024 HDD
			MB/sec	MB/sec	MB/sec
			ratio	ratio	ratio
100	64k	1	1%	-5%	-8%
100	128k	1	0%	2%	-10%
100	256k	1	2%	-14%	-15%
100	512k	1	-5%	6%	-1%
75	64k	1	4%	0%	1%
75	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 8. Percent delta, Sequential workloads, USP-V, 7d+1p to 3d+1p.

APPENDIX A. USP V Architectural Overviews

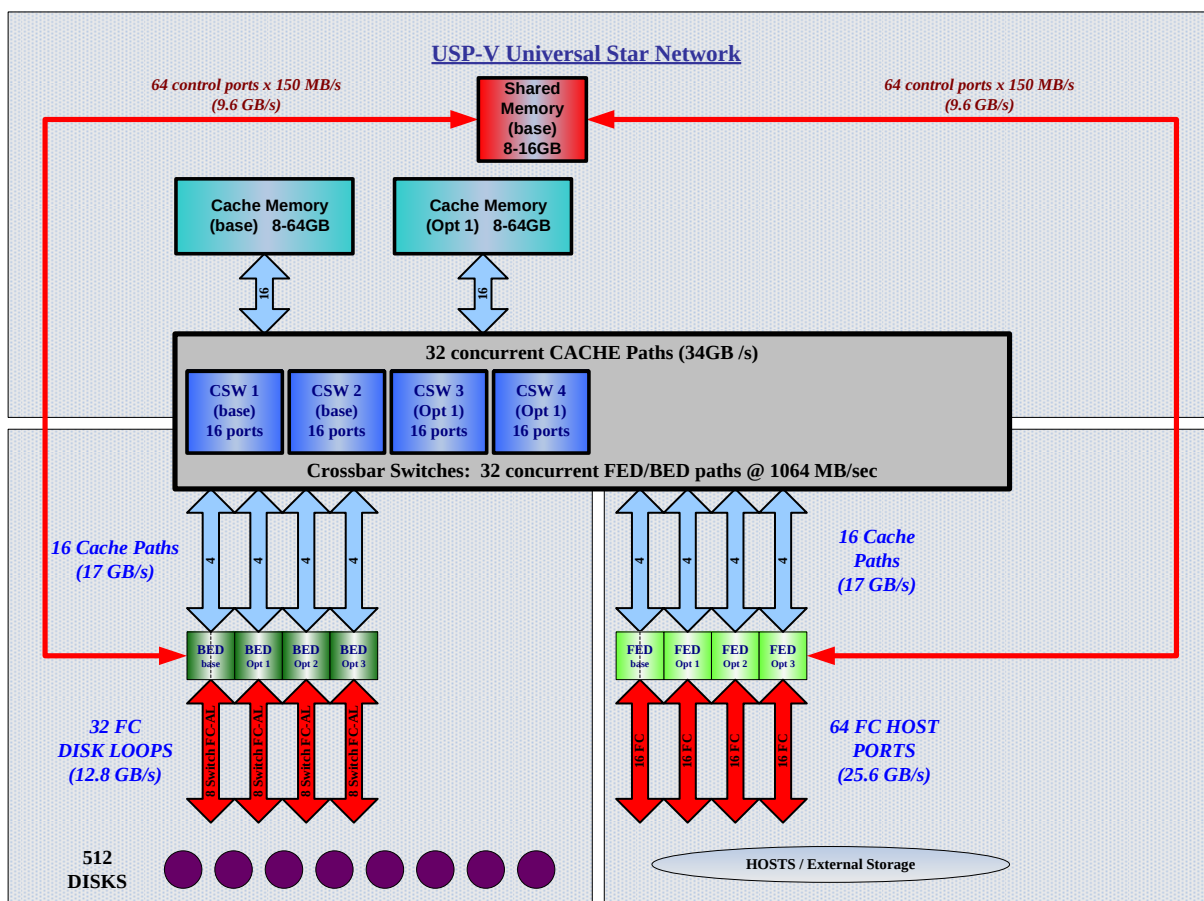
Small

This picture illustrates how a “small” (2 FED, 2 BED, 256 HDD) USP V could be configured. The test system had all of the Cache switches, Cache options, and Shared Memory options installed, but only 2 FED and 2 BED features were used.



Midsize

This picture illustrates how a “midsize” (4 FED, 4 BED, 512 HDD) USP V could be configured. The test system had all of the Cache switches, Cache options, and Shared Memory options installed, but only 4 FED and 4 BED features were used.



Large

This picture illustrates how a “large” (8 FED, 8 BED, 1024 HDD) USP V would be configured.

