



TagmaStore USP 4Gbit Backend Scalability by Model (2d+2d, 1-4 FEDs & BEDs, 1024 HDD)

A Performance Brief

Alan Benway
Senior Performance Consultant

Ron Lee
Performance Consultant

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

26 September 2006

Copyright 2006 Hitachi Data Systems Corporation, ALL RIGHTS RESERVED

Hitachi Data System Restricted Confidential

Table of Contents

1. PURPOSE OF THIS TESTING	2
2. VDBENCH INFORMATION	2
3. WORKLOAD CHARACTERISTICS	2
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
5. TEST RESULTS SUMMARY	5
RANDOM RESULTS SUMMARY	5
SEQUENTIAL RESULTS SUMMARY	5
6. TEST RESULTS	6
RANDOM I/O TRANSACTIONS	6
SEQUENTIAL I/O THROUGHPUT	11
APPENDIX A. USP1100 ARCHITECTURAL OVERVIEW	13
APPENDIX B. USP600 TEST CONFIGURATION OVERVIEW.....	14
APPENDIX C. USP100 TEST CONFIGURATION OVERVIEW.....	15

1. Purpose of this Testing

The purpose of this testing was to conduct a disk scalability analysis using the HDS TagmaStore USP1100 enterprise storage system. The first set of tests evaluated how a USP100 (1 FED, 1BED, 256HDD) could perform. The second set of tests measured a USP600 (2 FEDs, 2BEDs, 512HDD). The final set of tests measured a USP1100 (4 FEDs, 4BEDs, 1024HDD). The system was populated with 64GB of cache and 1024 146GB 15k RPM disks. The front-end paths were 4Gbit Fibre Channel. The RAID level used was RAID-10 (2+2). Both Random and Sequential workloads were used for these tests, with varying block sizes for each. These tests used up to 256 RAID Groups (2+2) and 64 FC ports. The workloads were specified to achieve a 0% cache hit ratio (100% seek per I/O). There were 64, 128, then 256 33GB LUNs created (one per Array Group), and these were mapped to alternate front-end ports (no shared FED processors).

Note that while the FEDs and BEDs components were scaled (by not using them), the cache and shared memory sections remained constant at the USP1100 levels. It would have made the testing much more complicated to remove hardware and reconfigure the storage system for each level. Hence the “100” and “600” each benefited from having many more paths to cache and shared memory than would be normal. Since there were two shared memory pairs (four blades), then all of the 32 shared memory paths on the BEDs came into use. When a configuration has only one shared memory pair, then only 16 of these paths are usable per BED.

These results will help answer questions about the kind of scalability to expect across the USP family. A fairly powerful IBM p595 server was used for these tests. The performance results are presented in the charts in Section 6 of this report. For convenience, the tests will be labeled as *100*, *600*, or *1100* even though there are many possible variations on these models.

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
September 2006	September 2006	Santa Clara CA, USA	Ron Lee

Host Configuration

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

Storage Configuration

Storage	Microcode	Number of Controllers	Cache Size
TagmaStore USP1100	50-07-64-00/00	4 FED, 64 FC ports (4 Gbit)	64GB

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

RAID Level	RAID Configuration	Number of Array Groups	Number of LUNs	Size of each LUN
RAID 10	2D + 2D	256	256	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	✓	✓	✓	✓
	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	✓	✓	✓	✓
	50 %			1	✓	✓	✓	✓
	75 %			1				
	100 %			1	✓	✓	✓	✓

The tests were set up to use alternate ports from each FED blade, such 1A, 3A, 1B, 3B, 1C, 3C, 1D, 3D.

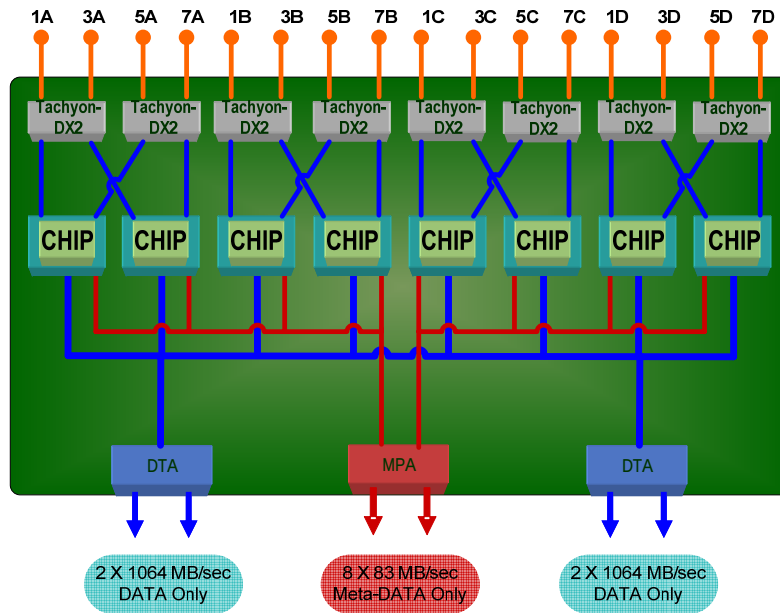


Figure 1. Overview of an FED cluster board with its port names.

5. Test Results Summary

Due to the large number of results to consider, all of the following charts present the maximum values achieved for a particular test. This greatly reduced 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%.

RANDOM RESULTS SUMMARY

It is important to keep in mind that a TPS result without the context of the target application is meaningless. If one is using a database with a 16KB block size, then the 8KB block size result must be reduced by a factor of two to compare it to the 16KB block size result. Then the apparent lower (random) TPS result for 16KB in this test will far outperform the "higher" 8KB TPS result.

The highest level of Random performance achieved is shown below in Table 1:

	1-FED/BED USP100		2 FED/BED USP600		4 FED/BED USP1100	
	TPS	RT	TPS	RT	TPS	RT
100% Read	26,878	19.2	54,068	18.9	98,104	20.9
75% Read	21,057	24.4	39,904	25.8	74,801	27.5
50% Read	15,429	33.3	30,368	34.0	57,643	35.7
100% Write	10,783	24.0	20,922	24.5	39,852	52.0

Table 1. Maximum Random workload results.

SEQUENTIAL RESULTS SUMMARY

The highest level of Sequential performance achieved is shown below in Table 2.

	1-FED/BED USP100	2 FED/BED USP600	4 FED/BED USP1100
	MB/sec	MB/sec	MB/sec
100% Read	2,949	6,048	9,412
50% Read	1,422	2,671	4,716
100% Write	1,300	2,360	4,047

Table 2. Maximum Sequential workload results.

6. Test Results

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% Read, 75% Read, 50% Read, and 100% Write)
- **Sequential I/O, Throughput**, at 64, 128, 256, and 512KB block sizes with 1 Thread (100% Read, 50% Read, and 100% Write)

Note that the X-axis is the block size under test. Also, be aware that *Transactions per Second* is not the same thing as *HDD IOPS*. A transaction is a completed user-level request, whereas there is one physical disk IOP per Read, and 3 (RAID-10) to 5 (RAID-5) physical disk IOPS for each transaction Write request.

Random I/O Transactions

The following table and charts show the maximum TPS results and Response Times by block size for the three system types.

Random I/O 0% cache hit ratio		1024 HDD 146GB 15k	USP100		USP600		USP1100	
Read %	Block	Thread	TPS	ms	TPS	ms	TPS	ms
100	2k	peak average	26,828	19.2	54,036	18.9	98,043	20.9
100	4k		26,828	19.2	54,068	18.9	98,104	20.9
100	8k		26,829	19.2	54,050	18.9	98,091	20.9
100	16k		26,878	19.2	54,061	18.9	97,999	20.9
			26,878	19.2	54,068	18.9	98,104	20.9
			26,841	19.2	54,054	18.9	98,059	20.9
75	2k	peak average	21,032	24.4	39,904	25.7	74,801	27.4
75	4k		21,057	24.4	39,765	25.8	74,628	27.5
75	8k		21,012	24.4	39,891	25.7	74,636	27.5
75	16k		21,057	24.4	39,815	25.7	74,558	27.5
			21,057	24.4	39,904	25.8	74,801	27.5
			21,039	24.4	39,844	25.7	74,656	27.5
50	2k	peak average	15,390	33.3	30,306	33.8	57,643	35.6
50	4k		15,429	33.2	30,368	33.8	57,513	35.7
50	8k		15,400	16.7	30,315	34.0	57,507	35.6
50	16k		15,413	16.7	30,266	33.9	57,408	35.6
			15,429	33.3	30,368	34.0	57,643	35.7
			15,408	25.0	30,314	33.9	57,518	35.6
0	2k	peak average	10,711	23.9	20,899	24.5	39,696	25.8
0	4k		10,684	24.0	20,902	24.5	39,852	51.6
0	8k		10,688	24.0	20,897	24.5	39,596	25.8
0	16k		10,783	23.8	20,922	24.5	39,468	52.0
			10,783	24.0	20,922	24.5	39,852	52.0
			10,716	23.9	20,905	24.5	39,653	38.8

Table 3. Summary of all peak random results.

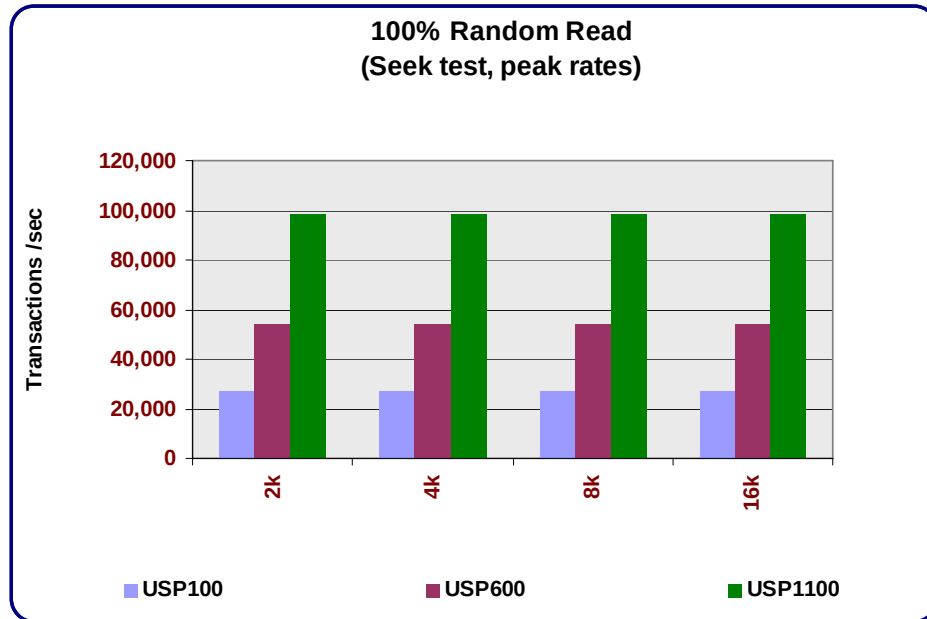


Figure 2. Peak 100% Random Read rates.

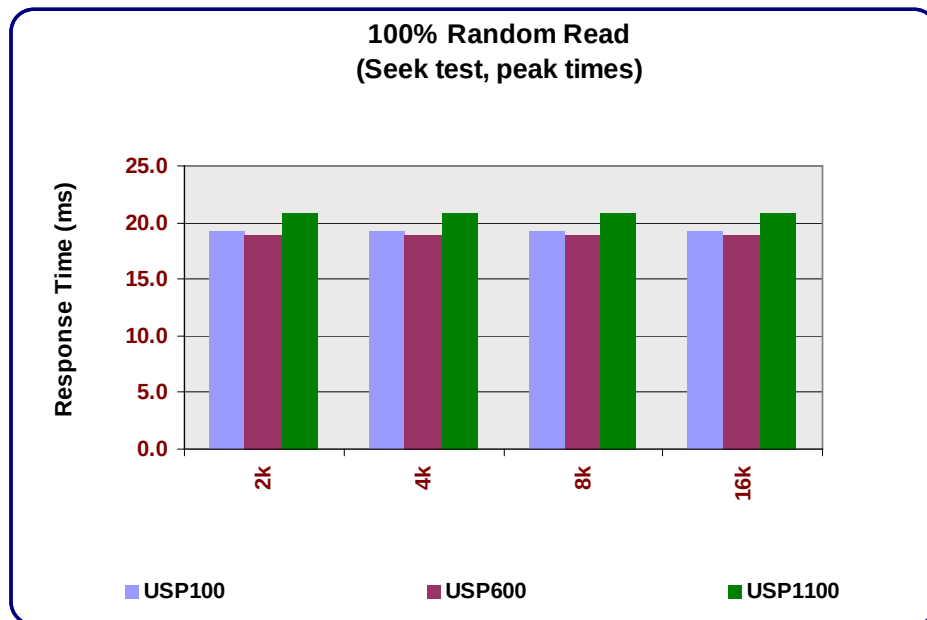


Figure 3. Peak 100% Random Read response times.

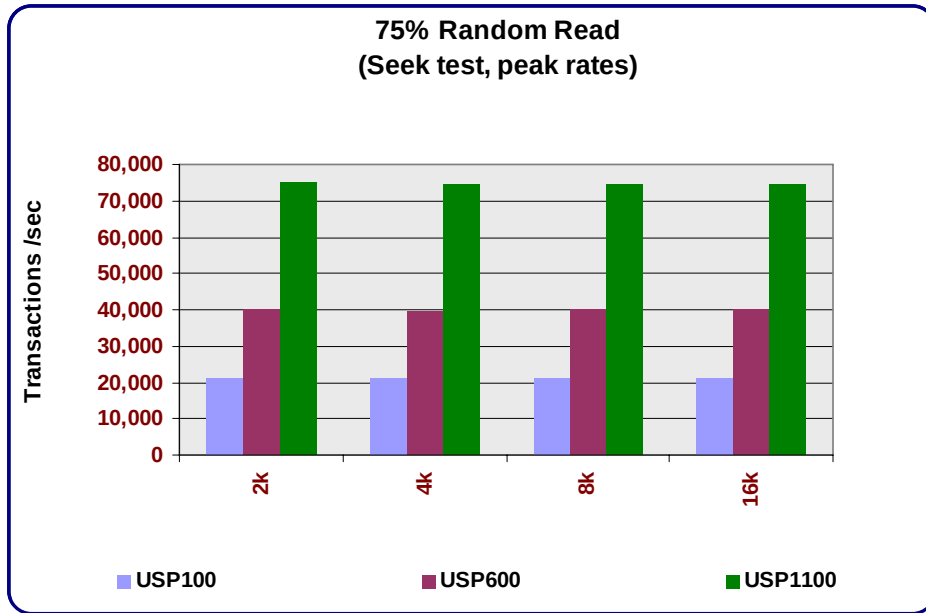


Figure 4. Peak 75% Random Read rates.

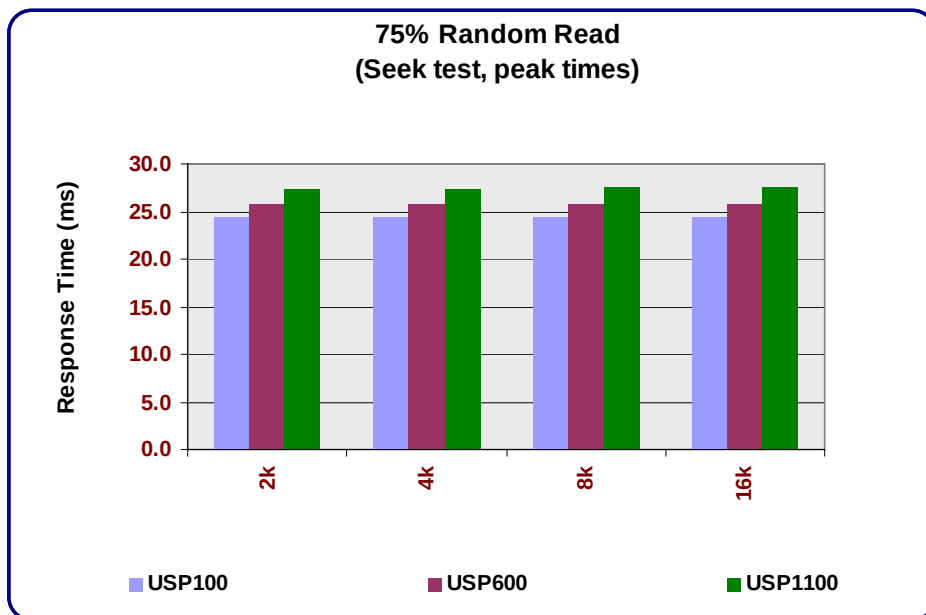


Figure 5. Peak 75% Random Read response times.

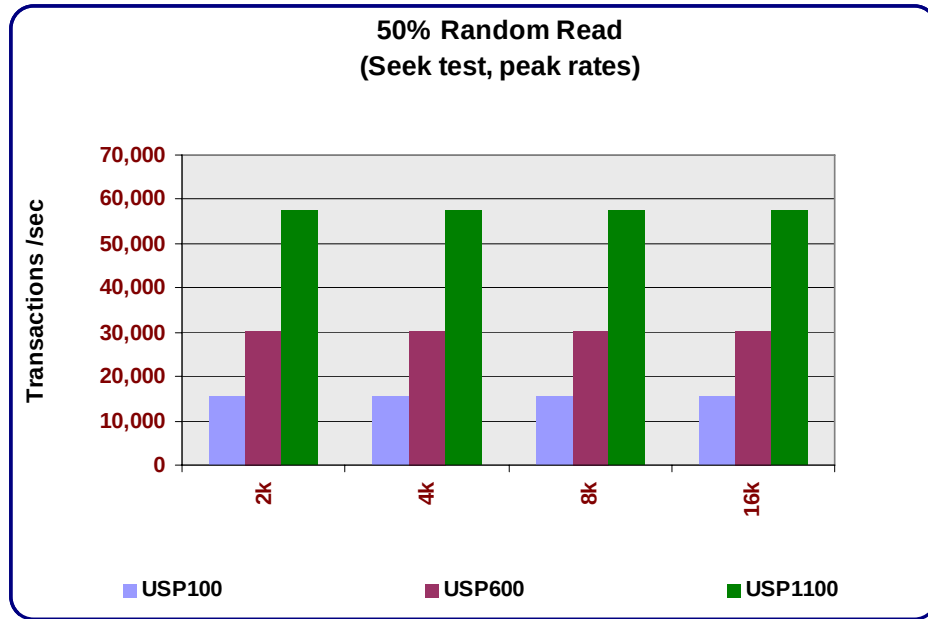


Figure 6. Peak 50% Random Read rates.

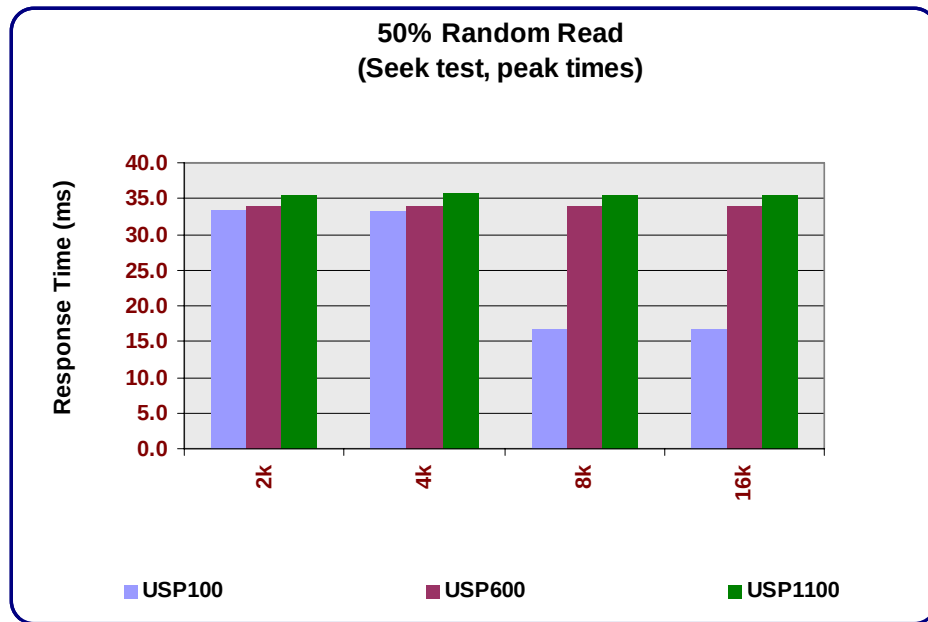


Figure 7. Peak 50% Random Read response times.

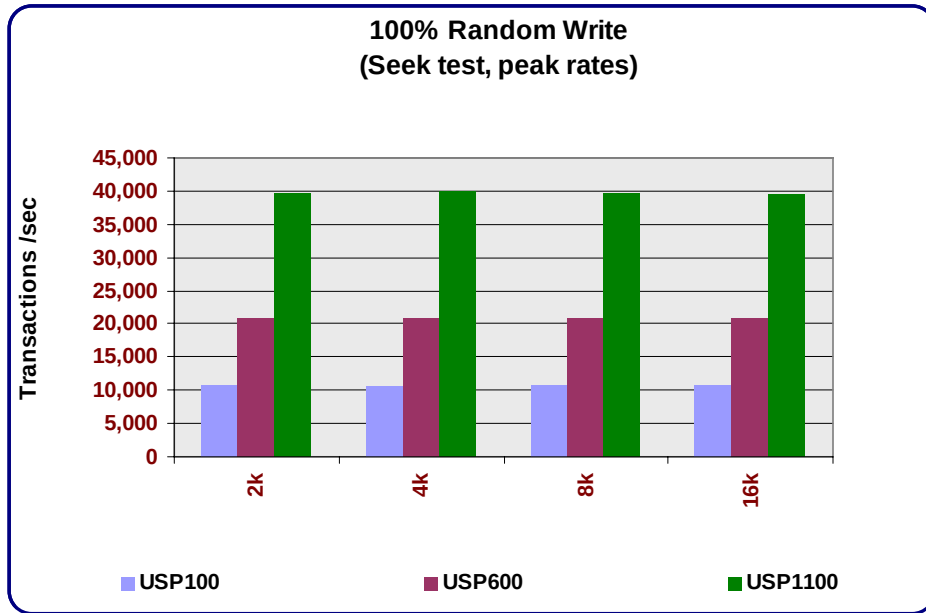


Figure 8. Peak 100% Random Write rates.

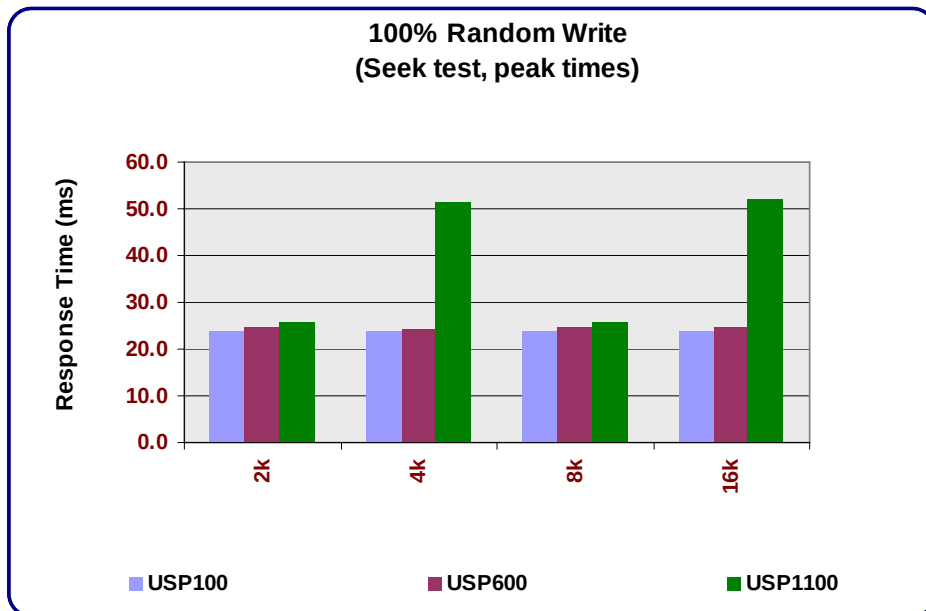


Figure 9. Peak 100% Random Write response times.

Sequential I/O Throughput

The following table and charts show the maximum Throughput results by block size for the three system types.

Sequential I/O			1024 HDD	USP100	USP600	USP1100
0% cache hit	2+2	146GB 15k				
Read %	Block	Thread		MB/sec	MB/sec	MB/sec
100	64k	1		2,629	6,048	9,386
100	128k	1		2,819	5,984	9,374
100	256k	1		2,899	5,659	9,404
100	512k	1		2,949	5,817	9,412
		peak		2,949	6,048	9,412
		average		2,824	5,877	9,394
50	64k	1		1,168	2,137	3,848
50	128k	1		1,367	2,574	4,628
50	256k	1		1,422	2,671	4,716
50	512k	1		1,421	2,661	4,716
		peak		1,422	2,671	4,716
		average		1,345	2,511	4,477
0	64k	1		1,298	2,360	4,047
0	128k	1		1,289	2,350	4,037
0	256k	1		1,300	2,342	4,045
0	512k	1		1,292	2,353	4,045
		peak		1,300	2,360	4,047
		average		1,295	2,351	4,044

Table 4. Summary of all peak Sequential results.

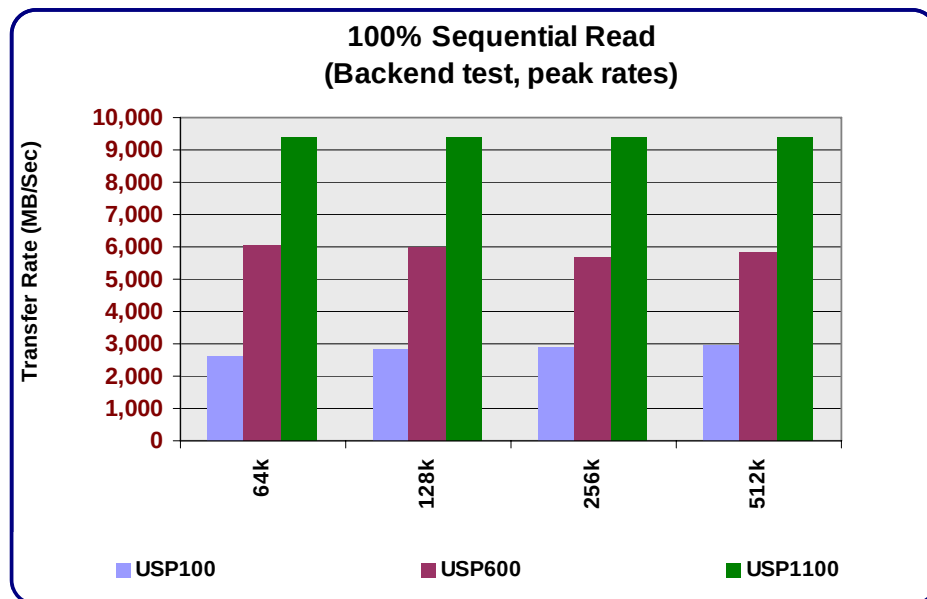


Figure 10. Peak 100% Sequential Read rates.

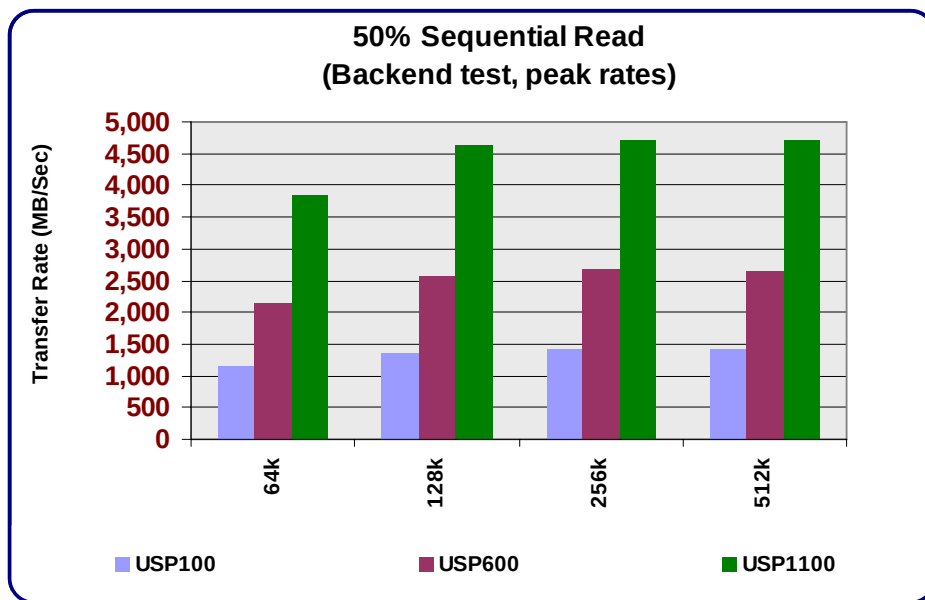


Figure 11. Peak 50% Sequential Read rates.

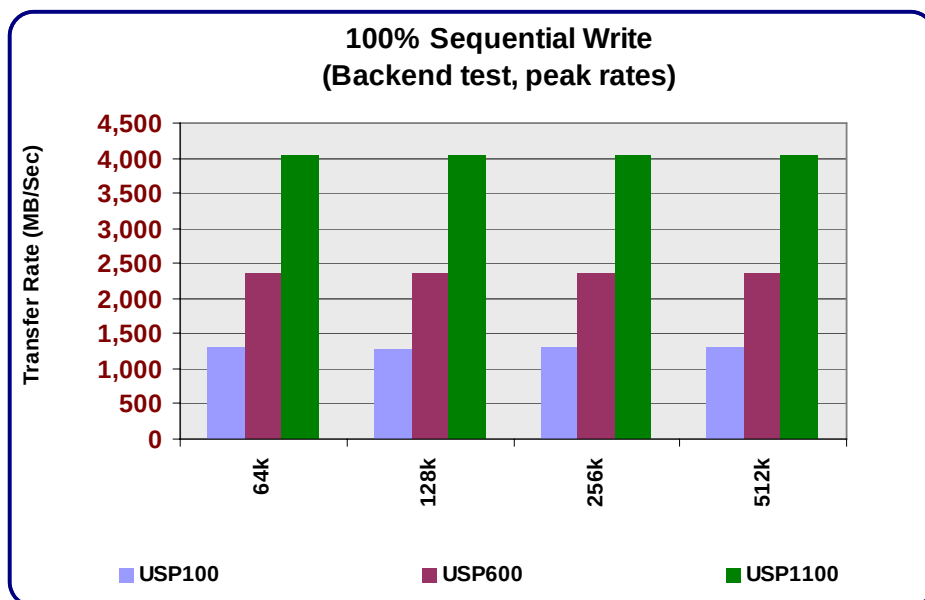
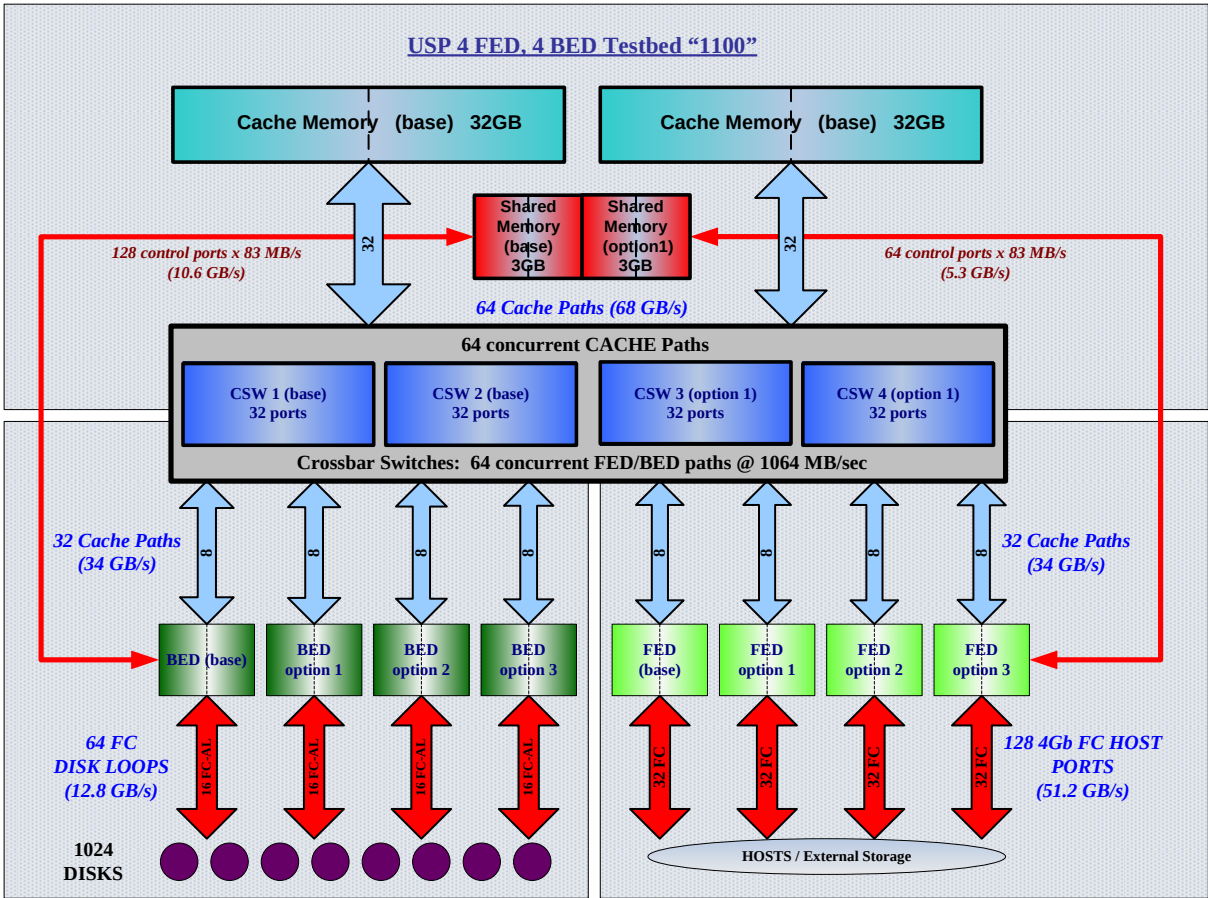


Figure 12. Peak 100% Sequential Write rates.

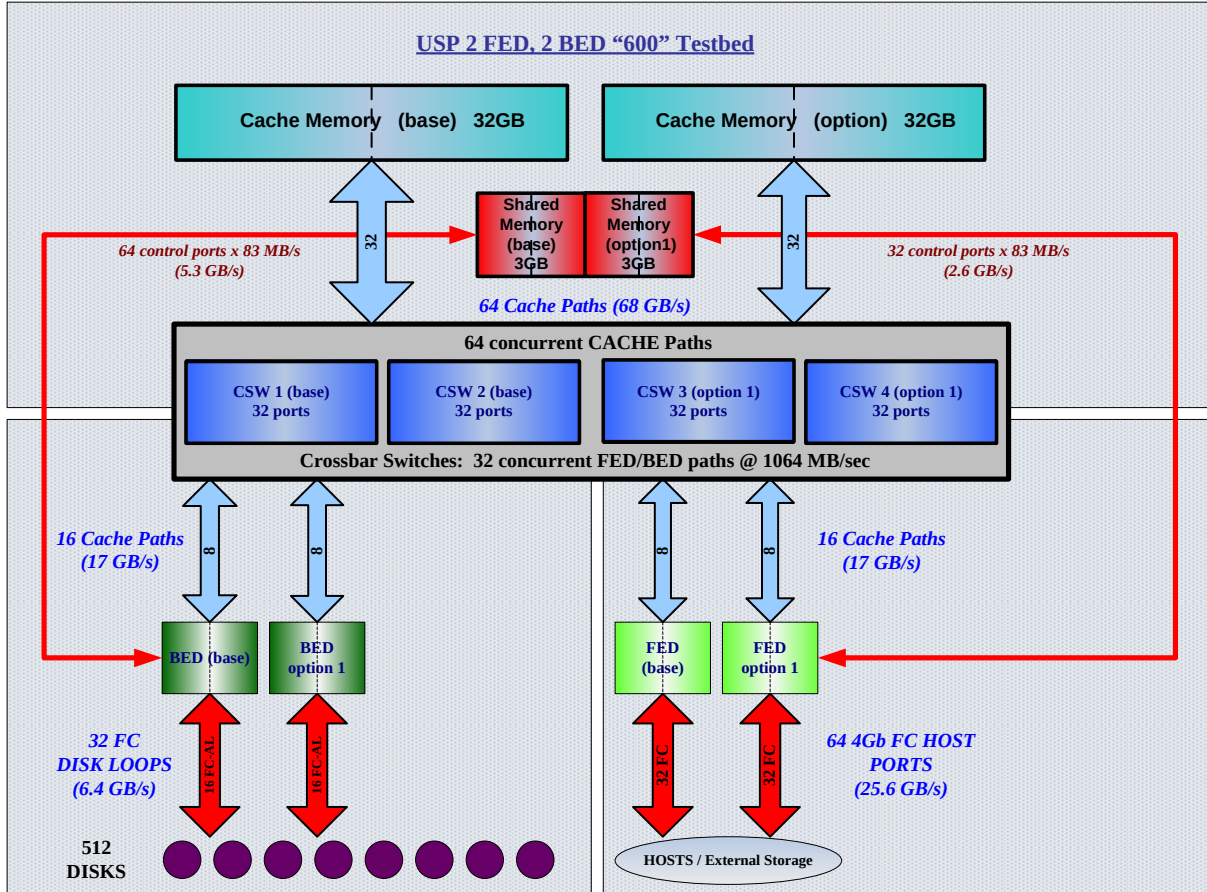
APPENDIX A. USP1100 Architectural Overview

This illustrates how the 4 FED 4 BED 1024 HDD “1100” test bed was laid out.



APPENDIX B. USP600 Test Configuration Overview

This illustrates how the 2 FED 2 BED 512 HDD “600” test bed was laid out.



APPENDIX C. USP100 Test Configuration Overview

This illustrates how the 1 FED 1 BED 256 HDD “100” test bed was laid out.

