



TagmaStore USP Port Design and Performance Analysis

A White Paper

Alan Benway
Senior Performance Consultant

Gil Rangel
Director, Performance Measurements Group

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

22 March 2006

Copyright 2006 Hitachi Data Systems Corporation, ALL RIGHTS RESERVED

Hitachi Data System Restricted Confidential

Table of Contents

OBJECTIVE.....	2
ARCHITECTURE DETAILS.....	2
TACHYON PROTOCOL CHIPS	2
CHIP – CHANNEL HOST INTERFACE PROCESSOR.....	2
CACHE MEMORY INTERFACE (DTA)	3
SHARED MEMORY INTERFACE (MPA).....	3
USP OPEN FIBRE 16-PORT FRONT-END-DIRECTOR	3
USP OPEN FIBRE 32-PORT FRONT-END-DIRECTOR	4
PORT SHARING BY CHIPS.....	5
5. TEST RESULTS SUMMARY	5
RANDOM RESULTS SUMMARY.....	6
SEQUENTIAL RESULTS SUMMARY	8
6. TEST RESULTS	8
RANDOM I/O TRANSACTIONS	9
SEQUENTIAL I/O THROUGHPUT	17
APPENDIX A. TAGMASTORE™ USP 100.....	20
APPENDIX B. TAGMASTORE™ USP 600.....	21
APPENDIX C. TAGMASTORE™ USP 1100	22
APPENDIX D. TEST CONFIGURATION OVERVIEW.	23
TEST INFORMATION	23
HOST CONFIGURATION	23
STORAGE CONFIGURATION	23
VDBENCH CONFIGURATION (RANDOM I/O BACKEND DISK TEST)	24
VDBENCH CONFIGURATION (SEQUENTIAL I/O BACKEND DISK TEST).....	24
APPENDIX E. ESCON FED BLADE.....	25
APPENDIX F. FICON FED BLADE.....	26

Objective

The purpose of this paper is to provide a detailed overview of the Open Systems Fibre Channel Front-end Director (FED) blades available for the TagmaStore Universal Storage Platform line of products. This includes port performance and scalability when using the VDbench workload generator. There are four types of USP FED blades:

- an 8-port ESCON blade
- a 4-port FICON blade
- an 16-port Open Fibre (or Fibre Channel, FC) Option
- a 32-port Open Fibre (or Fibre Channel, FC) Option

This analysis will only address the Open Fibre Option.

Architecture Details

A Front-end Director feature is actually a pair of blades identified as Cluster-1 and Cluster-2. There are two types of USP FC blades, these being the 8-port and 16-port versions. Hence, an FC FED feature will provide either 16 or 32 2 gigabit/sec ports. Each blade is composed of four major sections, these being:

1. Tachyon Chips
2. Processors (CHIP) and local memory
3. Interface to cache memory (DTA)
4. Interface to Shared Memory (MPA)

Tachyon Protocol Chips

The Tachyon chip is the device used to manage all aspects of encoding/decoding the Fibre Channel protocols (Fabric, Point-to-Point, or Arbitrated-Loop) used to communicate with host servers or externally attached storage. The USP FC blades use the DX-2 version of this chip, which auto-negotiates between 1gigabit/sec or 2 gigabit/sec connections. The DX-4 version adds 4 gigabit/sec support.

CHIP – Channel Host Interface Processor

These are the special purpose (as opposed to general purpose workstation CPUs) 64-bit MIPS 400MHz 10-pipeline superscaler microprocessors that control I/O requests to or from the hosts (or externally attached storage), as well as all software packages. These interact with the data cache, the shared memory regions, the Tachyon chips, the large local memory (per CHIP), and the internal network interconnection of all of the CHIPS and the Service Processor (SVP). The local memory is both a temporary data cache, a workspace and the location where both firmware and storage application software is installed.

Cache Memory Interface (DTA)

The FED blades are interconnected by two types of network connections that are included under the umbrella name of Universal Star Network. One of these networks is the **Cache Switched Network** (CSN, for data only) The DTA, or Data Adapter, is a special ASIC for communication between the CHIPS and Cache, using bi-directional serial links rated at 1,064 MB/sec. Unlike all competing designs, these cache paths (and cache itself) are only used for data – not any metadata, control signaling, or software. On each FED blade, there are four DTA switch ports (1,064 MB/sec each into Cache). All four of the DTA switch network ports per blade are simultaneously active (non-blocking), and they allow the much slower FC I/O pipelines (four-to-eight per blade) to run at full speed. These I/O pipelines are driven by the blade's CHIP processors.

Shared Memory Interface (MPA)

The FED blades are also interconnected by the **Shared Memory Network** (SMN, for metadata and control). The MPA, or Microprocessor Adapter, is a special ASIC for communication between the CHIPS and Shared Memory, using bi-directional 83MB/sec serial links for Metadata and Control signals. On each FED blade, there are eight such MPA Shared Memory network ports. All eight of the MPA ports per blade are simultaneously active (non-blocking).

USP Open Fibre 16-Port Front-End-Director

The 16-Port Front-End-Director consists of two blades, each with (8) 2gigabit/sec Open Fibre ports. Each blade has four 400MHz CHIP Processors. While the number of CHIPS is equivalent to the Lightning 9900V's 16-port Open Fibre Channel Adapter, the 9900V's Tachyon XL2 has been replaced by the Tachyon-DX2 and each CHIP processor has improved architecture and double the clock speed. In addition, ports are configured in 4 port packages on the PCB. Figure-1 is a high level diagram of an 8-port blade from the 16-Port Front-end Director feature.

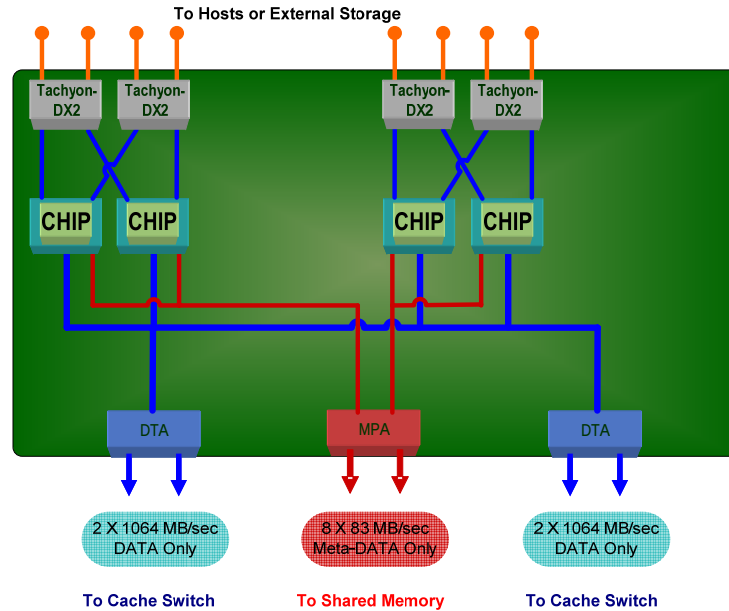


Figure 1. USP Open Fibre 16-port Front-end Director blade.

USP Open Fibre 32-Port Front-End-Director

The Open Fibre 32-Port Front-end Director feature consists of two blades, each with (16) 2Gbit/sec Open Fibre ports. Each blade has eight 400MHz 64-bit CHIP Processors. Ports are configured in 4 port increments. Figure-2 is a high level diagram of a 16-port blade from the 32-Port Front-end Director feature. This is the type used in these tests.

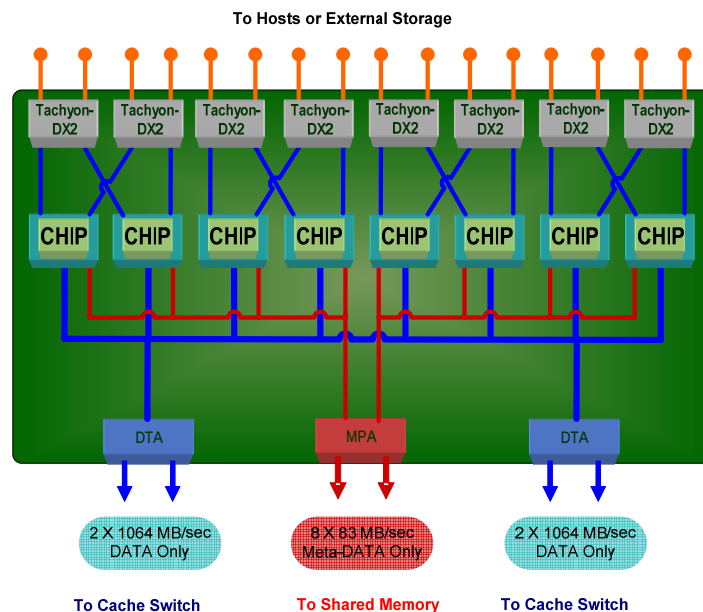


Figure 2. USP Open Fibre 32-Port Front-End-Director blade.

Port Sharing by CHiPs

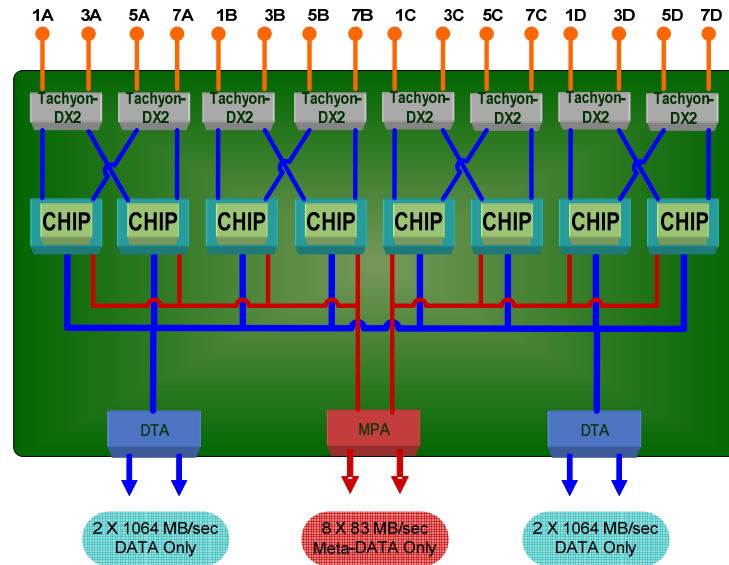


Figure 3. Overview of a 16-port blade (Cluster-1) with its port names.

When configuring a USP, one must be aware that (like all such storage systems) some of the host ports are sharing the same CHIP. Under some combinations of workloads, one has to be careful about mapping the internal LDEVs as LUNs to ports and hosts. Notice how there are two ports assigned to each MP, although these ports alternate in numbering. Ports 1A and 5A share CHIP-1, whereas port 3A (adjacent to 1A on the connector panel) does not. The other blade in the Cluster pair (Cluster-2) has a similar naming sequence, but using 2A, 4A, 6A, 8A, 2B, 4B, 6B, 8B, 2C, 4C, 6C, 8C, 2D, 4D, 6D, and 8D. The 8-port blades only use the first half of these ranges (1A-7B and 2A-8B).

Recall that each CHIP has local memory where it caches I/O blocks to some degree. Also be aware that the CHIP is where the detection of random versus sequential patterns of I/O occurs. A preferred configuration would avoid having LUNs with high workloads assigned to the pairs of ports that are assigned to a shared CHIP if possible. Also, avoiding a mixture of heavy workloads of both random and sequential nature into the same CHIP is desirable. In the next section, you can see how the workloads across the ports scale, with the effects noted above playing a part in the performance dips for tests with shared CHiPs.

5. Test Results Summary

These tests included the use of one 16-port FC blade (Cluster-1 from FED-1) and all four BEDs (Back-end Directors). The ports and associated LUNs were remapped onto the server for each test. The tests were set up to use the following nine host port combinations, where the notation means X-ports with Y-processors (using MP for CHIP):

Ports-MPs	Array Groups	LUNs per Port	FED (Cluster-1) Ports
1p-1mp	1-256	1-256	1A
2p-1mp	2-256	1-128	1A,5A
2p-2mp	2-256	1-128	1A,3A
4p-2mp	4-256	1-64	1A,3A,5A,7A
4p-4mp	4-256	1-64	1A,5A,1B,5B
8p-4mp	8-128	1-16	1A,3A,5A,7A,1B,3B,5B,7B
8p-8mp	8-128	1-16	1A,5A,1B,5B,1C,5C,1D,5D
16p-8mp	16-128	1-8	1A,3A,5A,7A,1B,3B,5B,7B,1C,3C,5C,7C,1D,3D,5D,7D

Figure 4. Port and CHIP test mapping.

There were up to (1024) 72GB 15K rpm disks used for this test. The RAID level used was the four disk RAID-10 (2d+2d). Both Random and Sequential workloads were used with varying block sizes for each. These were 2KB, 4KB, 8KB and 16KB for Random, and 64KB, 128KB, 256KB, and 512KB for Sequential workloads. The tests scaled up to 256 Array Groups (2d+2d) and 16 FC ports with eight processors. The first set of workloads were directed to attempt a 100% cache hit ratio (not always possible), and the second set used a 0% cache hit ratio. There were (256) 33GB LUNs created (one per Array Group), and these were mapped to various combinations of front-end ports for testing.

These results will help answer questions about the kind of scalability to expect across one 16-port blade. Note that the 32-port option for the TagmaStore USP has 16 2-Gbit FC ports and 8 MPs (also known as CHIPS) per blade (see Figure-3). We reported the results for just one of the blades of a cluster pair, since the additional results from the second blade (16p-16mp, 32p-16mp) showed that we were exceeding the I/O capacity of the PCI-X busses of the fairly powerful IBM server used for these tests. The single blade performance results are presented in the charts below.

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 reduces the number of data points to consider. Half of the tests were designed to have a 100% cache hit ratio, but that isn't always possible. Writes must be destaged at some point, and in the order they were received for a particular LUN. All workloads except 100% Read have a significant Write component.

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 **100% Random Read (100% cache hit)** TPS performance achieved for this single FED was 110,417 TPS (8p-8mp, 8KB block, 8 LUNs) at 0.6ms response time.

Notice how all block sizes reached a peak then fell, and these peaks were all located at different port-MP combinations.

The highest level of **100% Random Read (0% cache hit)** TPS performance achieved for this single FED was 39,921 TPS (8p-8mp, 8KB block, 128 LUNs) at 25.6ms response time. Notice how all block sizes showed a near identical profile. The dips at 4p-2mp, 8p-4mp, and 16-8mp are due to how much work the MPs could do when driving two ports

The highest level of **75% Random Read (75% cache hit)** TPS performance achieved for this single FED was 37,173 TPS (16p-8mp, 2KB block, 64 LUNs) at 13.8ms response time. Notice how all block sizes had nearly identical response profiles for this test. The dips at 4p-2mp, 8p-4mp, and 16-8mp are due to how much work the MPs could do when driving two ports.

The highest level of **75% Random Read (0% cache hit)** TPS performance achieved for this single FED was 31,463 TPS (16p-8mp, 8KB block, 128 LUNs) at 32.5ms response time. Notice how all block sizes had nearly identical response profiles for this test. The dips at 4p-2mp, 8p-4mp, and 16-8mp are due to how much work the MPs could do when driving two ports.

The highest level of **50% Random Read (50% cache hit)** TPS performance achieved for this single FED was 33,574 TPS (16p-8mp, 2KB block, 64 LUNs) at 15.2ms response time. Notice how all block sizes had nearly identical response profiles for this test. The dip at 16p-8mp was due to how much work the MPs could do when driving two ports.

The highest level of **50% Random Read (0% cache hit)** TPS performance achieved for this single FED was 26,778 TPS (16p-8mp, 8KB block, 64 LUNs) at 19.1ms response time. Notice how all block sizes had nearly identical response profiles for this test. The dip at 16p-8mp was due to how much work the MPs could do when driving two ports.

The highest level of **100% Random Write (100% cache hit)** TPS performance achieved for this single FED was 89,933 TPS (8p-8mp, 2KB block, 8 LUNs) at 0.7ms response time. The dips at 2p-1mp, 4p-2mp, 8p-4mp, and 16-8mp were due to how much work the MPs could do when driving two ports.

The highest level of **100% Random Write (0% cache hit)** TPS performance achieved for this single FED was 21,564 TPS (8p-8mp, 4KB block, 32 LUNs) at 11.9ms response time. As above, the dips at 2p-1mp, 4p-2mp, 8p-4mp, and 16-8mp were due to how much work the MPs could do when driving two ports.

SEQUENTIAL RESULTS SUMMARY

The highest level of **100% Sequential Read (100% cache hit)** performance achieved for this single FED was 1,357MB/sec (16p-86mp, 256KB block, 64 LUNs). Notice how all block sizes had nearly identical throughput profiles for this test.

The highest level of **100% Sequential Read (0% cache hit)** performance achieved for this single FED was 1,353MB/sec (16p-86mp, 512KB block, 64 LUNs). Notice how all block sizes had nearly identical throughput profiles for this test.

The highest level of **50% Sequential Read (100% cache hit)** performance achieved for this single FED was 1,674MB/sec (16p-8mp, 256KB block, 64 LUNs). Notice how all block sizes had similar throughput profiles for this test, though 64KB lagged the others somewhat.

The highest level of **50% Sequential Read (0% cache hit)** performance achieved for this single FED was 1,624MB/sec (16p-8mp, 256KB block, 64 LUNs). Notice how all block sizes had similar throughput profiles for this test, though 64KB lagged the others somewhat.

The highest level of **100% Sequential Read (100% cache hit)** performance achieved for this single FED was 1,780MB/sec (16p-8mp, 256KB block, 32 LUNs). Notice how all block sizes had nearly identical throughput profiles for this test. At that point the 64KB block test flattened out, though the others kept increasing at a reduced rate.

The highest level of **100% Sequential Read (0% cache hit)** performance achieved for this single FED was 1,728MB/sec (16p-8mp, 256KB block, 32 LUNs). Notice how all block sizes had nearly identical throughput profiles for this test.

6. Test Results

The following three sections present the results in this manner:

- **Random I/O, Transactions per second**, at 2, 4, 8, and 16KB block sizes with 8 Threads (100% Read, 75% Read, 50% Read, and 100% Write)
- **Random I/O, 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 host port and FED blade microprocessor combination from above. Also, be aware that Transactions Per Second is not the same thing as disk 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 Transactions/Sec charts have a polynomial trendline against the 8KB data points added to better visualize the performance increases by port and microprocessor increments. Both the 100% cache hit ratio and 100% Seek results are shown in each pair. The pairs of Transactions/Sec charts are followed by their associated Response Times charts.

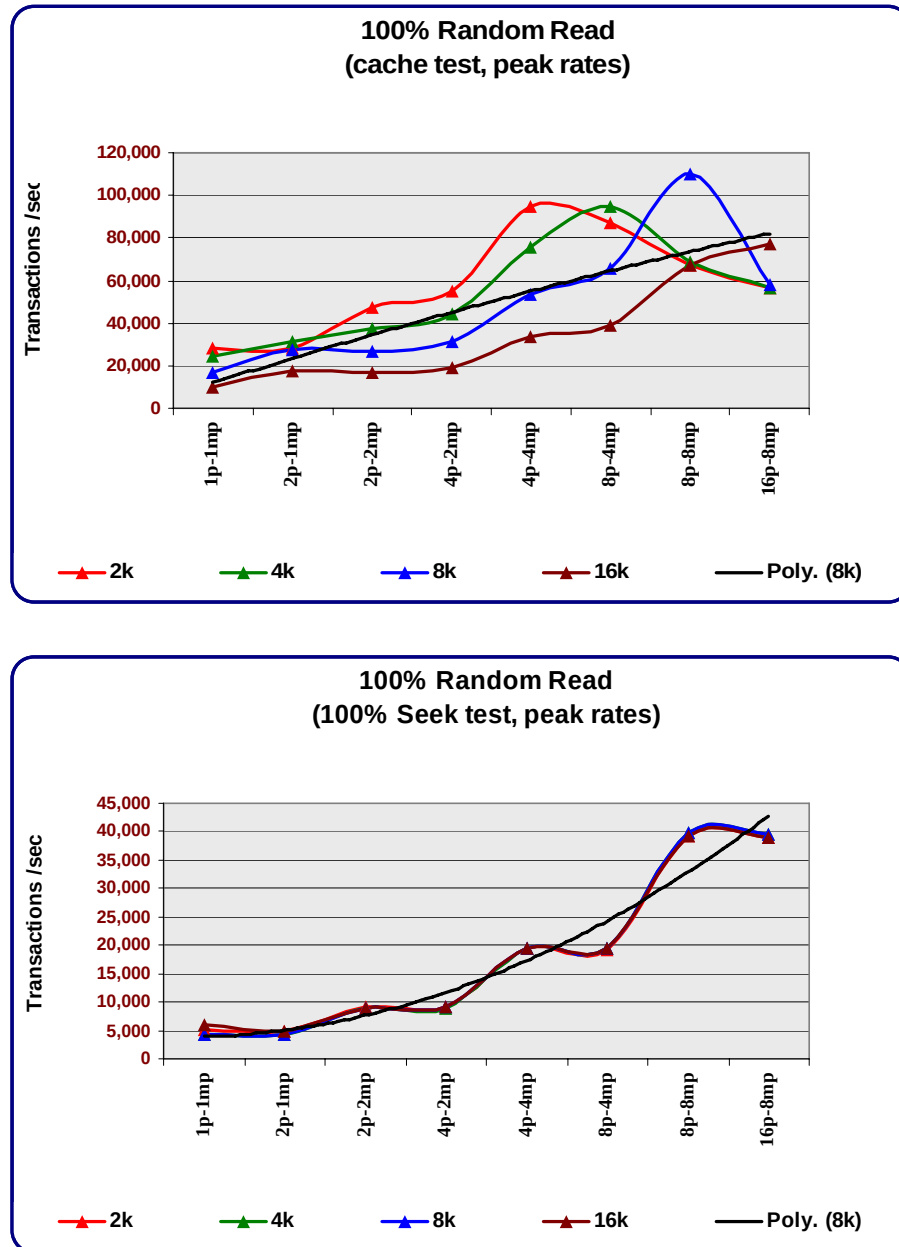


Figure 5. Peak 100% Random Read rates, by port and MP.

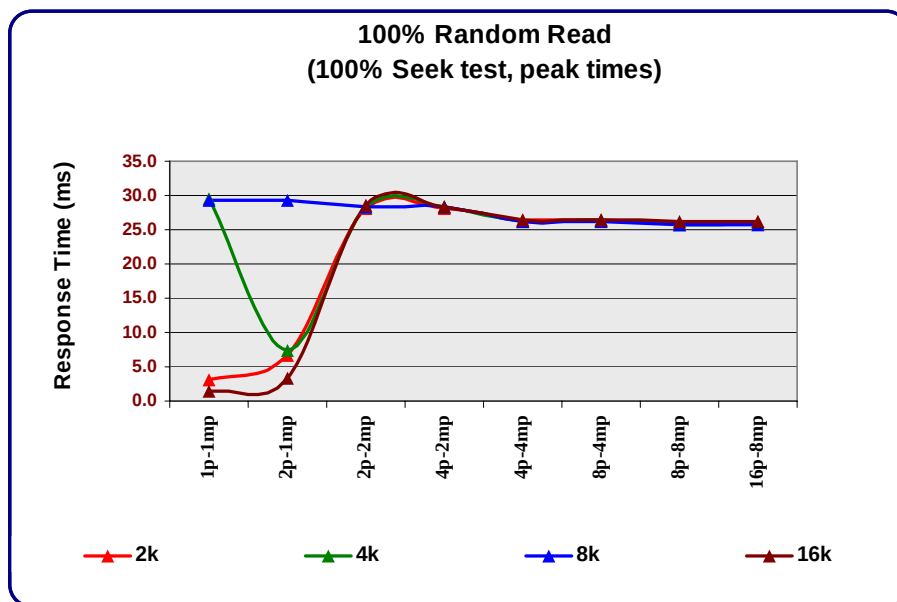
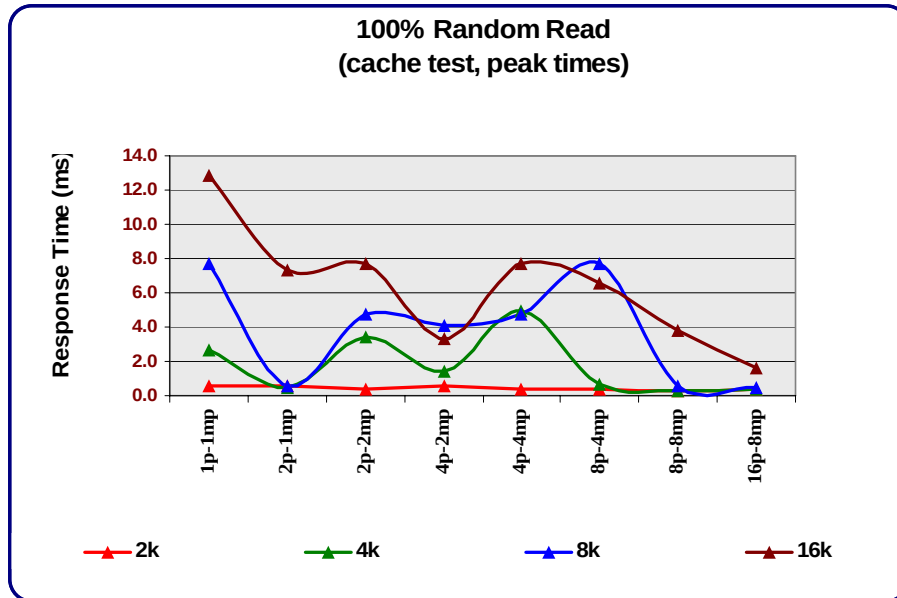


Figure 6. Peak 100% Random Read response times.

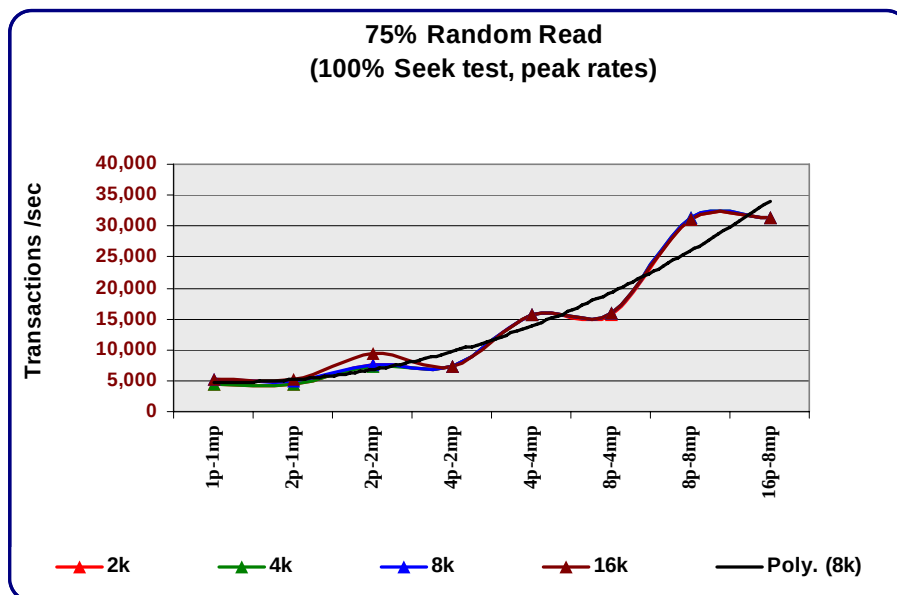
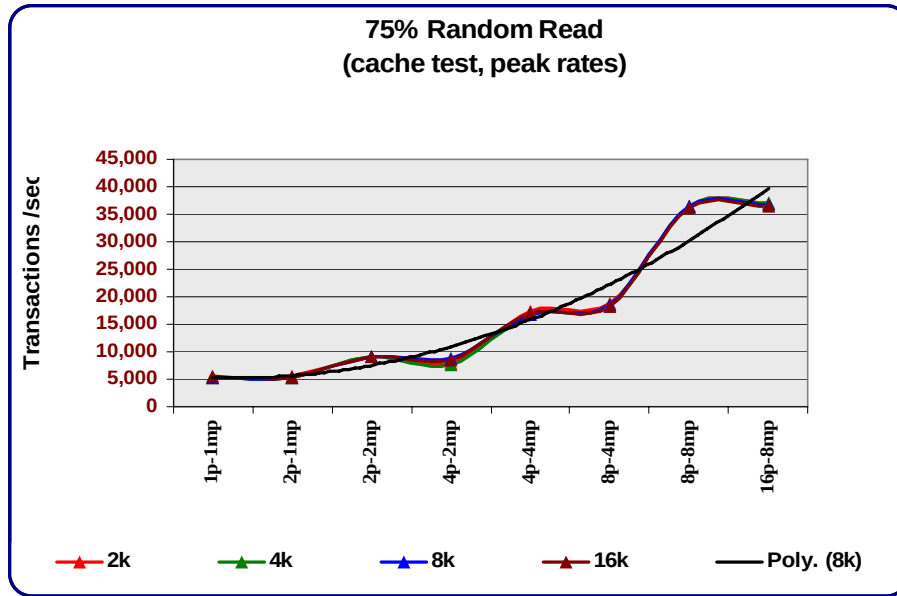


Figure 7. Peak 75% Random Read rates, by port and MP.

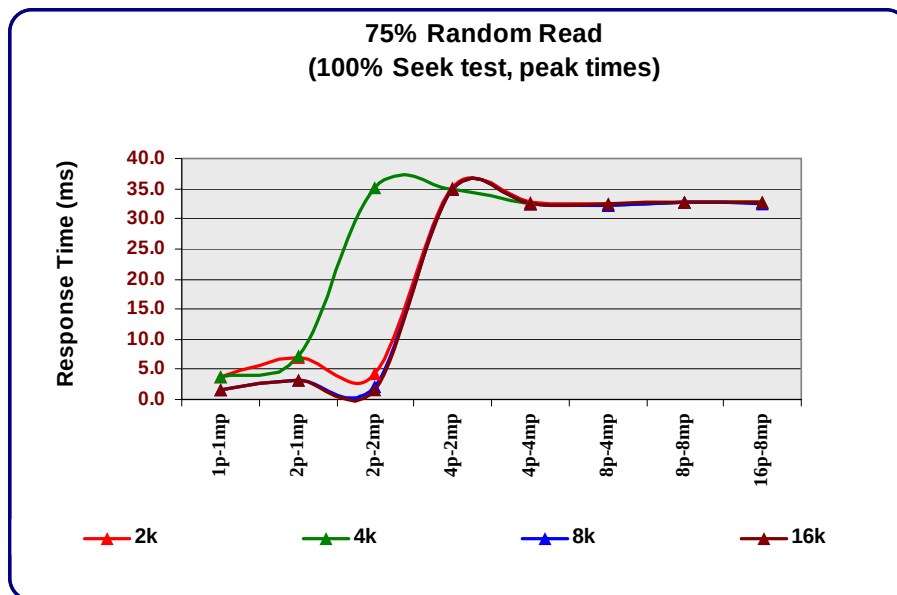
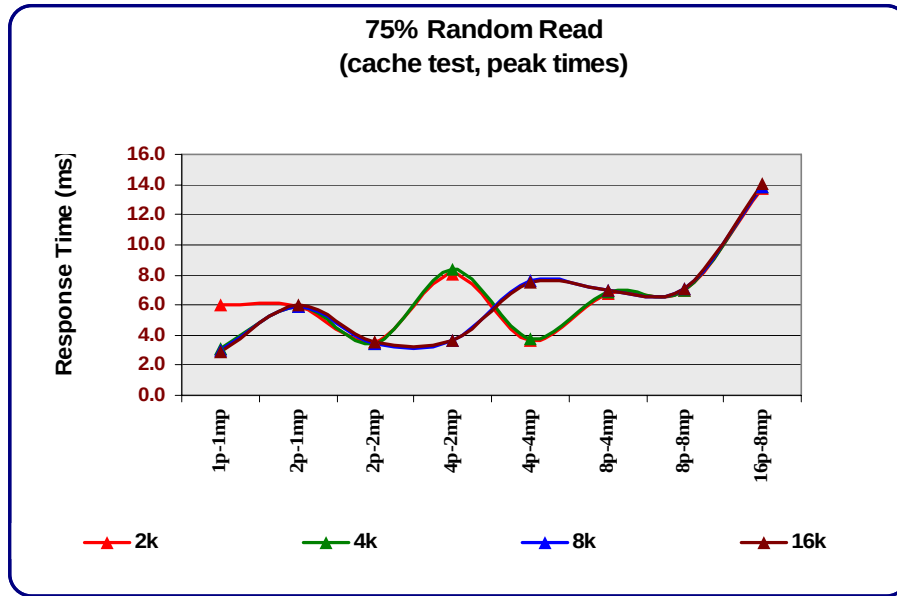


Figure 8. Peak 75% Random Read response times.

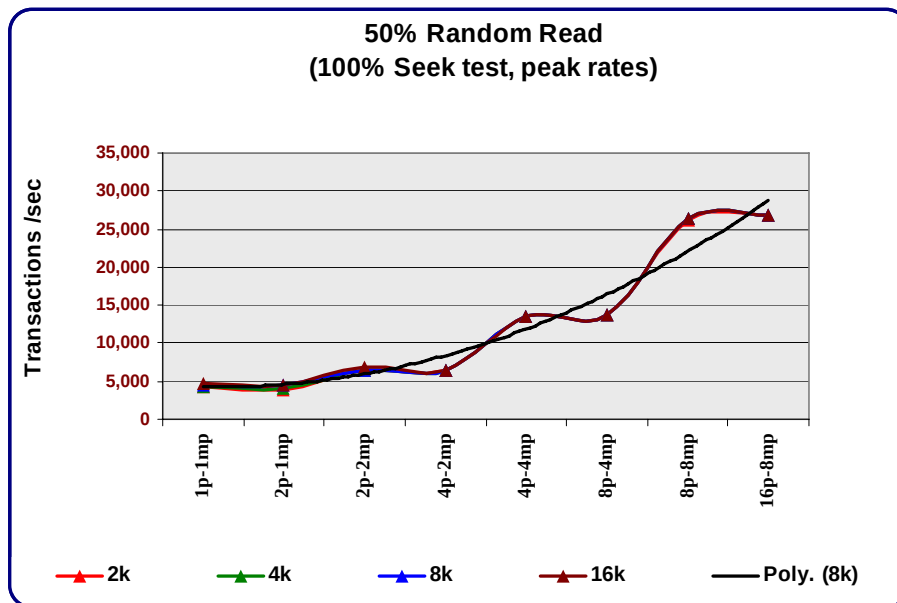
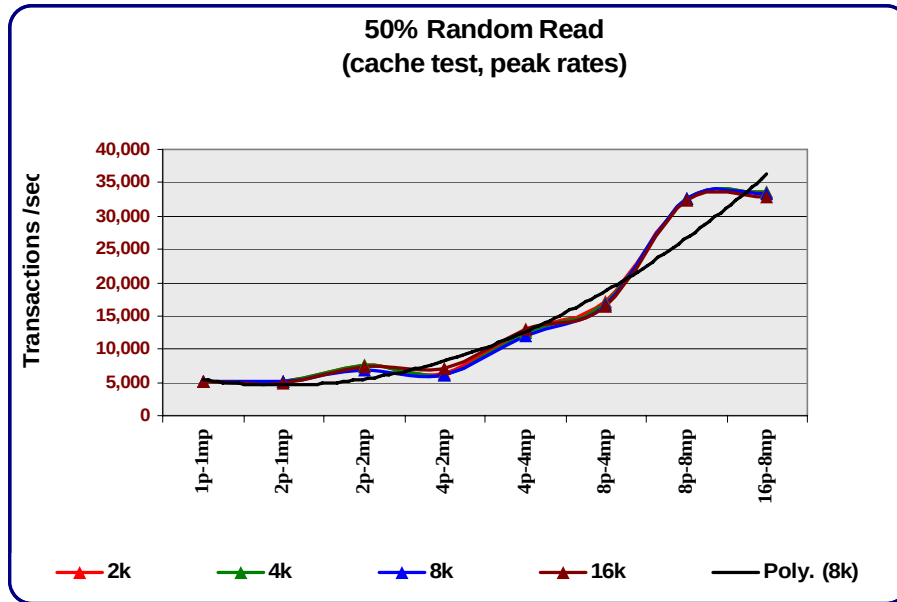


Figure 9. Peak 50% Random Read rates, by port and MP.

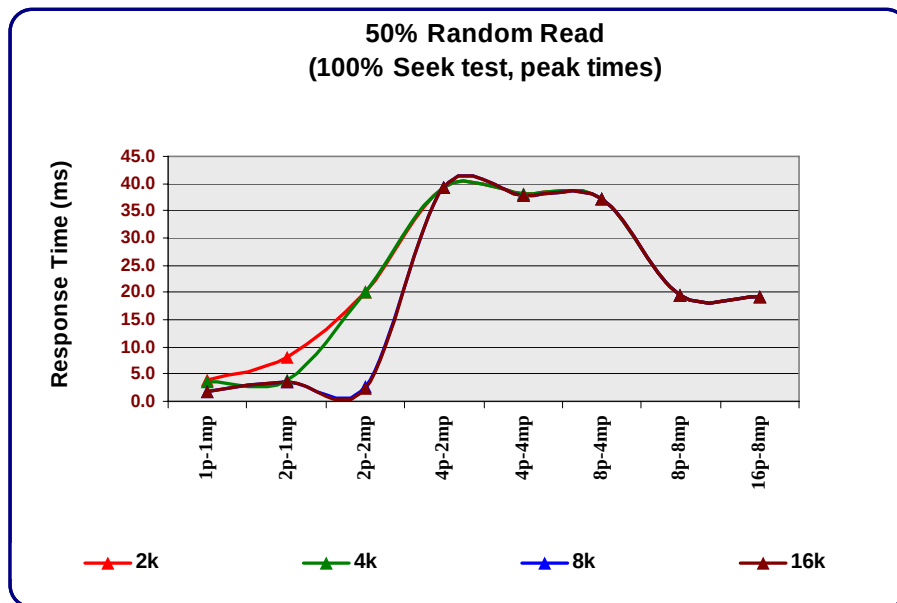
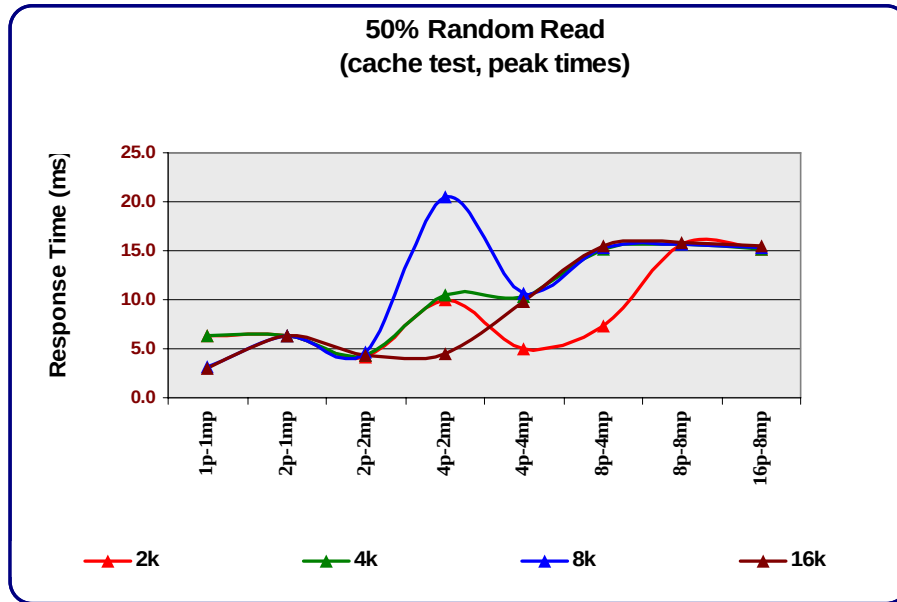


Figure 10. Peak 50% Random Read response times.

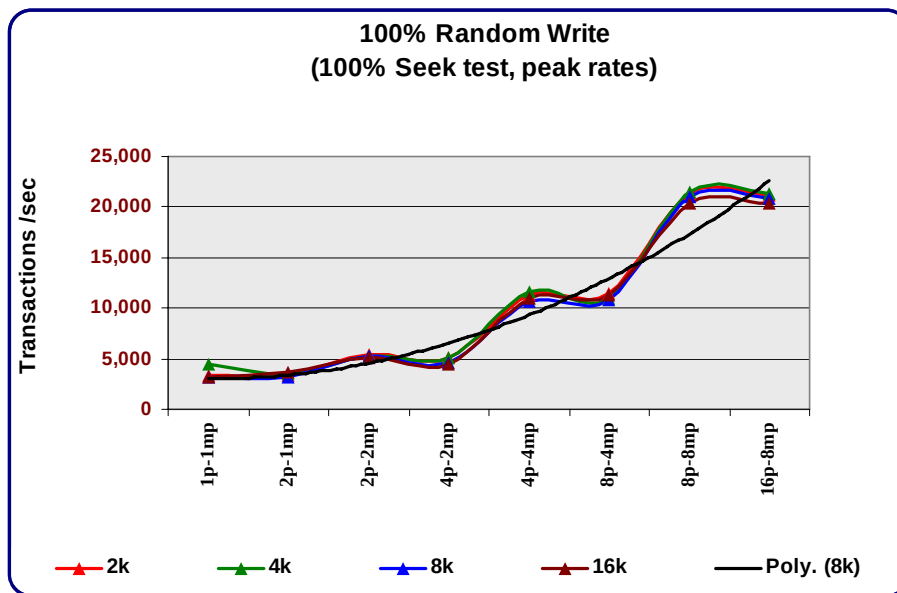
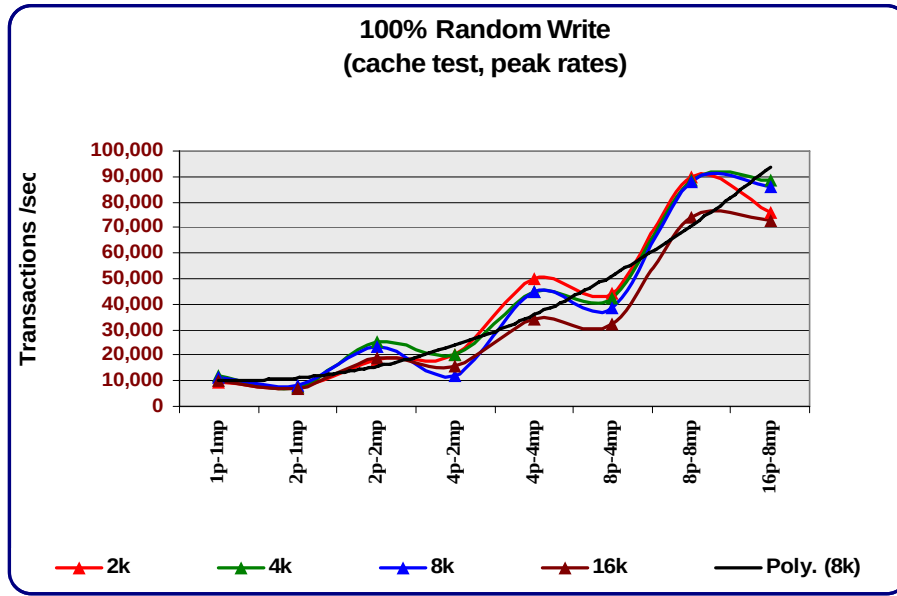


Figure 11. Peak 100% Random Write rates, by port and MP.

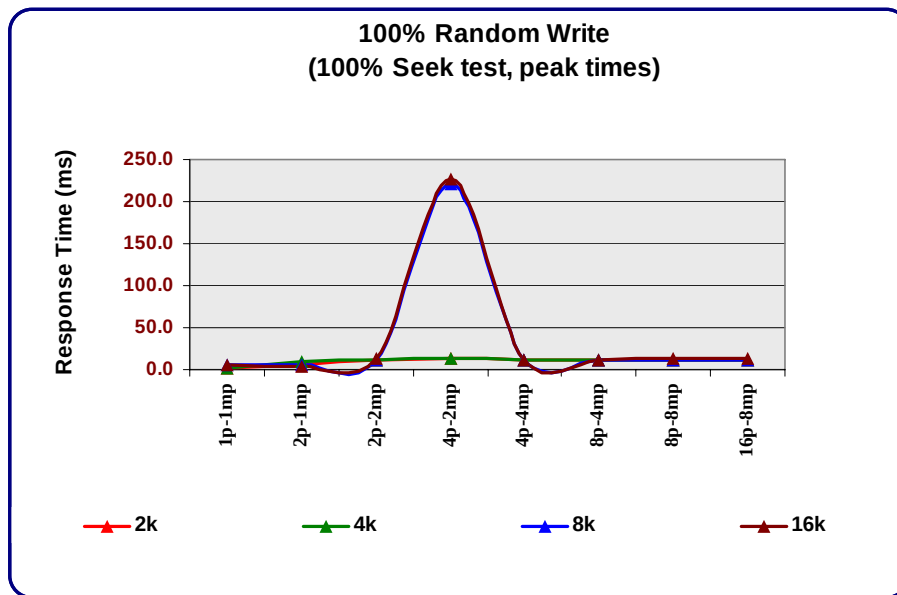
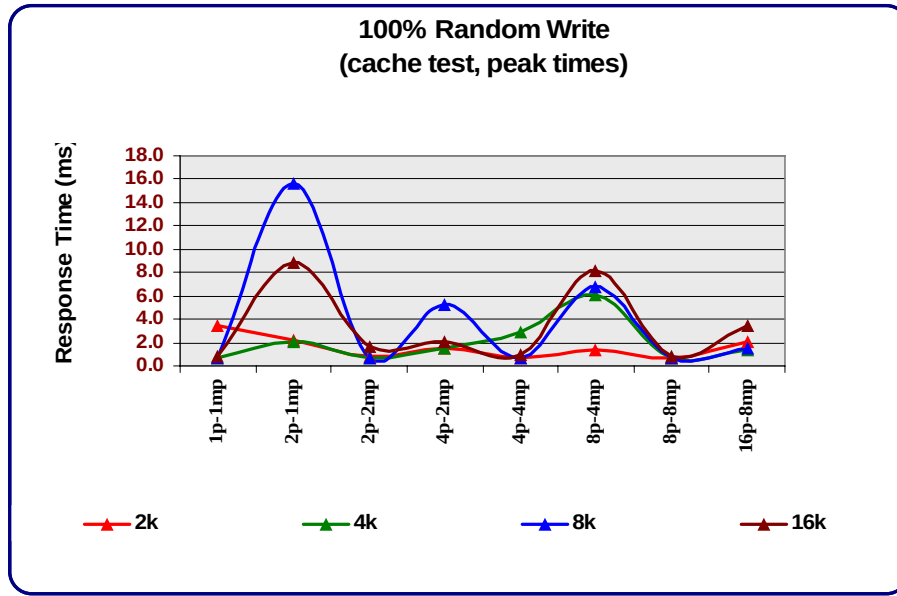


Figure 12. Peak 100% Random Write response times.

Sequential I/O Throughput

The following charts have a polynomial trendline against the 256KB data points added to better visualize the performance increases by port and microprocessor increments. Both the 100% cache hit ratio and 100% Seek results are shown in each pair.

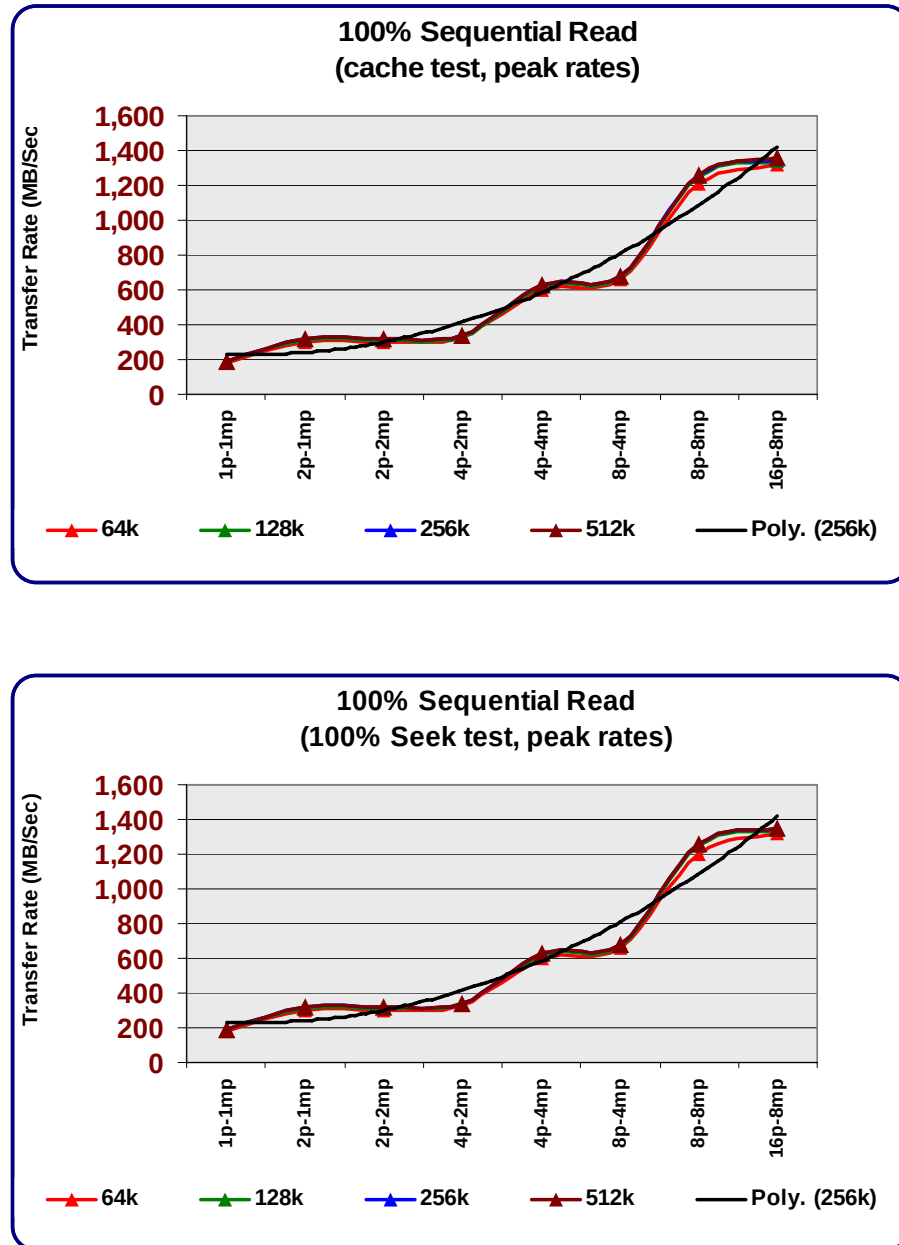


Figure 13. Peak 100% Sequential Read rates, by port and MP.

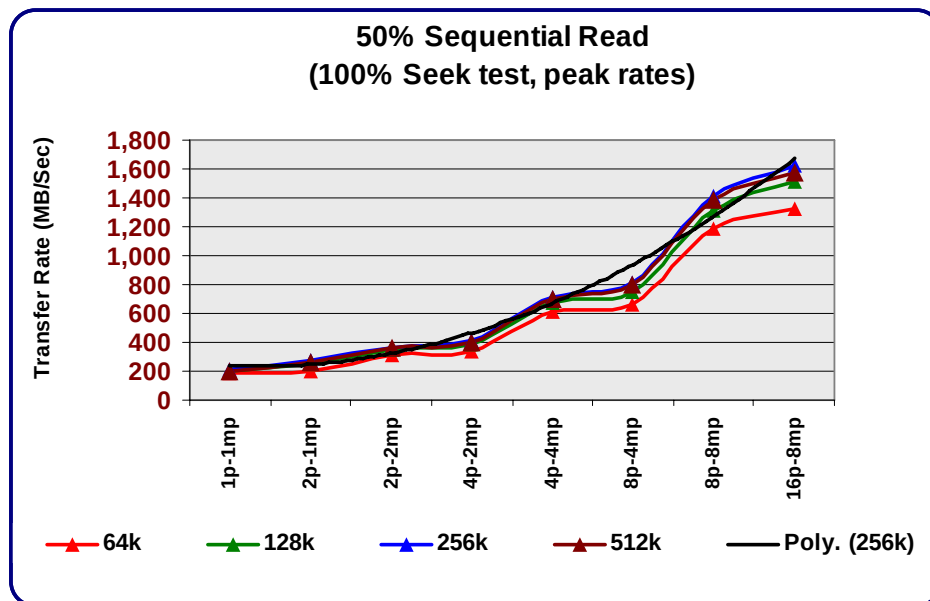
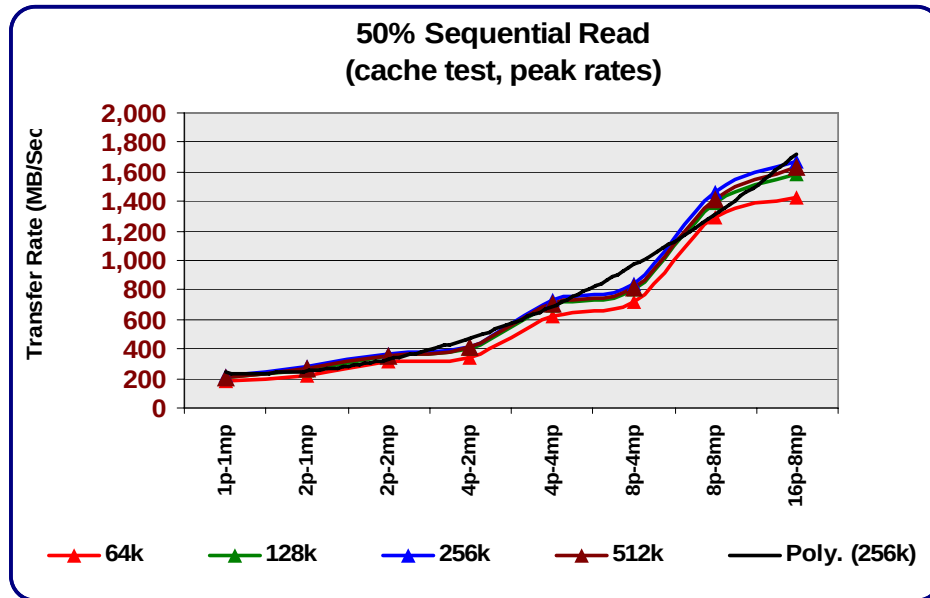


Figure 14. Peak 50% Sequential Read rates, by port and MP.

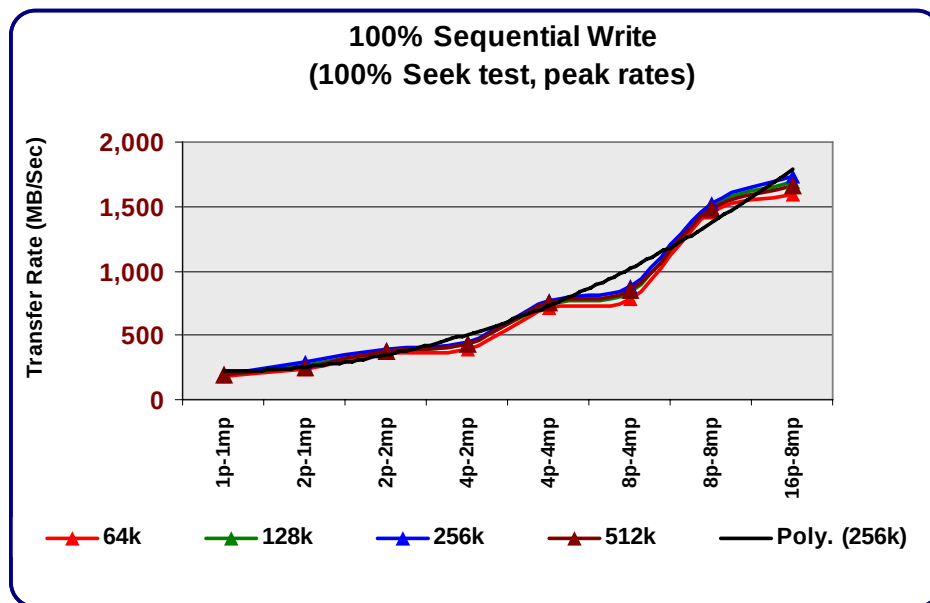
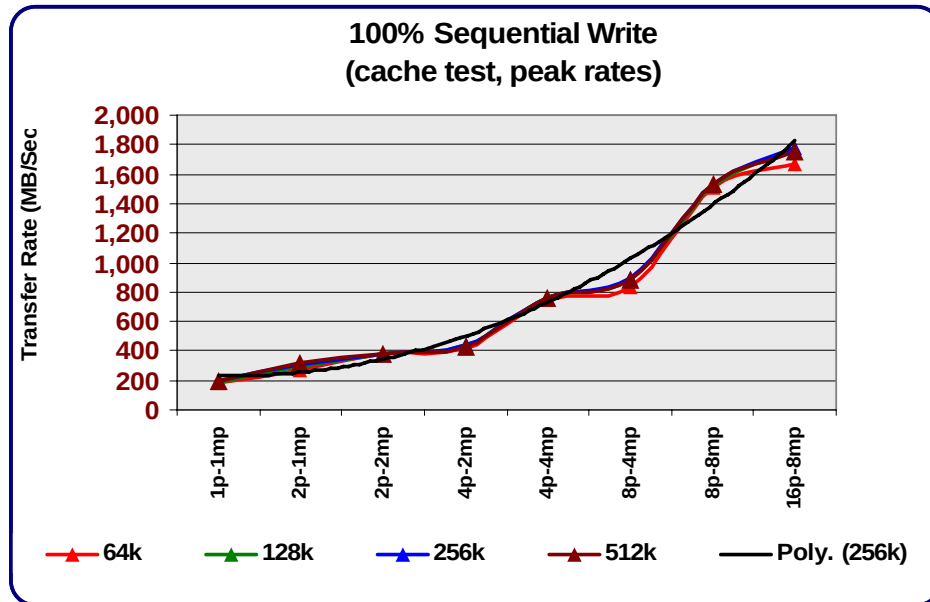
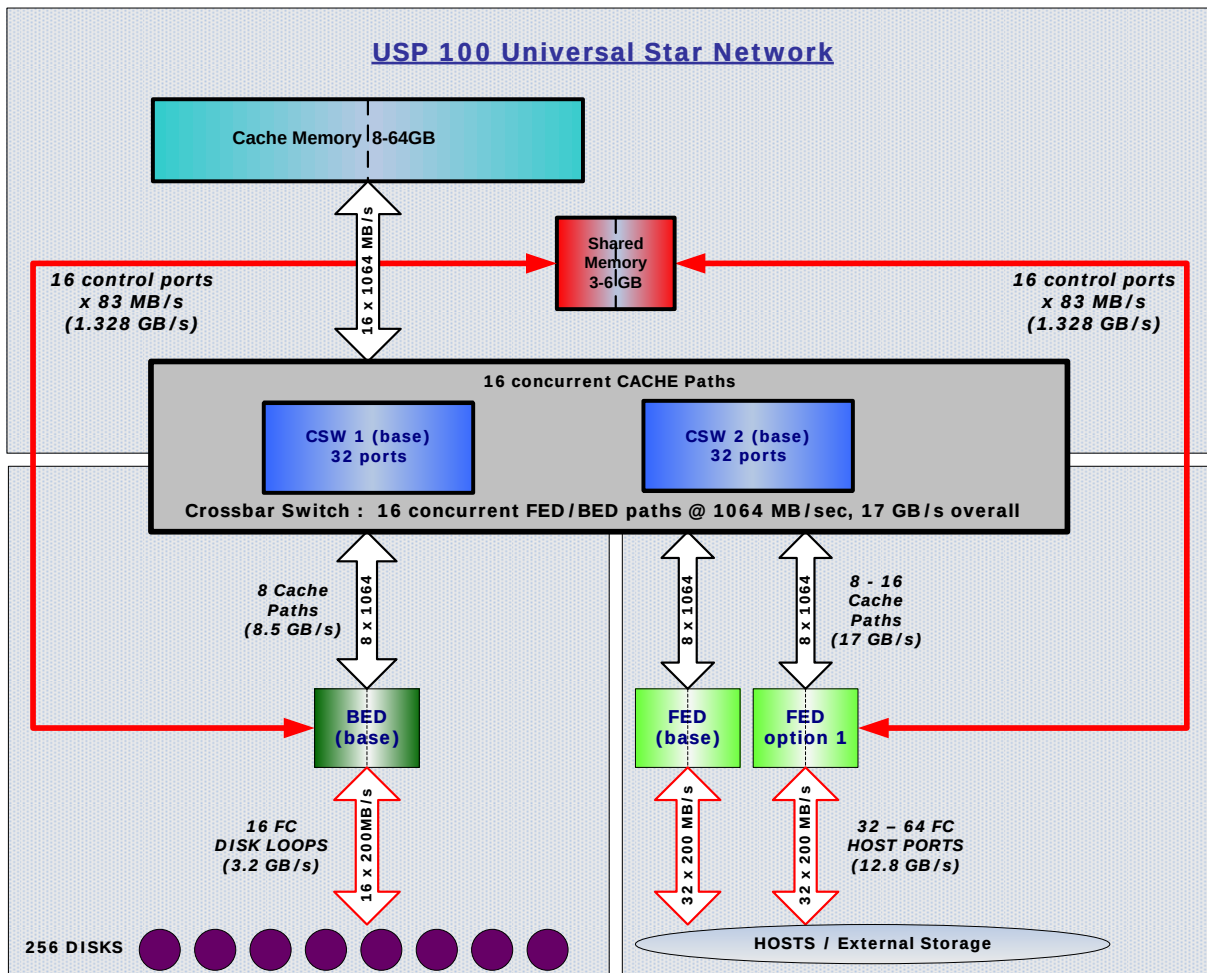


Figure 15. Peak 100% Sequential Write rates, by port and MP.

Appendix A. TagmaStore USP 100

The USP100 is comprised of 16 x 1064MB/sec Data paths, representing 17GB/sec of data bandwidth, and 32 x 83MB/sec Control or Metadata paths, representing 2.6GB/sec of control bandwidth. The overall bandwidth when compared to other solutions is **19.6GB/s**. The USP100 may include these features:

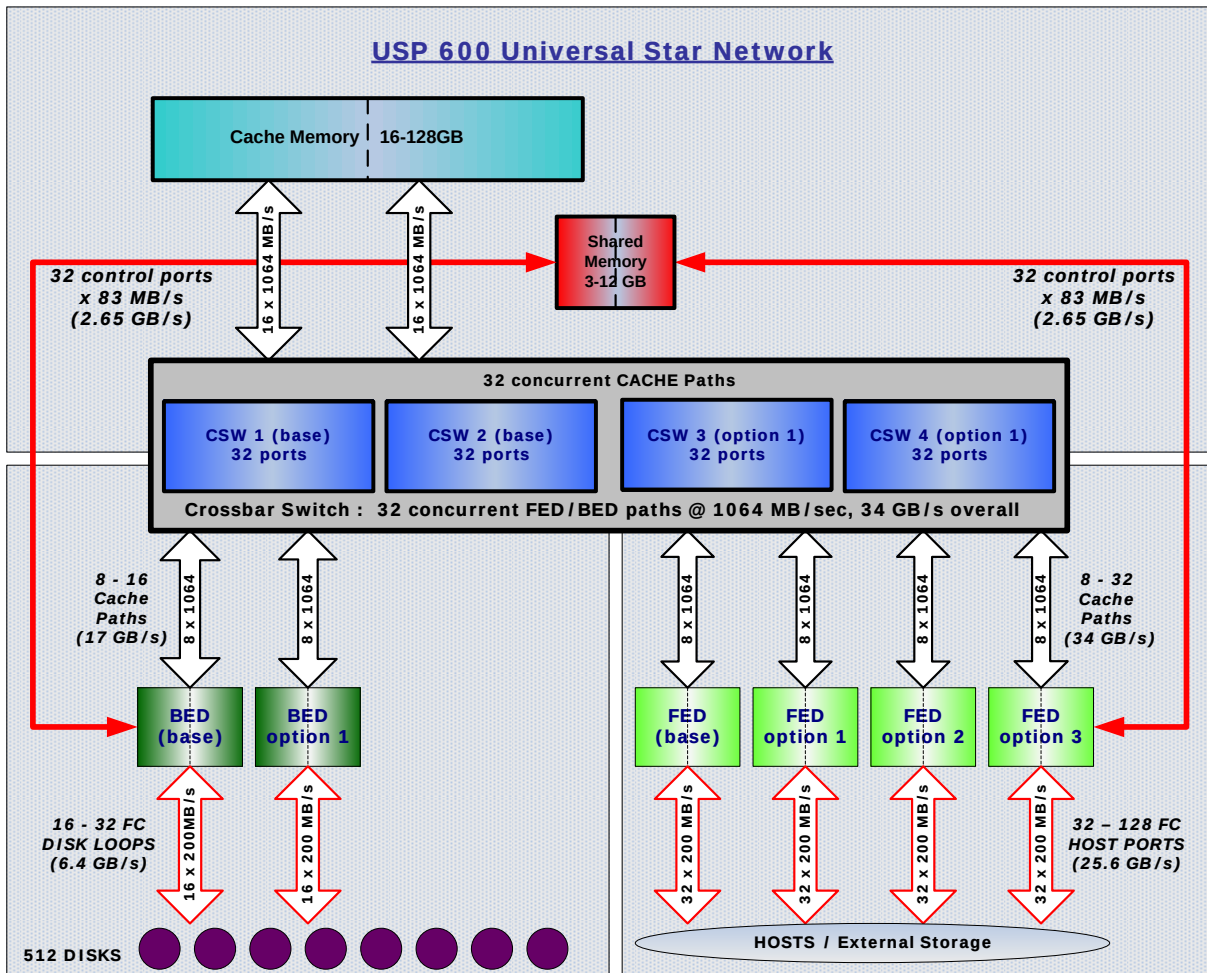
- Up to 256 Fibre Channel disks.
- One Back-end-Director Pair, consisting of 16 – 2Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 3.2GB/sec.
- One Cache Card Pair supporting up to 64GB of cache.
- Two Cache Switches with a total internal data bandwidth of 17GB/sec.
- One Shared Memory Pair with a total metadata bandwidth of 2.6GB/sec.
- Two Front-End-Director Pairs supporting either 32 – ESCON, 16 – 2Gbit/sec FICON, 32 – 2Gbit/sec Fibre Channel, or 64 – 2Gbit/sec Fibre Channel ports.



Appendix B. TagmaStore USP 600

The USP600 is comprised of 32 x 1064MB/sec Data paths, representing 34GB/sec of data bandwidth, and 64 x 83MB/sec Control or Metadata paths, representing 5.3GB/sec of control bandwidth. The overall bandwidth when compared to other solutions is **39.3GB/s**. The USP600 may include these features:

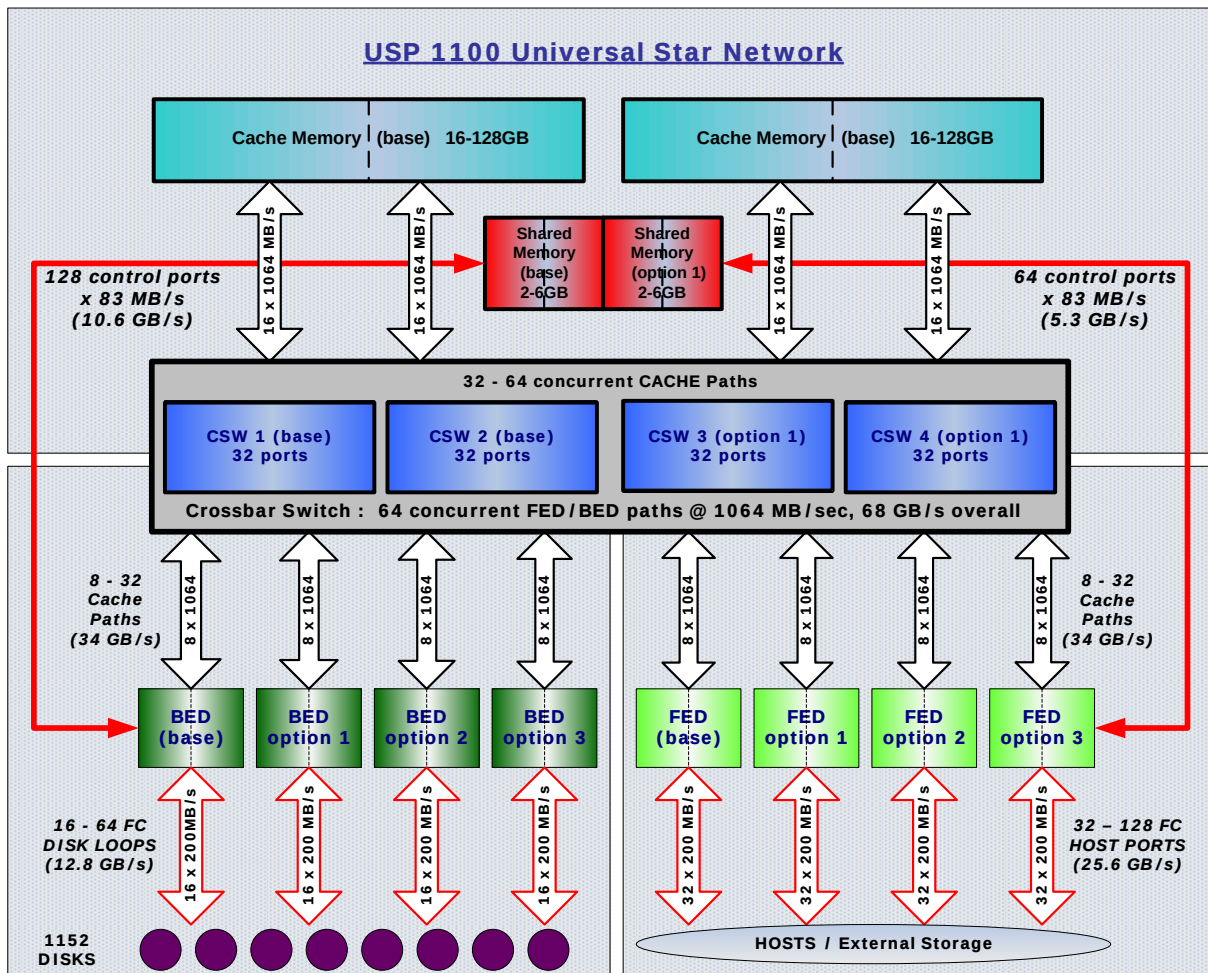
- Up to 512 Fibre Channel disks, 128 in the controller frame, 256 in the R1 frame and 128 in the R2 expansion frame.
- Two Back-end-Director Pairs, consisting of 32 – 2Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 6.4GB/sec.
- One Cache Card Pair supporting up to 64GB of cache.
- Four Cache Switches with a total internal data bandwidth of 34GB/sec.
- One Shared Memory Pair with a total metadata bandwidth of 5.3GB/sec.
- Four Front-End-Director Pairs supporting either 32 – ESCON, 16 – 2Gbit/sec FICON, 32 – 2Gbit/sec Fibre Channel, or 64 – 2Gbit/sec Fibre Channel ports.



Appendix C. TagmaStore USP 1100

The USP1100 is comprised of 64 x 1064MB/sec Data paths, representing 68GB/sec of data bandwidth, and 192 x 83MB/sec Control or Metadata paths, representing 15.6GB/sec of control bandwidth. The overall bandwidth when compared to other solutions is **83.6GB/s**. The USP1100 may include these features:

- Up to 1152 Fibre Channel disks, 128 in the controller frame, 256 in each R1, R2, L1, and L2 expansion frames.
- Four Back-end-Director Pairs, consisting of 64 – 2Gbit/sec Fibre Channel loops, for a total back-end bandwidth of 12.6GB/sec.
- Two Cache Card Pairs supporting up to 128GB of cache.
- Four Cache Switches with a total internal data bandwidth of 34GB/sec.
- Two Shared Memory Pairs with a total metadata bandwidth of 15.6GB/sec.
- Four Front-End-Director Pairs supporting either 32 – ESCON, 16 – 2Gbit/sec FICON, 32 – 2Gbit/sec Fibre Channel, or 64 – 2Gbit/sec Fibre Channel ports.



Appendix D. Test Configuration Overview.

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.

Test Information

Test Period	Report Date	Location	Tester
February/March 2006	March 2006	Santa Clara CA, USA	Performance Measurement Group

Host Configuration

Server	Operating System	CPU	Memory	HBA
IBM p595 (Scruggs)	AIX 5.3	8 x 1.65 GHz Power5	24GB RAM	32 x 2Gb QLogic HBA

Storage Configuration

Storage	Microcode	Number of Controllers	Cache Size
TagmaStore USP1100	50-05-24-00/00	1 FED, 32 FC ports (2 Gbit)	128GB

Type of Drive	Size of Drive	Number of Drives	RPM
FC-AL	72GB	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 %, 100 %	0 %, 100 %	8	✓	✓	✓	✓
	50 %			8	✓	✓	✓	✓
	75 %			8	✓	✓	✓	✓
	100 %			8	✓	✓	✓	✓

VDbench Configuration (Sequential I/O Backend Disk Test)

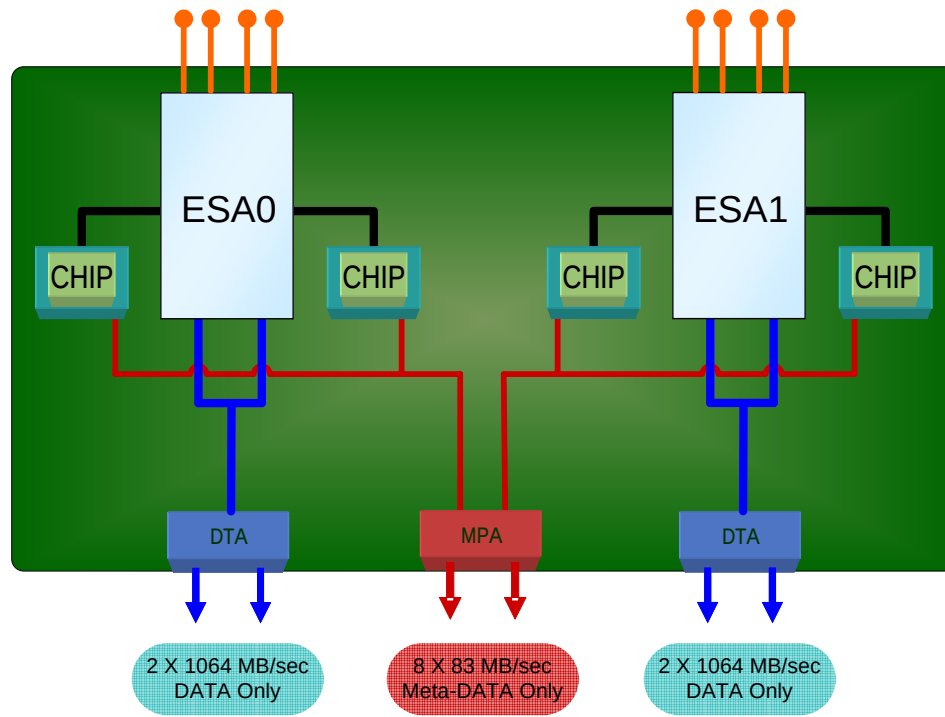
Seek %	Read %	Cache Read Hit %	Cache Write Hit %	Threads per LUN	64K Block	128K Block	256K Block	512K Block
100 %, 0 %	0 %	0 %, 100 %	0 %, 100 %	1	✓	✓	✓	✓
	50 %			1	✓	✓	✓	✓
	75 %			1				
	100 %			1	✓	✓	✓	✓

The tests were set up to use the following nine host port combinations, where the notation means X-ports with Y-processors (CHiPs):

Ports-MPs	Array Groups	LUNs per Port	FED (Cluster-1) Ports
1p-1mp	1-256	1-256	1A
2p-1mp	2-256	1-128	1A,5A
2p-2mp	2-256	1-128	1A,3A
4p-2mp	4-256	1-64	1A,3A,5A,7A
4p-4mp	4-256	1-64	1A,5A,1B,5B
8p-4mp	8-128	1-16	1A,3A,5A,7A,1B,3B,5B,7B
8p-8mp	8-128	1-16	1A,5A,1B,5B,1C,5C,1D,5D
16p-8mp	16-128	1-8	1A,3A,5A,7A,1B,3B,5B,7B,1C,3C,5C,7C,1D,3D,5D,7D

For the higher port count tests, the number of LUNs was reduced from a maximum of 256 down one step to 128. These extra tests would have taken a long time and added nothing useful to the information since the jump in response times indicated that we were mostly testing the back-end instead of the front-end at that point. Overall there were 600 Sequential tests and 800 Random tests conducted, each taking about 8 minutes to complete.

Appendix E. USP ESCON FED blade.



Appendix F. USP FICON FED blade.

