



TagmaStore USP Architecture Analysis

By Alan Benway
Senior Performance Consultant

Gil Rangel
Director of Performance Measurement

*Hitachi Data Systems
Performance Measurement Group,
Enterprise Labs, Technical Operations
750 Central Expressway
Santa Clara, CA 95050*

January 2006 Update

Copyright 2005 Hitachi Data Systems Corporation, ALL RIGHTS RESERVED

Notices and Disclaimer

Copyright© 2005 Hitachi Data Systems, Inc.

No part of this document may be reproduced or transmitted without written approval from Hitachi Data Systems, Inc.

This document has been reviewed for accuracy as of the date of initial publication. Hitachi Data Systems, Inc. may make improvements and/or changes in product and/or programs at any time without notice.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED “AS IS” WITHOUT ANY WARRANTY, EITHER EXPRESSED OR IMPLIED. Hitachi Data Systems, Inc. EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR PARTICULAR PURPOSE OR NON-INFRINGEMENT.

Lightning 9980V[®] and TagmaStore USP[®] are registered trademarks of Hitachi Data Systems, Inc. in the United States, other countries, or both.

Other company, product or service names may be trademarks or service marks of others.

Executive Summary

The intent of this document is to familiarize Hitachi Data Systems' sales personnel and technical support staff with a comparative architectural overview and derived real-world bandwidths of the TagmaStore USP.

Some points to take away:

- The TagmaStore USP uses bidirectional, fault tolerant data and metadata paths which are substantially faster than all competitive products.
- All write data is mirror protected in cache, which avoids the potential for data loss, should there be a catastrophic failure such as power loss to one cluster, or the failure of a cache board. Such protection is not available in some competitive products which attempt to use cache parity for protection.
- The Universal Star Network is Hitachi's third generation Crossbar switched network with separate data and metadata paths. All competitive products use a single shared cache space and common paths for data and metadata.
- All of the USP's Back-end Fibre Channel loops are simultaneously active. A major competitive product uses an active/passive design, thus losing use of half of these loops to fail-over mode.
- The TagmaStore USP uses purpose-built control processors and dedicated RAID processors. A competitive product uses general purpose PowerPC processors to manage storage functions in software.
- The TagmaStore USP has great flexibility in terms of intermixing RAID levels within and across Back End Directors. A competitive product does not.
- The TagmaStore USP models are easily upgraded by adding components, having no requirement for a "forklift upgrade".
- The TagmaStore USP allows multiple host types to share the same Front end Director ports. Most competitive products only allow one host type per port.

Table Of Contents

<i>Notices and Disclaimer</i>	<i>i</i>
<i>Executive Summary</i>	<i>ii</i>
<i>TECHNICAL SUMMARY</i>	<i>1</i>
<i>I. ARCHITECTURAL OVERVIEW.....</i>	<i>2</i>
<i>II. HARDWARE COMPONENTS</i>	<i>4</i>
<i>FRONT-END DIRECTORS</i>	<i>5</i>
<i>CACHE BOARDS</i>	<i>5</i>
<i>BACK-END DIRECTORS</i>	<i>6</i>
<i>DISK DRIVES</i>	<i>7</i>
<i>III BANDWIDTH OVERVIEW.....</i>	<i>9</i>
<i>Appendix A. The USP Cache Engine.....</i>	<i>10</i>
<i>Appendix B. USP600 Architecture Overview.....</i>	<i>12</i>
<i>Appendix C. USP100 Architecture Overview.....</i>	<i>13</i>
<i>Appendix D. Open Fibre 32-port FED blades</i>	<i>14</i>
<i>Appendix E. FICON 4-port FED blade</i>	<i>15</i>
<i>Appendix F. ESCON 8-port FED blade.....</i>	<i>16</i>

TECHNICAL SUMMARY

The intent of this document is to familiarize Hitachi Data Systems' sales personnel and technical support staff with an architectural overview and derived real-world bandwidths of the Hitachi Data System TagmaStore USP product. Some points to take away:

- The **TagmaStore USP** storage is first a collection (in the dozens to a couple hundred) of **hardware RAID disk groups** that are then partitioned into one or more LUNs. EMC's **DMX** storage is first a large (perhaps in the thousands) **JBOD pool** of small disk partitions that are then organized into software RAID groups layered on top of these randomly selected disk partitions (Hypervolumes). These designs are totally opposite in philosophy.
- **The USP has dedicated hardware RAID processors on the Back-End Director boards.** There are **no RAID processors** in any of EMC's Symmetrix products. EMC uses software RAID, where this work is managed in software by the four PowerPC processors on each Disk Director and inside the disks themselves.
- **With the USP, any mix of RAID protection modes may be chosen within each system.** Five choices of RAID modes are available. There is RAID-5 (3D+1P, 7D+1P), RAID-6 (6D+2P), and RAID-1+ (RAID-10 as 2D+2D, 4D+4D). One does not have to make a global selection of using either 4-disk or 8-disk groups as with EMC DMX.
- The **Universal Star Network** is a third generation Crossbar Switched Network with **separate data and metadata paths**. In this design, all of these paths are simultaneously active, bidirectional, and fast. In the USP1100, there is 64GB/sec of cache bandwidth (data only) and 15.9GB/sec of shared memory bandwidth (metadata only).
- The USP1100 has 32 independent banks of cache memory, and the EMC DMX3000 has 32 "memory domains" (banks of memory). The USP also has several banks of shared memory (metadata), while an EMC has none.
- The abrupt failure of a cache board will not **crash** the USP system, unlike the EMC DMX.
- The **USP's Back-End** boards operate in pairs for fault tolerance, with the two FC-AL loops per disk being connected to separate BED boards within a pair. However, these loops are operated as **Active/Active**, where either BED can access a disk, so that all FC-AL loops are in use. In an EMC DMX, half of these paths are sitting unused in fail-over (passive) mode.
- One can simply **add parts to upgrade** a USP system to grow from the USP100 to the USP600 and then into the USP1100.

- On the USP the **host emulation is at the LUN level**, using the idea of virtual ports (as *Host Storage Domains*), whereby multiple host types can attach to the same FED port via a SAN switch, but with security provided to isolate them from one another. With **DMX, the host emulation setting is at the port level**.

I. ARCHITECTURAL OVERVIEW

This perspective of this paper is a feature and bandwidth analysis of the TagmaStore USP architecture using derived bandwidth examples of fully configured systems. A distinction is made between physical paths and simultaneously usable paths or functions. This is the true measure of the bandwidth offered by a particular design. This routine type of analysis is commonly referred to by engineers as “Bandwidth Reduction Analysis”. Other impediments to the moving of user data blocks in or out of a storage system are also pointed out with estimates for those reductions where possible. Hidden limitations of features are also listed when applicable. The following sections delve into the board-level characteristics of each system, working from front to back (host to disk). There are detailed drawings in the Appendix to further illustrate the hardware layouts and overall bandwidths.

The Hitachi Data Systems ***TagmaStore USP architecture*** is a large scale enterprise-grade storage system that supports Open Systems and mainframe environments. The largest TagmaStore USP solution scales up to 1,152 300GB disks (about 345 TB raw), 128GB of system-wide cache, and 128 Open Fibre host connections. The TagmaStore USP controller frame uses supports a variety of boards, these being:

1. **Front-end Director (FED, or host controller board)**
2. **Back-end Director (BED, or disk controller board)**
3. **Cache Switch boards (CSW or crossbar switch)**
4. **Cache Boards**
5. **Shared Memory Boards**

All of these boards are installed in pairs, with each board located on a separate power boundary (separate power supply rails). The individual boards within the pair are referred to as *Cluster-1* or *Cluster-2* depending on which power boundary is associated with the specific chassis slot.

TagmaStore USP supports three types of data protection modes, these being:

- **RAID-10** (2D+2D, 4D+4D)
- **RAID-5** (3D+1P, 7D+1P)
- **RAID-6** (6D+2P)

Note that the “D” signifies “data” and P means “Parity”. RAID-10 means that you have mirrored disks (RAID-1) assembled into a volume (RAID-0), where the data blocks are striped across the elements. In the case of 4D+4D, there would be a single RAID-0 stripe over 4 mirrors. Note that RAID-0+1 is very different from RAID-10, and highly vulnerable to failure. RAID-0+1 means that there are two 4-disk RAID-0 stripe sets (no

data protection) that are in a mirrored configuration. If one disk in one RAID-0 stripe set fails, then the mirror fails and you are left with one unprotected RAID-0 volume.

Storage is managed internally as **LDEV's** (Logical DEVICES), these being the set of partitions (user selectable size) allocated across RAID-protected groups of four or eight disks. One can choose any mixture of protection types on any BED pair for any set of four or eight disks. These LDEV's appear as a LUN's (Logical Unit) on a host system once assigned to a FED path to an attached host (for Open Systems connections) and as Logical Devices for FICON and ESCON connections.

The type of host Operating System is specified on a **per-LUN** basis. On the USP the host connection emulation is at the LUN level, using the idea of virtual ports (Host Storage Domains), whereby multiple host types can attach to the same FED port via a SAN switch, with security provided to isolate them from one another.

OVERVIEW SUMMARY	HDS USP-100	HDS USP-600	HDS USP-1100
a. Chassis board slots	14	22	28
b. Front-end Director board pairs (FED)	2	4	4
- Open Fibre ports (2-Gbit)	64	128	128
- FICON ports (2-Gbit)	16	32	32
- ESCON PORTS	32	64	64
- NAS GigE ports	16	32	32
c. Cache board pairs	1	1	2
Min-max cache (GB)	8-64	8-64	16-128
d. Shared Memory board pairs	1	1	2
Min-max Shared Memory (GB)	3-12	3-12	3-12
e. Cache Switch board pairs (CSW)	1	1	2
f. Back-end Director board pairs (BED)	2	4	4
- FC-AL 2Gbit loops	16	32	64
g. Disks	256	512	1152
h. Total DTA cache ports	16	32	64
i. Total MPA Shared Memory ports	32	64	196

Table 1. Overview of TagmaStore USP configuration options.

The FED and BED boards are interconnected by two types of network connections that are included under the umbrella name of Universal Star Network. These two networks include the **Cache Switched Network** (CSN, for data only) and the **Shared Memory Network** (SMN, for metadata and control). The DTA, or Data Adapter is a special ASIC for communication between the Directors and Cache Switch, which use bi-directional serial links rated at 1064 MB/sec for Data movement only. The MPA, or Microprocessor Adapter, is a special ASIC for communication between the Directors and Shared Memory Adapters, which use bi-directional 83 MB/sec serial links for Metadata or Control data movement only.

There are 4 DTA switch ports on every FED or BED board, and 16 on each Cache board. There are 32 DTA ports per Cache Switch. There are 8 MPA ports on each FED board, 16 on each BED board, and up to 192 on the Shared Memory boards. There is also a private Ethernet network on and among the FED and BED boards. This is how the four-to-eight processors (MP's) per board communicate with the Service Processor (SVP).

New firmware is downloaded to each MP over this network, and MP run time metrics are collected here by the SVP.

The NEC MIPS-based 10-pipeline superscaler processor (MP) used on TagmaStore FED and BED boards is a 64-bit special purpose processor developed for control systems. Each MP on each board has a large amount of local RAM used to hold firmware, software packages, and local short-term data caching.

II. HARDWARE COMPONENTS

Shown below in Figure-1 is an overview of the major sub-components and maximum useable raw bandwidths of a model USP1100. If you used eight disks for global spares, then the number of RAID groups would range from 143 (8-disk types only) to 286 (4-disk types only) on a system with 1152 disks. Any mix of RAID group types is allowed. See Appendix B & C for overviews of the USP600 and USP100.

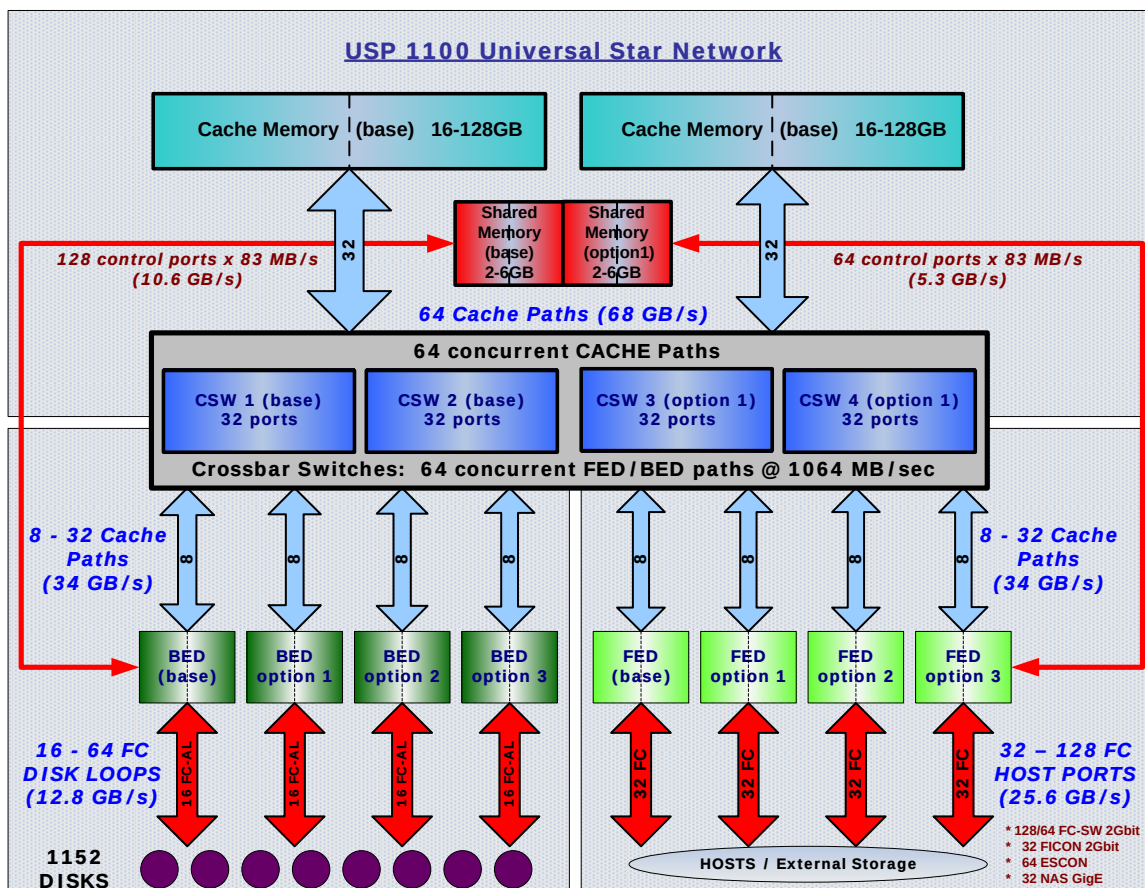


Figure 1. Overview of a fully optioned USP1100 (Open Fibre FEDs).

FRONT-END DIRECTORS

Shown below in Table 2 is a summary of details for the Front-End Director boards. These must be installed as pairs, where each board is located in a chassis slot that uses different power domains. The boards of a pair are referred to as Cluster-1 or Cluster-2. On each FED board, there are four DTA switch ports (1064 MB/sec each into the Cache Switches) and eight MPA Shared Memory network ports (83 MB/sec each). All four of the DTA switch network ports per board are simultaneously active, and allow the four-to-eight I/O pipelines per board to run at full speed. These pipelines use the NEC processors (MP's) to drive them.

There are four types of FED boards: an 8-port ESCON board (four MP's), a 4-port FICON board (eight MP's), an 8-port Open Fibre board (four MP's), and a 16-port Open Fibre board (eight MP's). On the Open Fibre boards, since each pair of host ports share an MP, only half of the host paths may be simultaneously active. See Appendix d, E, and F for illustrations of these boards.

FRONT-END DIRECTOR DETAILS	USP-100	USP-600	USP-1100
FED pairs	2	4	4
- total host paths (w/ FC 2Gb)	64	128	128
active host paths (concurrent)	32	64	64
bandwidth to hosts (MB/s)	6,400	12,800	12,800
- total DTA cache ports	16	32	32
active DTA ports (concurrent)	16	32	32
data bandwidth to cache (MB/sec)	17,024	34,048	34,048
- total MPA Shared Memory ports	16	32	32
metadata bandwidth (MB/s)	1,328	2,656	2,656

Table 2. Overview of TagmaStore USP Front-end Directors.

CACHE BOARDS

Shown below in Table 3 is an overview of the details of the Cache and Shared Memory boards. There are sixteen DTA switch ports per cache board, and all are simultaneously active. Each DTA port has a bidirectional bandwidth of 1064 MB/sec. There are eight banks of DDR RAM per board, and each RAM bank has more than twice the bandwidth of the individual DTA ports. With port buffering to handle the small transaction bursts, the 8-banks of RAM can keep all 16 ports running at full speed. See *Appendix A* for details.

CACHE DETAILS	USP-100	USP-600	USP-1100
Cache board pairs	1	1	2
- cache size (max GB)	64	64	128
- total useable DTA cache ports	16	32	64
active DTA ports (concurrent)	16	32	64
data bandwidth in-out of cache (MB/sec)	17,024	34,048	68,096
Shared Memory board pairs	1	1	2
- Shared memory size (max GB)	12	12	12
- total SM (MPA) Shared Memory ports	32	64	196
data bandwidth in-out of control memory (MB/sec)	2,656	5,312	16,268

Table 3. Overview of the TagmaStore USP Cache and Shared Memory engine.

BACK-END DIRECTORS

Shown below in Table 4 is a summary of details regarding the Back-End Disk Directors (BED's). These are FC-AL (Fibre Channel Arbitrated Loop) disk controllers, with sixteen FC-AL loops per BED pair. These are used as eight pairs of redundant loops, where each disk is attached to a pair of loops (from different BED boards) for redundancy. Also, these loops are operated in an Active/Active mode for parallel access to the disks. Thus, all of the backend loops are working, so the full 16-path bandwidth is available.

On each BED board, there are four DTA switch ports (4,256 MB/sec into the Cache Switches) and sixteen MPA network ports (1,328 MB/sec into Shared Memory Adapters). All four of the DTA switch network ports per board are simultaneously active, and these allow the eight I/O pipelines per board to run at full speed. These pipelines use eight 64-bit NEC processors (MP's) and eight dedicated Data Recovery & Regeneration (DRR) ASIC's to manage the I/O operations between disks and global cache. RAID functionality is managed by the DRR ASIC's, including all mirroring and XOR parity functions (eXclusive OR, a digital logic function used to generate parity blocks).

BACK-END DIRECTOR DETAILS	USP-100	USP-600	USP-1100
BED pairs	1	2	4
- total FC loops (2-Gbit) to disk	16	32	64
active loops to disk (concurrent)	16	32	64
peak bandwidth to disks (MB/s)	3,200	6,400	12,800
- total DTA cache ports	8	16	32
active DTA ports (concurrent)	8	16	32
data bandwidth in-out of cache (MB/sec)	8,512	17,024	34,048
- total MPA Shared Memory ports	16	32	128
metadata bandwidth (MB/s)	1,328	2,656	10,624

Table 4. Overview of TagmaStore USP Back-End Directors.

Figure-2 below illustrates the manner in which disks are attached to the 16 FC-AL loops behind each BED pair, and how they could be organized into the various types of RAID groups. Note that as these RAID groups are isolated behind each BED pair, one may practice "performance management by workload isolation" by locating production RAID groups behind different BED pairs than used by general purpose storage RAID groups. The DMX does not allow for such workload isolation. Notice how the 4-disk RAID groups (i.e. RAID-10, 2D+2D; RAID-5 3D+1P) and 8-disk RAID groups (4D+4D, 7dD+1P) are situated behind a single BED cluster pair.

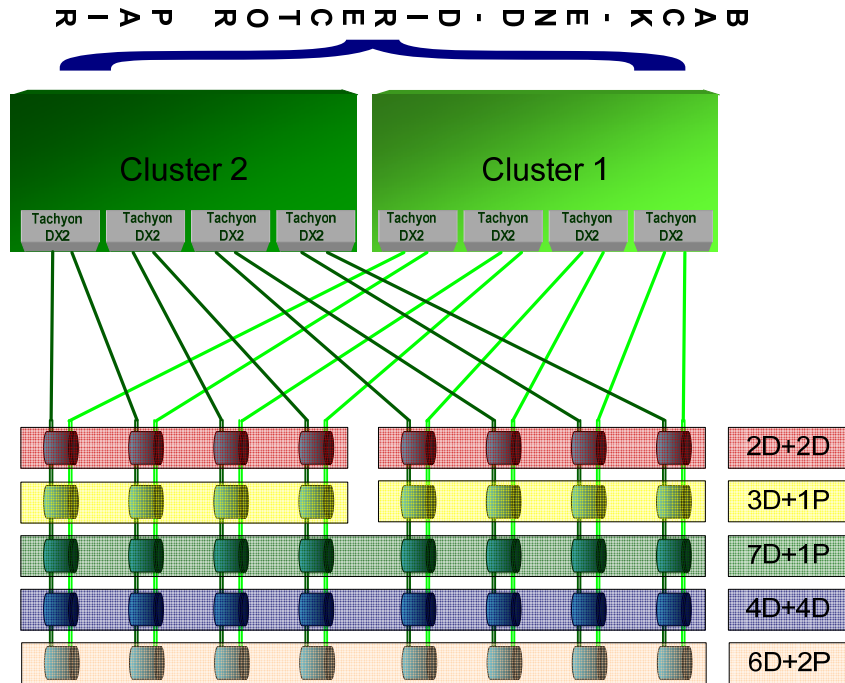


Figure 2. Typical RAID group layouts behind a BED pair.

DISK DRIVES

The 2Gbit FC disk choices include: 73GB 15k rpm, 146GB 10k & 15k rpm, and 300GB 10k rpm. It is important to note that the disks behind a BED can be independently configured as RAID-10 (2+2, 4+4), RAID-5 (3+1, 7+1), or RAID-6 (6+2).

For the IOPS calculations, note that these assume the block to be updated has already been read. It takes two BED writes for RAID-1, and three BED I/O's (one read and two writes) for RAID-5 parity operations.

DISKS (73GB: 10k & 15k; 146GB: 10k & 15k; 300GB: 10k)	USP-100	USP-600	USP-1100
Disks, total (max)	256	512	1,152
Disks per pair of Active/Active FC-loops	32	32	36
simultaneous disk transfers (concurrent)	16	32	64
Maximum available disk IOPS (8kb blocks, 10k rpm disks):			
- RAID-10, 4+4 writes	15,360	30,720	69,120
- RAID-10, 4+4 reads	30,720	61,440	138,240
- RAID-5, 7+1 writes	8,960	17,920	40,320
- RAID-5, 7+1 reads	26,880	53,760	120,960

Table 5. Overview of TagmaStore USP disk drives.

Table 6 below provides an overview of the association of disks and BED pairs. References such as “loop-A” are for the 16 FC-AL loops of a BED pair. The numeric references (i.e. 1:1-6) are the hardware names of the parity groups to be found in that section. Each rack has a front and rear section. There are sixteen HDU enclosures (each with 16 drives) per expansion rack, with only eight in the DKC controller rack.

L2		L1		USP 1100		R1		R2	
BED-4 (18: 1-16)	BED-3 (16: 1-16)	BED-4 (10: 1-16)	BED-3 (8: 1-16)	BED-1 (2: 1-16)	BED-1 (1: 1-16)	BED-2 (6: 1-16)	BED-1 (4: 1-16)	BED-2 (14: 1-16)	BED-1 (12: 1-16)
8 loops-D 64 HDD	8 loops-C 64 HDD	8 loops-D 64 HDD	8 loops-C 64 HDD	8 loops-A 64 HDD	8 loops-A 64 HDD	8 loops-B 64 HDD 4 HDUs (rear)	8 loops-A 64 HDD 4 HDUs (front)	8 loops-B 64 HDD	8 loops-A 64 HDD
BED-4 (17: 1-16)	BED-3 (15: 1-16)	BED-4 (9: 1-16)	BED-3 (7: 1-16)	1152 HDD * 12 GB Shared Memory * 128 GB Cache * 64 FC to disks (4 BED) (4 sets of 16 loops) (BED1: 48 HDD) (BED2,3,4: 32 HDD)		BED-2 (5: 1-16)	BED-1 (3: 1-16)	BED-2 (13: 1-16)	BED-1 (11: 1-16)
8 loops-D 64 HDD HDU Box	8 loops-C 64 HDD HDU Box	8 loops-D 64 HDD HDU Box	8 loops-C 64 HDD HDU Box			8 loops-B 64 HDD 4 HDUs (rear)	8 loops-A 64 HDD 4 HDUs (front)	8 loops-B 64 HDD	8 loops-A 64 HDD
dka		dka		dkc		dka		dka	

Table 6. Overview of USP1100 racks and disks.

III BANDWIDTH OVERVIEW

In Table 7, the bandwidth numbers from Tables 2, 3, and 4 have been extracted for easier consideration. A value called ‘limit’ has been added to show the lowest point of bandwidth, this being the true limiting factor in I/O to the host. These numbers reflect actual data movement, since data blocks are the only elements in the global CSN cache. There is no cache space or bandwidth lost for overhead (metadata, control tables, or software packages). The TagmaStore USP design has a separate Shared Memory cache space (with a private connection network) where all metadata and control tables reside.

BANDWIDTH SUMMARY	HDS USP-100	HDS USP-600	HDS USP-1100
100% cache miss Read operation bandwidth (RAID-10): MB/sec			
1. channels to host	6,400	12,800	12,800
2. FED from cache	17,024	34,048	34,048
3. BED to cache	8,512	17,024	34,048
4. disks to BED	3,200	6,400	12,800
limit (MB/s):	3,200	6,400	12,800
Write operation bandwidth (RAID-10): MB/sec			
1. channels from host	6,400	12,800	12,800
2. FED to cache (duplexed)	8,512	17,024	17,024
3. BED from cache (duplexed)	4,256	8,512	17,024
4. disk director to disk (mirrored)	1,600	3,200	6,400
limit (MB/s):	1,600	3,200	6,400

Table 7. TagmaStore USP raw usable bandwidth overview.

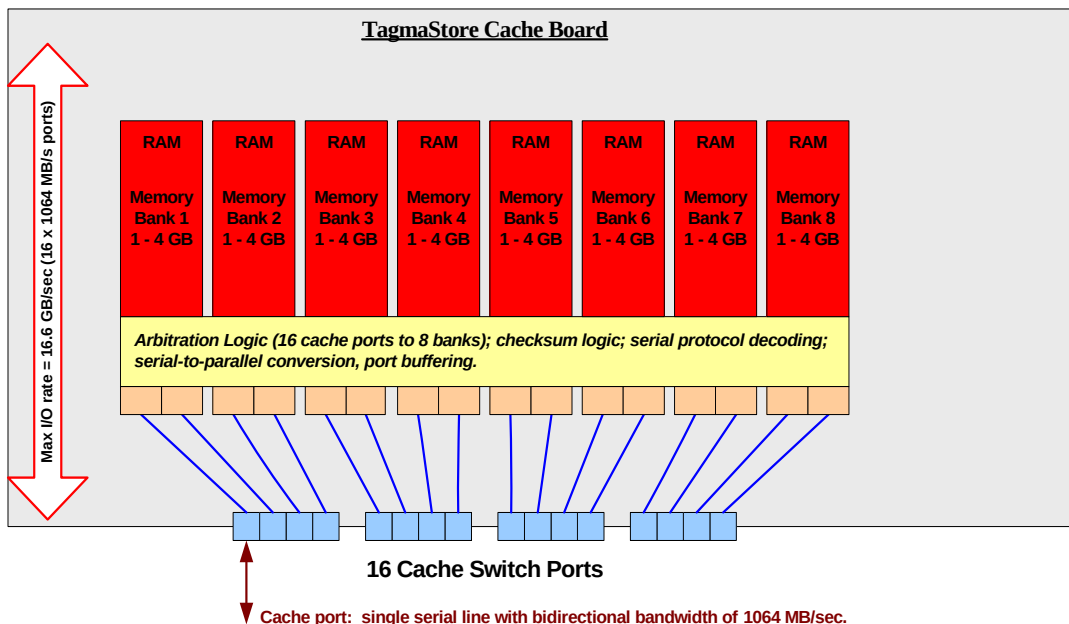
Appendix A. The USP Cache Engine

The TagmaStore USP cache design is one of triple parallelism just as in the Lightning design. There is a primary cache system (CSN - Cache Switched Network, data only), a secondary cache system (SMN - Shared Memory Network, metadata & control only), and a large local memory per Front-end board processor (MP). All of the space in the CSN data cache is strictly used for user data blocks. The metadata associated with these blocks, the system-wide control tables and block maps are all stored within the SMN Shared Memory cache. HDS software products are installed into the local RAM space local to each Front-End Director MP.

The USP cache is organized into cache slots and segments. In the case of CKD mainframe usage, the cache slot is 58KB and has one segment. For Open Systems and the Open-V LDEVs, the cache slot is 256KB with four segments (64KB each). On the Lightning product, the Open Systems cache slot is 48KB (with one segment). Multiple slots are allocated to Sequential I/O transfers, whereas individual segments from any cache slot are allocated for small block I/O requests. Therefore, the use of cache space for user data blocks is highly efficient.

The USP uses duplexing (dynamic mirroring) of write buffers (or cache lines) across two cache boards to protect write operations against the failure of a cache board. This is not dedicated cache board level mirroring, which is far simpler and wasteful.

The Lightning's cache system (for data and metadata) had an aggregate raw bandwidth of 15,936 MB/sec (10,624 + 5,312 MB/sec); in the USP this is 84,362 MB/sec (68,096 + 16,268 MB/sec). In order to actually achieve these rates, the I/O patterns would have to be evenly distributed across all of the front-end and back-end paths and all disk drives.

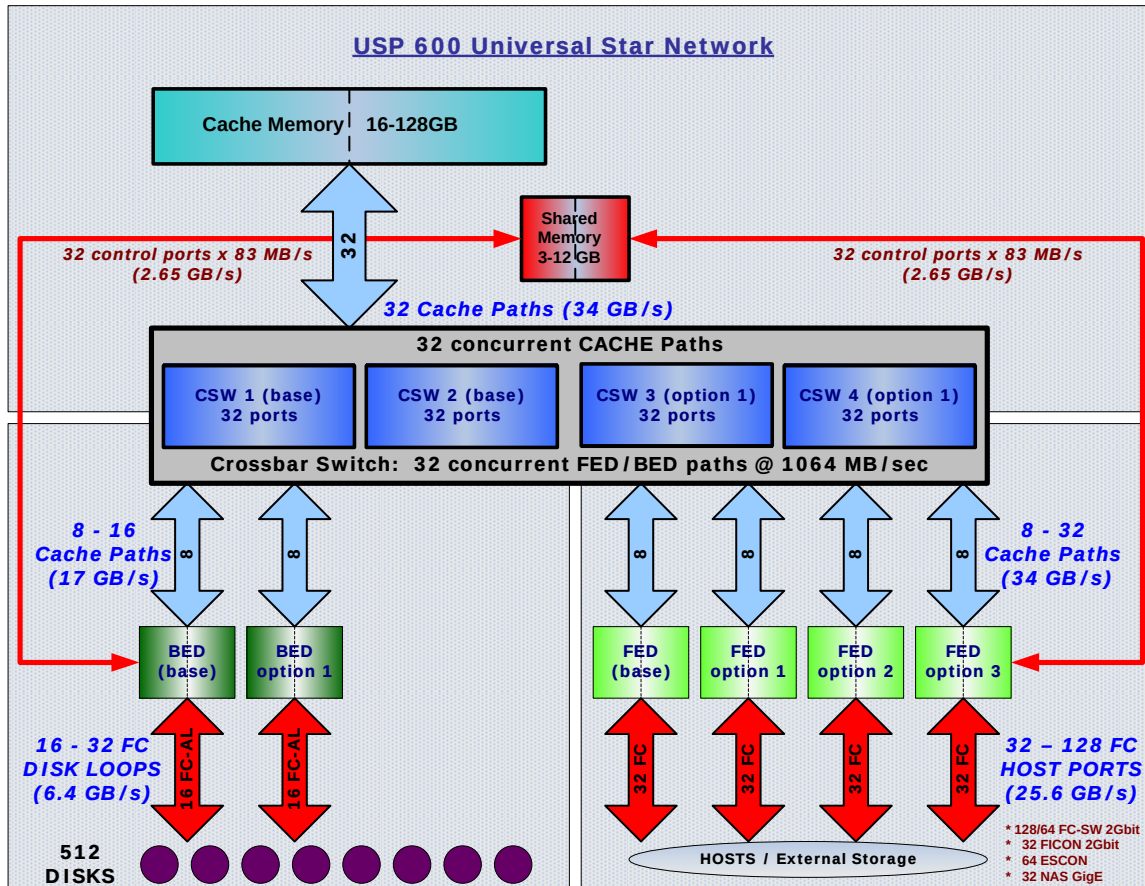


Summary of TagmaStore USP Cache board features:

- 16 DTA cache switch ports (used by data only) per board.
- 8 banks of RAM per Cache board, but all 16 DTA cache ports are kept running at full speed due to port buffering and a 2x faster DDR RAM speed as compared to port speed.
- Each DTA port is a bidirectional 1064 MB/sec serial link.
- All dynamically allocated write buffers are duplexed (mirrored) across pairs of cache boards (controlled by the FED's) for data integrity in the event of a failure.
- There can be 8, 16, or 32 GB of DDR RAM per cache board (red blocks).
- The global cache space is used only for user data blocks (no metadata, control tables, or software).
- If a cache board should abruptly fail, the system will continue to operate and no writes will be lost. All blocks to be written are duplexed across two cache boards by the Front-End Director processors, and there are no metadata or system control tables maintained in the CSN cache.

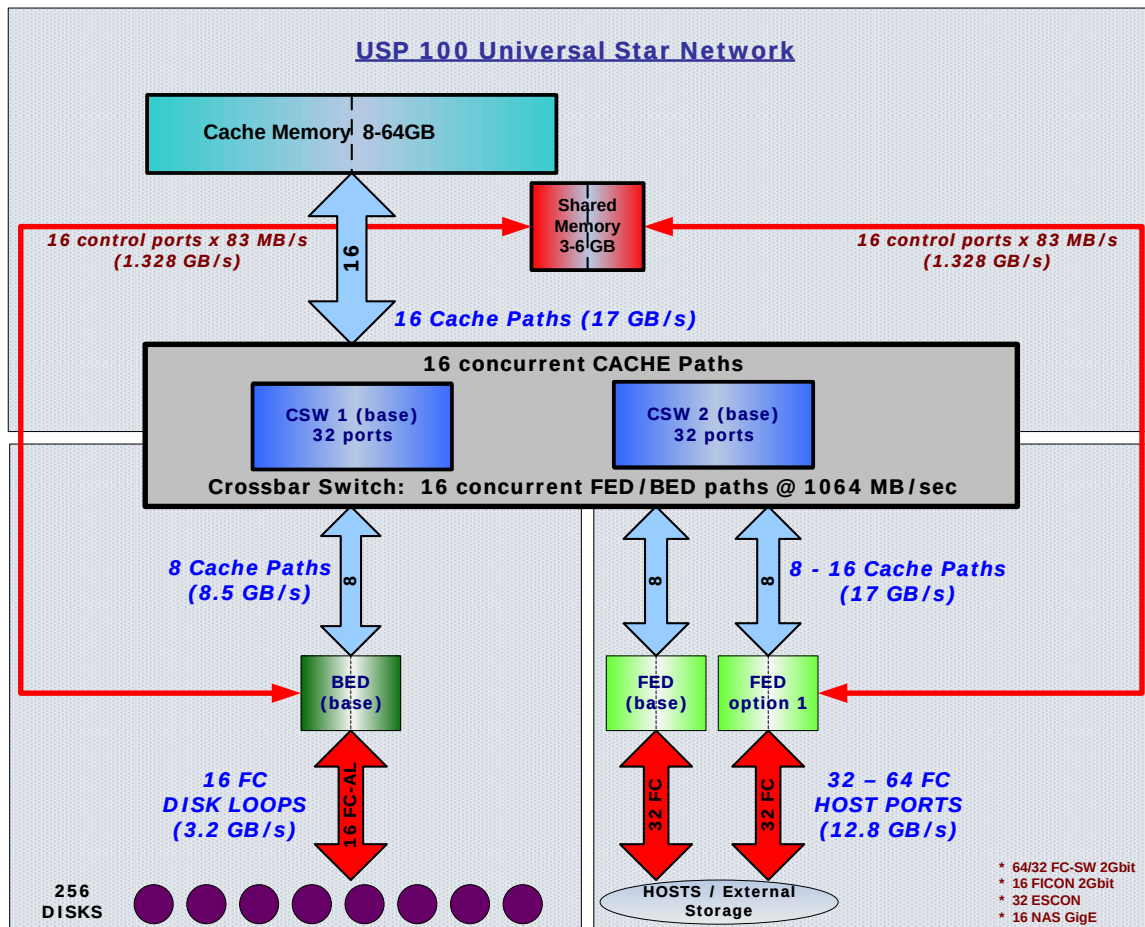
Appendix B. USP600 Architecture Overview

Shown below is an overview of the major sub-components and maximum usable raw bandwidths of a model USP600. Other combinations of BED/FED are available. If you used eight disks for global spares, then the number of RAID groups would range from 63 (8-disk types only) to 126 (4-disk types only) on a system with 512 disks. Any mix of RAID group types is allowed.

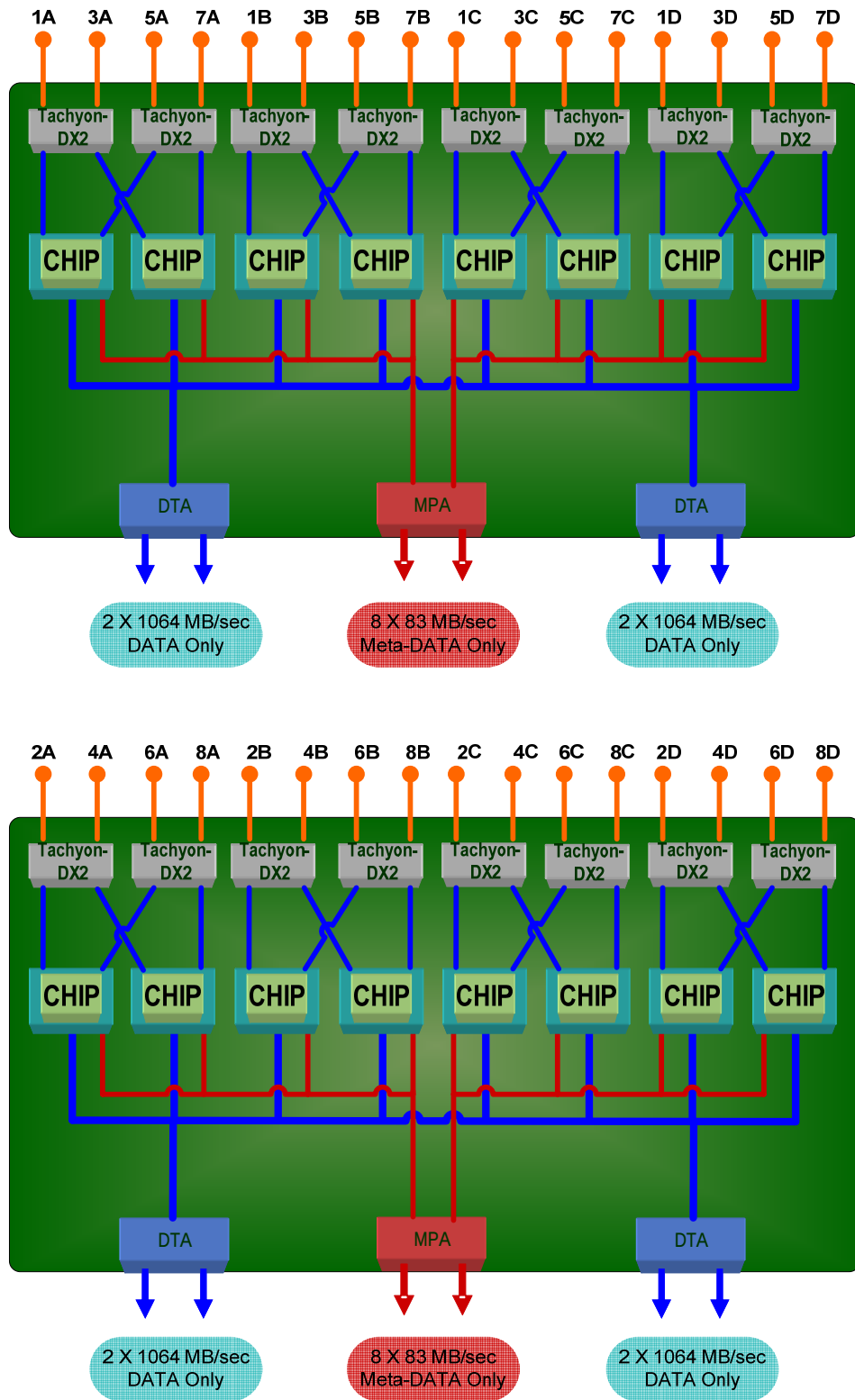


Appendix C. USP100 Architecture Overview

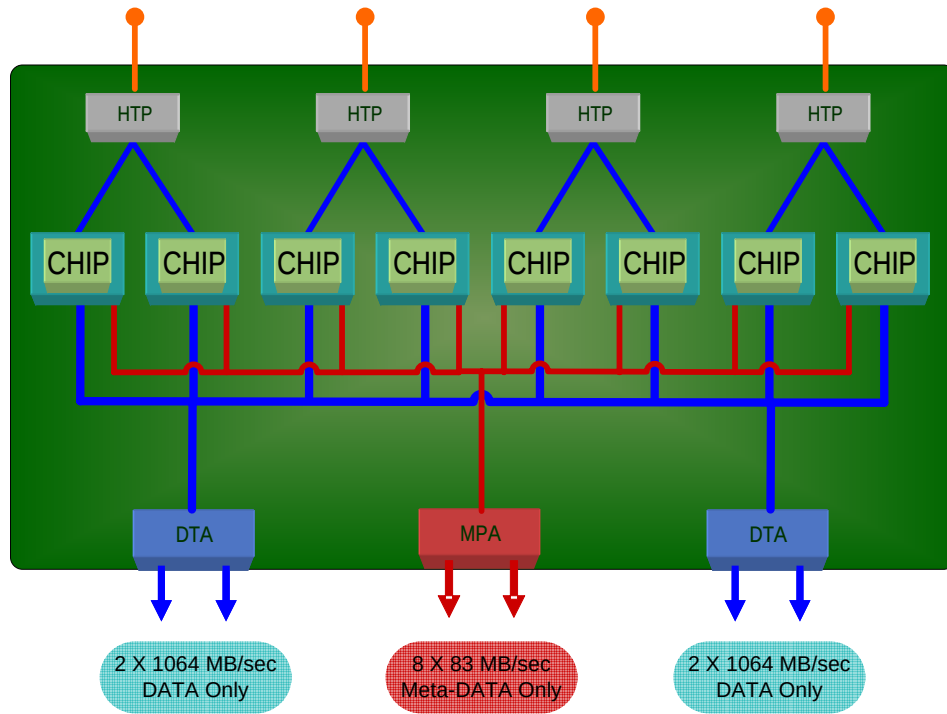
Shown below is an overview of the major sub-components and maximum useable raw bandwidths of a model USP100. If you used eight disks for global spares, then the number of RAID groups would range from 31 (8-disk types only) to 62 (4-disk types only) on a system with 256 disks. Any mix of RAID group types is allowed.



Appendix D. Open Fibre 32-port FED blades



Appendix E. FICON 4-port FED blade



Appendix F. ESCON 8-port FED blade

