May 7, 2001

XP Performance
including
Performance Advisor

A White Paper

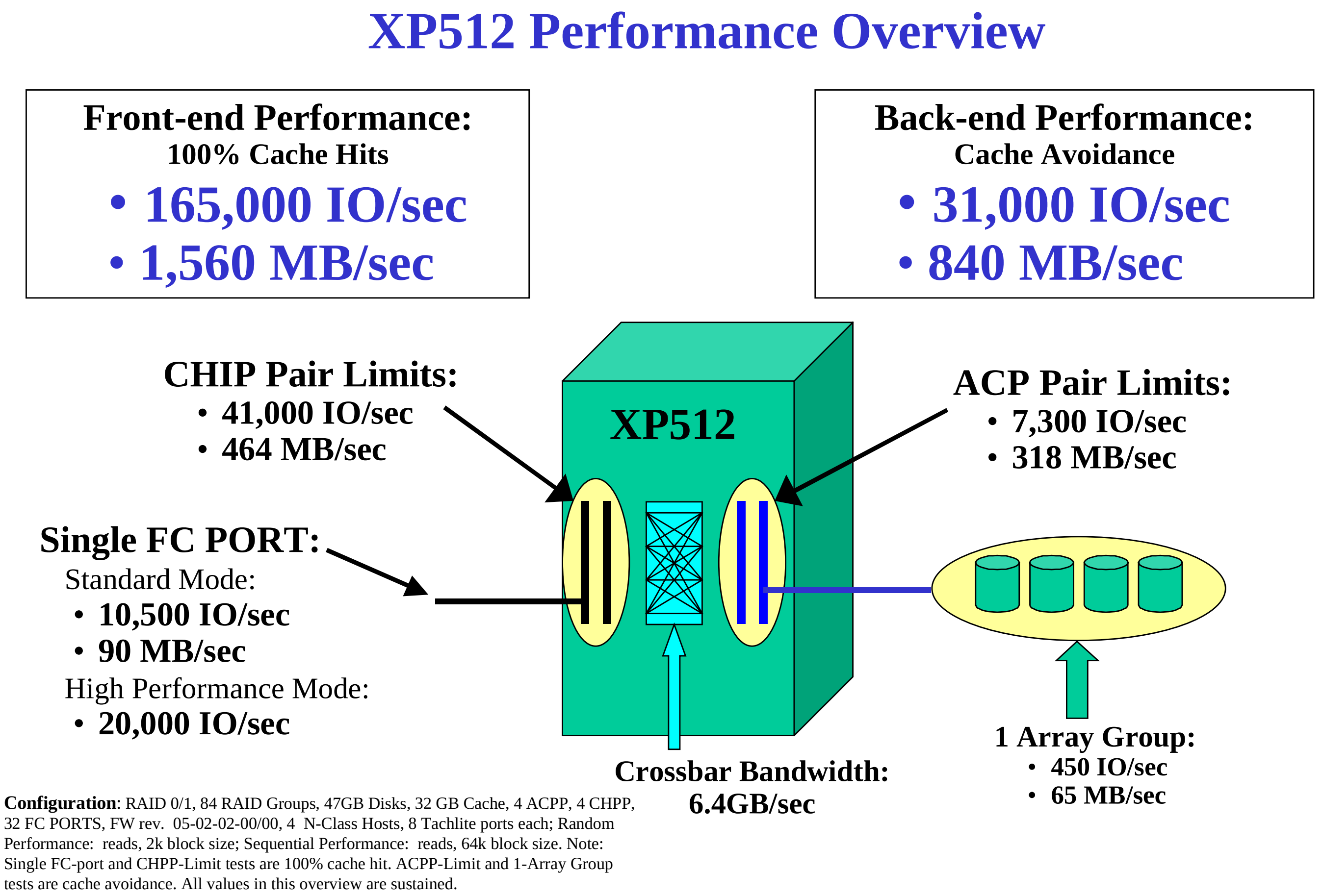
Revision 1.0

Consolidated Solutions Unit Lab
Performance Team
Roseville, CA

Hewlett-Packard

Introduction

The HP SureStore E Disk Array XP family of products provides high capacity, high speed mass storage, continuous data availability, ease of service, and great scalability and connectivity. This paper discusses the basic architecture of the XP48 and XP512 as well as the performance characteristics of the array. In addition, the Performance Advisor software application is discussed. This paper will show how Performance Advisor can be used to view the utilization of the various components of the array. Figure 1 shows maximum performance numbers measured on an XP512 array. The remainder of this paper will discuss each of these numbers in more detail, and show where to observe these numbers using Performance Advisor.



XP512 Performance Overview

Front-end Performance:
100% Cache Hits
- 165,000 IO/sec
- 1,560 MB/sec

Back-end Performance:
Cache Avoidance
- 31,000 IO/sec
- 840 MB/sec

CHIP Pair Limits:
- 41,000 IO/sec
- 464 MB/sec

ACP Pair Limits:
- 7,300 IO/sec
- 318 MB/sec

Single FC PORT:
Standard Mode:
- 10,500 IO/sec
- 90 MB/sec

High Performance Mode:
- 20,000 IO/sec

Crossbar Bandwidth:
6.4GB/sec

1 Array Group:
- 450 IO/sec
- 65 MB/sec

Configuration: RAID 0/1, 84 RAID Groups, 47GB Disks, 32 GB Cache, 4 ACPP, 4 CHPP, 32 FC PORTS, FW rev. 05-02-02-00/00, 4 N-Class Hosts, 8 Tachlite ports each; Random Performance: reads, 2k block size; Sequential Performance: reads, 64k block size. Note: Single FC-port and CHPP-Limit tests are 100% cache hit. ACPP-Limit and 1-Array Group tests are cache avoidance. All values in this overview are sustained.

Figure 1 XP512 Performance Overview

XP512 Hardware Overview

Architecture Overview

The physical components of the XP512 consist of the following major hardware components:

- The DKC (DisK Controller frame). This is the central part of the XP disk array. It contains the control panel, Fibre Channel connection hardware, service processor (SVP), and control components for the disk array. All control frame components can be repaired or replaced without interrupting access to user data. The DKC is the disk subsystem's basic unit. It also contains the user control panel to activate or disable host connect ports.

- The DKU (DisK Unit or Disk Array Frame). These cabinets contain the physical disk drives, including disk groups and the dynamic spare disk drives. The XP512 can be configured with up to six DKU's. In the XP48, there are no DKUs because disks fit into the DKC due to other component being limited in number.

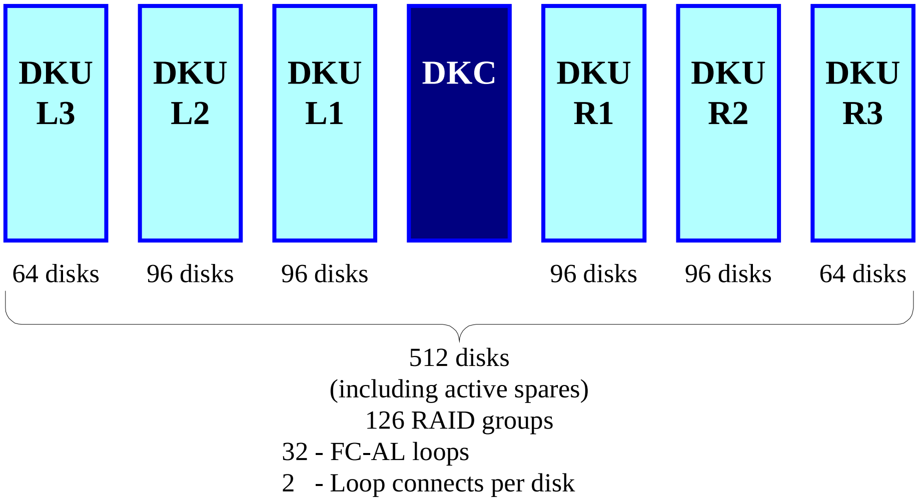


Figure 2. XP512 Block Diagram

The internal architecture of the XP contains 5 major blocks:

- The CHIP (Client Host Interface Processor) pairs that support connections from the host servers to the array.
- The ACP (Array Control Processor) pairs, which supports the physical disk access.
- The internal crossbar provides the interconnect and high bandwidth links between the CHIPs, ACPs, and cache.
- The Shared Memory and Data Cache.
- The Canisters

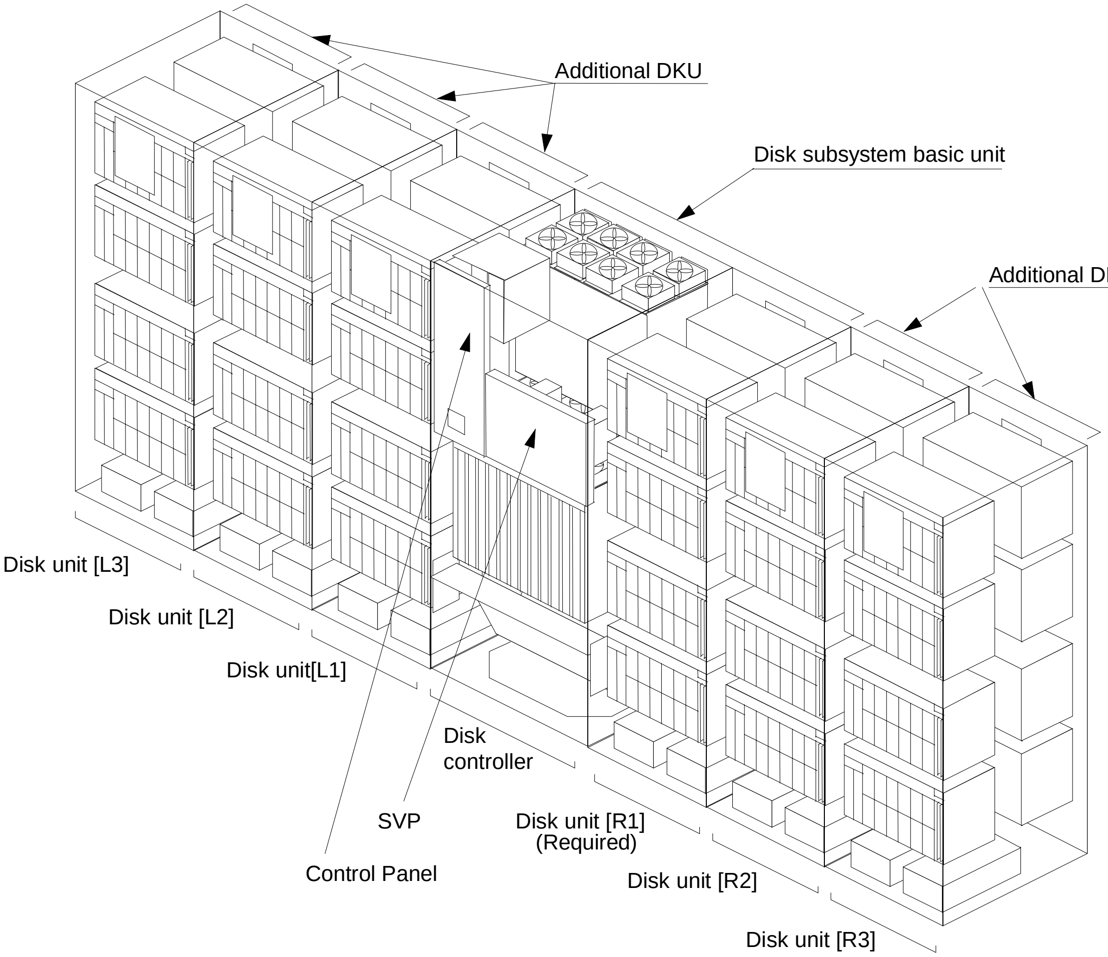


Figure 3. XP512 Internal View

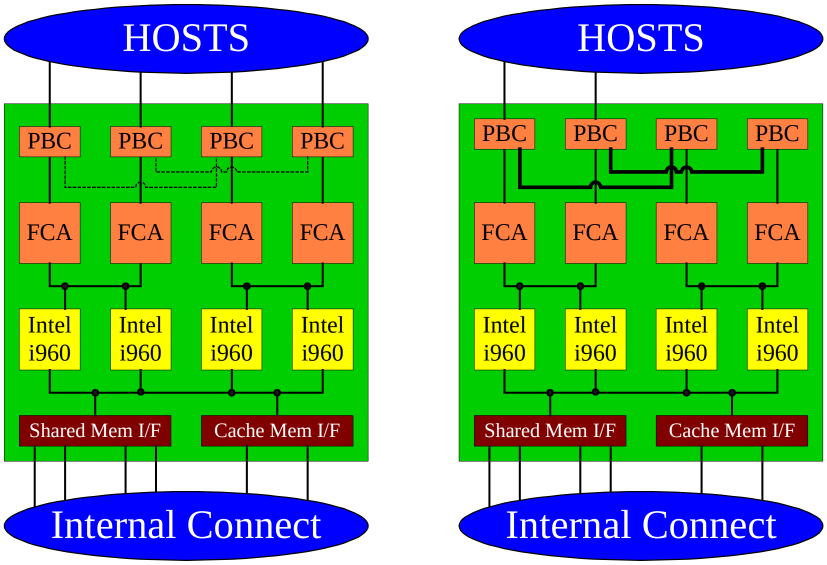
Most everything in the XP512 is provided in pairs. The hard disks are actually provided in groups of four called ‘array group’ or ‘parity group’. With all parts associated in pairs or quads, this provides the XP with no single point of failure. There are dual data and control paths all the way from the host connect to where the data is physically stored on the disk.

CHIP (Client Host Interface Processor)

The CHIPS primary function is to process host commands and signal the ACP to read/write memory to or from the disks. Additional CHIP functions are to access and update the cache track directory, monitor data access patterns, and to emulate host device types. The CHIP provides a connect point for the hosts connectivity to the array. Typically, a host will have a dual port connection to two different CHIP boards to provide multiple paths required for high availability. The CHIP boards provided in pairs are powered by independent power domains. For hosts that have multiple connections to the array in a high availability (HA) configuration: in the event of an internal power supply failure, or a CHIP board failure, the second CHIP board in the pair provides an alternate path from the host to the physically stored data.

.

Each CHIP provides four host connections except the 2 port ECSON CHIP board. The host connection goes through the FCA (Fibre Channel Adapter) interface and converts the host commands to internal commands. The FCA interface consists of a port bypass circuit (PBC) and an Agilent Tachyon TS (TacLite) Fibre Channel interface IC. The port bypass circuit is used to convert from ‘Standard Mode’ to ‘High Performance Mode’. The host connection through the FCA is controlled by it’s own i960 microprocessor that are connected in pairs. Each i960 backs the other i960 up in case of failure, and they are also used together for load sharing between two input ports in ‘Standard Mode’. The pair of i960’s automatically will split the incoming transaction into even and odd Logical Units (LUNs) to increase the throughput through the CHIP to cache memory. Each i960 processor also manages the host interface along with the point-to-point connections to shared and cache memory.



(Standard Mode)

(High Performance Mode)

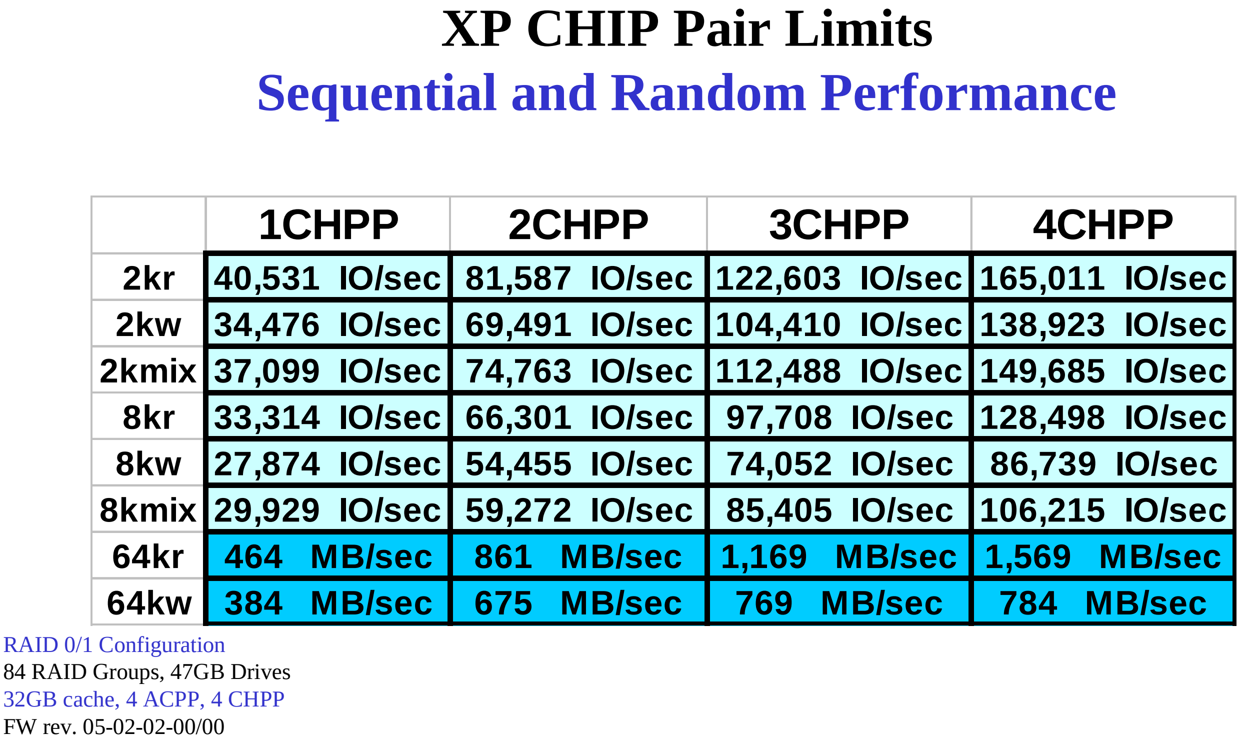
Figure 4. XP512 CHIP Block Diagram (4 port FC)

With the XP512, a total of four CHIP pairs are available (8 total boards) for host interconnect. With the XP48, a total of three CHIP pairs are available (6 boards total). These boards are always added in pairs and are ordered as pairs.

The four-port Fibre Channel CHIP board can be configured into two operating modes: either in ‘Standard Mode’ or ‘High Performance Mode’. In ‘Standard Mode’ the CHIP is logically split into two halves with two i960’s used to control two Fibre Channel ports. In “High Performance Mode’ the CHIP is operating as a single unit. In “High Performance mode’ the port bypass circuit is enabled and two of the four external Fibre Channel host connect ports are disabled. The circuitry on the CHIP then uses all four i960’s to control just two Fibre Channel host ports. This mode provides a very high throughput through the two host ports. “High Performance Mode” cuts the number of available host connect ports in a CHIP in half.

Figure 5, below, shows the performance limitations of a CHIP Pair for various workloads. Keep in mind that the XP48 is limited to three CHIP pairs. The workload is described in the first column as follows:

- 2kr = 2k random reads
- 2kw = 2k random writes
- 2k mix = 2k 60% random reads, 40% random writes
- 8k = same as 2k but with an 8k blocksize
- 64kr = 64k sequential reads
- 64kw = 64k sequential writes



XP CHIP Pair Limits

Sequential and Random Performance

	1CHPP	2CHPP	3CHPP	4CHPP
2kr	40,531 IO/sec	81,587 IO/sec	122,603 IO/sec	165,011 IO/sec
2kw	34,476 IO/sec	69,491 IO/sec	104,410 IO/sec	138,923 IO/sec
2kmix	37,099 IO/sec	74,763 IO/sec	112,488 IO/sec	149,685 IO/sec
8kr	33,314 IO/sec	66,301 IO/sec	97,708 IO/sec	128,498 IO/sec
8kw	27,874 IO/sec	54,455 IO/sec	74,052 IO/sec	86,739 IO/sec
8kmix	29,929 IO/sec	59,272 IO/sec	85,405 IO/sec	106,215 IO/sec
64kr	464 MB/sec	861 MB/sec	1,169 MB/sec	1,569 MB/sec
64kw	384 MB/sec	675 MB/sec	769 MB/sec	784 MB/sec

RAID 0/1 Configuration
84 RAID Groups, 47GB Drives
32GB cache, 4 ACPP, 4 CHPP
FW rev. 05-02-02-00/00

Figure 5. XP512 CHIP Pair Limits

The best way to look at performance of a CHIP board is on a per-processor basis. Each i960 can sustain approximately 5000 IO/sec. Note that the processors can be considered to go in “pairs.” For each pair of processors, one will handle all the odd LUNs and one will handle all the even LUNs. In “Standard Mode,” a single port will “share” two of the processors. If that port has configured half of its LUNs odd and half of them even, then that single port can sustain 10,000 IO/sec. Note that in standard mode two ports share those two processors, so the combined performance of those two ports will have a maximum of 10,000 IO/sec.

In “High Performance Mode” all four processors on a CHIP board are accessible to each port. Therefore, a single port can attain the full 20,000 IO/sec (4 processors at 5000 IO/sec each). Remember though, that this 20,000 IO/sec is split between both ports if they are both being utilized.

One final performance note, a single FC fibre is limited to approximately 90 MB/sec, and a single CHIP board to approximately 200 MB/sec.

Internal Crossbar Switches

The internal architecture of the XP48 and XP512 are very different from the XP256's shared bus based backplane. The XP256 uses a backplane architecture that has multiple CHIP's and ACP's sharing resources on the four internal buses (2 data and 2 control). The XP512 and XP48 use dedicated point-to-point and crossbar interconnect technologies for it's internal connections.

Point-to-point and crossbar backplanes are today's industry leading technologies used in high performance computer system designs. A few years ago point-to-point and crossbar technologies were only found on a few high-end, high-performance data center computing platforms like the HP 9000 high-end Enterprise Servers that require very high internal bus bandwidths to meet its performance requirements. Recent VLSI technology advances have allowed crossbar and point-to-point technologies to be put into a broader range of computing applications. Today most high-end server platforms use some sort of crossbar or point-to-point technologies to maximize on the performance throughput of data paths through their product while many mid-range server platforms continue to use shared bus based architectures. The XP family of arrays is the first storage array family to provide the advantage of point-to-point and crossbar technologies, creating a new class of high-end enterprise-class storage.

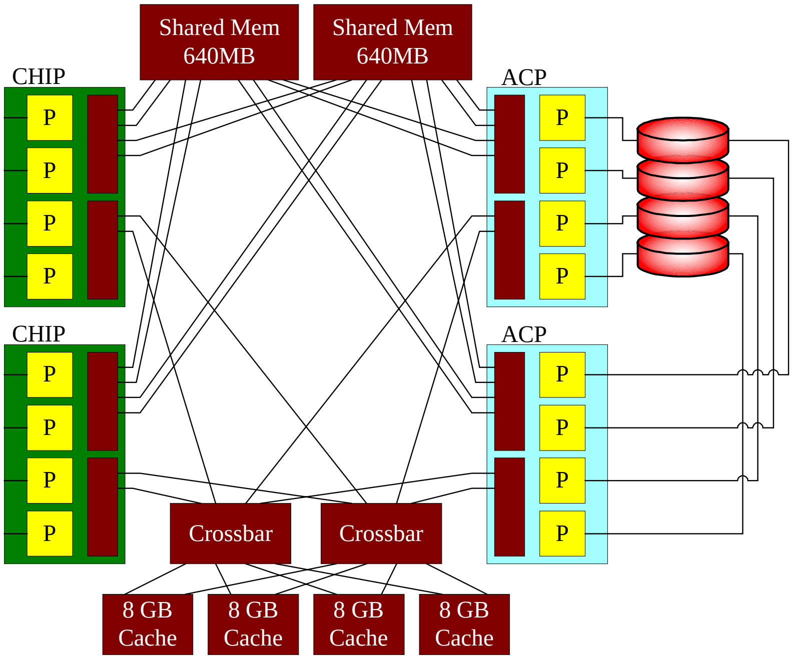


Figure 6. XP512 Internal Architecture

The XP512 and XP48 use point-to-point interconnects between both the CHIP's and ACP's to shared memory. There are four direct connect port connections between shared memory and each CHIP (and ACP) board. The XP512 can be configured with up to 64 ports to shared memory. The XP48 can be configured with up to 32 ports to shared memory. The connection to the shared memory is a point-to-point connection.

The connection from the CHIP's (and ACP's) to cache memory is a point-to-point connection that goes through a crossbar switch. Each CHIP (and ACP) has 2 ports that connect to the crossbar switch. Each crossbar switch is also on a separate power domain. The XP512 can be configured with 32 ports to cache memory (the XP48 can be configured with 16 ports). The other side of the crossbar switch connects directly to the cache memory that is grouped into four blocks (two blocks on the XP48) in two power domains. To achieve the full cache bandwidth between the crossbar switches and cache memory, all the cache boards (four of them – two in the base product, and two in the additional platform board) need to be installed. To maximize the bandwidth of cache, the cache memory modules should be distributed evenly across all cache boards installed.

The XP512 array internal backplane supports 4 CHIP pairs and 4 ACP pairs along with 4 cache expansion PCBs. The XP512 array backplane is designed actually as a mid-plane in the DKC cabinet with ACP pairs plugging into one side and the CHIP pairs plugging into the other side. The XP48 array backplane was redesigned due to the reduction in connections required from the reduced number of CHIP pairs, ACP pairs, and cache expansion PCBs supported. The number of connections into the backplane of the XP48 array is exactly ½ of the connections that the backplane on the XP512 array supports. By only supporting ½ of the expansion cards the backplane in the XP48 array is no longer required to be a mid-plane with cards plugged into both sides. This allows the control portion of the XP48 array to occupy the front half of the cabinet. The rear portion of the XP48 array cabinet contains the 4 canisters of disks, connected to the ACP pair.

Cache

The cache is one of the most critical portions of any disk array subsystem. All read or write operations move data through the cache. The cache must be robust to ensure that there are no data integrity problems, and the cache needs to be fast enough to not become a bottleneck in system performance. The XP has optimized both these requirements by providing a duplex crossbar interface to the cache. The crossbar interface to cache allows multiple simultaneous data paths in and out of cache. The duplex write feature between different power domains of the cache provides the robustness required to ensure no data loss will occur between the host and physical disk storage mechanisms.

The XP48, and XP512 architectures have separate and independent shared memory from the cache memory. Up to 1.28GB of shared memory is supported in the XP512, while 1.0GB is supported on the XP48. The shared memory is used for system configuration tables. The configuration tables are used for system components, physical to logical disk mapping of LUNs, and for identification of RAID levels for any given LUNs. The shared memory also keeps track of the cache hit and miss rates and is used to control cache pointers that allow for virtual contiguous access of the cache to either the CHIP or ACP interface. The connection between shared memory and the CHIP’s (and ACP’s) is a direct point-to-point connection. The data crossbar provides the data path between the host connection and physical disk drives through the cache. Having separate shared and cache memory allows the cache to entirely be used for data transactions to and from the physical disk drives.

Up to 32GB of cache is supported in the XP512, and 16GB in the XP48. A dynamic duplexed write cache is used for data transfers. All cache writes are duplexed between two portions of the cache that are on different power boundaries. This maximizes data integrity between the time that the data leaves the host and data is written onto a physical drive.

Four basic cache operations exist on the XP512:

- Read Hit
- Read Miss
- Fast Write
- Deferred Write

A Read-hit occurs when the data requested from the host exists in cache. The CHIP initiates a search on the cache directory in shared memory. A read hit is acknowledged and the requested data is immediately transferred to the host. The directory cache is updated to reflect the most recently used data.

A Read-miss occurs when the data requested from the host does not exist in cache. Like the read hit scenario, the first step is to initiate a search on the cache directory in shared memory. A read miss is signaled and the requested data is transferred from the disk to the read cache. The requested data is then transferred to the host. Also like the read hit, the cache directory in shared memory is updated to reflect the most recently used data. Unlike the write data, there is only one copy of the read data kept in cache. An algorithm that uses parameters of what physical disk the data was being read from determines which side of the duplex cache the read data is stored.

A fast write occurs when the cache is not full and does not need to be destaged to the disk before the write can occur. The CHIP initiates a search on the cache directory in shared memory to identify if an old copy of the data to be written is still in cache and if the cache space is available. The data is transferred from the host to the cache and duplexed to both Cache A and Cache B. The cache directory in shared memory is modified to reflect the most recently used data. The host is notified of an I/O completion. The data in cache is destaged to the disk in the background. The reason for writing the data to both cache areas is that data could potentially be lost if a cache error occurs before the data has been written to the physical disk, and there was only a single copy of the data.

A deferred write occurs if the duplex write cache is at the write limits and cannot accept the new data before destaging a cache block to the disk. In this situation the CHIP initiates a search on the cache directory in shared memory and identifies that the cache is full. The least recently used data is identified and destaged to disk. After the least recently used data is destaged, the data is transferred from the host to the cache and duplexed to both Cache A and Cache B. The cache directory is updated to reflect the most recently used data, and the host is notified of the I/O completion. The data in the cache is destaged to the disk in the background.

The XP adds data integrity codes to the host data at various points in the path from the host I/F to the physical disk. These data integrity codes are appended by hardware to allow maximum data transfer rates through the subsystem interfaces. The parity bits and integrity codes are analyzed, if the analysis identifies an error, then a cache error has occurred. The failing data in cache is identified as bad and the cache page becomes non-operative. After the failing cache page becomes non-operative the XP's Continuous Track XP "Phone Home" capability is used to alert an HP field service representative of the failure, and a notification is sent to the Command View XP remote console.

ACP (Array Control Processor)

The ACP functions are to deal with read/writes to disk, read miss staging and write de-staging from the cache. The ACP's also handle the task of media protection. Media protection is done by dynamic spares, mirrored storage (RAID 0/1), dynamic data rebuild, and hardware RAID 5 parity generation.

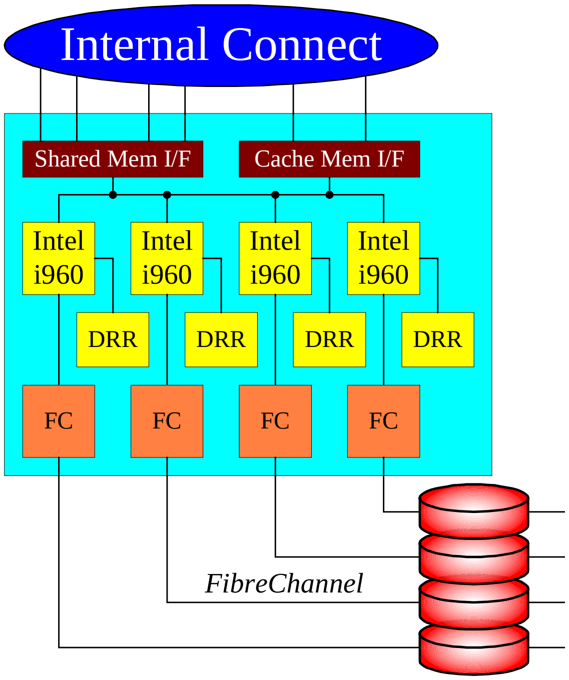


Figure 7. XP512 ACP Block Diagram

To accomplish the task of hardware parity generation and controlling the data flow to the physical drives, the ACP has four I960 controllers each linked to a DRR (Data Recovery and Restore) parity generator chip. The I960 controllers communicate with the shared and cache memory interface, and the Agilent Tachyon TL (TacLite) Fibre Channel I/F chips. The DRR's are the hardware RAID 5 parity generators that do the work of generating and checking the parity for RAID 5 groups, and duplicating data copies for mirrored storage (RAID 0/1). Each Fibre Channel interface has a dedicated I960 and parity circuit.

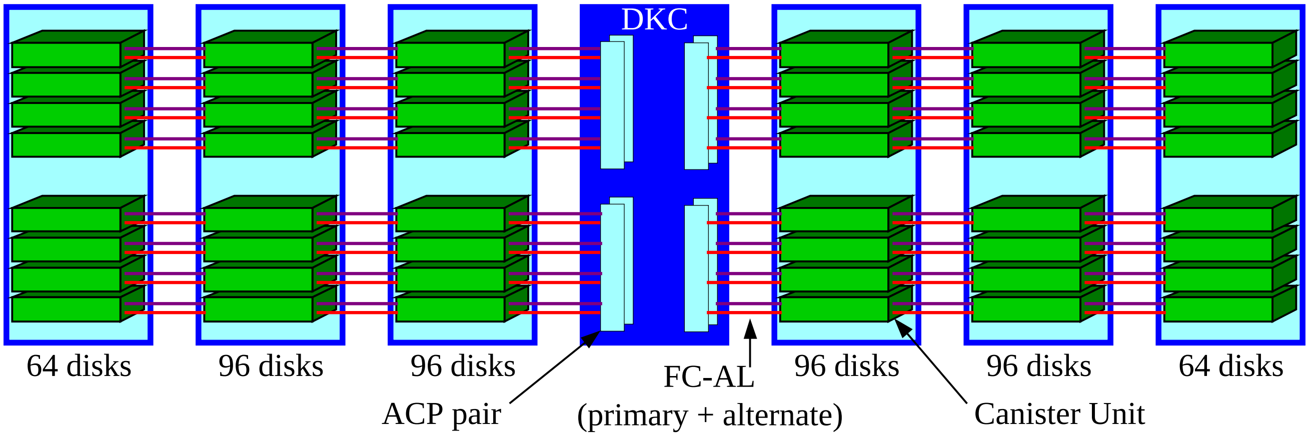


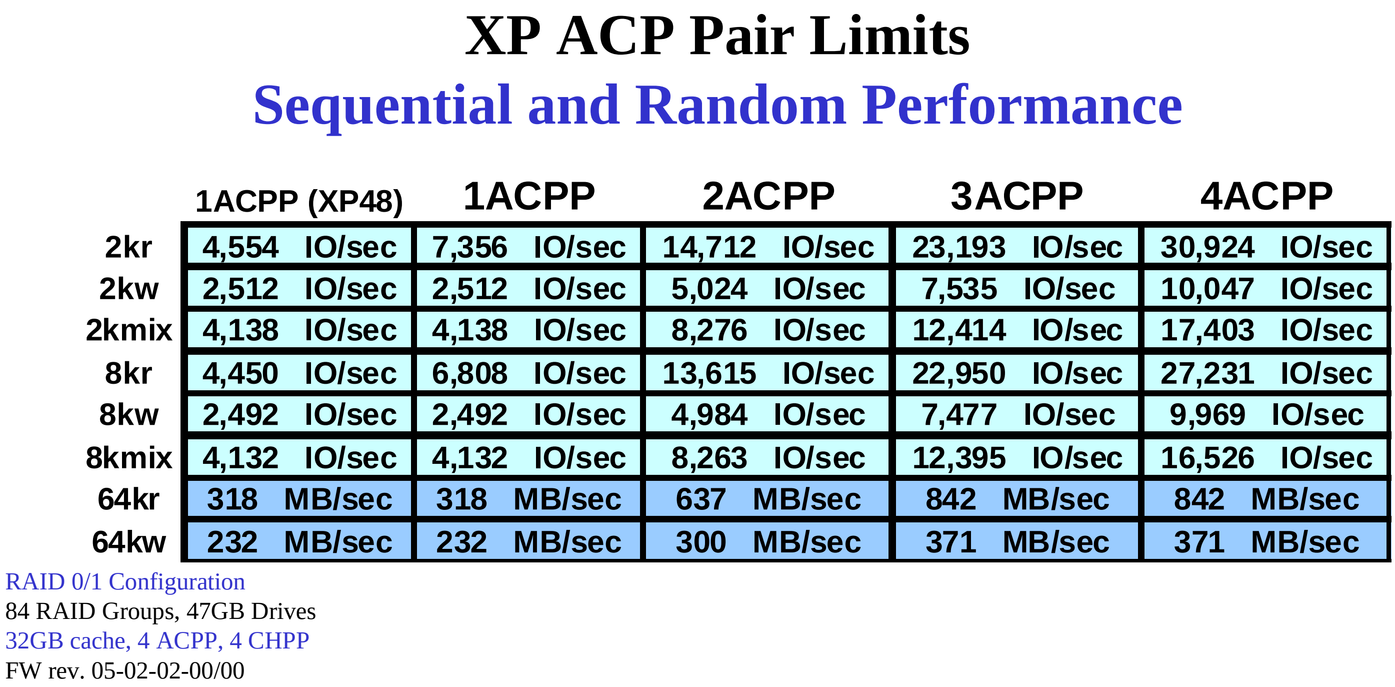
Figure 8. XP512 ACP connections to DKU

A minimum XP512 system requires a single ACP pair. The two ACP boards reside on separate power domains similar to the CHIP pairs, and cache memory. In the event of an ACP failure the second ACP board will take over all transactions to the physical disks. During normal operation both ACP's are used to load balance the data transfer between the cache and the physical disks. Each ACP pair can control up to 128 physical disks including both data and dynamic spare disks on eight Fibre Channel arbitrated loops. To get 128 physical disks on a single ACP pair, the disks need to be populated in three DKU's on a single side of the DKC. In a full XP512 system configuration, up to four ACP pairs can be installed into the DKC.

The XP48 supports only a single ACP pair. The single ACP pair in the XP48 has 8 FC-AL loops that connect to the disk drives in the canisters. The XP48 is required to have 4 canisters to attach all of the 8 FC-AL loops. The ACP pair in the XP48 is the same assembly that is used in the XP512 array.

Figure 9, below, shows the performance limitations of an ACP Pair for various workloads. Note the slight difference for the XP48 due to its ACP Pair only being able to hold 48 disks. The workload is described in the first column as follows:

- 2kr = 2k random reads
- 2kw = 2k random writes
- 2k mix = 2k 60% random reads, 40% random writes
- 8k = same as 2k but with an 8k blocksize
- 64kr = 64k sequential reads
- 64kw = 64k sequential writes



XP ACP Pair Limits

Sequential and Random Performance

	1ACPP (XP48)	1ACPP	2ACPP	3ACPP	4ACPP
2kr	4,554 IO/sec	7,356 IO/sec	14,712 IO/sec	23,193 IO/sec	30,924 IO/sec
2kw	2,512 IO/sec	2,512 IO/sec	5,024 IO/sec	7,535 IO/sec	10,047 IO/sec
2kmix	4,138 IO/sec	4,138 IO/sec	8,276 IO/sec	12,414 IO/sec	17,403 IO/sec
8kr	4,450 IO/sec	6,808 IO/sec	13,615 IO/sec	22,950 IO/sec	27,231 IO/sec
8kw	2,492 IO/sec	2,492 IO/sec	4,984 IO/sec	7,477 IO/sec	9,969 IO/sec
8kmix	4,132 IO/sec	4,132 IO/sec	8,263 IO/sec	12,395 IO/sec	16,526 IO/sec
64kr	318 MB/sec	318 MB/sec	637 MB/sec	842 MB/sec	842 MB/sec
64kw	232 MB/sec	232 MB/sec	300 MB/sec	371 MB/sec	371 MB/sec

RAID 0/1 Configuration
84 RAID Groups, 47GB Drives
32GB cache, 4 ACPP, 4 CHPP
FW rev. 05-02-02-00/00

Figure 9. XP ACP Pair Limits

Canisters

The canisters are where the physical disk drives mount. A single canister can hold up to 12 physical disks. The disks in a single canister are connected to a separate Fibre Channel arbitrated loop to both the primary and secondary ACP channels. Four canisters within the same DKU are required to configure an array group. A single array group (or parity group) requires four disks; one disk is located in each of the four canisters. A total of eight canisters are installed in a DKU cabinet, and up to 24 array groups can be added into a single cabinet. The two lower canisters on both the front and rear of the DKU cabinet are a part of a single ACP pair. Likewise the two upper canisters on the front and back of the DKU cabinet are also part of a single ACP pair.

The DKU units can be combined on a single side of the DKC cabinet. A single ACP pair with three DKU cabinets will have three canisters connected on a single FC-AL loop pair (Primary + Alternate). The canisters in either the R3 or L3 DKU unit will only allow eight physical disks to be connected in the canister. A total of 32 disks in three canisters can be connected on a single FC-AL loop pair. At this time there are hardware limitations in the XP512 that limit the maximum number of disks on the ACP's FC-AL loop to 32.

The XP48 array supports just 4 canisters with no additional expansion capabilities.

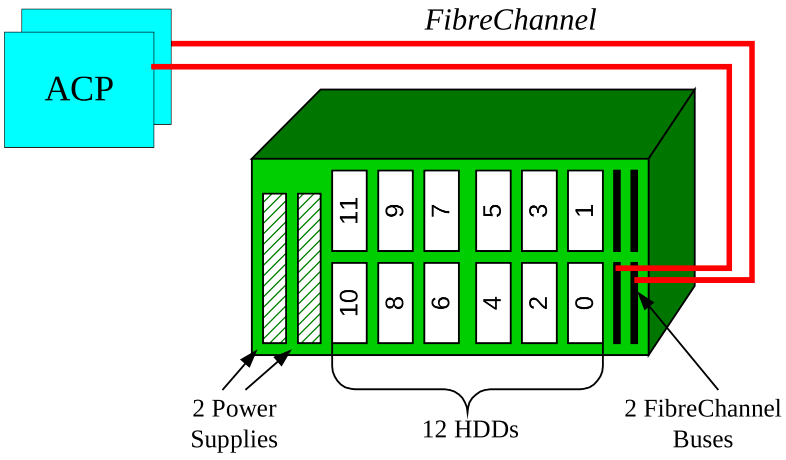


Figure 10. XP512 Canister Block Diagram

Disks Array Groups

The XP512 basic disk configuration is know as the ‘array group’ or ‘parity group’. An array group depends on the RAID level that the user plans on implementing. The objectives of the RAID technology are low cost, high reliability, and high I/O performance of disk storage devices. To achieve these objectives the XP512 disk array supports two different modes of RAID control. These are RAID 0/1, and RAID 5. All RAID groups in the XP48 and XP512 require four physical disks.

Which RAID level the user would want depends on the type of data storage solution they require. In most situations, RAID 0/1 will have better performance than RAID 5, but RAID 5 will provide more useable capacity. If the host application is read intensive there should not be much, if any, performance difference between either of the two RAID levels. If the host application is more write intensive, this is where the user will see some performance differences between the two RAID levels. The downside to RAID 0/1 over RAID 5 is that there is a 50% overhead on storage redundancy associated with RAID 0/1. RAID 0/1 is often referred to as mirrored storage because the host data is physically duplicated in the array. The user will get more overall useable storage out of the array with a RAID 5 implementation. RAID 5 achieves the storage redundancy by generating parity data based on the data stored. RAID 5 only requires a 25% storage overhead.

RAID Level	Disk Usage	RAID HA Overhead
RAID 0/1	2D + 2D	50%
RAID 5	3D + 1P	25%

Table 1. RAID Array Usage and Overhead

RAID 0/1 (2D+2D)

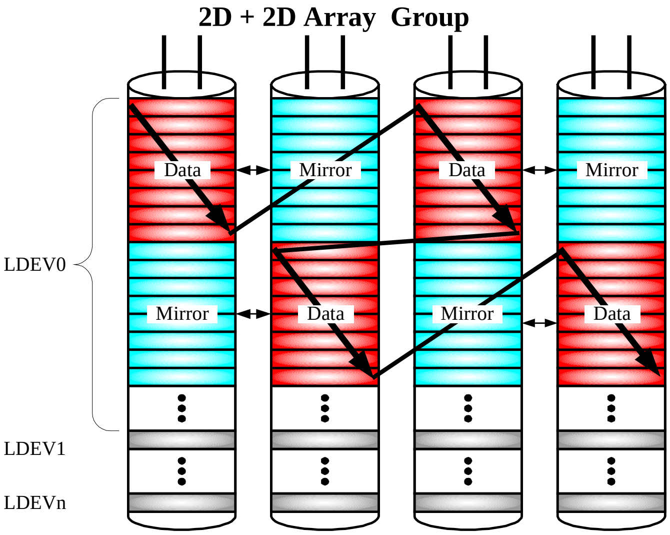
RAID 0/1 uses both RAID 0 and RAID 1 technologies implemented together to achieve high reliability and high I/O performance.

RAID 0 produces a striped drive volume. Striped data means that the stream of data from the host is split and distributed onto two or more disk devices on a block basis. RAID 0 produces a very high performance I/O disk subsystem where fault tolerance is not required.

RAID 1 uses a disk-mirroring algorithm that requires at least two disk drives. RAID 1 produces a mirrored drive volume. RAID 1 is the simplest of RAID technologies because of the nature of mirrored storage. The RAID controller writes the data to both duplex areas of cache. After the first write to the primary disk is completed, the secondary copy of data in cache is written out to the secondary disk. After the writes to both disks have completed the write cache is released. The RAID 1 is fault tolerant, but it has lower performance than RAID 0.

As mentioned earlier RAID 0/1 adds high I/O performance features of RAID0 striping to the high reliability features of mirrored storage with RAID 1. RAID 0/1 uses all four disks in an array group for the primary storage path. The I/O performance improvements are achieved from the fact that the I/O data streams are split onto all four disks in the array group. From the diagram below one can see how the primary data is physically written across all four drives. The diagram also shows how the secondary portion of the storage array is used for the mirrored copy.

Writes to the array in RAID 0/1 mode are done the same way explained above for RAID 1. The primary side is destaged first. After the primary data has been successfully destaged, the secondary data is destaged to the mirror drive mechanisms.



RAID 0/1 is a striped and mirrored copy within a ACP pair

Figure 11. XP512 RAID 0/1

RAID 5

RAID 5 requires four disks for an array group. RAID 5 on the XP512 allows the data and parity to be split and distributed onto four disk drives using striping. Parity data for the group is created and stored on the parity stripe. Data can easily be recovered if a device in the parity group becomes inoperative or causes a read error.

In RAID 5 the striping size is set to that of a multiple of the block size that is transferred from cache. This allows the RAID controller to access each disk for a single stripe equivalence of data and allows the RAID controller to perform I/O operations on other disks in parallel, therefore increasing I/O performance substantially. In small scale or random I/O applications the data transfer rate remains the same as conventional RAID systems. In large or sequential I/O applications, RAID 5 permits the blocks in the same parity group to be processed in parallel, resulting in an increase in the data transfer rate. For small writes to single blocks, RAID 5 requires extra reads from the data and parity disks before a small block can be written to the disk. Since the parity data is distributed on all disks in the group, it still allows parallel I/O processing of multiple blocks.

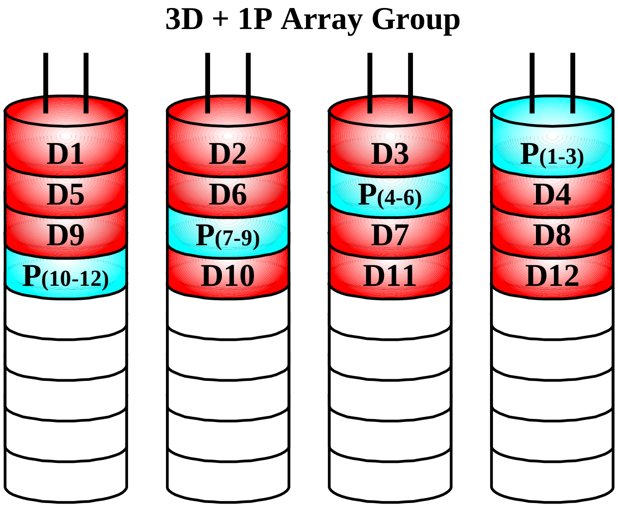
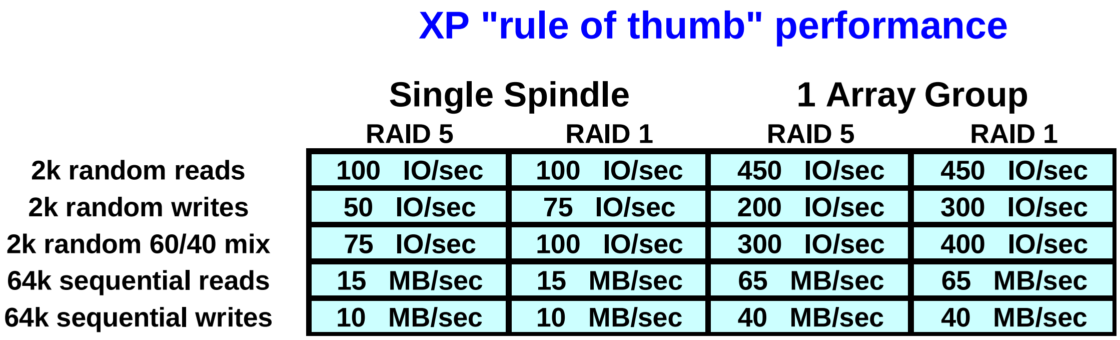


Figure 12. XP512 RAID 5

Array Group Performance

XP single array group performance is summarized in the following, Figure 13. Included in the table are guidelines for single-spindle performance as well.



XP "rule of thumb" performance

	Single Spindle		1 Array Group	
	RAID 5	RAID 1	RAID 5	RAID 1
2k random reads	100 IO/sec	100 IO/sec	450 IO/sec	450 IO/sec
2k random writes	50 IO/sec	75 IO/sec	200 IO/sec	300 IO/sec
2k random 60/40 mix	75 IO/sec	100 IO/sec	300 IO/sec	400 IO/sec
64k sequential reads	15 MB/sec	15 MB/sec	65 MB/sec	65 MB/sec
64k sequential writes	10 MB/sec	10 MB/sec	40 MB/sec	40 MB/sec

Figure 13. XP Spindle and Array Group Performance

LUNs and LDEVs

Without getting into specific details about the software tools that surround configuring LUNs and Logical Devices (LDEVs), this section will explain the use of terms for LUNs, LDEVs and open volumes.

The LUN Configuration Manager XP is the software tool that enables the administrator to port/address pairings (LUNS) to raw disk space of open volumes represented by LDEVs on the array. LUNs are what the host server's see as physical storage space.

A user first needs to determine what RAID level of storage that they want to set up along with what types of open volumes need to be set up. An array group is divided into open volumes of all the same size. Open volumes have a fixed size depending on their type. This size is customizable using the LUSE (Logical Unit Size Expansion) tool or CVS (Custom Volume Size). These tools are part of the LUN Configuration Manager XP product. Once the open volumes (LDEVs) have been created, the LUSE tool will allow you to combine LDEVs together to create a single larger volume, which can then be "mapped" to a host port as a single LUN. It is possible to combine up to 36 LDEVs into a single LUN with LUSE. The CVS tool allows the user to configure custom size volumes (CVs) that are smaller then normal volumes. Only one type of open volume emulation type can be defined for the entire array group.

LDEVs (Logical DEVices) are created out of physical disks. The LDEV provides the mapping of a physical open volume in a specific array group. The process of creating a LDEV out of array groups is managed by CUs (Control Units). A CU is nothing more than an internal table holding the backend configuration of the array groups. A CU can only support a single RAID level. There are 16 CU's in one XP512, each of them supporting a maximum of 256 devices.

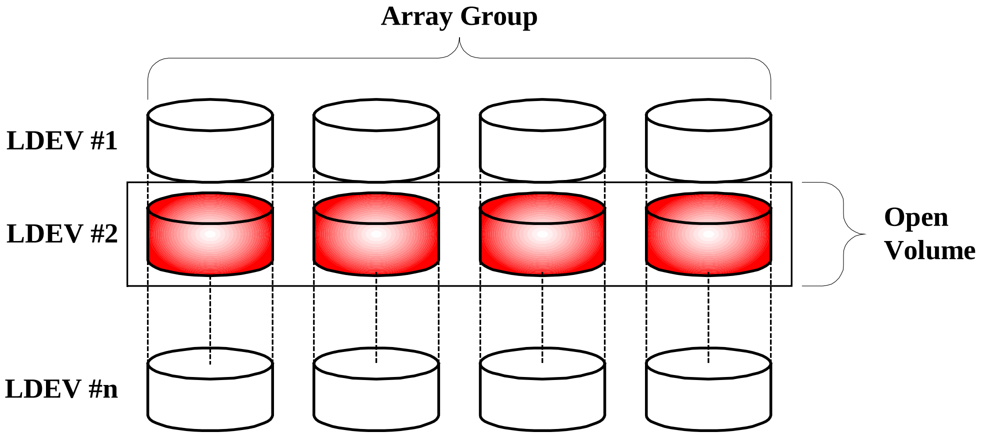


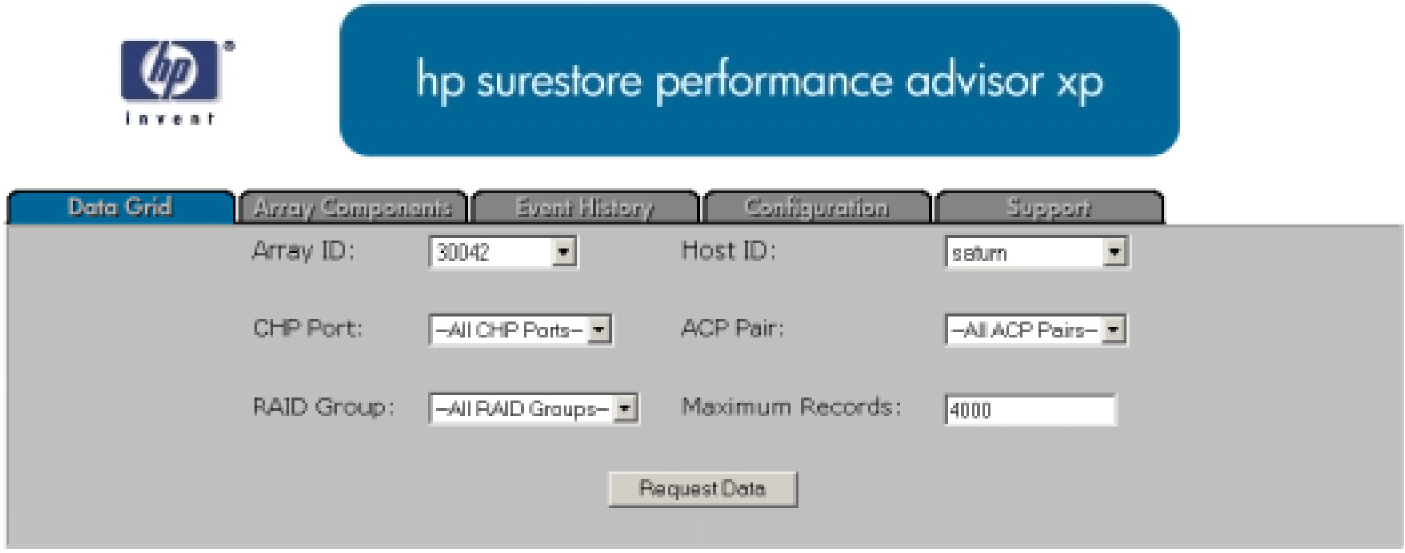
Figure 14. Open volumes in an array group

LUNs are the logical addresses that correlate CHIP port/address pairings to back-end logical devices (LDEVs). Remember for a HA configuration that the XP can have a dual path through the CHIP's all the way to the physical disk. LUN Configuration Manager XP is where two different ports on the CHIP's can be mapped to a common LDEV in the array. This mapping provides the failover link to the common storage space.

Performance Advisor

Home Page

The HP Surestore Performance Advisor XP product gives the user visibility into the performance and utilization of various XP array components. These are the same components discussed in the earlier sections of this paper. Let’s begin with the *Performance Advisor Home Page* screen, Figure 15.



Array ID: 30042
Host ID: saturn
CHP Port: --All CHP Ports--
ACP Pair: --All ACP Pairs--
RAID Group: --All RAID Groups--
Maximum Records: 4000

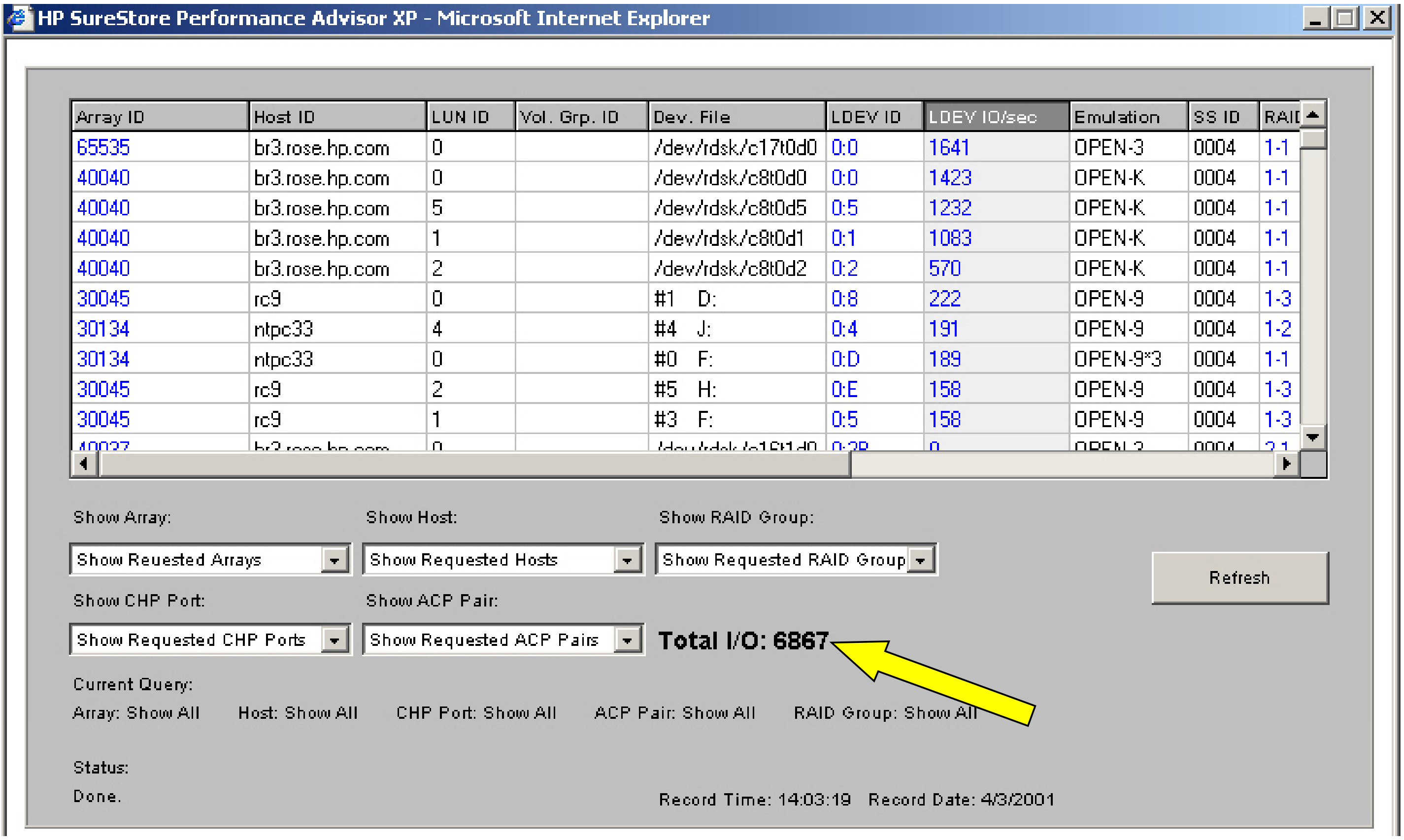
Figure 15. Performance Advisor – Home Page

As you can see, from this screen you can choose to request data based on array, CHIP Ports, Array Groups, Host ID, or ACP Pairs. Once the data is requested, you then get the *Data Grid* screen, Figure 16.

Data Grid

The *Data Grid* screen displays not only the I/O activity seen on each ldev, but also the Total I/O activity for all the components selected (including the entire array if *all* components are selected. The yellow arrow on Figure 16 shows where the “Total I/O” is displayed. To aid in determining which ldevs have the most I/O activity, you can sort the display by clicking on the “LDEV IO/sec” label in the top row.

The “Total I/O” value can be useful if you limit the selection of components. For example, if you were to look at only a single ACP Pair, you can then compare to the ACP Pair Limit numbers in Figure 9.



HP SureStore Performance Advisor XP - Microsoft Internet Explorer

Array ID	Host ID	LUN ID	Vol. Grp. ID	Dev. File	LDEV ID	LDEV IO/sec	Emulation	SS ID	RAI
65535	br3.rose.hp.com	0		/dev/rdsk/c17t0d0	0:0	1641	OPEN-3	0004	1-1
40040	br3.rose.hp.com	0		/dev/rdsk/c8t0d0	0:0	1423	OPEN-K	0004	1-1
40040	br3.rose.hp.com	5		/dev/rdsk/c8t0d5	0:5	1232	OPEN-K	0004	1-1
40040	br3.rose.hp.com	1		/dev/rdsk/c8t0d1	0:1	1083	OPEN-K	0004	1-1
40040	br3.rose.hp.com	2		/dev/rdsk/c8t0d2	0:2	570	OPEN-K	0004	1-1
30045	rc9	0		#1 D:	0:8	222	OPEN-9	0004	1-3
30134	ntpc33	4		#4 J:	0:4	191	OPEN-9	0004	1-2
30134	ntpc33	0		#0 F:	0:D	189	OPEN-9*3	0004	1-1
30045	rc9	2		#5 H:	0:E	158	OPEN-9	0004	1-3
30045	rc9	1		#3 F:	0:5	158	OPEN-9	0004	1-3
40037	br3.rose.hp.com	0		/dev/rdsk/c16t1d0	0:2B	0	OPEN-3	0004	2-1

Show Array: Show Reuested Arrays

Show Host: Show Requested Hosts

Show RAID Group: Show Requested RAID Group

Show CHP Port: Show Requested CHP Ports

Show ACP Pair: Show Requested ACP Pairs

Total I/O: 6867

Current Query:

Array: Show All Host: Show All CHP Port: Show All ACP Pair: Show All RAID Group: Show All

Status:

Done.

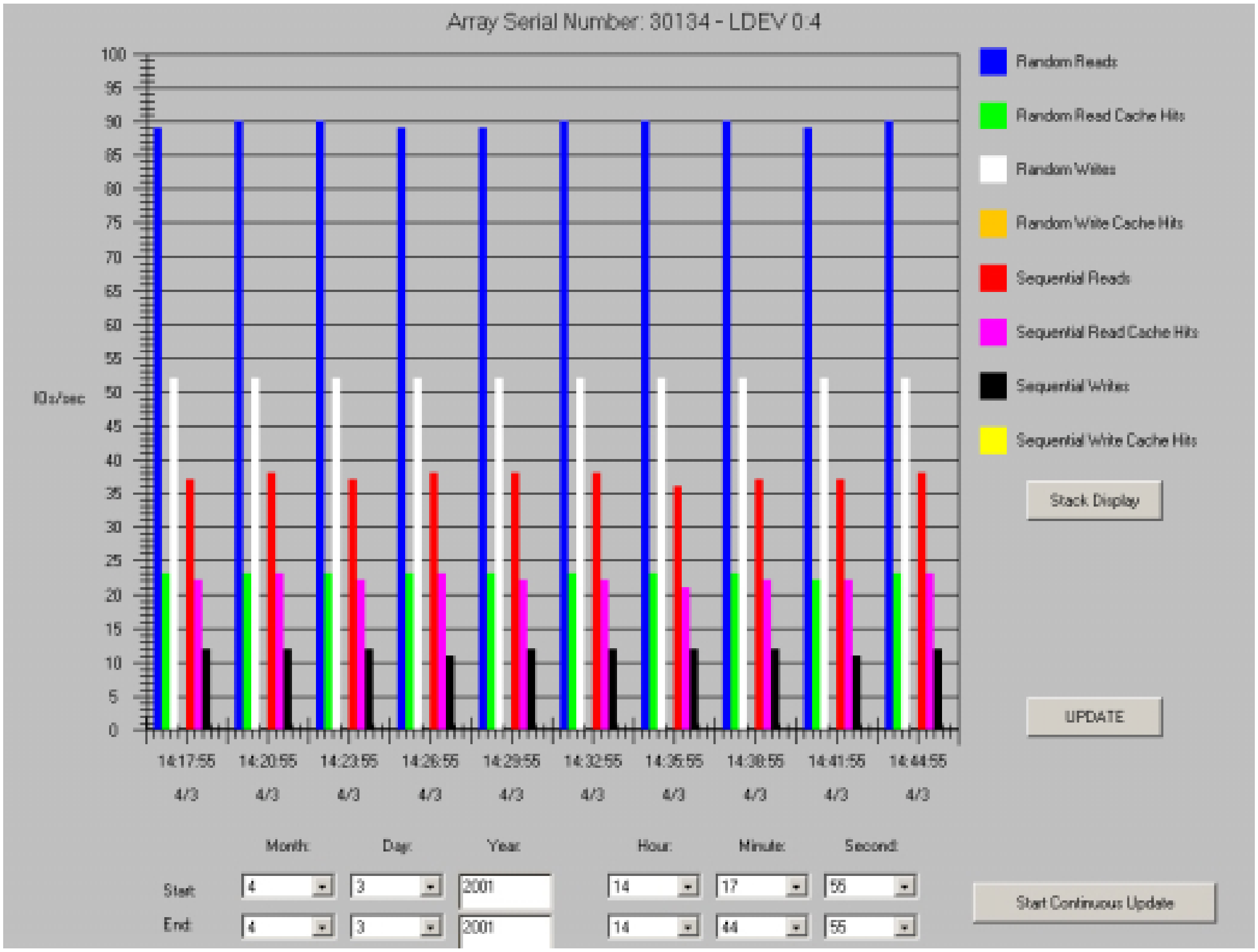
Record Time: 14:03:19 Record Date: 4/3/2001

Figure 16. Performance Advisor – Data Grid

LDEV History Graphs

From the *Data Grid* screen you can delve deeper into the performance characteristics of a single ldev by simply clicking on a particular "LDEV ID." This gives you a screen as shown in Figure 17. The different column labels tell you what type of workload the array has recognized for this particular ldev at the times noted on the x-axis.

Note that the "Stack Display" tab (which you will see on many of the PA screens) allows you to display the average of all the bars on a single bar. The display in the figures shown is the "Parallel Display" which separates out each individual component into its own bar.



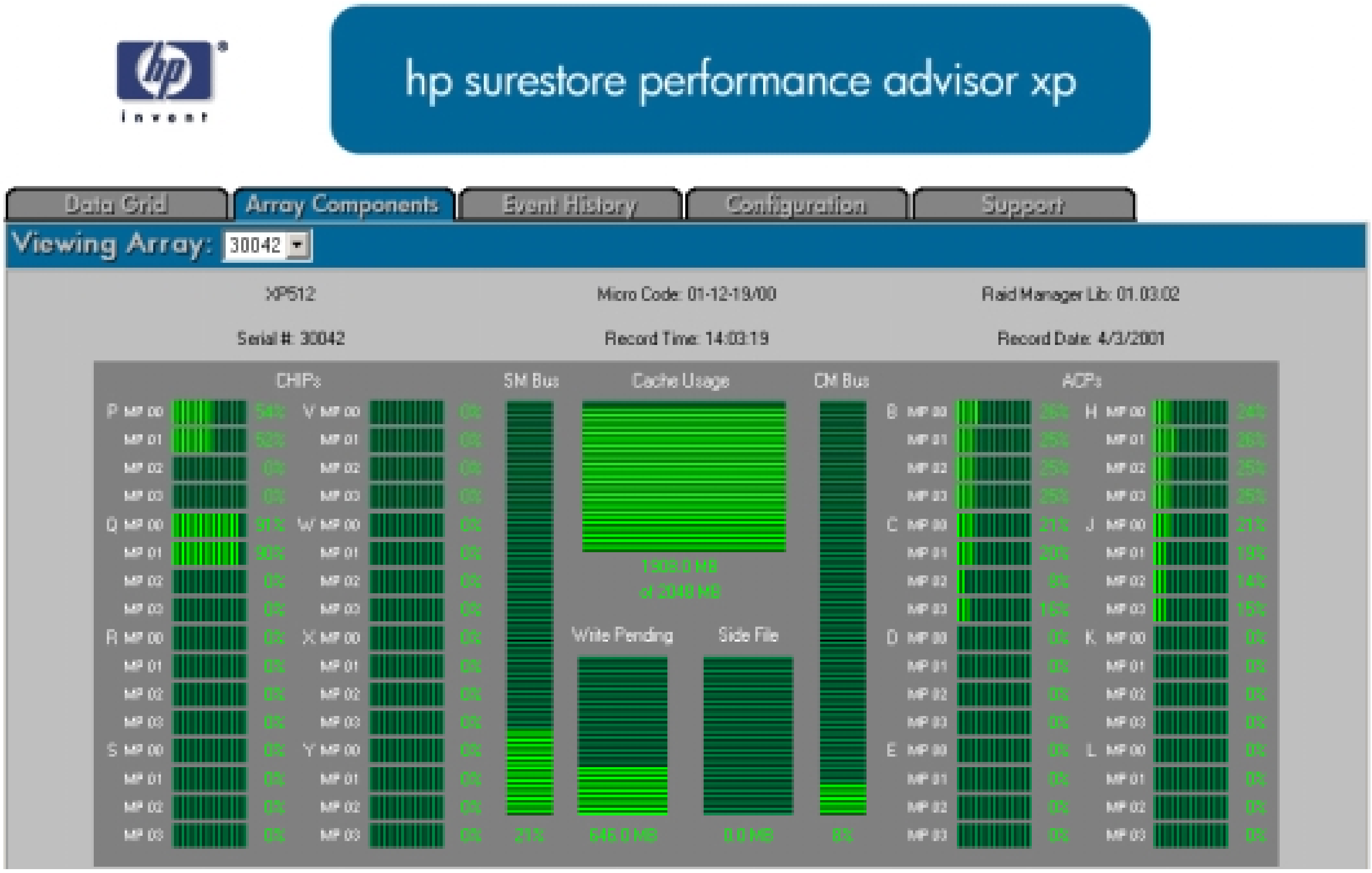
	Month:	Day:	Year:	Hour:	Minute:	Second:
Start:	4	3	2001	14	17	55
End:	4	3	2001	14	44	55

Figure 17. LDEV History Graph

Array Components

Also from the *Data Grid* screen, you can access the *Array Components* screen by clicking on an Array ID from the first column. This gives you the screen shown in Figure 18. This is probably the best screen for giving you an overall feeling for how the array is performing. On the left you can see each CHIP processor, which was discussed in detail beginning on page 5. In the middle section of the screen you can get information regarding the bus utilization of the shared memory bus and the cache memory bus. Also, the middle section shows the cache usage statistics. In a future revision of this White Paper, the cache, including its utilization, will be discussed in greater detail.

On the right side of the screen are the ACP Pairs. The utilization of each ACP processor is given here.



XP512 | Micro Code: 01-12-19/00 | Raid Manager Lib: 01.03.02

Serial #: 30042 | Record Time: 14:03:19 | Record Date: 4/3/2001

CHIPs

P	MP 00	54%	V	MP 00	0%
	MP 01	52%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%
Q	MP 00	91%	W	MP 00	0%
	MP 01	90%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%
R	MP 00	0%	X	MP 00	0%
	MP 01	0%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%
S	MP 00	0%	Y	MP 00	0%
	MP 01	0%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%

SM Bus: 21%

Cache Usage: 1908.0 MB of 2048 MB

Write Pending: 646.0 MB

Side File: 0.0 MB

CM Bus: 8%

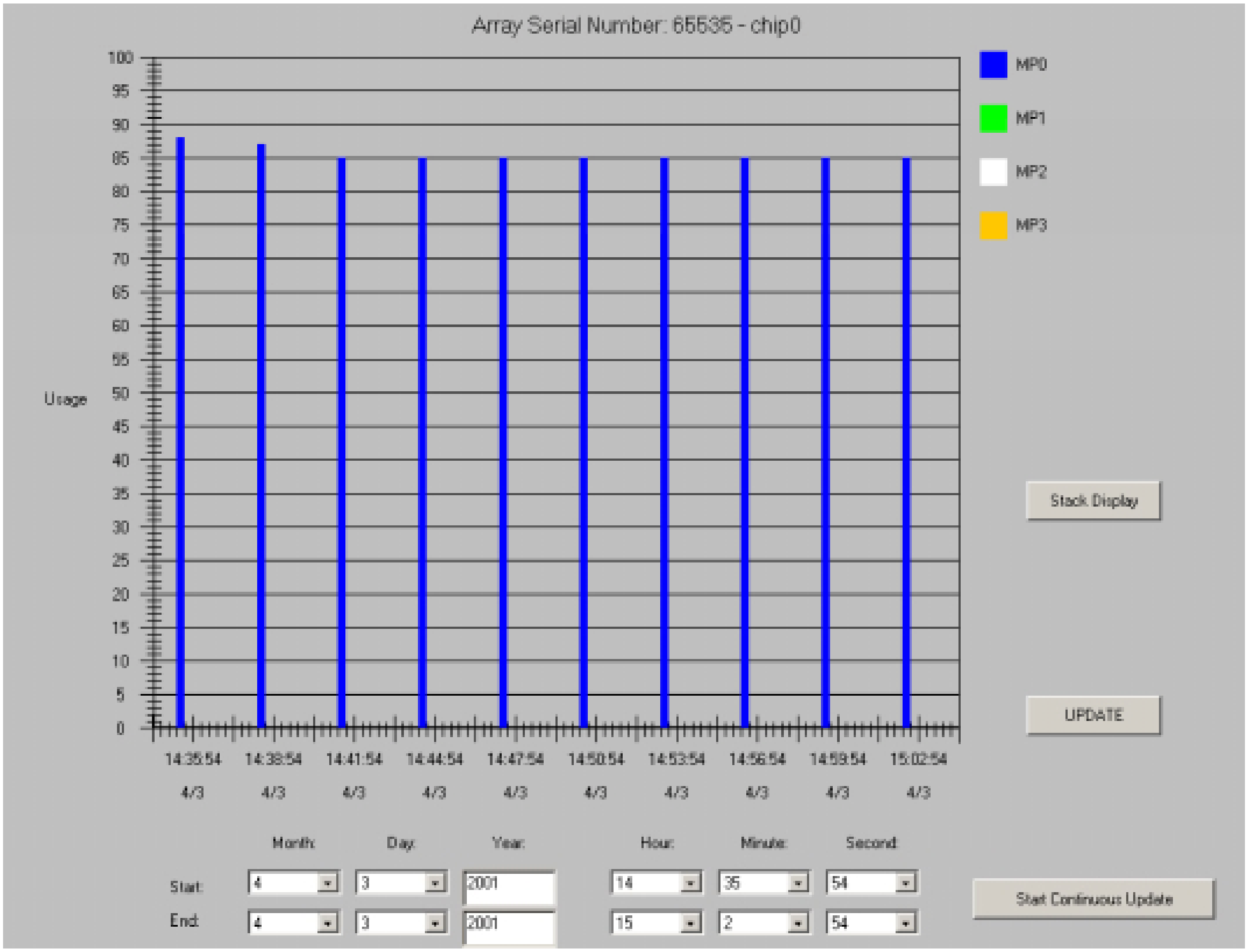
ACPs

B	MP 00	26%	H	MP 00	24%
	MP 01	25%		MP 01	26%
	MP 02	25%		MP 02	25%
	MP 03	25%		MP 03	25%
C	MP 00	21%	J	MP 00	21%
	MP 01	20%		MP 01	19%
	MP 02	8%		MP 02	14%
	MP 03	16%		MP 03	15%
D	MP 00	0%	K	MP 00	0%
	MP 01	0%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%
E	MP 00	0%	L	MP 00	0%
	MP 01	0%		MP 01	0%
	MP 02	0%		MP 02	0%
	MP 03	0%		MP 03	0%

Figure 18. Array Components

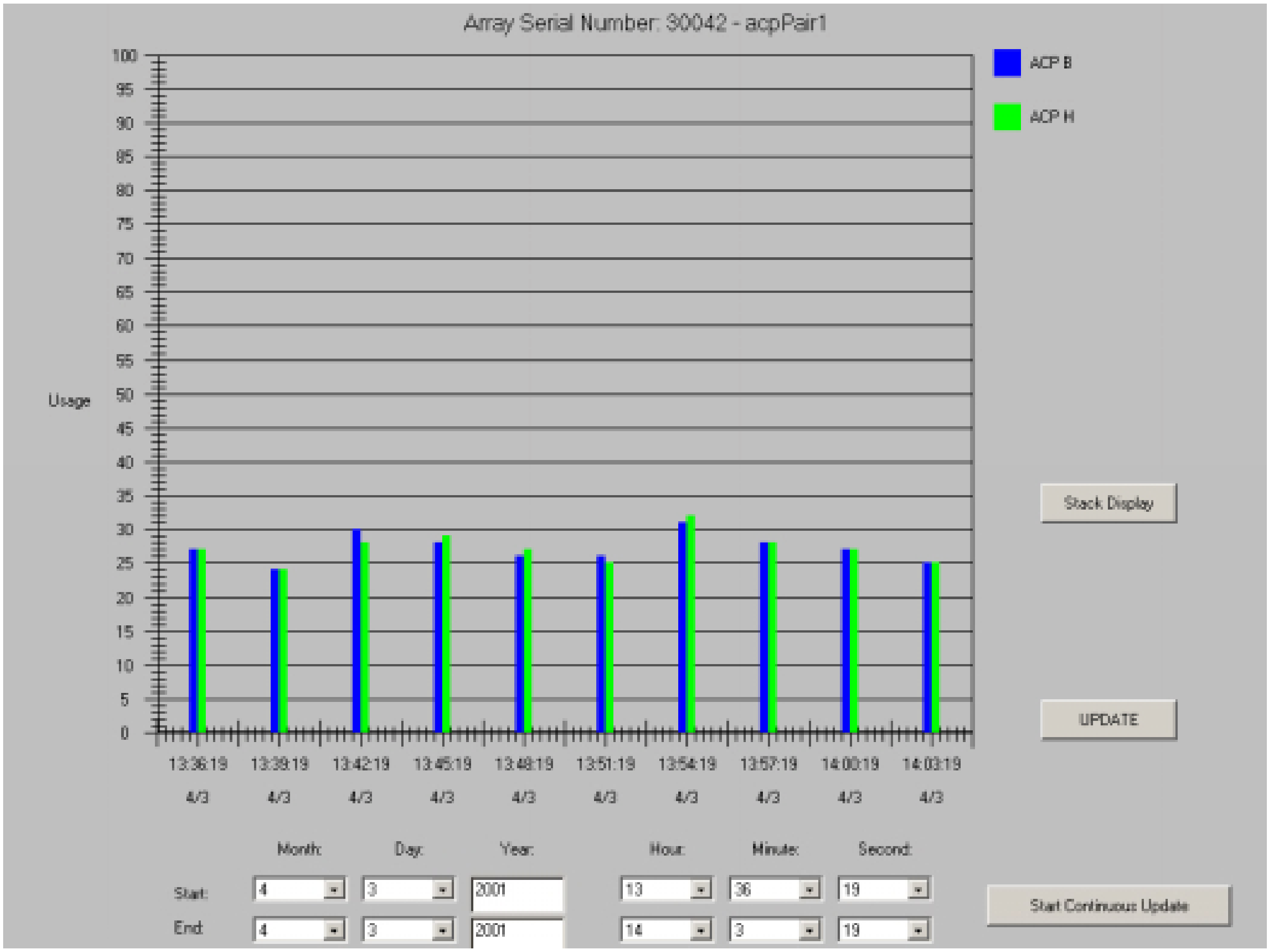
By clicking on any CHIP processor or ACP processor on the "Array Components" screen you can get more detailed information and history regarding the utilization of these components – see Figure 19 and Figure 20.

Note the letters to the left of the processor labels (on both the CHIPs and ACPs). These letters identify the slot into which the board is plugged in. These slots are labeled on the XP array, so you can match up the board (and therefore the ports) with the processor utilization shown in this display.



	Month	Day	Year	Hour	Minute	Second
Start:	4	3	2001	14	35	54
End:	4	3	2001	15	2	54

Figure 19. CHIP Processor Utilization

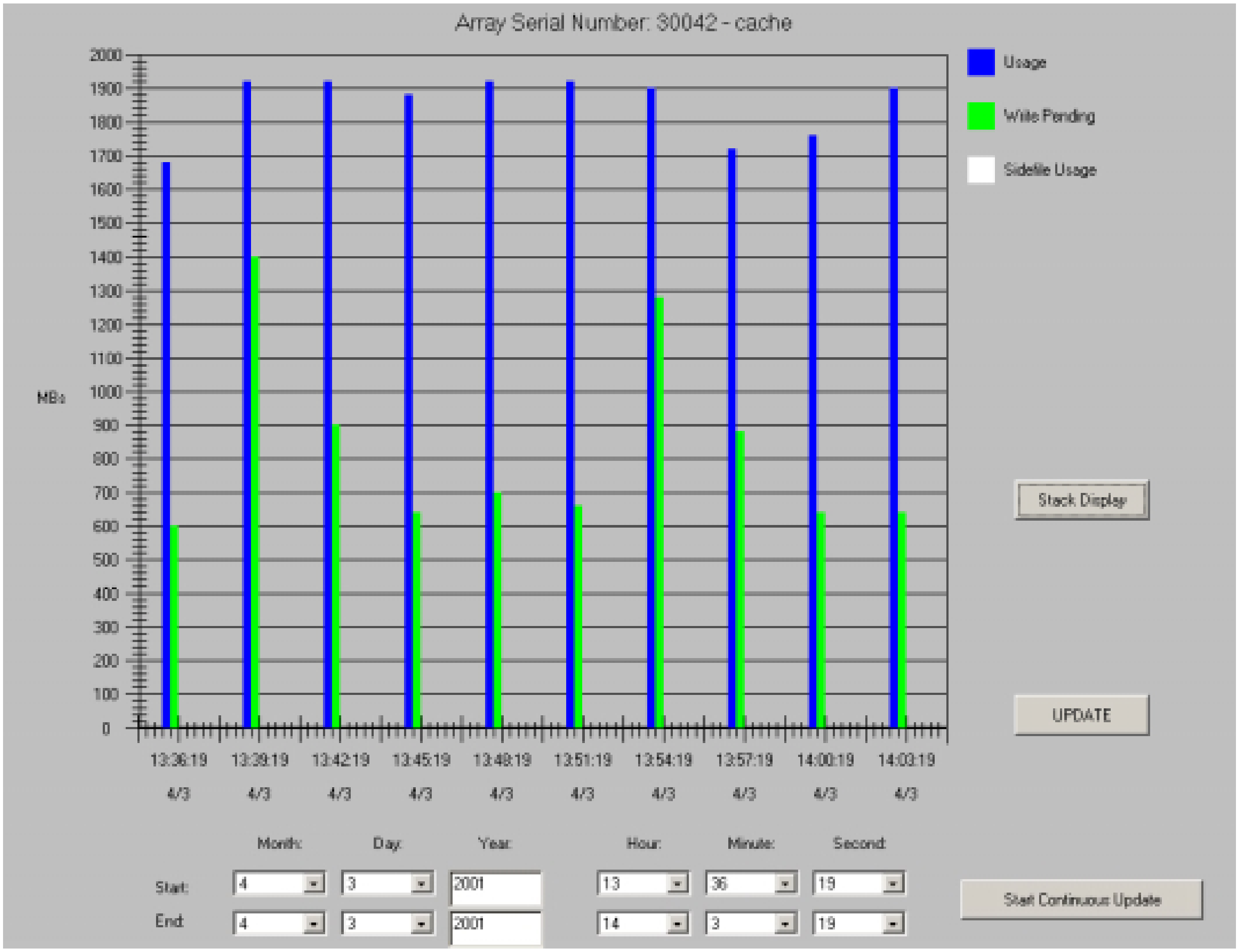


	Month:	Day:	Year:	Hour:	Minute:	Second:
Start:	4	3	2001	13	36	19
End:	4	3	2001	14	3	19

Figure 20. ACP Processor Utilization

Cache

Finally, clicking on the “Cache” from the “Array Components” screen will give you the view as shown in Figure 21.



	Month:	Day:	Year:	Hour:	Minute:	Second:
Start:	4	3	2001	13	36	19
End:	4	3	2001	14	3	19

Figure 21. Cache Usage

Basic Configuration Guidelines

There are a few basic configuration guidelines to consider when configuring an XP array for optimal performance:

- ❑ Understand the workload
- ❑ Spreading out the I/Os across all components
  - ❑ ACP Pairs
  - ❑ CHIP Pairs
  - ❑ Ports
  - ❑ Array Groups
  - ❑ Odd/Even LUNs

Understanding the Workload

This is the most critical piece in obtaining optimal performance. Though it may be difficult to assess (without first running the application and using ***Performance Advisor*** to analyze the workload), any information regarding the application workload will aid in configuring the application optimally across all the components of the array.

Spreading out the I/Os across all components

In an ideal world, where each ldev is identical in the application, you would want to ensure that you have spread out all the I/O activity across all the ACP pairs, CHIP pairs, Ports, and Array Groups evenly. Because real-world applications don't behave in this ideal manner, the optimal performance configuration can be obtained by getting as close to this ideal configuration as possible. Once an array is running an application, you can use ***Performance Advisor*** to monitor the usage of all these components, and make adjustments to the configuration as necessary. The use of ***Performance Advisor*** to identify "hot" components is highly recommended. In addition, continued usage and monitoring with ***Performance Advisor*** can aid in determining when potential performance bottlenecks may be occurring (this can be done by baselining the performance of your application and monitoring for changes).