Preliminary (off) white paper on high availability disk products for the HP3000

» Return to original page

by Walter McCullough

Version Date May 20, 1997

This (short) paper addresses the current need to describe to customers and field personnel the features and performance of 3 disk products. The paper will try to set customer expectations so that these products can be successfully integrated as high availability or data center management solutions. Data is being collected for a more detailed paper that will be released in the near future.

High Availability Storage System (Jamaica Enclosure)

The Jamaica enclosure is just that, a low cost cabinet that can hold 4 full height or 8 half height disk drives that is supported on MPE/iX release 5.0. The drives come in 2Gb, 4Gb, and 9Gb capacities. The enclosure has a hot swap capability, but is only supported with the FastWide configuration. *(Hot Swapping on SingleEnded devices may/will cause aberrant signals on the SCSI bus that could cause disk corruption and/or the destruction of the interface card.)*

The enclosure does not support a RAID configuration but is ideal for implementing Mirrored Disk/iX, a software mirroring solution. To take advantage of the high availability configuration, the mirrored partners should be configured to ensure pathway redundancy. Disk failures are recognized by the Mirrored Disk/iX software and messages are then sent to the system console.

A UPS is required due to a new disk technology called "Write on Arrival". The UPS will guarantee write atomicity in case of power failure, ensuring that the MPE/iX Transaction Manager will successfully recover pertinent data. *For more information on this topic please refer to the ESP document with keyword "DISK POWERFAIL".*

Model 10 & 20 Disk Array (Nike)

The Nike Array is another high availability disk solution costing more than the Jamaica enclosure but with greater functionality. It is supported on MPE/iX release 5.5 with a FastWide SCSI adapter firmware release level of 3636 or greater and PowerPatch 1. The cabinet comes in multiple models with the largest supported model containing twenty 4Gb disk drives.

The Nike cabinet supports several RAID protocols and can be used to protect the LDEV 1. RAID-1 is the only protection mode supported for LDEV 1, which is hardware mirroring. The cabinet supports a dual SCSI connector that can be used for redundancy, but is currently not supported under MPE/iX. Automatic failover (or Auto-Trespass) to the other connector is currently under development with release scheduled for 1998. The automatic failover will give the Nike pathway redundancy, increasing its high availability feature set by reducing the likelihood that any of the five points of failure will deny access to user data.

The RAID configurations perform slightly slower than generic disk drives or JBOD (Just a Bunch Of Disks) due to the added complexity of the protection algorithms, all done by the Nike processor. These other protection algorithms may need to "look" at the data before it goes to disk to create parity or "repair" bits that are stored across or on other disk drives within the cabinet.

The Nike Array also supports up to 64Kb of cache used to manage data internally. Use of this cache has not shown any performance improvement for MPE/iX.

This device is configured through a separate terminal connected to a serial port on the back panel of the Nike cabinet. The interface allows the user to define the level of (RAID) protection for each disk drive within the cabinet. The cabinet also supports Hot Standby drives. The Nike processor can detect disk failures and automatically switch to the Hot Standby drive and use that drive until repairs are made and the failed drive is replaced.

EMC Symmetrix 3000

The Symmetrix 3000 is currently the high end disk product offered by Hewlett-Packard Company, in conjunction with EMC Inc., particularly for the HP3000 and HP9000 high availability and data center management solutions. The Symmetrix 3000 provides a unique data capacity, protection and configuration of disk drives.

The Symmetrix 3000 is supported on MPE/iX release 5.5 with a FastWide SCSI adapter firmware release level of 3636 or greater and PowerPatch 1 or MPE/iX release 5.0 and PowerPatch 6. The cabinet contains slots for 96 disk drives in 4Gb and 9Gb capacities. The Symmetrix 3000 manages these disk drives as an array of disk space that can be partitioned ,by the EMC technician, independently of the physical disk capacity. This ability to partition means that four 4Gb drives can be configured to "present" a view that eight 2Gb drives are connected. This allows the System Administrator and EMC technician the flexibility to configure groups of disk partitions as LDEVs and assign them to different computers.

The Symmetrix cabinet supports up to a 4Gb cache to help reduce the number of physical I/Os to the disk drives. The affect of the cache on system performance can vary depending on the application's I/O characteristics and other factors listed later in this paper.

The feature sets for this product are listed below:

Environmental

> The footprint and disk capacity of the cabinet is impressive. 96 units ***** 4Gb gives approximately 3.84 Terabytes of data storage.

Heterogeneous Hardware Support

> The cabinet can be shared among multiple machines running MPE/iX or HP-UX as well as other non HP operating systems. ***(Shared hardware not shared data****)*

High Availability Features

> The protection of data using RAID protection algorithms along with EMC's proprietary RAID-S mode. Hot standby of disk drives allow for automatic replacement of failed components.

Support

> The Symmetrix cabinet is configured to "call home" when errors are detected by the Symmetrix processor. The "call home" feature alerts EMC to get in touch with the customer to walk them through any problems they may encounter or they can make some adjustments remotely. They will also dispatch a CE if a problem needs immediate attention.

Other features that are not yet certified for the HP3000 are the SRDF (Symmetrix Remote Data Facility). SRDF provides remote shadowing of data to another Symmetrix cabinet that can be installed at a remote data center. This feature allows the customer to design a disaster tolerant solution.

Another feature not yet certified is SMMF (Symmetrix Multiple Mirror Facility). SMMF provides mirroring of data within the same cabinet but to drives connected to other machines. This is much like the SRDF feature but internal to the cabinet.

The configuration of the EMC Symmetrix 3000 requires a trained EMC technician. Once configured, the system administrator can use the familiar VOLUTIL utility to create User Volumes or to add members to existing volume sets.

Disk Subsystem Performance

Scenario For a Successful Implementation

The performance of the disk subsystem depends greatly on a number factors. Current data shows that 10-20 physical disk I/Os per second per drive is about the limit before performance starts to degrade because of the I/O bottleneck. The number of physical devices that the adapter can handle and the amount of memory installed in the HP3000 can also impact performance. The application's I/O characteristics, serial vs. random, can greatly affect performance due to the pre-fetching algorithms used by either MPE/iX or by the Symmetrix 3000 processor. All these factors should be considered before reorganizing the disk subsystem either by moving data from smaller to larger disks (JBOD or arrays) or combining LDEVs per physical device (Symmetrix).

Below are two case studies where the first case shows a properly configured disk subsystem where a customer improved performance and the second case shows an improperly configured disk subsystem that exposed some problems:

> Case1 This customer has an extensive amount of batch processing. The user's previous configuration was below the 10-20 I/Os per second per device threshold and they had a good understanding of their application's needs (I/O resources). They effectively used MPE/iX's User Volumes to segregate major application data onto different volume sets. They also migrated their data from 4Gb stand-alone disk drives (JBOD) to 4Gb Symmetrix drives, without reducing the number of spindles or LDEVs.
>
> Analysis This was a successful integration of a new disk technology, like the EMC Symmetrix 3000. The serial nature of the application lends itself to the caching algorithms done by the Symmetrix processor. The Symmetrix processor pre-fetches more data than the MPE/iX memory manager, allowing for a better hit ratio to the Symmetrix cache. A cache miss causes an increase in the total time it takes to do an I/O, from the start of the request to its completion, because the I/O request is satisfied by the slower disk mechanism instead of the much faster cache hit.
>
> Case2 A site containing a large number of 1Gb and 2Gb disks with 3-4 User Volumes and a 4 member System Volume Set was consolidated into a Symmetrix cabinet and reconfigured and reloaded onto one System Volume Set with disk drives partitioned into 4Gb volumes. This user was at the I/O threshold of 20 I/Os per second per LDEV before the implementation and was clearly over it after consolidating the data onto the Symmetrix. This customer also expected a 20 to 40 percent performance improvement because of the Symmetrix cache but was very disappointed after collecting performance data on his application's very random I/O characteristics.
>
> Analysis This new configuration reduces the number of LDEVs and also reduces the number of pathways to the data. By reducing the number of pathways to the data, MPE/iX takes longer to fetch all the data that is required for an operation to take place. MPE/iX tries to asynchronously gather this data

but if there is only one pathway to the data then it will queue I/O requests for the drive mechanism and wait for each I/O completion until proceeding.

The next problem with this configuration is a major reduction in throughput caused by MPE/iX's use of the Transaction Manager. This mechanism keeps track of changes to MPE/iX internal data structures and user data bases, such as KSAM and Image/SQL files. The Transaction Manager's (XM's) log file lives on the master volume of a volume set. If there is one volume set then all activity will be to that volume set's master.

As the many applications did their random I/O to their data files, the Symmetrix internal cache was of no use because of the lack of good hit rate. The customer's random access I/O characteristics showed a hit rate of only 30%. A cache hit rate of 75% must be achieved before any I/O performance gains are realized when four 1Gb stand-alone disk drives are consolidated into one 4Gb Symmetrix drive.

"Rules of Thumb"

For Improving I/O Performance and Migrating Data to New Disk Drive Technologies

1. Limit the disk size or partition size from 1Gb to 2Gb and have several pathways (LDEVs) for the system volume set. MPE/iX uses opened files on disk as extensions to its memory and does "file operations" through its memory manager. This operating system works best when there are multiple pathways to the working set (transient space) and permanent data.

2. Restore the application's data to User Volumes. Create as many User Volume Sets as make sense. This reduces the XM bottleneck to the master volume of the volume set, adds more pathways to the data and adds fault containment. This will limit the amount of data to reload in case of catastrophic disk failure.

3. Gather data on the user's application I/O characteristics. There are several tools available like GLANCE/XL that can help collect this data.

4. Do not exceed 10-20 I/O operations per second per physical device. Consolidating a large number of physical devices to a fewer number of larger capacity devices might cause you to exceed the recommended I/O rate and cause an I/O bottleneck. For better performance throughput it is also recommended that NO MORE THAN 4 DRIVES BE CONNECTED TO THE SINGLE ENDED BUS and 8 FOR FastWide. *(Your mileage will vary depending on driving habits)*

5. Purchase a high availability disk drive technology, like the Nike Array or the Symmetrix 3000, for its feature set and its ability to protect data. Performance of these products is dependent on a clear understanding of how it is used, I/O characteristics of the application, and the way it is configured.

» Return to original page

Privacy statement Using this site means you accept its terms Feedback to webmaster