Performance Implications of PCI on the New HP e3000 Servers

By
Stephen F. Macsisak
Hewlett-Packard Company
19447 Pruneridge Avenue. MS 47UA
Cupertino, CA 95014
408.447.5851 Direct
408.447.4278 Fax
Paper ID#- 282

In 7.0, Hewlett-Packard will introduce significant changes in the HP e3000 I/O subsystem. The new system will support faster cards and higher speed buses, thus changing some of the old rules about devices per card and number of cards per I/O channel. The typical system will have fewer card slots, but support more disk connectivity.

The HP e3000 system manager must determine the right mix of cards, buses and physical disks to get the best performance and minimize costs. The system manager must choose between Raid 1, Raid 5 or software mirroring to achieve "24X7" objectives that users require. This session discusses how best to achieve high performance, low cost and high availability with the new I/O subsystems on the HP e3000.

The following discussion focuses on the new HP e3000 features without getting lost in the gory detail. For more detailed information, access the following web sites:

- For new features of the N- and A- class servers, access communicator articles, new installation and configuration documents at http://docs.hp.com.
- For details on the new HP e3000 business servers, access http://www.hp.com/products1/mpeixservers.
- Access the *HP e3000 Ordering and Configuration Guide* at http://www.hp.com/products1/mpeixservers/info_library/config_order/index.html.
- As always, you can find many important documents on the standard HP web site, http://hp.com.

Summary

The new N-class and A-class systems have much faster I/O buses. The N-class system supports up to a I/O bandwidth of 6.4 GB/second; the A-class system supports up to a I/O bandwidth of 1.9GB/second.

The system uses Peripheral Component Interconnect (PCI) adapter cards to communicate with disks, tapes and networks. The new machines have from 2-12 PCI slots per system. The maximum number of slots is similar to the capability of the Series 9x9 system with an I/O expander. Only Series 997/xxx systems have the ability to have more slots than the N-Class. Unlike all previous HP e3000 systems, each PCI slot - except two slots in the A400 - has its own high speed Turbo or Twin Turbo I/O channel running at about 250 MB/second or 500 MB/second. The new system also supports a ULTRA2 SCII card capable of 80 MB/second connections to disk, unlike the current maximum of 20 MB/second connections to disks.

Another new feature of the N-class PCI slots is the ability to support multi-port cards that allow the equivalent of two independent SCSI buses on single adapter card. This allows a N-class system to have up to 24 SCSI ports, each capable of having fifteen disks for a total of 360 disks, not counting the core I/O card included with each system.

In the past, some systems managers had low numbers of devices per adapter card because of concerns about performance. The faster PCI bus and support for Ultra2 SCSI cards that can transfer data at 80 MB/second effectively allows more devices per card without fear of performance degradation caused by bandwidth limitations. The typical current high-end OLTP HP e3000 has an aggregate I/O bandwidth of 2-5

MB/second that is the equivalent of 250 – 620 physical disk I/Os per second at 8 KB per I/O. The new N-class systems have an I/O bandwidth capability over one thousand times larger then current high-end HP e3000 systems , take advantage of. In contrast, the current NIO based systems have a maximum of 4-8 I/O channels at a maximum of 32 MB/second per channel. The N-class has a maximum of 6.4 GB/second I/O bandwidth capability versus the current NIO systems maximum of 256 MB/second I/O bandwidth in an eight HP-PB channel Series 997/XXX. Otherwise, the limit is four channels and a maximum I/O bandwidth of 128 MB/second.

In a heavy OLTP environment, the savvy HP e3000 system manager knows that it is more important to have as many physical disks as possible than to worry about the bytes per second bandwidth of the HP e3000 I/O subsystem. The number of physical spindles has the greatest overall effect on disk performance. The new N-class and A-class systems certainly makes I/O bandwidth much less of a performance issue.

Software Compatibility with 7.0 and 7.0 Express One

The 7.0 Communicator cites several differences in 6.5 versus 7.0 features, where the focus was support for the N-class and the A-class. All of those features are minor (in my opinion), with the major features of 6.5 included in 7.0. Express One 7.0 release adds all of those minor features to the release, plus adds support for multi-way N-class and more than 2GB memory on the N-class and the A-class. Also included in 7.0 are large file support, memory manager changes to support more the four GB of memory and the other scaling features of 6.5.

Any application or tool that was tested on 6.5 almost certainly works on 7.0 because the PCI support is at the hardware device level not at the intrinsic level (unlike 6.5 where several tools needed changes to support large mapped files). A detailed list of products/tools that have been tested/certified for 7.0 can be found at http://www.hp.com/products1/mpeixservers/news_events/E3000000VSM.html.

New HP e3000 Acronyms

HVD High voltage differential – electrical signaling used with SCSI Devices on the HP e3000. Sometimes called Fast/Wide differential SCSI and what most of you use for disks on the HP e3000. Normally runs at 20 MB/second.

LVD Low voltage differential – electrical signaling used with some of the SCSI device adapters offered on the N-class and the A-class HP e3000.

SE Single Ended – electrical signaling used with SCSI device adapters. Generally used with slower devices but LVD cards also support this type of connections. If SE devices are on some bus (wire) as LVD devices with a LVD type adapter, then the bandwidth is reduced to 5 MB/second.

Ultra SCSI (+WIDE) - protocol used to talk between disks and adapter cards. When used with 16(WIDE) bit connections can provide up to 40 MB/second bandwidth between devices and system. This is what the HVD

Ultra SCSI cards on the N-class and A-class support, but MPE/iX is certified for 20 MB/second support only at this time.

Ultra2 SCSI (+WIDE) – protocol used to talk between devices and adapter cards. When used with 16(WIDE) bit connections, Ultra2 SCSI can provide up to 80 MB/ second bandwidth between devices and system. These cards are only supported on N-class and A-class HP e3000 servers.

Twin Channel – PCI slot in N-class and A-class servers that can support up to 250 MB/ second of bandwidth.

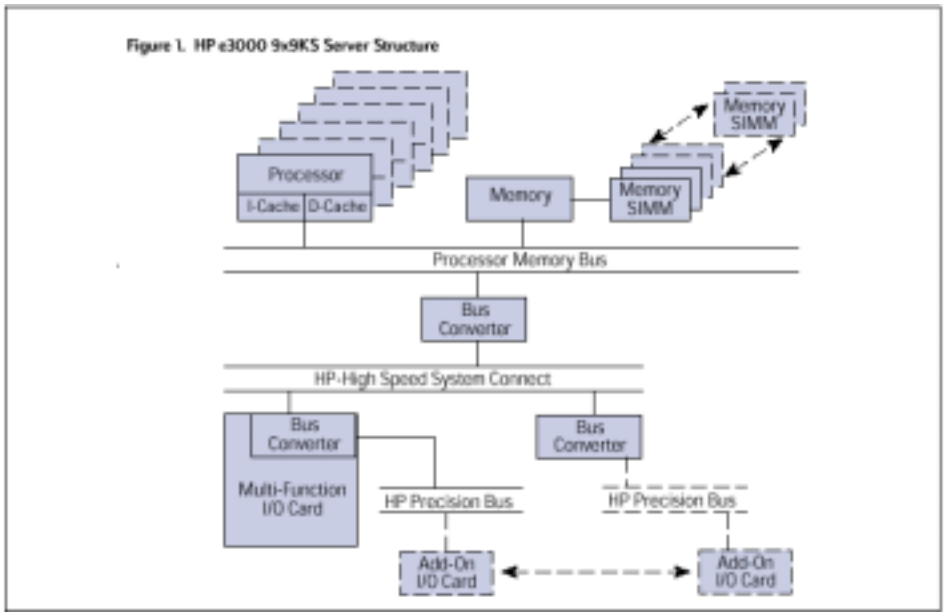
Twin Turbo Channel – PCI slot in N-class and A-class servers that can support up to 500 MB/ second of bandwidth.

Multi-Port Adapter – adapter card that supports two independent SCSI buses, equivalent to two separate adapter cards and can support up to 30 devices. Only supported on N-class systems. This type of adapter card allows the N-class server to support up to (12*2*15) or 360 Disks with twelve PCI slots.

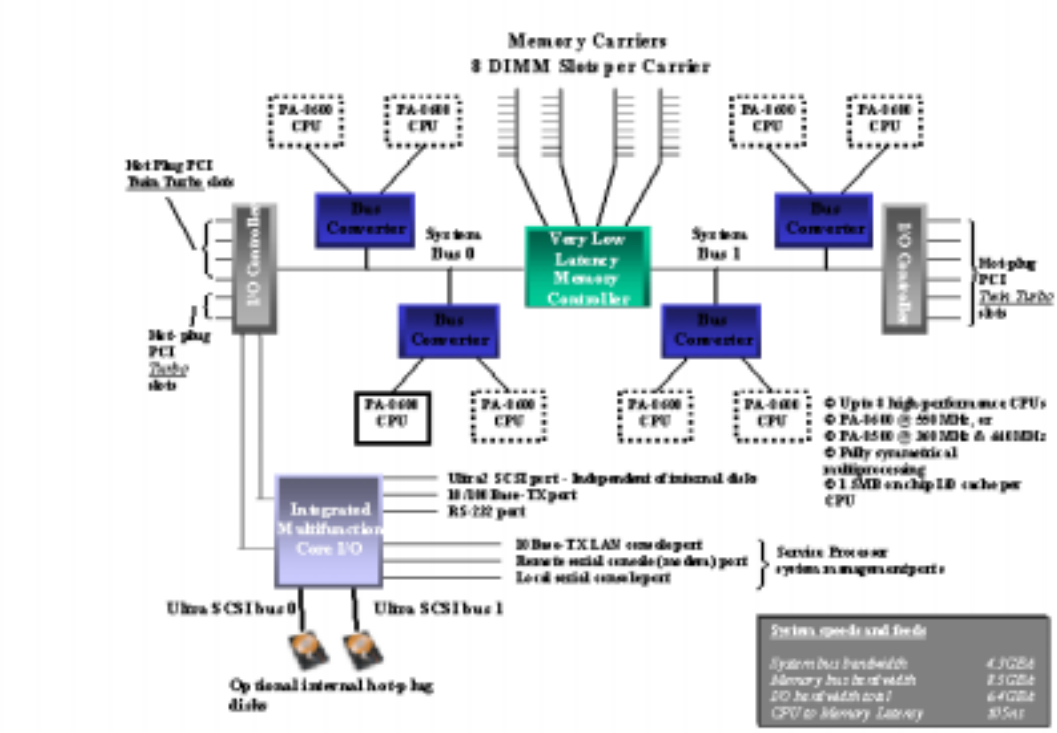
HP-PB Hewlett-Packard precession bus – the bus used to link adapter cards to the next higher performance system bus in HP 3000 systems.

Comparison of PCI with HP-PB

The current systems have I/O buses that are similar to the following figure. There are differences between systems, but all systems have an HP-PB bus into which multiple adapter cards are plugged. The maximum speed of the HP-PB bus is 32 MB/ second.



The new N-class has a much higher performance I/O architecture. The following figure shows some of the different types of buses. It is important to be aware that each of the N-class PCI slots have more I/O bandwidth capability then the HP-PB design.



Changes to the MPE/iX I/O Software Subsystem

The Device Adapter software was rewritten to support the new cards and run on the N-class and the A-class. The changes are at a very low level and should not present any incompatibility with application software and/or software tools. Our tests have shown that running on 7.0 should be “no problem” if the software already runs on 6.5 or was changed to work with large files if mapped file accesses are being made.

The following picture shows how the MPE/iX I/O software is designed. Again, the general organization of the MPE/iX I/O system has the I/O software *mirroring* the I/O hardware. There is a specific manager for each type of hardware module, and an *instance* of each manager for each *instance* of hardware module. There are several reasons for this design:

> This concept provides for a simple manager design. The manager only needs to know about one hardware module and doesn't need to know anything internal about any other piece of hardware in its 'path' (from device to CPU).
>
> This concept provides for a portable manager design. For example, a disk DM can operate either through the CIO SCSI DAM/DA or through the NIO SCSI DAM/DA without changes to the disk DM

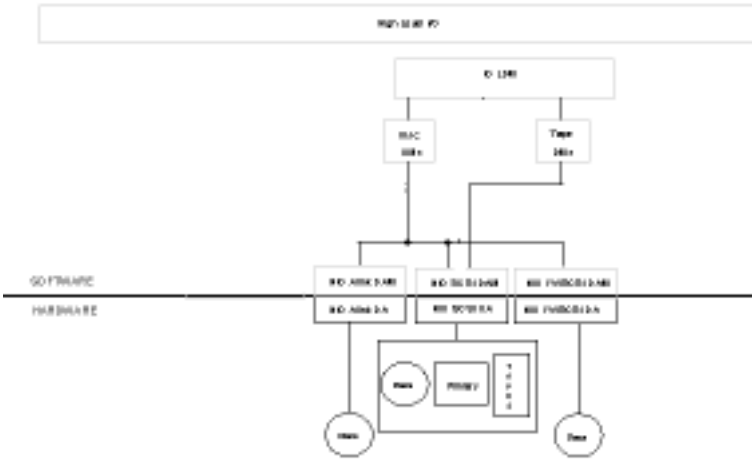
code.

System configuration is very flexible. Additional hardware and software modules are easily added to the system.

Hardware-specific changes are localized. If a particular hardware module has a change that requires software modification, only the software controlling that hardware needs to change (we hope!).

The I/O software is made more reliable. Conflicts and inconsistencies can occur if multiple pieces of software control a single piece of hardware. Since only one piece of software controls a particular piece of hardware, it is clear who is in control and responsible. Software maintenance is centralized and, therefore, more reliable. A logical change may only need to be made in one place rather than spread throughout several pieces of software.

I/O Module Relationships



Definitions for I/O module Relationship figure.

HLIO -- On the upper part of the slide, the High Level I/O (HLIO) layer provides an abstract LDEV based interface to upper layers of MPE/iX (file system type managers and storage managers). Information flows from HLIO to the LDMs.

LDM -- The Logical Device Manager Module accepts requests from the HLIO module and processes information in a non-device specific way. It then passes information to the DM.

DM -- Each Device Manager Module is devoted to a specific kind of device. Current DM Module Server Procedures include, but are not limited to:

SCSI_disc_dm for Single Ended (SE) SCSI Disks.

SCSI_tape_dm for SCSI DDS Tapes

SCSI_disk_and_array_dm for FastWide (FW) SCSI Disks and Disk Arrays, etc.

The DM does device specific processing and then passes information to the DAM.

DAM -- Each Device Adapter Manager Module is devoted to a specific kind of Device Adapter Card. Current DAM Modules Server Procedures include, but are not limited to:

SCSI_dam for the SE SCSI DA.

fwSCSI_dam for the FWSCSI DA.

HvdSCSI_dam for the new PCI DA.

LvdSCSI_dam fro the new PCI DA, etc.

NIO DAMs will interact directly with the hardware bus.

NIO (Native I/O) is the Generic PA-RISC Bus. This bus is deemed "Native" as it complies with PA-RISC I/O Architecture definitions. GSC+ (General System Connection) is a high performance native I/O bus. Currently, this bus is used in some systems as a system bus and in other systems, as an interconnection bus between the System Main bus and NIO busses on other systems.

Performance Considerations When Using PCI Connected Devices

Each PIC slot has a very high bandwidth capability. In the past, standard configuration guidelines said we supported fifteen SCSII devices (disks) on a single FWD card, but Hewlett-Packard recommended ten disks for best performance. In some configuration tables, only ten devices were recommended; sometimes, the documents implied only ten devices were supported per FWD card. When using SCSI target addressing, the number is fifteen disks. When using multi-LUN addresses, we can actually support something like 254 devices per card. Currently, the only devices that support multi-LUN addressing on MPE/iX are the HP SureStore E Disk Array XP256, XP512 and XP48 subsystems.

You should be aware that all the warnings about performance were based on heavily loaded subsystems doing sequential copies designed to overload the card for diagnostic testing. You can do 20 MB/second with the HVD SCSI cards supported on the N-class. If you have fifteen drives doing forty I/Os per second of 8kb I/Os, you would have a bandwidth load of about four-five MB/second, which is much lower than the 20 MB/second the FWD cards support. The PCI slot into which the adapter card is plugged now supports 250 MB/second – 500 MB/second of bandwidth unlike the HP-PB bus that is only 32 MB/second. The new LVD Ultra2 SCSI adapter cards on the N-class can do up to 80 MB/ second. Bottom line is that you don’t need to worry much about how many disk devices you have per SCSI adapter card. At a minimum, you can go up to ten devices per card without any performance degradation. In an OLTP environment, you will bottleneck on the number of physical spindles well before the number of devices per card is a factor.

Best Disk Performance

From a system standpoint, the major cause of disk bottleneck issues is the number of physical spindles on a system. Disks are mechanical devices and while MPE/iX does a great job of reducing disk I/O because of its GLOBAL LRU memory polices, each disk I/O takes 5-25 milliseconds. Direct memory access times are in the nanosecond range VS milliseconds. Since disks are slow, the best way to optimize disk performance from a system point of view is to access the disk in parallel.

If performance is still unacceptable, add more disks to achieve more parallelism. The optimum number of disk I/Os per physical spindle is about 25 percent of a fully utilized disk. For example, if the average access time of a disk I/O is 9.5 milliseconds, then the highest I/O rate per drive should be (1000/9.5)*.25 or 26 I/Os per second. At I/O rates higher than 25 percent utilization of a disk, wait queues start to build for each disk and OLTP transactions times are elongated. If possible, all drives on the system should have about the same I/O rate and multiple programs should be accessing the data on all drives. A single threaded batch program typically reads or pre-fetches data from only one drive at a time, so there is little added benefit to having the data spread over thirty-five disks. A fifty online user OLTP and batch environment would benefit greatly if the number of physical spindles the data is spread over were doubled or tripled. The disk industry is building larger, cheaper disk drives, but the average access times don’t significantly change every time the disk storage capacity doubles.

In the past, system mangers purchased their disk storage by determining how much storage capacity they needed, then worried about performance and high availability. Today, system mangers need to make sure they have enough disk space and enough disk performance. A system with one or two high speed 72 GB drives could have better performance, but if you traded in ten or fifteen 4 GB 5400 RPM Fast/Wide drives, your users will notice. Two 72GB drives with an average access time of 7 milliseconds can do 1000/7*.25* 2 or 70 I/Os per second at 25 percent utilization. Ten 4 GB drives with an average access time of 25 milliseconds (absurdly slow) can do 1000/25*.25*10 or 100 I/Os per second at 25 percent utilization. If your I/O needs are lower than 70 I/Os per second, then it doesn’t matter what you do. But if you are I/O bottlenecked on physical spindle access today, then the extra bandwidth supported on the new PCI based systems will not help much.

The new N-class and A-class servers have all new SCSI adapters. A single and a dual port, Ultra2 LDVD/SE SCSI adapter capable of up to 80 MB/sec bandwidth. A single and a dual port, Ultra HVD(Fast\Wide) SCSI adapter capable of 40 MB/second bandwidth but only certified on MPE/iX at 20 MB/second. The dual port adapters are not supported on the A-class. Access hp.com for more information about the new adapter cards.

In summary

The new N-class and A-class systems support much higher I/O bandwidth capability. One Twin Turbo PCI slot on the N-class has up to 500 MB/second of I/O bandwidth capability which is almost twice as much as eight HP-PB channels on the Series 997/xxx. There should be no performance problems with the number of disk drives per SCSI adapter on the New PCI based systems. The same rules for disk drives apply, in OLTP environments (all things being equal) it is better for system performance to as many physical spindles as possible.