Working with Arrays of Inexpensive EIDE Disk Drives

David Sanders, Chris Riley, Lucien Cremaldi, and Don Summers
University of Mississippi - Oxford

Don Petravick
Fermilab

Abstract:

In today's marketplace, the cost per Terabyte of disks with EIDE interfaces is about a third that of disks with SCSI. Hence, three times as many particle physics events could be put online with EIDE. The modern EIDE interface includes many of the performance features that appeared earlier in SCSI. EIDE bus speeds approach 33 Megabytes/s and need only be shared between two disks rather than seven disks. The internal I/O rate of very fast (and expensive) SCSI disks is only 50 per cent greater than EIDE disks. Hence, two EIDE disks whose combined cost is much less than one very fast SCSI disk can actually give more data throughput due to the advantage of multiple spindles and head actuators. We explore the use of 12 and 16 Gigabyte EIDE disks with motherboard and PCI bus card interfaces on a number of operating systems and CPUs. These include Red Hat Linux and Windows 95/98 on a Pentium, MacOS and Apple's Rhapsody/NeXT/UNIX on a PowerPC, and Sun Solaris on a UltraSparc 10 workstation.

Computing in High Energy Physics Conference (CHEP' 98)
August 31 - September 4, 1998
Hotel Inter-Continental
Chicago, Illinois, USA

Introduction

In today's marketplace, the cost per Terabyte of disks with EIDE (Enhanced Integrated Drive Electronics) interfaces is about a third that of disks with SCSI (Small Computer System Interface). Hence, three times as many particle physics events could be put online with EIDE. The modern EIDE interface includes many of the performance features that appeared earlier in SCSI. EIDE bus speeds approach 33 Megabytes/s and need only be shared between two disks rather than seven disks. The internal I/O rate of very fast (and expensive) SCSI disks is only 50 percent greater than EIDE disks. Direct Memory Access (DMA), scatter/gather data transfers without intervention of the Central Processor Unit (CPU), elevator seeks, and command queuing are now available for EIDE, as well as support for disks larger than 8.4 Gigabytes. PCI (Peripheral Control Interface) cards allow the addition of even more EIDE interfaces, in addition to those already on the motherboard.

Motivation

There are a number of High Energy Physics Experiments that have produced Terabytes of data[1]. A few examples as of 12/95 are:

Experiment	Data set (Terabytes)
FNAL-E791	50
FNAL-D0	40
FNAL-CDF	10
HERA-ZEUS	5
CESR-CLEO	5
LEP-Delphi	~5
LEP-L3	3.4
HERA-H1	2.5
LEP-Aleph	1.7
LEP-OPAL	1.5

The efficiency of data analysis is greatly enhanced by using disk based files of filtered Data Summary Tapes (DSTs) rather than continually loading the files from tapes. As shown in the following two tables, EIDE disks cost much less than SCSI disks.

Big Disks

EIDE Disk Model	Bigfoot	Deskstar	Deskstar	Diamond
Manufacturer	Quantum	IBM	IBM	Maxtor
Capacity (Gigabytes)	12	16.8	14.4	17.2
Max. Int. I/O (Mbits/s)	142	162	174	
Avg. Seek Time (ms)	12.0	9.5	9.5	9.0
RPM	4000	5400	7200	5400
Unit Street Cost	$241	$407	$407	$410
Cost $/Terabyte	$20000	$24000	$28000	$24000

SCSI Disk Model	Cheetah	Ultrastar
Manufacturer	Seagate	IBM
Capacity (Gigabytes)	18.2	18
Max. Int. I/O (Mbits/s)	231	180
Avg. Seek Time (ms)	5.7	6.5
RPM	10000	7200
Unit Street Cost	$1242	$1000
Cost $/Terabyte	$68000	$55000

Tests Performed

For this paper we tested two of the large capacity EIDE disks with six different operating systems and a PCI EIDE disk controller card. The six operating systems are Mac OS 8.1, Apple Rhapsody DR2, Sun Solaris 2.6, Windows 95b, Windows 98, and RedHat LINUX 5.1 (kernel 2.0.34). The two disk drives and the disk controller card are described below:

- Quantum Bigfoot™ TX[2] 12 GB, 4000 RPM, 142 Mbits/sec Maximum internal data rate, 12 ms average seek time.
- The IBM Deskstar™ 16GP[3] 16.8 GB, 5400 RPM, 162 Mbits/sec Maximum internal data rate, 9.5 ms average seek time.
- Promise Technologies Ultra 33™ PCI EIDE controller card[4] Supports 4 drives, Ultra ATA/EIDE/Fast ATA-2. Cost: $50.

Both the Quantum Bigfoot™ TX 12 GB and the IBM Deskstar™16GP 16 GB disks were successfully tested with the following systems:

System	Notes
Mac OS 8.1 on a Macintosh G3[5] rev. 2 motherboard	With HFS+ and both Master/Slave.
Windows 95b on a Dell Dimension XPS 350 computer[6] with PhoenixBIOS	Ok, depending on the BIOS*. Use FAT 32.
Windows 98 on a Dell Dimension XPS 350 computer[5] with PhoenixBIOS	Ok, depending on the BIOS*. Use FAT 32.
RedHat LINUX 5.1 (kernel 2.0.34) on a Dell Dimension XPS 350 computer[5] with PhoenixBIOS	Ok, depending on the BIOS*.
Promise Technologies Ultra 33 on a Dell Dimension XPS 350 computer[5] with PhoenixBIOS	Ok with Windows 95b and Windows 98. However, a patch† was needed for Red Hat LINUX.

Ten Terabyte EIDE Disk Architecture

The recipe for a simple 10 Terabyte EIDE Disk Architecture is as follows:

- Attach eight 16GB EIDE disks to each of 75 CPUs with the help of Promise PCI controller cards.
- Since EIDE cables have a maximum length of 18", it is easier to run extra DC power cables into a computer tower than to run EIDE cables out.
- Load data on these disk arrays.
- Plan to usually run analysis jobs on the same machine as the data.
- Use fast Ethernet switches to allow for remote jobs at a modest level.

Future

Future plans may include testing the drives with Apple Rhapsody, Sun Solaris and testing drives with Red Hat LINUX 5.1 – kernel version 2.0.35. (The 8 GB limit seen so far on Rhapsody DR2 and Solaris 2.6, however the commercial release of Rhapsody, MacOS X Server, is scheduled for fall 1998.) Also new technologies that are worth investigating include both “Lazy RAID” and Firewire™.

---

* See “Getting beyond the ATA 8.4 GB limit”
http://www.storage.ibm.com/hardsoft/diskdrdl/library/8.4gb.htm
http://www.storage.ibm.com/techsup/hddtech/welcome.htm
http://www.storage.ibm.com/hardsoft/diskdrdl/prod/deskstar.htm
http://www.storage.ibm.com/hardsoft/diskdrdl/prod/ultrastar.htm
and “8.4 GB Barrier”
http://www.quantum.com/src/whitepapers/8.4barrier.html
http://www.quantum.com/products/hdd/bigfoot_tx/
and “IDE Hard Drive Capacity Barriers”
http://www.maxtor.com/technology/whitepapers/capbar0.html

† Patch available from http://pobox.com/~brion/linux/promise34.gz,
but support is included in kernel 2.0.35

Lazy RAID

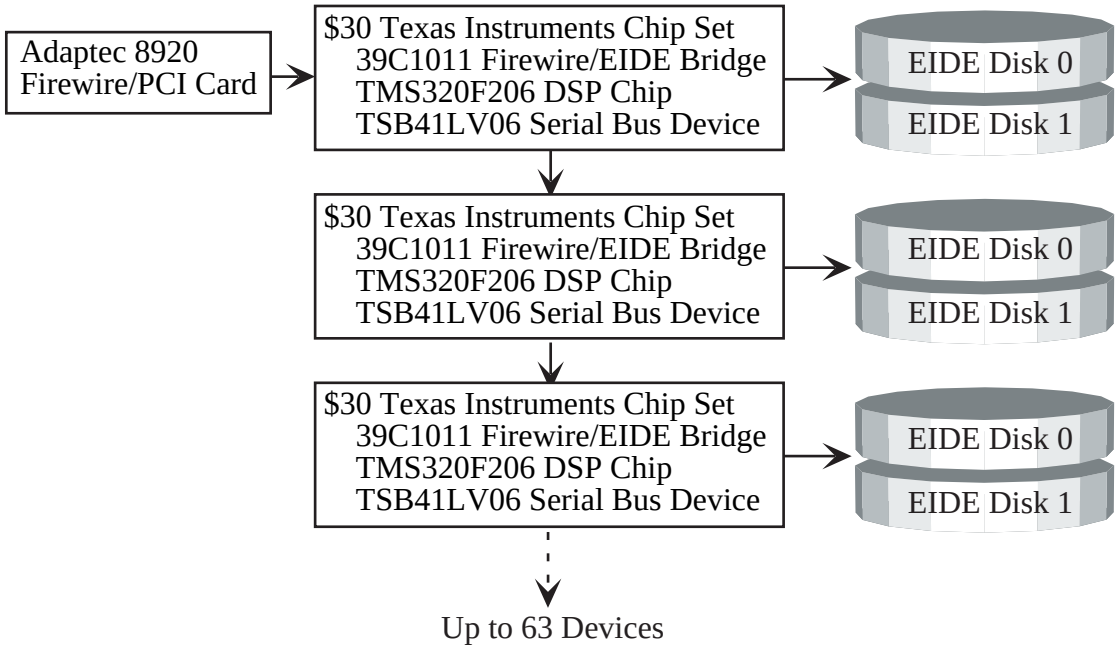
Lazy RAID (Redundant Array of Inexpensive Disks) is an idea for using disk arrays that offers protection for scratch disks in the event of catastrophic failure of one disk in the array. This system uses a number of data disks (say 7) plus one parity disk. Therefore, if one disk dies the parity disk would allow the recovery of data from the dead disk. One could use the RAW DEVICE interface to calculate parity with the CPU. If a disk fails then the operator would swap out the dead EIDE drive and reconstitute the dead disk drive onto the replacement drive using the parity disk and the remaining data disks. This system is well suited for use as scratch disks where a filtered DST is placed on disk once and read and analyzed many times. Using this scheme the one parity disk is updated only when a file is written to (or erased from) a disk.

Firewire

Firewire IEEE 1394 Specifications[7]:

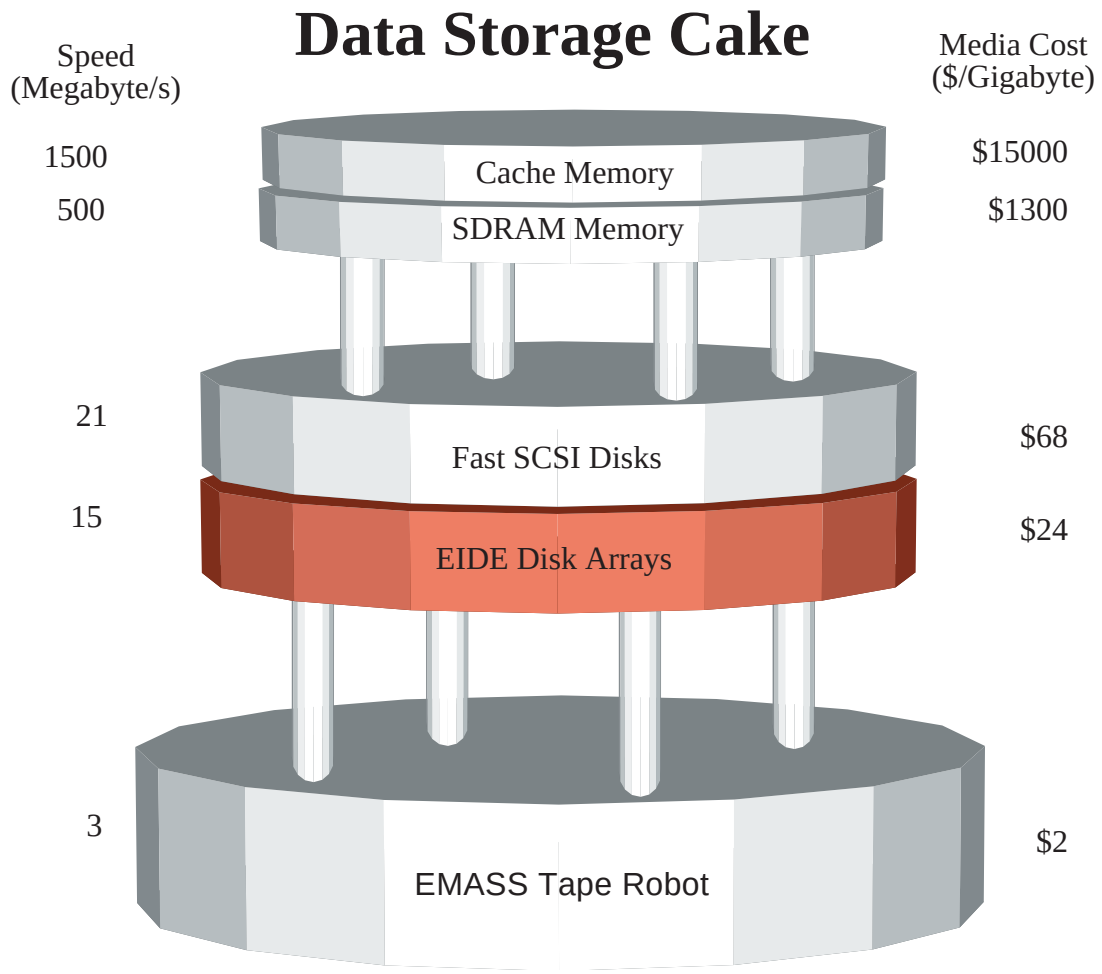
- Up to 25 or 50 Megabyte/s.
- Up to 63 devices per interface.
- Uses two twisted pair data lines.
- "Fairness" bus arbitration.
- Supported by MacOS and Windows 98.

A printed circuit board and DSP driver software would have to be developed using the TI chip set[8]. Shown below is a Firewire to EIDE Disk Block Diagram that might allow one Terabyte Per PCI Slot:



Conclusion

EIDE disk arrays are an inexpensive way to add large amounts of disk space to both single Workstations (and PCs) and multiprocessor computing farms. They provide an additional layer to the data storage "cake".



---

1 S. Bracker, K. Gounder, K. Hendrix, and D. Summers, *A Simple Multiprocessor Management System for Event-Parallel Computing*, IEEE NS-43 (1996) 2457.

2 http://www.quantum.com/products/hdd/bigfoot_tx/

3 http://www.storage.ibm.com/hardsoft/diskdrdl/desk/1614data.htm

4 http://www.promise.com/html/sales/Ultra33.html

5 http://www.apple.com/powermac/g3/

6 http://www.dell.com/products/dim/xpsr/specs/index.htm

7 http://www.skipstone.com/info.html

8 http://www.ti.com/sc/docs/news/1998/98029.htm;
http://www.ti.com/sc/docs/dsps/details/43/flash.htm;
http://www.ti.com/sc/docs/msp/1394/41lv0x.htm;
http://www.ti.com/sc/docs/storage/products/cont.htm