Multi-Terabyte EIDE Disk Arrays running Linux RAID5

by

University of Mississippi: David A. Sanders, Lucien M. Cremaldi, Vance Eschenburg, Romulus Godang, Michael D. Joy and Donald J. Summers and Fermilab: Donald L. Petravick

Presented at Computing in High Energy Physics 2004 (CHEP04) 27th September - 1st October, 2004, Congress Centre, Interlaken, Switzerland

Introduction

- $2000 per Terabyte Storage is Available
- 18 times ceaper than Sun StorEdge 6120
- Scalable for use at both Small and Large Institutions — From 2 TB to 250 TB, the same as a $ Million tape silo.
- 250 TB is 1 month of LHC data
- Redundant — RAID5
- Commodity Hardware

Hardware — Base System

- System Disk — 120 GB Western Digital
- 24" EIDE Cables
- CPU — Dual 2 GHz AMD Athlon
- Motherboard — MSI K7D Master MPX
- 1024 MB DDR memory
- Second Power Supply (15A at 12V)
- Gigabit Ethernet

RAID5 Test Box

8 Disk Array

12 Disk Array — upper bay

Results — Software RAID5

- Base write speed 29 MB/s
- Speed of 24 MB/s for 2 concurrent writes (17 % overhead)
- 37 MB/s read (using cp to system disk)
- 33 MB/s write (using cp from system disk)
- 10-15% CPU-overhead but on other CPU.

High Energy Physics Data Analysis Strategy

- Use Parallel Processing
- Split data and store on many RAID5 PCs
- Analysis for a subset of data takes place locally on the PC where the data resides
- Network is only used to combine results
- Or use NFS to mount RAID5 array on many PCs (Less efficient due to network overhead)

Definitions

- RAID — Redundant Array of Inexpensive Disks
- RAID level 0 — Concatenation
- RAID level 1 — Mirroring
- RAID level 4 — Parity
- RAID level 5 — Striped-Parity
- EIDE — Enhanced Integrated Drive Electronics

Additional Hardware

Software RAID5

- Eight 250 GB Western Digital disks or eight 300 GB Maxtor disks
- 2 Promise Ultra133 PCI cards

Hardware RAID5

- Twelve 250 GB Western Digital disks
- 3ware 12 disk RAID controller 7506-12

RAID5 Boxes for CERN and SLAC

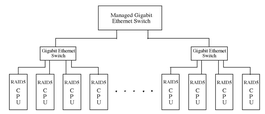
Results — Hardware RAID5

With 9 250 GB disks RAID5 (2 TB Array)
- Base write speed 41 MB/s

With RAID0 of two 1.2 TB RAID5 arrays (2.4 TB Array)
- Base write speed 29 MB/s
- Speed of 25 MB/s for 2 concurrent writes (14 % overhead)

With two 1.2 TB RAID5 arrays
- 33 MB/s write (using cp from system disk)
- 37 MB/s write (using cp from array1 to array2)
- 1-5% CPU-overhead (Journaling).

High Energy Physics Cluster



Why Use Commodity Hardware?

"Frankly sir, we're tired of being on the cutting edge of technology."

Disks

Disk	RPM	$/GB	GB/platter	Amps@12V	Warranty
300 GB Maxtor Maxline II (2MB Cache)	5400	0.75	75	0.59	3 Yrs.
250 GB Western Digital (8MB cache)	7200	0.61	80	0.43	3 Yrs.
250 GB Maxtor Maxline Plus II	7200	0.69	80	0.92	3 Yrs.
300 GB Maxtor Maxline III SATA	7200	0.77	100	0.63	3 Yrs.
200 GB Seagate Barracuda SATA	7200	0.63	100	1.13	5 Yrs.

Problems and Solutions

Problems	Solutions
Not enough power with stock power supply	Used second stock power supply
Could only use 2 Promise EIDE controller cards. (Not enough PIRQs)	Used Hardware RAID disk controller (3ware Escalade 7506-12)
3ware 7506-12 controller has 2 TB limit	Used combination of software RAID-0 of 2 Hardware RAID5 sets
2 TB size limit for 2.4 kernel	Used 2.6 kernel or only 8 disks
EIDE ribbon cables take too much space	Used Round EIDE cables

Results — Gigabit Network

- Base internal write speed 29 MB/s

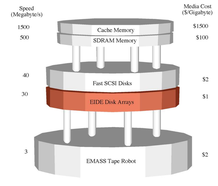
NFS
- 13 MB/s write (synchronous mount)
- 18 MB/s write (asynchronous mount)
- 23 MB/s write (using cp over asynchronous mount)
- 23-38% overhead — NFS+network

FTP
- 22 MB/s
- 24% network overhead

Grid Disk Cache

Caching avoids Grid-Lock

Data Storage Cake



Tests

- Measured speed and CPU-overhead of Software RAID5
- Measured speed and CPU-overhead of Hardware RAID5
- Measured speed of network file transfers
- Tested journaling file systems: ext3 and ReiserFS

Future Recommendations

Hardware:
- Use SATA Disks
- Use 3ware Escalade 9500S-12 (or 9500S-12MI) SATA card
- Use Hot-Swappable drives
- CPU — 2.0 GHz AMD Athlon (or better)
- Gigabit Ethernet Card/built-in
- Fiber Channel Arbitrated Loop (FCAL)
- Second Power Supply (15A at 12V)

Software:
- Use Linux Kernel 2.6
- Use ReiserFS Journaling File system

Commercial Systems

	Capacity	Size	Price*	Price/GB
Apple Xserve RAID	3.5 TB	3U	$10,999	$3.14
Dell EMC CX200	2.1 TB	3U	$18,999	$9.05
HP StorageWorks 1000	2.1 TB	3U	$23,925	$11.39
IBM FASt200 3542-1R	2.1 TB	3U	$51,895	$24.71
Sun StorEdge 6120	2.04 TB	Two 3U	$74,600	$36.57

*Based on suggested retail prices on December 10, 2003 From Apple document L301297A_XserveRAID_TO.pdf

Summary

- $2000 per Terabyte RAID5 arrays of EIDE Drives tested, without tape backup.
- They are Scalable —Cost less/TB than a tape silo, but scalable down to 2 TB.
- Uses Commodity Hardware.
- Tested Hardware and Software RAID with 250 and 300 GB hard disks
- Tested Gigabit networking

Supported by the U.S. Department of Energy under DE-FG05-91ER40622 and DE-AC02-76CH03000.