Low-cost BYO Mass Storage Project

James Cizek
Unix Systems Manager
Academic Computing and Networking Services
The Problem

- Reduced Budget
- Storage needs growing
- Storage needs changing (Tiered Storage)
- I NEED MORE DISK SPACE! (DBA’s!!)
- Current commercial offerings are not addressing this problem without major budget implications
Projected Needs (2009 Survey)

Research Data Storage Need (TBytes)

Now: 219
2 Years: 1,384
5 Years: 4,914

November 17, 2010

Subnet Managers Meeting
The Goal

- Find a mass storage solution that won’t break the bank
- CSU attempted NSF grant to meet this need ($1 million for 500 TB x 2), but were not awarded the grant (1,000 1 TB drives!!!)
- Vendors sell high-speed, costly systems (suitable for Amazon, Google, etc.), but we want slower, low-cost
  - Looking at vendor offerings, we decided to “roll our own”
- Maximize TB/$ with reasonable assurance that data are redundant and safe
Some Understandings

• Project approached as “Secondary” or “Tier 2” type storage, not intended to replace extremely fast, ultra-reliable, expensive disk systems

• Device management, support, and component failure need to be addressed
A starting point

- Online backup company “Backblaze” open-sourced their storage pod design, see https://www.backblaze.com/petabytes-on-a-budget-how-to-build-cheap-cloud-storage.html
- Starting with a proven design would eliminate many unknowns and speed up our design process
- Turned out to be helpful, but ran into many of our own headaches
The BackBlaze design
BackBlaze vs. CSU design goals

• Realized that the BackBlaze design didn’t exactly meet our requirements
• No redundant power supplies
• Cheap SATA cards didn’t take advantage of performance available by having large number of spinning hard drives
• Case too small to accommodate server-class motherboard
• Single “system” hard drive is single point of failure.
• Realized the need to over-engineer cooling and vibration reduction (2 major contributors to drive failure)
• Chassis was red instead of CSU green!
CSU design changes

- Lengthened case by 3 inches to accommodate dual CPU server-class motherboard
- Added more RAM for file system buffering (6 GB compared to BackBlaze 4GB)
- Added larger, redundant power supplies - individual supply can run entire case
- Used “Enterprise” grade drives instead of consumer grade, after much research
  - Drives selected have vibration sense / damping
- Replaced cheap SATA cards with high-performance PCI-e cards
CSU chassis nearing completion
CSU chassis nearing completion
Costs

• Case: $700
• 1 TB Drives: $100 x 45 ($4,500)
  – Drives were purchased earlier this year, now 1.5TB for $100
• Motherboard / Processors / Memory: $900
• Power Supplies: $200
• SATA cards: $300
• Ethernet card with iSCSI offload: $350
• SATA Multipliers: $45 x 9 ($405)
• Fans/Cables/Hardware/DVD/Mounts/etc.: $1,000

Total: 45 Raw TB for $8,355!

November 17, 2010
Subnet Managers Meeting
Testing Environment

• Testing was done with both small files and large files (Larger than largest memory buffer)

• Same data was used for all tests. Allowed us to validate results from various benchmark utilities against each other

• All RAID configurations were done in multiples of 3 to spread load across as many backplanes as possible

• All test data below assume worst case (Reads all random, writes all continuous)

• Highest recorded temperature (excluding CPU exhaust fan) under full load is 100F (ambient office temperature at input, should see even more improvement in datacenter)
Initial Performance

• Internal performance (using dd)
  – 11GB dataset using 18 drive Raid6:
    • Read: 472 MB/s
    • Write: 162 MB/s

• Over 4GB Fibre Channel connection
  – 11GB dataset using 18 drive Raid6:
    • Read: 115 MB/s
    • Write: 98 MB/s

• RAID sets less than 6 drives showed degraded performance, RAID sets above 18 drives showed only small performance benefits
Cost / Performance Comparison

• We are using IBM DS4300 and DS4700 Fibre channel disk systems as Tier 1 disk in the unix environment. These use 18U and nearly $100K

• We are using Equallogic (various models) iSCSI arrays for Tier 1 disk in windows environment. P6500E model hold 48 TB but runs near $80K

• We are using “Jetstor” SATA based products for Tier 2.
  – 16TB capacity for $8000 Although Fibre channel capable, have no ability to present disk space standalone (i.e. must be connected to a server)

• DIY disk box is 45 (67) TB for $8300 in 4U
Configuration

Much was learned during testing
  – RAID levels, 5 & 6 tested, 5 faster, but not enough to disregard the added safety of 6. 1 & 10 not considered
  – Operating system – Debian 64bit Server
  – Performance testing – unix DD, IOMeter, IOzone
    • Consistent data obtained from all tools
  – Connection offerings (Fibre, iSCSI, NFS, AOE)
    • Fibre
    • iSCSI
    • NFS (SLOW!!!)
    • AOE (Working out kinks)
Challenges ahead

• Support management (What happens when a disk fails?)
• Backup and protection of stored data
  – Mirroring units
  – Avoid backing up to enterprise backup system
• Data storage and protection policies
• Parallel file system
Where will this be useful?

- Library digital repository
- Research computing
- HPC, tier 2
- Campus wide “Cloud” storage
- Second or Third Tier storage for your Enterprise backups
- Email/File archiving
- Database “snapshots” kept for long term (LMS)
What other possibilities?

• Very large JBOD (Directly attached to server)
• Linux server offering CIFS/NFS
  – NAS capability
  – iSCSI target
  – Direct connection via FibreChannel (HPC)
• VMWare ESXi
  – Standalone VM cluster with massive attached storage
  – 4 U server running Windows/Linux/FreeNAS/OpenFiler
Where are we today?
Next steps at CSU

• Collection of final “parts list” for a complete build
• Documentation
• Put it into “semi production” and see how it performs under real-world situations
Resources

- [http://www.ctcustomfab.com](http://www.ctcustomfab.com) (Cases)
- [http://www.chyangfun.com](http://www.chyangfun.com) (SATA multipliers)

- [http://www.colostate.edu/curtisb/mass_stORAGE](http://www.colostate.edu/curtisb/mass_stORAGE) (Wiki on CSU progress)
Questions?

• james@colostate.edu