Green Dirt-Cheap-Storage

"Saving bits -- zillions at a time"

Fernando J. Pineda

Draft May 02. 2014
Dept. of Molecular Microbiology & Immunology
Dept. of Biostatistics
Director, Joint High Performance Computing Exchange
Johns Hopkins Bloomberg School of Public Health
Fernando.pineda@jhu.edu
www.pinedalab.org

Rationale for storage strategy

- Our Faculty were desperate for big storage, not too concerned about performance. We have ample local-scratch space on compute nodes.
- Unable to find storage vendors with PB-scale solutions for less than about $350/raw-TB.
- After cost, Power/cooling is THE limiting resource.
- Vendors and SysAdmins agree: DIY is too hard, too risky, too scary
  - But my position is that the potential savings are too compelling ($100/raw-TB compared to >$350/raw-TB)
  - At the PB scale the savings can pay for an FTE
- The technology strategy
  - ZFS file system (modern file system designed for 'unreliable' drives)
  - New breed of "small-NAS" drives. Low-power, slow random-access performance, ~$50/raw-TB, 7sec TLER, vibration control & slow speed.
- The finance strategy
  - Cost sharing with PIs, but satisfy OMB regulations
  - Up front buy-in and then very modest yearly charge

Risk Mitigation

- Talked extensively to groups that had done this before
  - Alainna White, JHU/Physics & Astronomy (now at Cray)
  - Scott Roberts JHU HLT/COE
  - Avi Berman, Buck Institute (now at BioTeam).
- Used the same system integrator (Seneca) that built the Physics and Astronomy system
  - Seneca integrated system and burned-in disks at their site
  - They shipped and installed on our site
  - Hardware warranty
  - We bought extra disks & spare power supply
- Brought in additional ZFS/Linux expertise via consulting contract with BioTeam
- Marvin Newhouse (our Computing system manager) managed the design and coordinated all the moving parts

The technology

nothing fancy or formal, mostly back-of-the-envelope considerations

ZFS

- Copy-on-write serializes random i/o. This has two (mostly forgotten) implications:
    1. The disk drive head moves less, so less power is consumed.
    2. We don't care if disk drive has poor random i/o performance on standard benchmarks because ZFS shifts the balance of i/o towards serial i/o.
- ZFS on linux
  - Our team knows linux
  - Good timing: first stable release was March 2013
  - Gain experience for future mad schemes: lustre

Small-NAS drives

- Small-NAS drives are a new sector in the Hard Drive Market -- but great for big NAS as well!
- Western digital Red drives
    1. 7 second TLER
    2. Vibration control
    3. Sucky random-access performance (We don't care)
    4. Low RPMs (~5400RPM?)
    5. 5400RPM + little arm motion because of ZFS = low power (~5W/TB).
    6. Low cost: $165 (currently $135) for 3TB drive.

The finances

The PI value proposition per formatted TB
$105 down-payment + $36/year

- Development cost for our first system (staff salaries)
  - Absorbed into rates $43,000
- Sold 580 formatted TB @ $105/TB
  - Some TB not sold and kept in reserve
  - Sponsored budgets $54,877.75
  - Non-sponsored Budgets $5249.96
- Service Center Capital Equipment purchase
  - 5 year recovery $57,544.46
- What we will charge the stakeholders on a yearly basis
  - Cap. Equip. recovery $57,544.46/580TB = $19.84/TB-year
  - All other yearly expenses = $16/TB-yr
  - Total charges = $36/TB-year
  - In our initial prospectus, we told stakeholders to expect charges of $50/TB-year so they are very happy!

Financial engineering

- Minimize load on the school's limited capital
  - Avoids entanglement in political competition for financial resources.
  - Speeds innovation, by escaping from yearly financial planning cycle.
  - Makes the Dean happy

- Stay out of Leavenworth by satisfying OMB regulations. We buy & recover the infrastructure, users buy their disks in proportion to the storage they need
  - Capital equipment (5 year recovery from fees) $56,590
  - 345 drives (PI Sponsored budgets, i.e. grants) $54,878
  - 33 drives (PI non-sponsored budgets) $5249

Estimated cost of our 1st storage device: $166K

$117,671	The system	
	Includes: on site Installation, warranty & spare parts	
$43,000	JHPCE personnel (estimate)	
	M. Newhouse	6 mo @33% effort
	J. Yang	6 mo @ 10% effort
	B. Mohr	3 mo @20% effort
	F. Pineda	6 mo @ 5% effort
$1,500	Consulting (estimates)	
$162,171	**Total cost of the prototype**	

DCS01 vs deeply discounted ZFS enterprise storage appliance from major vendor

	ZFS appliance	DCS01
Cost	$161,876	$162,171
Raw_TB	492	1080
Formatted_TB	394	670
$/Raw_TB	$329	$150*
$/formatted_TB	$411	$242
Power dissipation	4kW (?)	3.5kW
Watts/formatted-TB	10W/TB(?)	5.2W/TB

* $108/raw-TB exclusive of development costs

Conclusions: 3TB WD Red Drives are ideal

- < 1% failure rate in first 6 months of operation (after preproduction burn-in)
- Our experience consistent, if not better than, HD reliability analysis conducted by Backblaze of 27,134 consumer-grade drives*.
- 4TB WD Red drives would be even more cost effective, but need to be qualified.

* http://blog.backblaze.com/2014/01/21/what-hard-drive-should-i-buy/

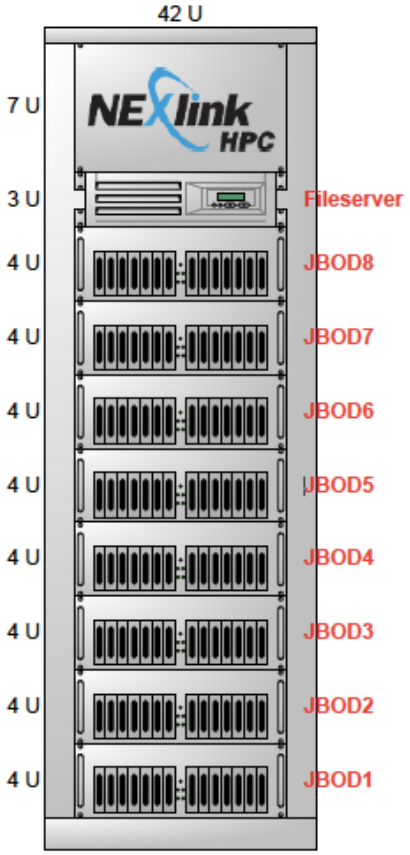
Conclusions

- Could now replicate similar PB-storage system for less than $100/raw-TB
- Half the power consumption of anything available from vendors (3.5kW rack).
- Significantly reduced impact on the school's capital budgets
- ZFS is the enabler!
  - Enables use of new class of "small-NAS" drives
  - 3TB WD Red drive: $165 when built ($135 as of May)
- System in production for 7 months.
  - 3 disk failures after initial burn-in
  - No performance issues
  - Nearly full
  - Stakeholders clamoring for another one!
- Our sense:
  - Storage is where linux clusters were 15 years ago
  - Only the brave venturing into building their own
  - Vendors have not caught up with "Big Data" needs of the Biomedical research community, in particular genomics
  - No vendors can (could?) provide 3.5kW/raw-Petabyte
  - A paradigm shift is coming!

Parts list

3U -- Dual Xeon File Server (256GB)

1	Seneca, Nexlink 3U 8-Bay Dual Xeon, 920W HSR 80+ PSU (3YR)
2	2.4GHz Intel Xeon E5-2665 8-Core 20MB Cache
16	16GB DDR3 1600MHz ECC Registered Memory (256GB total)
2	300GB 15K RPM SAS HDD (Raid 1 - OS Mirror)
4	100GB STEC ZeusIOPS Gen4 SAS SLC SSD (JBOD)
2	800GB STEC S840E eMLC SSD Drive (Raid 1)
6	3.5" to 2.5" Hotswap Tray Kit
1	Integrated Intel Dual Port GbE
2	Chelsio, Dual Port SFP+ 10Gbase-SR w/ Optical Transceivers N320E 2
2	LSI 9201-16e, 16-Port Ext SAS HBA PCIe 2.0 2
2	Internal 8087 SAS to 4-Port SATA Cable



8 x 4U -- 45-Bay JBOD - SATA

8	Nexlink 4U 45-Bay JBOD, 1400W HSR 80+ PSU (New Micro)
9	Internal 8087 to 8087 SAS Cable
1	1M 8088 to 8088 External SAS Cable
5	2M 8088 to 8088 External SAS Cable
2	4M 8088 to 8088 External SAS Cable

Rack & PDUs

1	APC, 42U Rack Cabinet, Wide
2	APC, 120/208V 3-Phase Input 208V Output, 30Amp, Switched
2	10M LC-LC Duplex 10Gb Multimode 50/125 OM3 Fiber Optic Patch Cable
18	2FT 208V C13 Power Cables 18

Disks

378	3TB WD30EFRX5400RPM HDD (Western Digital Red Drives, includes spares)