Dirt Cheap Storage (DCS)

Fernando J. Pineda

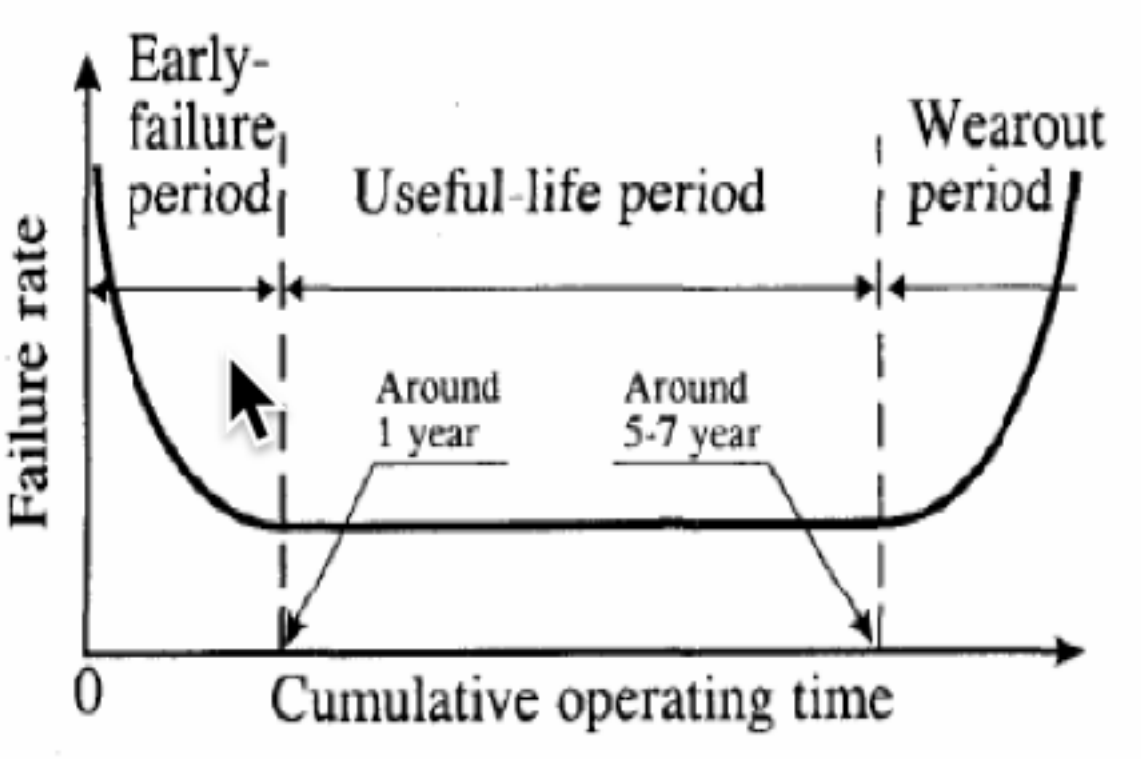
Sep. 9 2015
Update: July. 12, 2017
Dept. of Molecular Microbiology & Immunology
Dept. of Biostatistics
Director, Joint High Performance Computing Exchange
Johns Hopkins Bloomberg School of Public Health
Fernando.pineda@jhu.edu
www.pinedalab.org

- HDD considerations
- Hardware details
- Some financials

What do we really know about disk drive reliability and where can we get useful information?

Nothing formal or even particularly well thought out
--mostly just observations, our data and hypotheses

Life cycle failure pattern for hard drives



J. Yang and F.-B. Sun. "A comprehensive review of hard-disk drive reliability", in *Proc. of the Annual Reliability and Maintainability Symposium,* 1999.
Schroeder & Gibson, "DIsk failures in the real world: What does an MTTF of 1,000,000 mean to you?"
Fast '07: 5th USENIX Conference on File and Storage Technologies

A lot was learned in 2007 about HDD reliability

- The Google study (Pinheiro et al., 2007)
  - Disk MTBF numbers significantly understate failure rates
  - not much correlation between disk workload and failure rates
  - higher temperatures are not associated with higher failure rates.
- The CMU study (Schroeder & Gibson, 2007)
  - "Field replacement rates ...significantly larger than we expected based on datasheet MTTFs"
  - Young drives replacement rate 2-5 x expected, older drives (>5-8yrs) replacement rate a factor of 30 times higher than datasheet MTTF suggested.
  - "...a need for a better understanding of what disk failures look like in the field."
- Both studies...
  - agreed that datasheet MTTF (MTBF) underestimated replacement rates in the field
  - found replacement rates for "consumer" drives no worse than replacement rates for SCSI or FC disks"
- Can anyone point me to current studies?

Drive “failure” in the field is an ill-defined concept

- “Manufacturers and end-users often see different statistics...since they use different statistics when computing failures” (Pinheiro et al. 2007)
- “Since failures are sometimes the result of a combination of components...it is no surprise that a good number of drives that fail for a given user could still be considered operational in a different test harness” (Pinheiro et al. 2007)
- Data suggests that “... disk independent factors... affect replacement rates more than component specific factors” (Schroeder & Gibson, 2007)
- We could find no compelling reason to pay the premium for enterprise drives. Instead we chose Western Digital’s 3TB Red Drive.

Small-NAS drives

- In 2013 small-NAS drives were a new sector in the HDD Market. They looked ideal for ZFS.
- 3TB Western Digital Red drives (WD30EFRX)
    1. 7 second TLER enabled
    2. Designed for 24/7 operation
    3. Vibration damped
    4. Performance
        1. Slow: 5400RPM?
        2. Reasonable sequential-access performance * (in the "green drive" class)
        3. Sucky random-access performance * (But we were not too concerned because ZFS serializes writes)
    5. Low power consumption

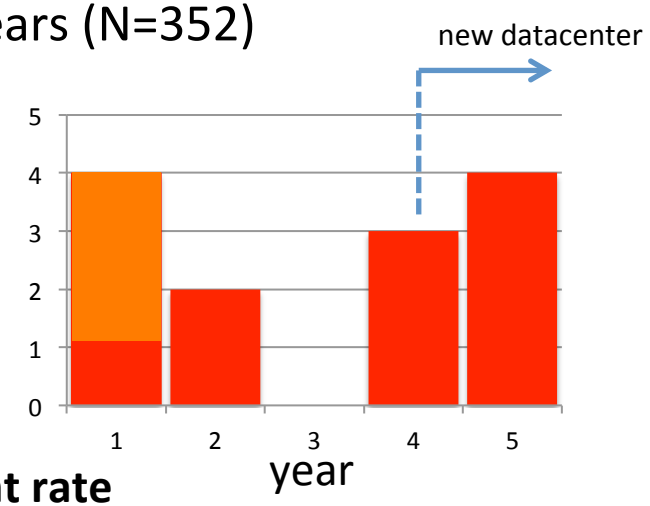
       5400RPM + reduced arm motion (because of ZFS) = low power (~5W/Drive).
    6. Cheap!

       $165 when we built it (currently < $109) .

* http://www.storagereview.com/western_digital_red_nas_hard_drive_review_wd30efrx

Reliability of WD 3TB Red drive (in our hands)

- Our drive replacement over 3 years (N=352)
  - 2013 4 = 1+ 3 (at integrator)
  - 2014 2
  - 2015 0
  - 2016 3
  - 2017 4



0.74% ± 0.2% annual replacement rate

- Backblaze replacement data (N 1024 - 1045)
  - Cumulative failure rate from Dec 31 2013 – March 31 2015 is 7.9%
  - Cumulative failure rate from Dec 31 2013 – March 31 2017 is 5.63% ± 0.8%
  - In our hands the WD 3TB drive is more reliable than Backblaze
  - Why?

Hard Drive Annualized Failure Rates

Reporting period April 2013 - March 31, 2017

MFG	Model	Size	Drive Count	Annualized Failure Rate	Confidence Interval	
					Low	High
HGST	HDS5C3030ALA630	3TB	4,380	0.84%	0.7%	1.0%
HGST	HDS723030ALA640	3TB	974	1.96%	1.5%	2.5%
Toshiba	DT01ACA300	3TB	46	3.89%	1.6%	8.0%
WDC	WD30EFRX	3TB	1,025	5.63%	4.8%	6.5%
HGST	HDS5C4040ALE630	4TB	2,624	0.88%	0.7%	1.1%
HGST	HMS5C4040ALE640	4TB	8,482	0.64%	0.5%	0.8%
HGST	HMS5C4040BLE640	4TB	15,339	0.68%	0.5%	0.8%
Seagate	ST4000DM000	4TB	34,540	3.00%	2.9%	3.1%
Seagate	ST4000DX000	4TB	170	7.51%	5.6%	9.8%
Toshiba	MD04ABA400V	4TB	146	1.50%	0.4%	3.8%
WDC	WD40EFRX	4TB	46	2.28%	0.5%	6.7%
Toshiba	MD04ABA500V	5TB	45	2.34%	0.3%	8.4%
Seagate	ST6000DX000	6TB	1,891	1.30%	0.9%	1.7%
WDC	WD60EFRX	6TB	443	5.59%	4.2%	7.2%
HGST	HUH728080ALE600	8TB	45	2.10%	0.3%	7.6%
Seagate	ST8000DM002	8TB	9,861	1.60%	1.2%	2.0%
Seagate	ST8000NM0055	8TB	2,459	2.38%	0.2%	7.4%
		Totals	82,516	2.07%	2.0%	2.1%

BACKBLAZE

Reliability of WD 3TB Red drive (in our hands)

- Perhaps it's just what the 2007 publications led us to expect: it's not the drive, so much as it's the environment/harness

	File system	JBOD	use case
JHU	ZFS	Supermicro 847E16-RJBOD1	HPC (life science)
Backblaze	custom	Backblaze POD	cloud storage

- Hypotheses
  - Not enough statistics?
  - Something about the JBODs?
  - Copy on write => less head motion => less vibration?
  - ZFS repairs errors rather than flagging them as irrecoverable?
  - Backblaze got a bad batch?

ZFS

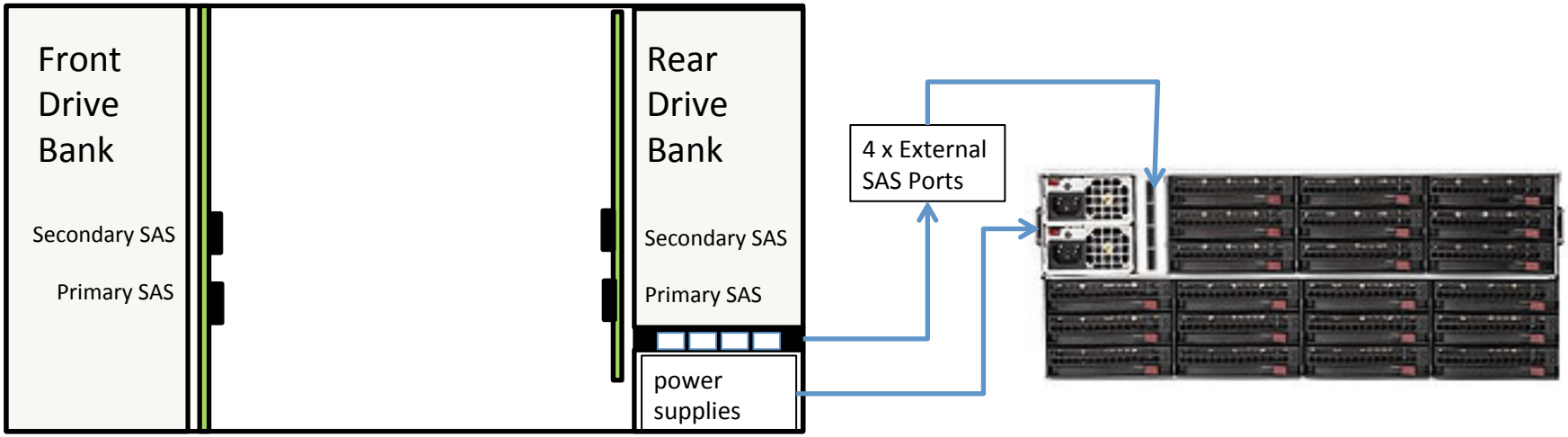
- Copy-on-write serializes random writes. This has several (likely) implications:
    1. The disk drive head moves less (at least on writes)
        - expect less power to be consumed
        - expect less wear and tear on the drive
    2. Poor random i/o performance on standard benchmarks is likely mitigated somewhat because ZFS copy-on-write shifts the balance of i/o towards serial

- Why we went with ZFS-on-linux in 2013?
    - Our team knew linux
    - Lucky timing: first stable release of ZFS-on-linux was March 2013
    - Wanted to gain experience for future mad schemes: lustre

Selected hardware details

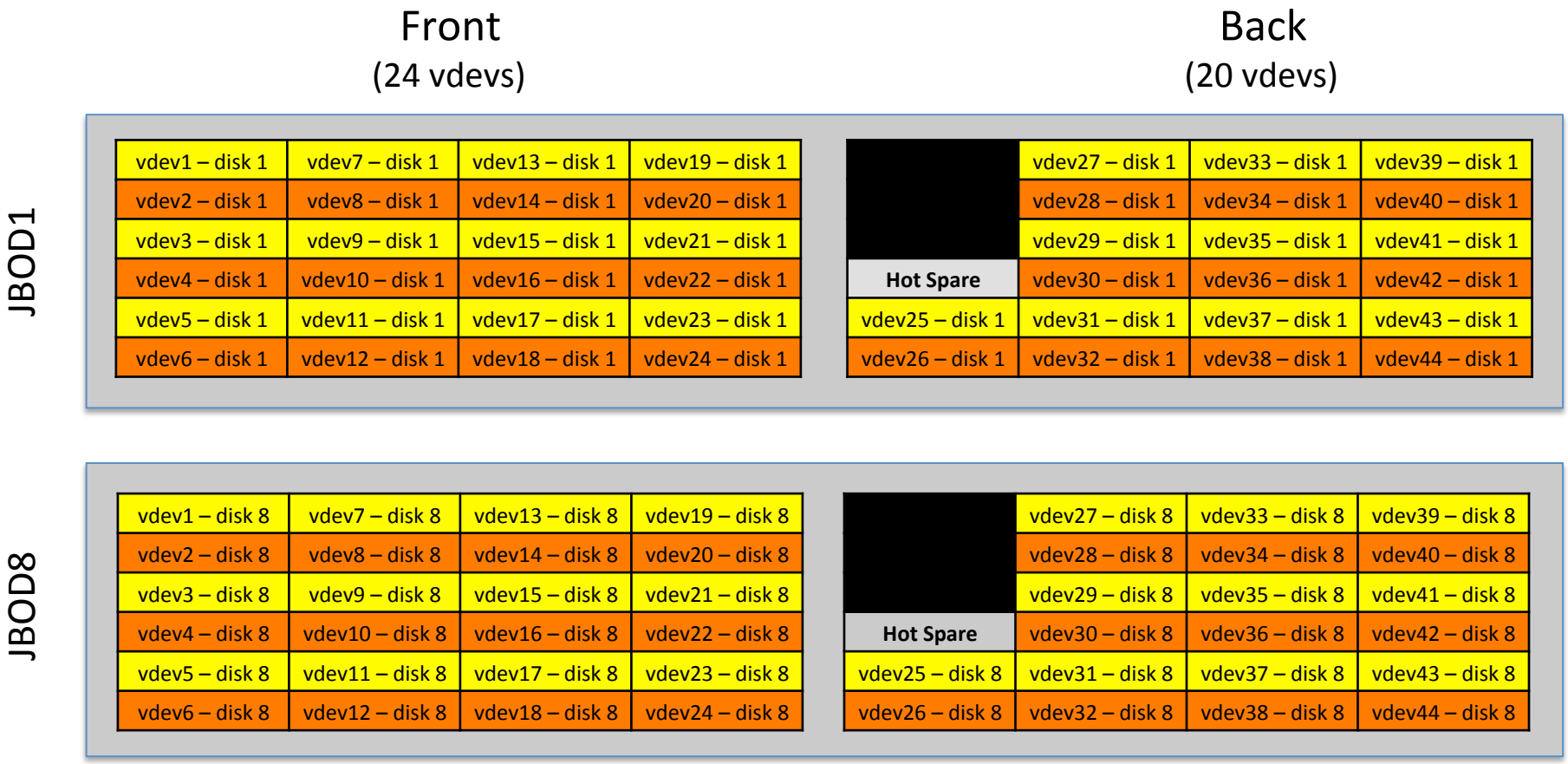
JBOD

SuperMicro SuperChassis 847E16-RJBOD1



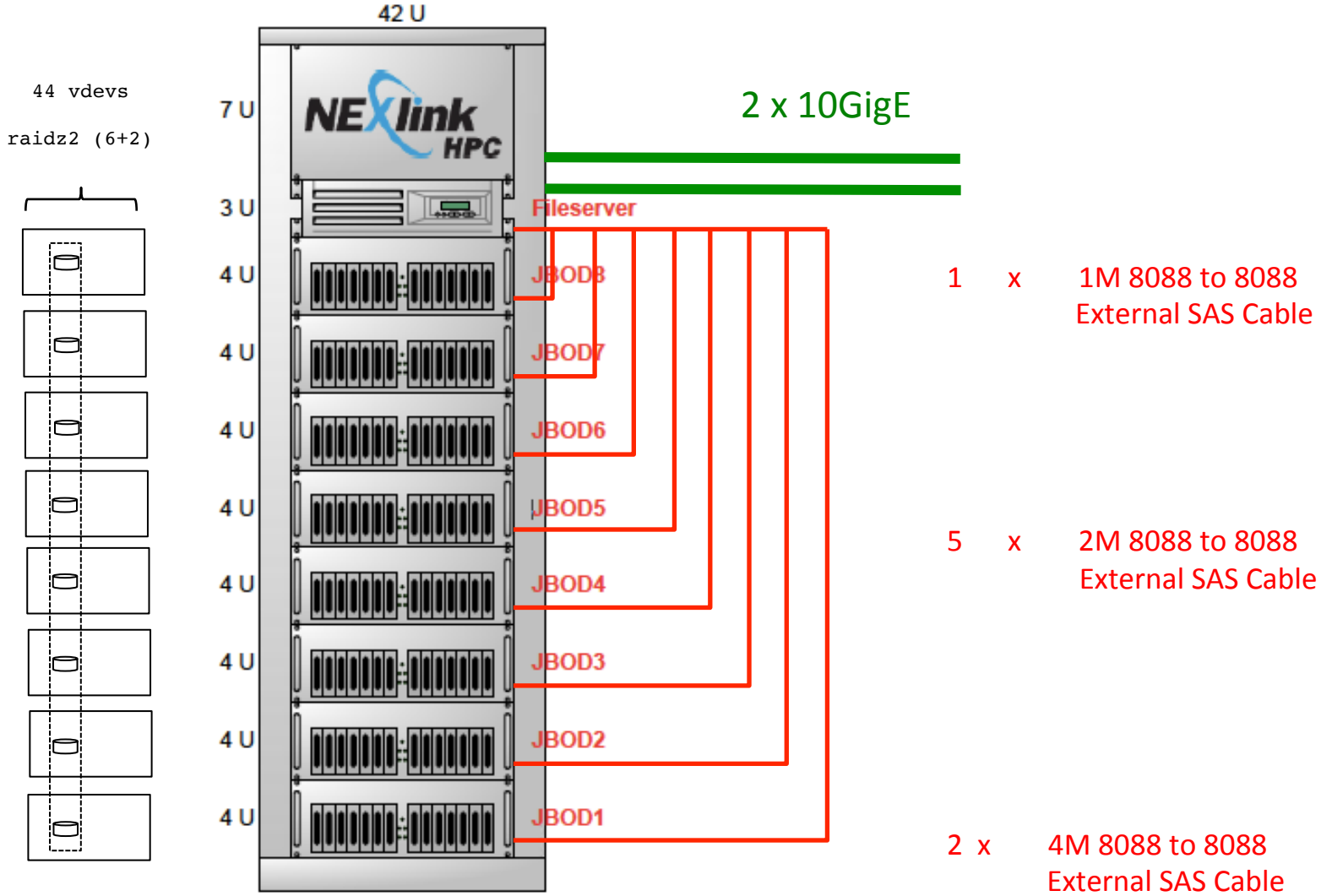
JBOD disk/vdev layout

SuperMicro SuperChassis 847E16-RJBOD1



- 44 vdevs total, 1056TB RAW, 688TB formatted
- vdevs – raidz2 {6+2} , 24 raw-TB/vdev (can lose up to 2 disks per vdev w/o data loss)
- Can lose up to 2 JBODS w/o data loss

DCS01 JBOD to server connections



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

$117,671	The system Includes: on site Installation, warranty & spare parts
$43,000	Development (3.6 man-months)
$1,500	Consultants
$162,171	**Total cost of the prototype**

DCS01 vs deeply discounted ZFS enterprise storage appliance from major vendor (ca. 2013)

	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

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

- Development cost for our first system (staff salaries)
  - Absorbed into current 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!

DCS01 Parts list (2013)

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

1	Seneca, Nexlink 3U 8-Bay Dual Xeon, 920W HSR 80+ PSU (3YR) (Super Micro - SYS-6037R-72RF)
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

8 x 4U -- 45-Bay JBOD, HBA & Cables

8	Nexlink 4U 45-Bay JBOD, 1400W HSR 80+ PSU (Super Micro – CSE-847E16-RJBOD1)
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
2	LSI 9201-16e, 16-Port Ext SAS HBA PCIe 2.0 2
2	Internal 8087 SAS to 4-Port SATA 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	Western Digital 3TB Red drive (WD30EFRX5400RPM HDD) (352 spinning + 8 hot spares + 18 boxed spares)

Conclusions

- Half the power consumption of anything available from vendors (3.5kW rack).
- ZFS is the enabler
  - ZFS likely contributes to low drive replacement rate
  - ZFS Enables use of of “small-NAS” drives

- System in production for 4 years
  - 13 disk failures after initial burn-in
  - Lost a JBOD in 2015
  - Few performance issues, (after ZFS update it stopped tipping over when hit by large multithreaded i/o jobs from the cluster)
  - Stakeholders clamored for more so we built DCL