

Dirt Cheap Lustre (DCL)

Fernando J. Pineda

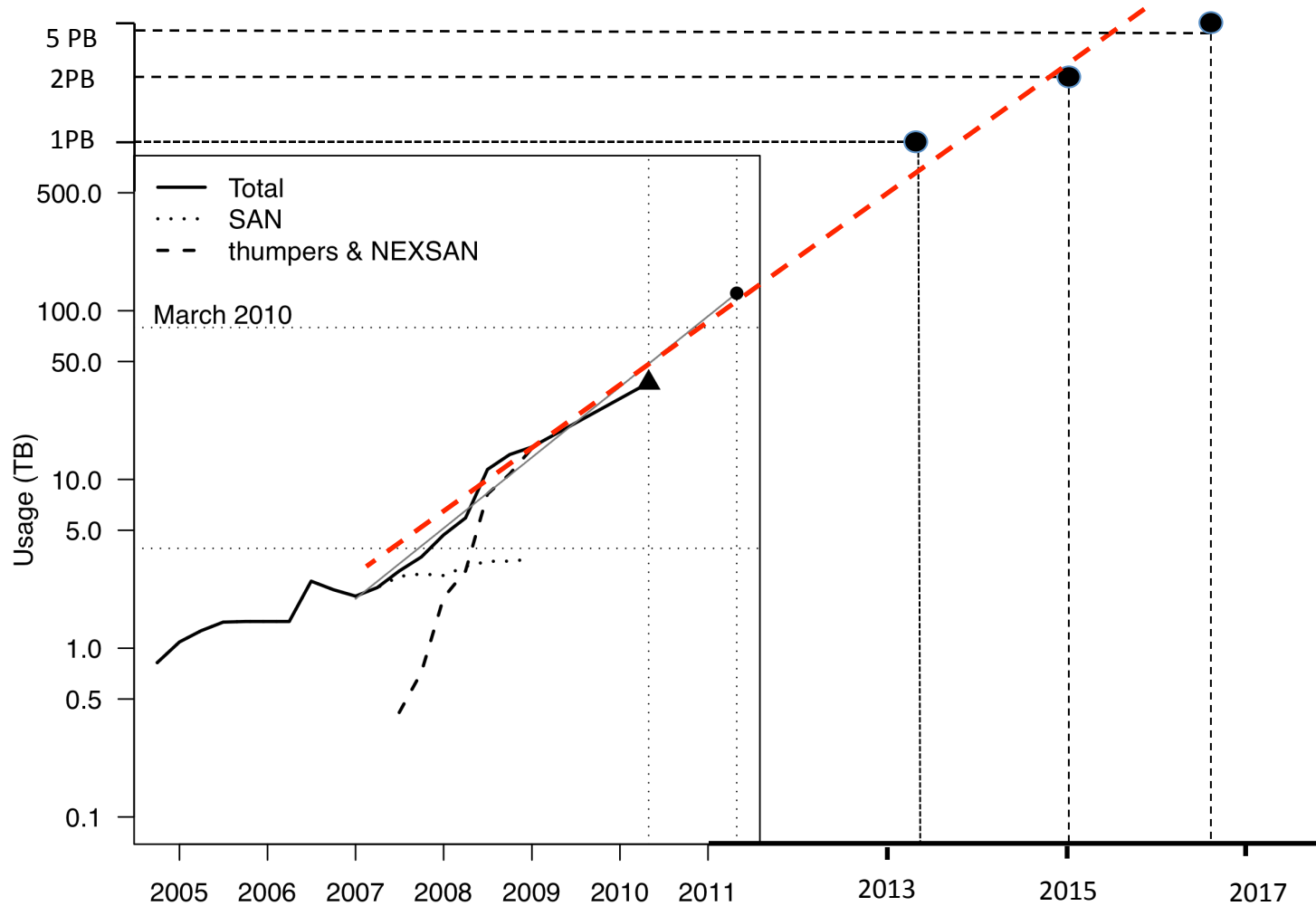
July. 14, 2017

Dept. of Molecular Microbiology & Immunology
Dept. of Biostatistics, Division of Medical Informatics
Director, Joint High Performance Computing Exchange
Johns Hopkins Bloomberg School of Public Health

Fernando.pineda@jhu.edu

www.pinedalab.org

Our storage capacity doubles every 12-18 months (since 2007)

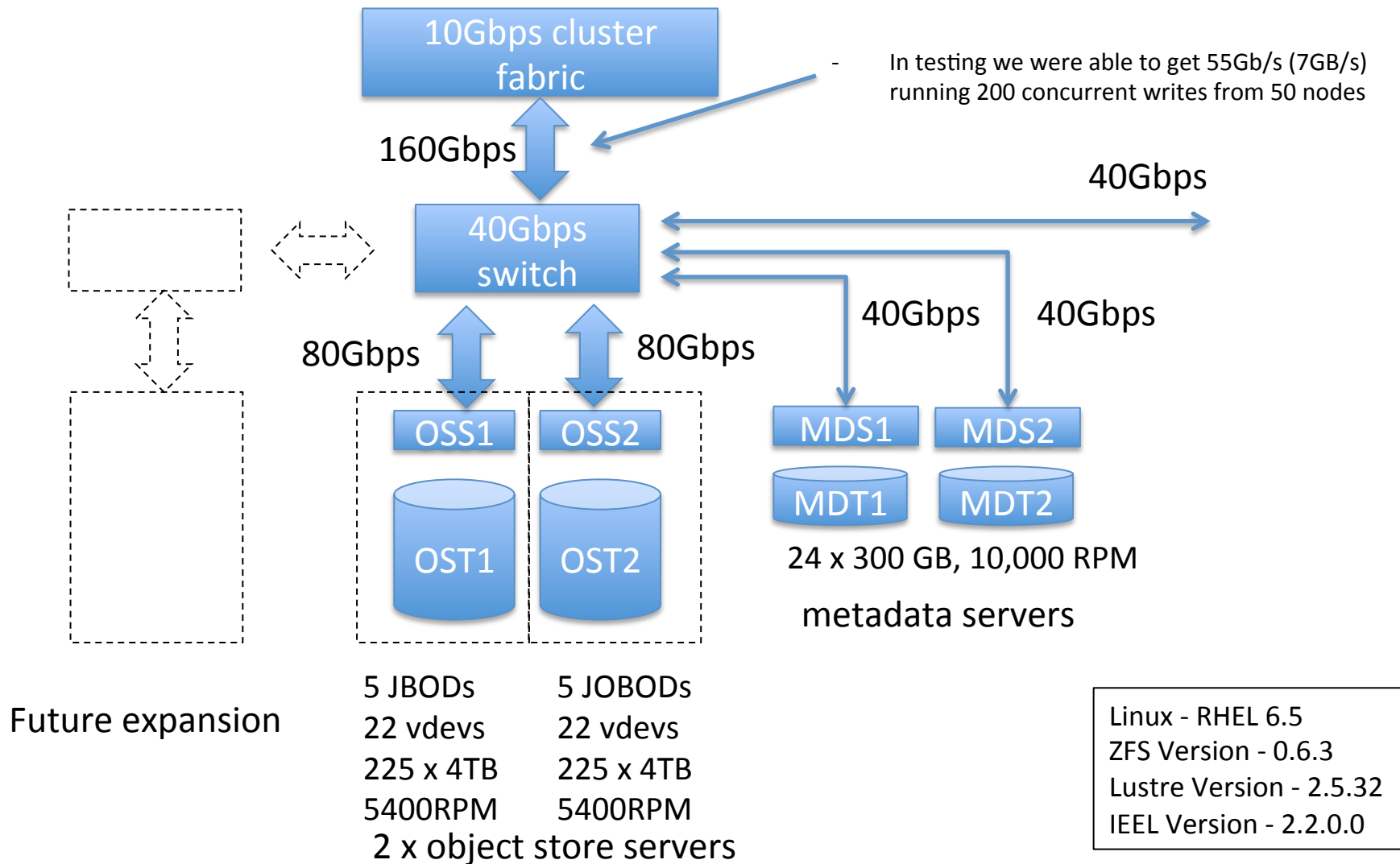


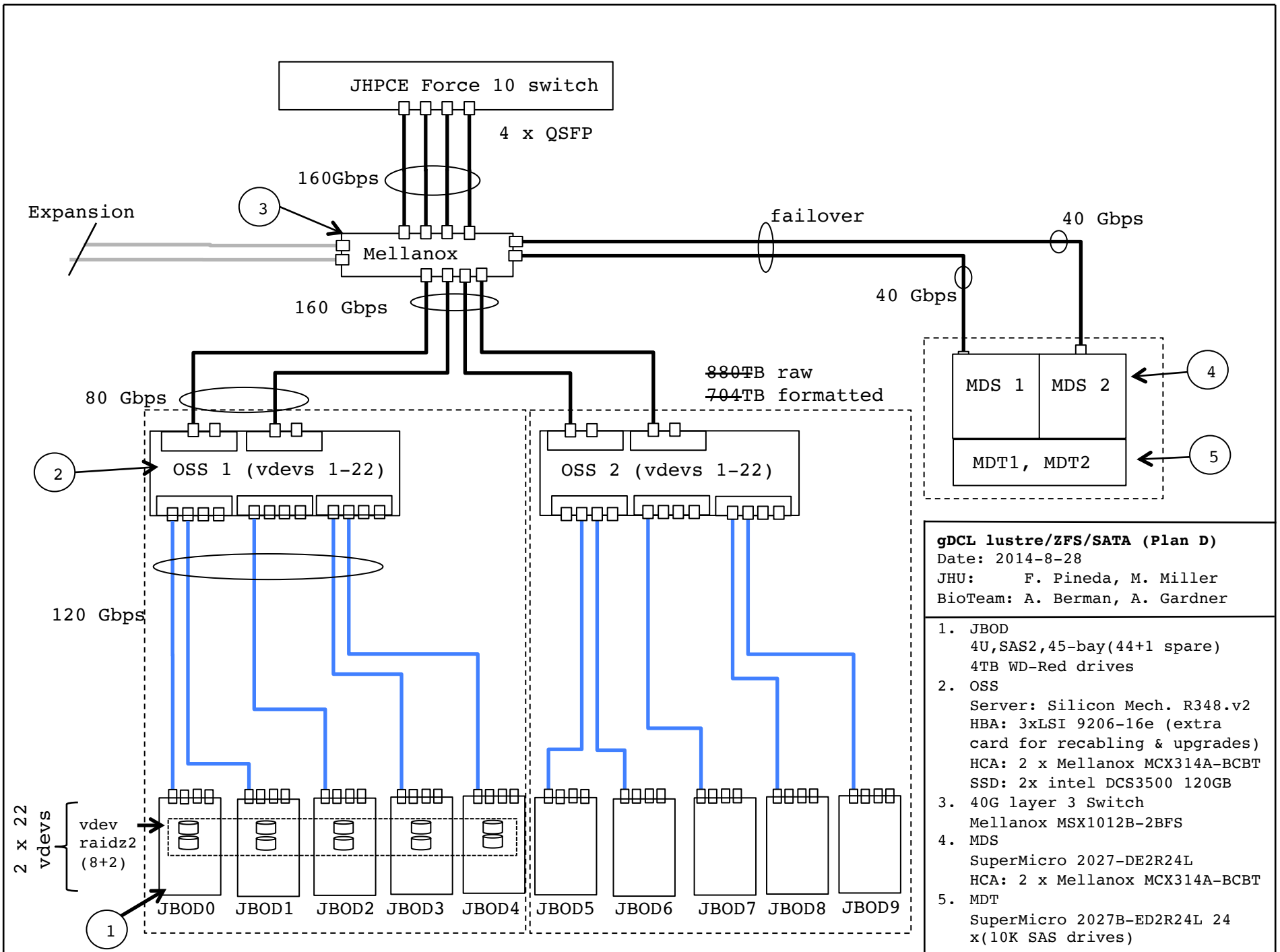
design considerations

- based on performance as the dependent variable
- optimize scientific throughput
- Capacity trumps performance
- Scalability
- one-9 so (HA not a priority)
- Fit the business model

Lustre system overview

BioTeam/PinedaLab





- gDCL lustre/ZFS/SATA (Plan D)**
 Date: 2014-8-28
 JHU: F. Pineda, M. Miller
 BioTeam: A. Berman, A. Gardner
1. JBOD
 4U, SAS2, 45-bay (44+1 spare)
 4TB WD-Red drives
 2. OSS
 Server: Silicon Mech. R348.v2
 HBA: 3xLSI 9206-16e (extra card for recabling & upgrades)
 HCA: 2 x Mellanox MCX314A-BCBT
 SSD: 2x intel DCS3500 120GB
 3. 40G layer 3 Switch
 Mellanox MSX1012B-2BFS
 4. MDS
 SuperMicro 2027-DE2R24L
 HCA: 2 x Mellanox MCX314A-BCBT
 5. MDT
 SuperMicro 2027B-ED2R24L 24 x(10K SAS drives)

We use WD 4TB HDD

- WD 3TB Reds performing very well in DCS01, so we chose WD 4TB for DCL

	JHPCE	Backblaze
File system	ZFS	EXT4 + no raid+proprietary Reed Solomon encoding+sharding
JBOBs	SuperMicro SuperChassis 847E16-RJBOD1	storage-pod chassis
use-case	cluster storage	cloud storage for backup
HDD	Western Digital Red 3 TB (WDC WD30EFRX)	Western Digital Red 3 TB (WDC WD30EFRX)
N	352	1045
replacement rate (2013-2017)	1%	5.6%

DCL01 JBOD/vdev Layout

Front
(24 vdevs)

Back
(20 vdevs)

JBOD1

vdev1 – disk 1	vdev4 – disk 1	vdev7 – disk 1	vdev10 – disk 1
vdev1 – disk 2	vdev4 – disk 2	vdev7 – disk 2	vdev10 – disk 2
vdev2 – disk 1	vdev5 – disk 1	vdev8 – disk 1	vdev11 – disk 1
vdev2 – disk 2	vdev5 – disk 2	vdev8 – disk 2	vdev11 – disk 2
vdev3 – disk 1	vdev6 – disk 1	vdev9 – disk 1	vdev12 – disk 1
vdev3 – disk 2	vdev6 – disk 2	vdev9 – disk 2	vdev12 – disk 2

	vdev14 – disk 1	vdev17 – disk 1	vdev20 – disk 1
	vdev14 – disk 2	vdev17 – disk 2	vdev20 – disk 2
	vdev15 – disk 1	vdev18 – disk 1	vdev21 – disk 1
Hot Spare	vdev15 – disk 2	vdev18 – disk 2	vdev21 – disk 2
vdev13 – disk 1	vdev16 – disk 1	vdev19 – disk 1	vdev22 – disk 1
vdev13 – disk 2	vdev16 – disk 2	vdev19 – disk 2	vdev22 – disk 2

JBOD8

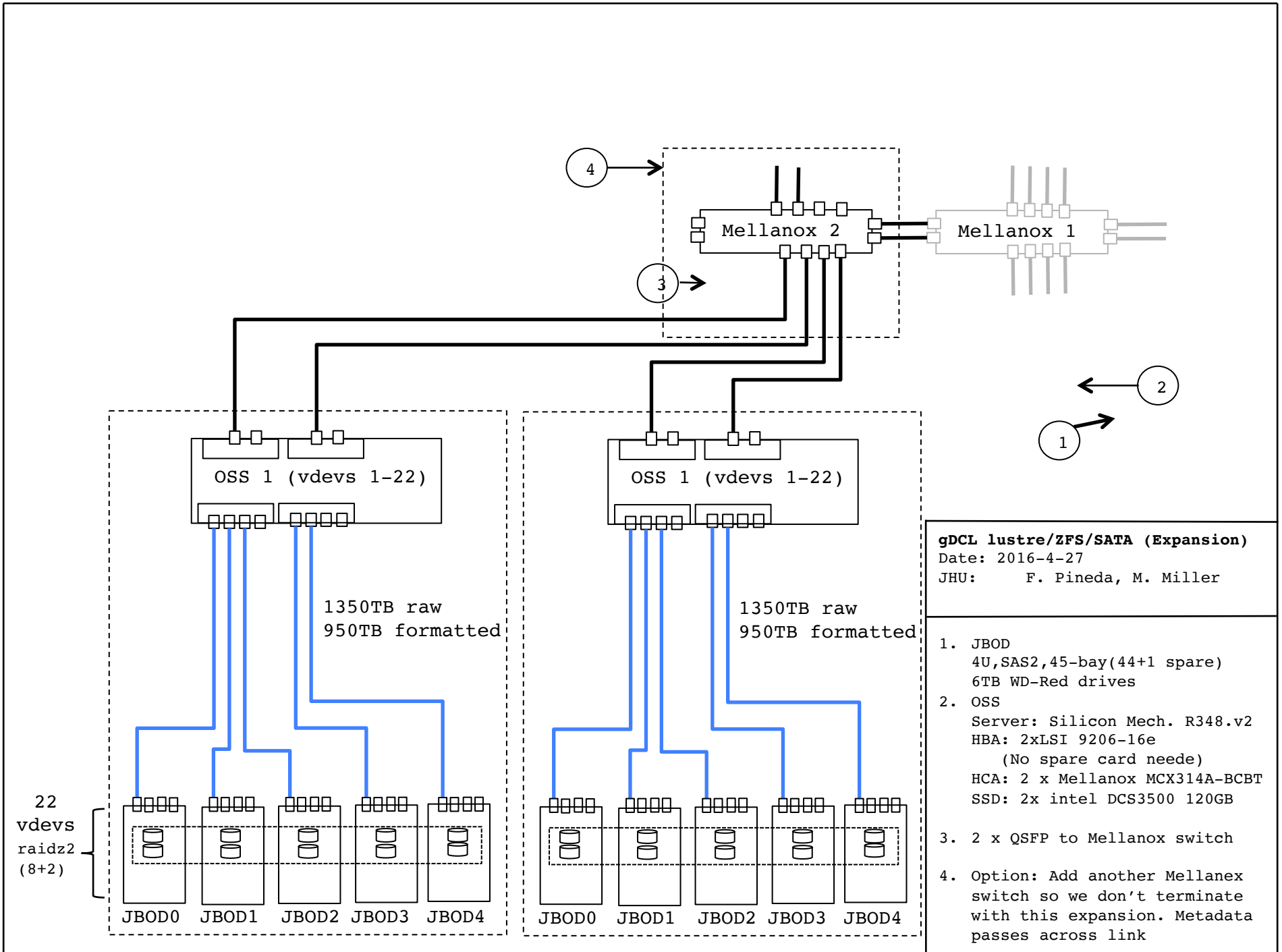
vdev1 – disk 9	vdev4 – disk 9	vdev7 – disk 9	vdev10 – disk 9
vdev1 – disk 10	vdev4 – disk 10	vdev7 – disk 10	vdev10 – disk 2
vdev2 – disk 9	vdev5 – disk 9	vdev8 – disk 9	vdev11 – disk 9
vdev2 – disk 10	vdev5 – disk 10	vdev8 – disk 10	vdev11 – disk 2
vdev3 – disk 9	vdev6 – disk 9	vdev9 – disk 9	vdev12 – disk 9
vdev3 – disk 10	vdev6 – disk 10	vdev9 – disk 10	vdev12 – disk 10

	vdev14 – disk 9	vdev17 – disk 9	vdev20 – disk 9
	vdev14 – disk 2	vdev17 – disk 2	vdev20 – disk 2
	vdev15 – disk 9	vdev18 – disk 9	vdev21 – disk 9
Hot Spare	vdev15 – disk 2	vdev18 – disk 2	vdev21 – disk 2
vdev13 – disk 9	vdev16 – disk 9	vdev19 – disk 9	vdev22 – disk 9
vdev13 – disk 10	vdev16 – disk 10	vdev19 – disk 10	vdev22 – disk 10

- vdevs – raidz2, (8+2) , 10 disks/vdev, 40TB spinning, 32TB *
- 22 vdevs total, 880TB spinning, 704TB *
- * unformatted

Cost?

- \$165,527 for 1.2PB = \$138/TB
 - \$89,776 servers, 40G switch, etc.
 - \$77,280 460 drives (WD Red 4TB)
- Value proposition to users:
 - \$90/TB up-front buy-in (10G min)*
 - < \$50/TB-year fees (mostly salaries)



gDCL lustre/ZFS/SATA (Expansion)

Date: 2016-4-27

JHU: F. Pineda, M. Miller

1. JBOD
4U, SAS2, 45-bay (44+1 spare)
6TB WD-Red drives
2. OSS
Server: Silicon Mech. R348.v2
HBA: 2xLSI 9206-16e
(No spare card needed)
HCA: 2 x Mellanox MCX314A-BCBT
SSD: 2x intel DCS3500 120GB
3. 2 x QSFP to Mellanox switch
4. Option: Add another Mellanex switch so we don't terminate with this expansion. Metadata passes across link

DCL01 JBOD/vdev Layout

Front
(24 vdevs)

Back
(20 vdevs)

JBOD1

vdev1 – disk 1	vdev4 – disk 1	vdev7 – disk 1	vdev10 – disk 1		vdev14 – disk 1	vdev17 – disk 1	vdev20 – disk 1
vdev1 – disk 2	vdev4 – disk 2	vdev7 – disk 2	vdev10 – disk 2		vdev14 – disk 2	vdev17 – disk 2	vdev20 – disk 2
vdev2 – disk 1	vdev5 – disk 1	vdev8 – disk 1	vdev11 – disk 1		vdev15 – disk 1	vdev18 – disk 1	vdev21 – disk 1
vdev2 – disk 2	vdev5 – disk 2	vdev8 – disk 2	vdev11 – disk 2	Hot Spare	vdev15 – disk 2	vdev18 – disk 2	vdev21 – disk 2
vdev3 – disk 1	vdev6 – disk 1	vdev9 – disk 1	vdev12 – disk 1	vdev13 – disk 1	vdev16 – disk 1	vdev19 – disk 1	vdev22 – disk 1
vdev3 – disk 2	vdev6 – disk 2	vdev9 – disk 2	vdev12 – disk 2	vdev13 – disk 2	vdev16 – disk 2	vdev19 – disk 2	vdev22 – disk 2

JBOD8

vdev1 – disk 9	vdev4 – disk 9	vdev7 – disk 9	vdev10 – disk 9		vdev14 – disk 9	vdev17 – disk 9	vdev20 – disk 9
vdev1 – disk 10	vdev4 – disk 10	vdev7 – disk 10	vdev10 – disk 2		vdev14 – disk 2	vdev17 – disk 2	vdev20 – disk 2
vdev2 – disk 9	vdev5 – disk 9	vdev8 – disk 9	vdev11 – disk 9		vdev15 – disk 9	vdev18 – disk 9	vdev21 – disk 9
vdev2 – disk 10	vdev5 – disk 10	vdev8 – disk 10	vdev11 – disk 2	Hot Spare	vdev15 – disk 2	vdev18 – disk 2	vdev21 – disk 2
vdev3 – disk 9	vdev6 – disk 9	vdev9 – disk 9	vdev12 – disk 9	vdev13 – disk 9	vdev16 – disk 9	vdev19 – disk 9	vdev22 – disk 9
vdev3 – disk 10	vdev6 – disk 10	vdev9 – disk 10	vdev12 – disk 10	vdev13 – disk 10	vdev16 – disk 10	vdev19 – disk 10	vdev22 – disk 10

- vdevs – raidz2 (8+2), 10 disks/vdev, 60TB spinning, 48 TB*
- 22 vdevs total, 1320TB spinning, 1056TB *
- * formatted

1.9PB Expansion

- \$171,742 for 1.9PB = \$90/TB
- Proposition to users:
 - 1900.80 usable capacity
 - \$86.67 buy-in per TB
 - \$55.19 annual operating expenses (salaries+service contract)
 - \$72.53 Annual TCO (annual operating expenses + 1/5 of buy-in)
- Pre-fab sold 1.2PB to stakeholders
 - unsold 100TB for fast scratch
 - unsold 600TB for future stakeholders*

Take aways

- Purchases driven by PI demand. No reading of tea leaves
- Design driven by
 - Business model
 - Capacity trumps performance
 - one-9 availability

- Storage Vendors not selling the systems that we need

- Disk manufactures not providing the disks we dream of:

WD Red 5400RPM with SAS interface

Disk vendors could build it cheaply, but ... probably they won't.