Lustre / ZFS at Indiana University

HPC-IODC Workshop, Frankfurt

June 28, 2018

Stephen Simms

Manager, High Performance File Systems

ssimms@iu.edu

Tom Crowe

Team Lead, High Performance File Systems

thcrowe@iu.edu

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

Greetings from Indiana University

Indiana University has 8 Campuses

Fall 2017 Total Student Count 93,934

IU has a central IT organization

- UITS
- 1200 Full / Part Time Staff
- Approximately $184M Budget
- HPC resources are managed by Research Technologies division with $15.9M Budget

Lustre at Indiana University

2005 – Small HP SFS Install, 1 Gb interconnnects

2006 – Data Capacitor (535 TB, DDN 9550s, 10Gb)

2008 – Data Capacitor WAN (339 TB, DDN 9550, 10Gb)

2011 – Data Capacitor 1.5 (1.1 PB, DDN SFA-10K, 10Gb)

PAINFUL MIGRATION using rsync

2013 – Data Capacitor II (5.3 PB, DDN SFA12K-40s, FDR)

2015 – Wrangler with TACC (10 PB, Dell, FDR)

2016 – Data Capacitor RAM (35 TB SSD, HP, FDR-10)

2016 DC-WAN 2

We wanted to replace our DDN 9550 WAN file system

- 1.1 PB DC 1.5 DDN SFA10K
  - 5 years of reliable production, now aging
- Recycled our 4 x 40 Gb Ethernet OSS nodes
- Upgrade to Lustre with IU developed nodemap
  - UID mapping for mounts over distance

We wanted to give laboratories bounded space

- DDN developed project quota were not available
- Used a combination of ZFS quotas and Lustre pools

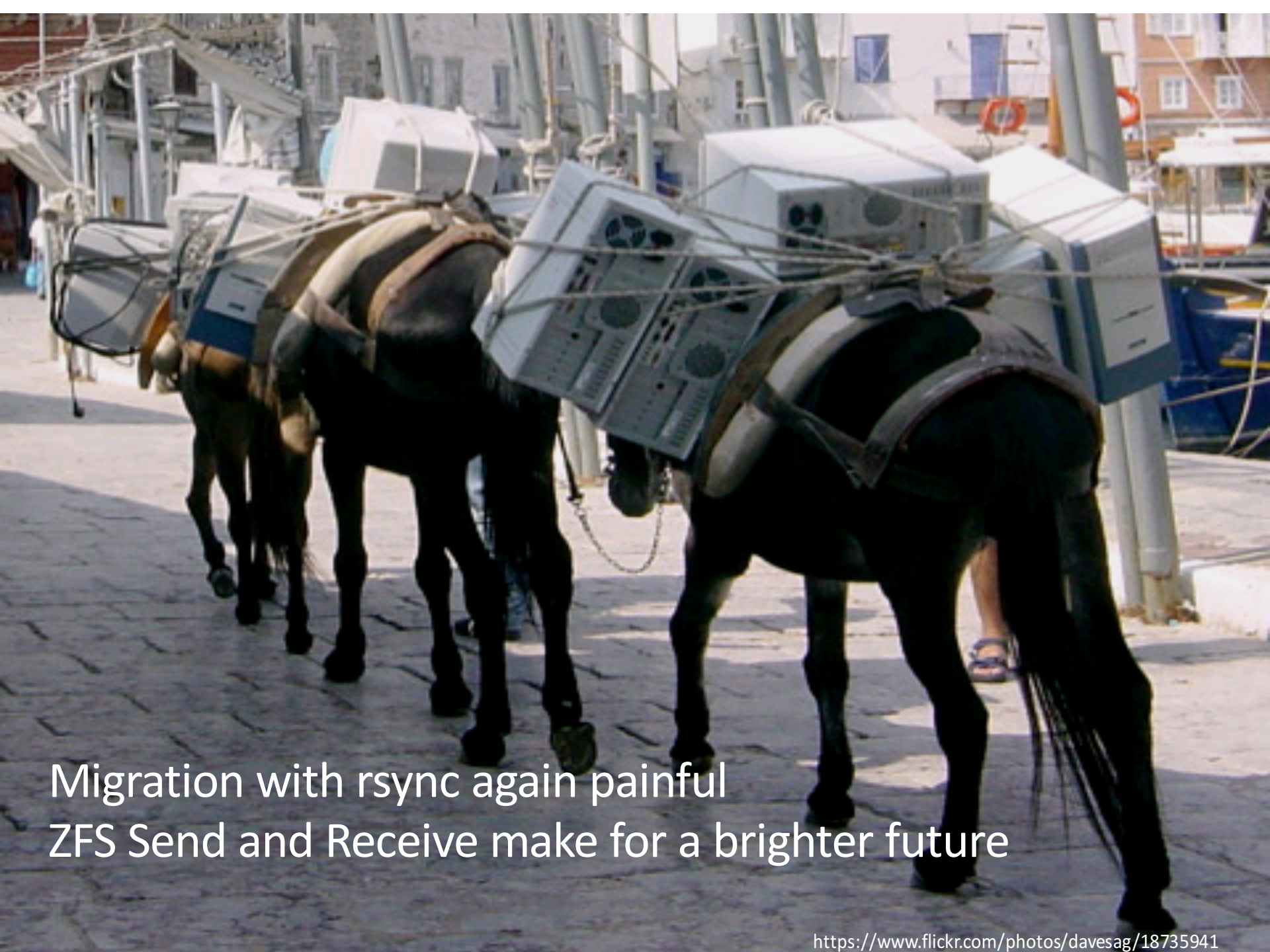
We wanted some operational experience with Lustre / ZFS

Why ZFS?

- Extreme Scaling
  - Max Lustre File System Size
    - LDISKFS – 512 PB
    - ZFS – 8 EB
- Snapshots
  - Supported in Lustre version 2.10
- Online Data Scrubbing
  - Provides error detection and handling
  - No downtime for fsck
- Copy on Write
  - Improves random write performance
- Compression
  - More storage for less hardware

Lessons Learned

- At that point metadata performance using ZFS was abysmal
  - We used LDISKFS for our MDT
- Multi-mount protection is a good thing – became available in ZFS 0.7
  - We would use manual failover
  - DDN SFA10K presentation feature kept us safe
- lz4 compression works without significant performance loss
  - Some users flustered when the same file shows different sizes
  - ls -l uncompressed size
  - du shows compressed size
  - du --apparent-size shows uncompressed size



https://www.flickr.com/photos/davesag/18735941

Need for Persistent High Performance Storage

- IU Grand Challenges
  - $300M over 5 years for interdisciplinary work to address issues
    - Addiction
    - Environmental Change
    - Precision Health

- From DC-WAN 2
  - Users liked having high performance storage without a purge policy
  - File system for individual use to be called Slate
  - Project and chargeback space to be called Condo

- IU issued an RFP for a Lustre / ZFS file system
  - After lengthy evaluation, IU chose DDN
  - As partners we have worked to create DDN's EXAScaler ZFS

Slate and Condo Specs

Slate

Lustre 2.10.3

4 PB RAW

2 MDS (active – active)

- 3.2 GHz / 512 GB RAM

2 MDT

- 10 x 1.92 TB SSD

8 OSS (configured as failover pairs)

- 2.6 GHz / 512 GB RAM
- 1 Mellanox ConnectX-4 card
- Dual 12 Gb SAS / enclosure

16 OST

- 40 x 8 TB drives

Condo

Lustre 2.10.3

8 PB RAW

2 MDS (active – active)

- 3.2 GHz / 512 GB RAM

2 MDT

- 10 x 1.92 TB SSD

16 OSS (configured as failover pairs)

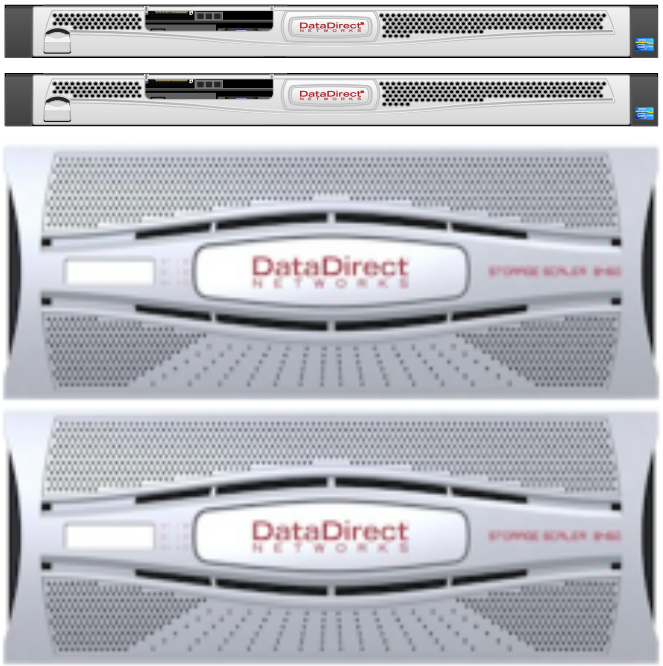
- 2.6 GHz / 512 GB RAM
- 1 Mellanox ConnectX-4 card
- Dual 12 Gb SAS / enclosure

32 OST

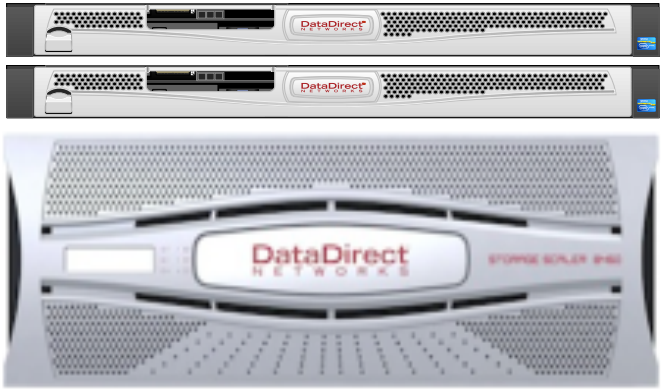
- 40 x 8 TB drives

Functional Units or Building Blocks

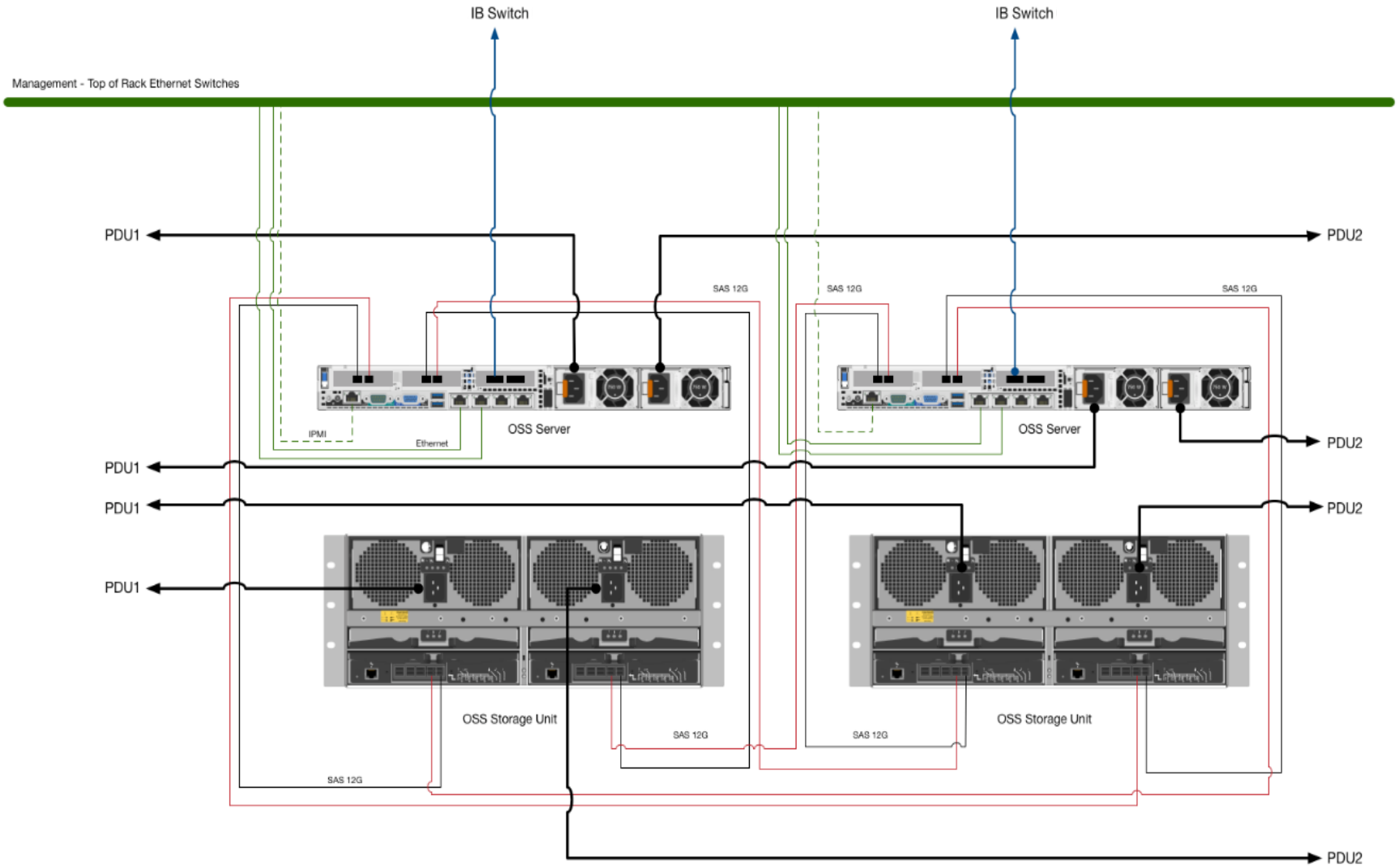
OSS



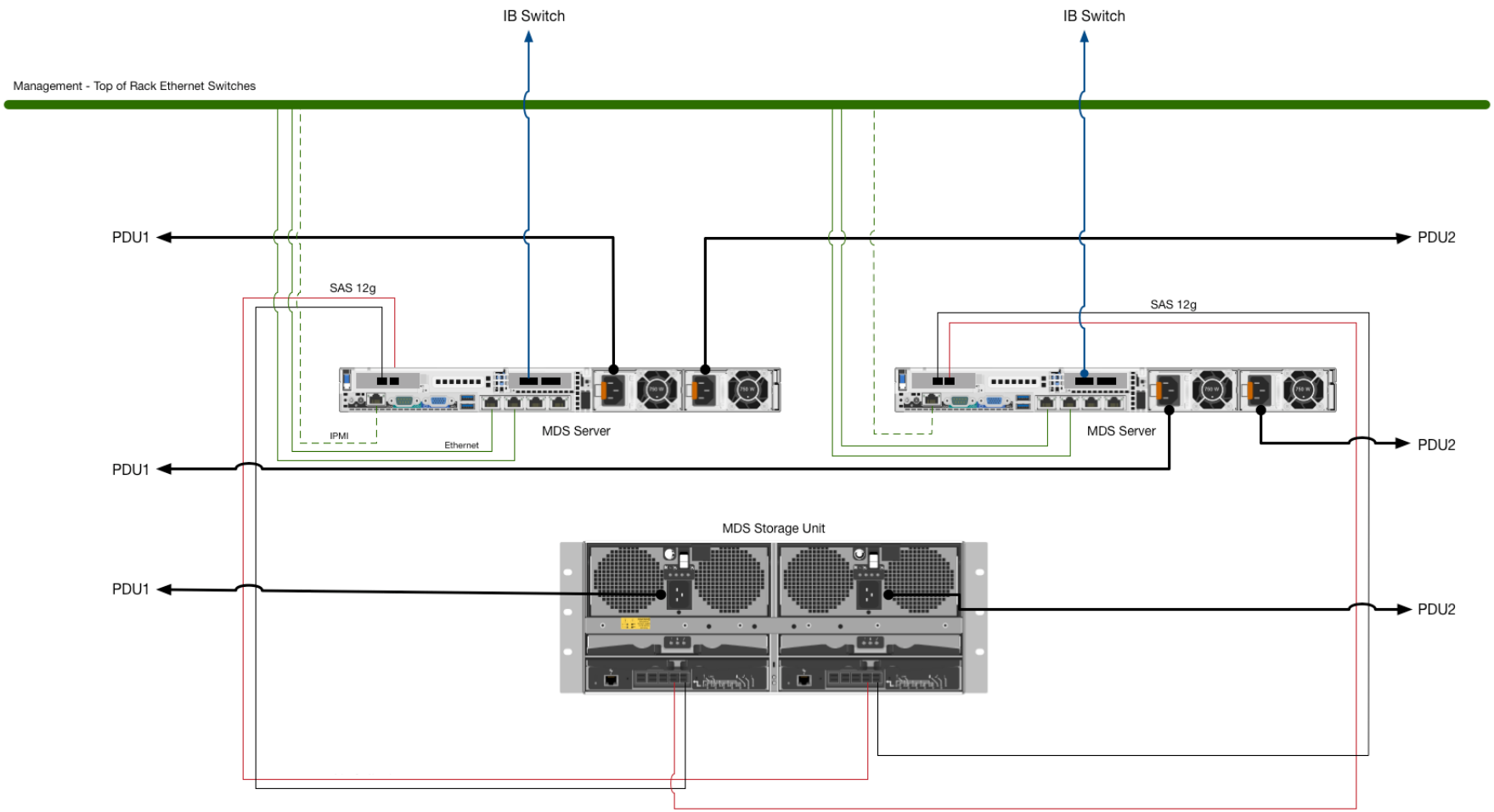
MDS



OSS Building Block Details

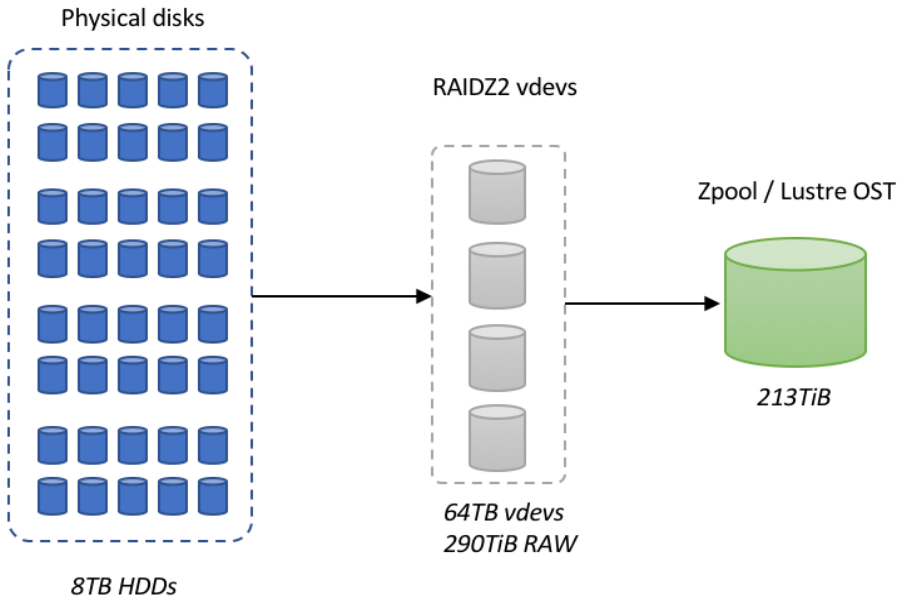


MDS Building Block Details



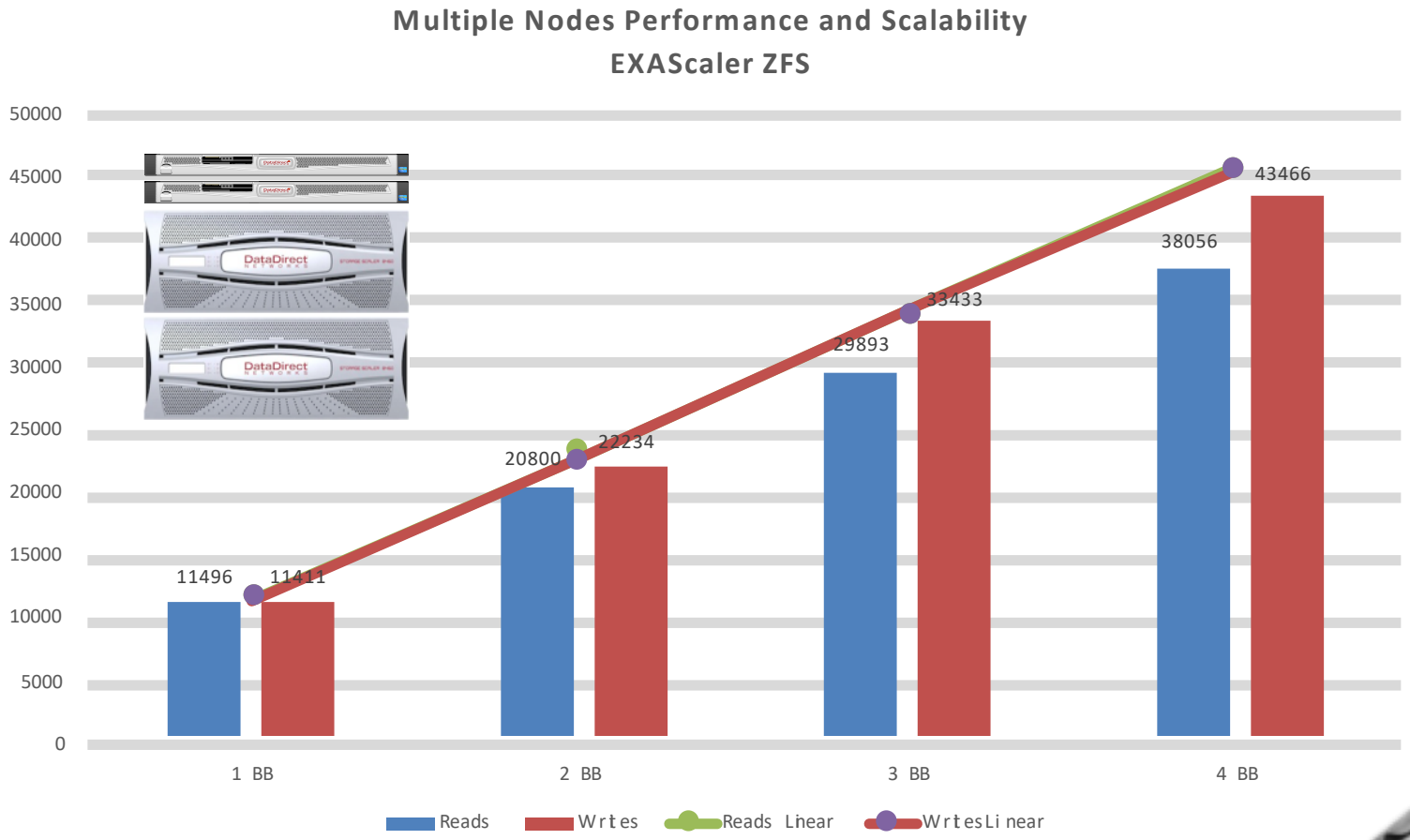
ZFS Configuration

- ZFS Version 0.7.5
- OSTs created from a zpools consisting of 4 (8 + 2) RAIDZ2 vdevs

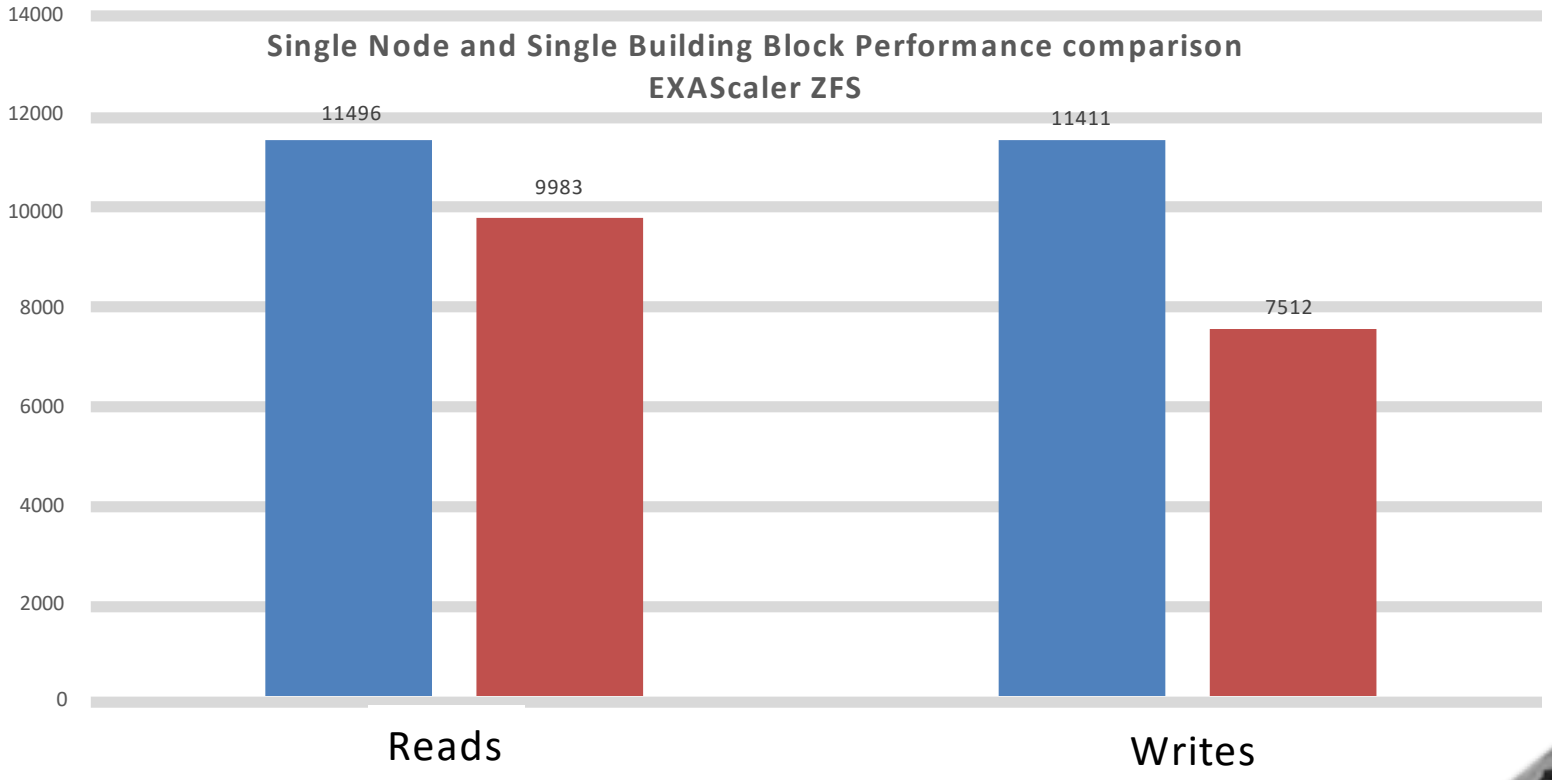


- With a ZFS record size of 1M each drive receives 128K on writes
- For production we will have lz4 compression turned on

Slate Building Blocks Scale

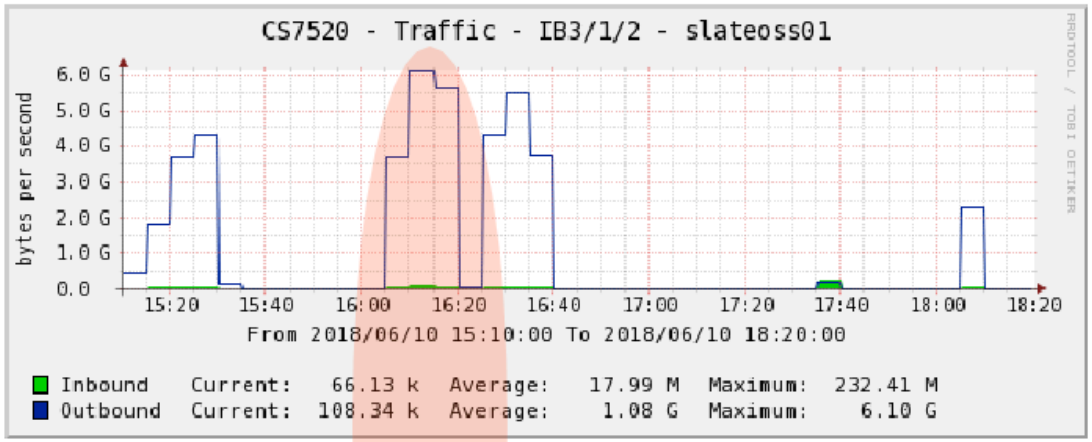


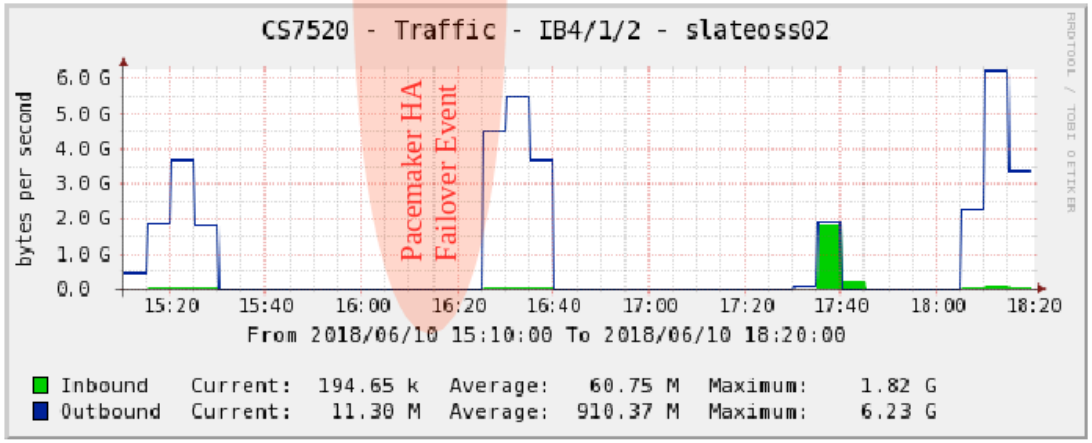
Single Building Block Performance in Blue Dead OSS Failover Performance in Red



Failover During IOR Testing

CACTI Graph of IB Throughput





IOR Stonewall write tests:

We'd lost OSS02

16:07:05 - 38.44 GB/sec

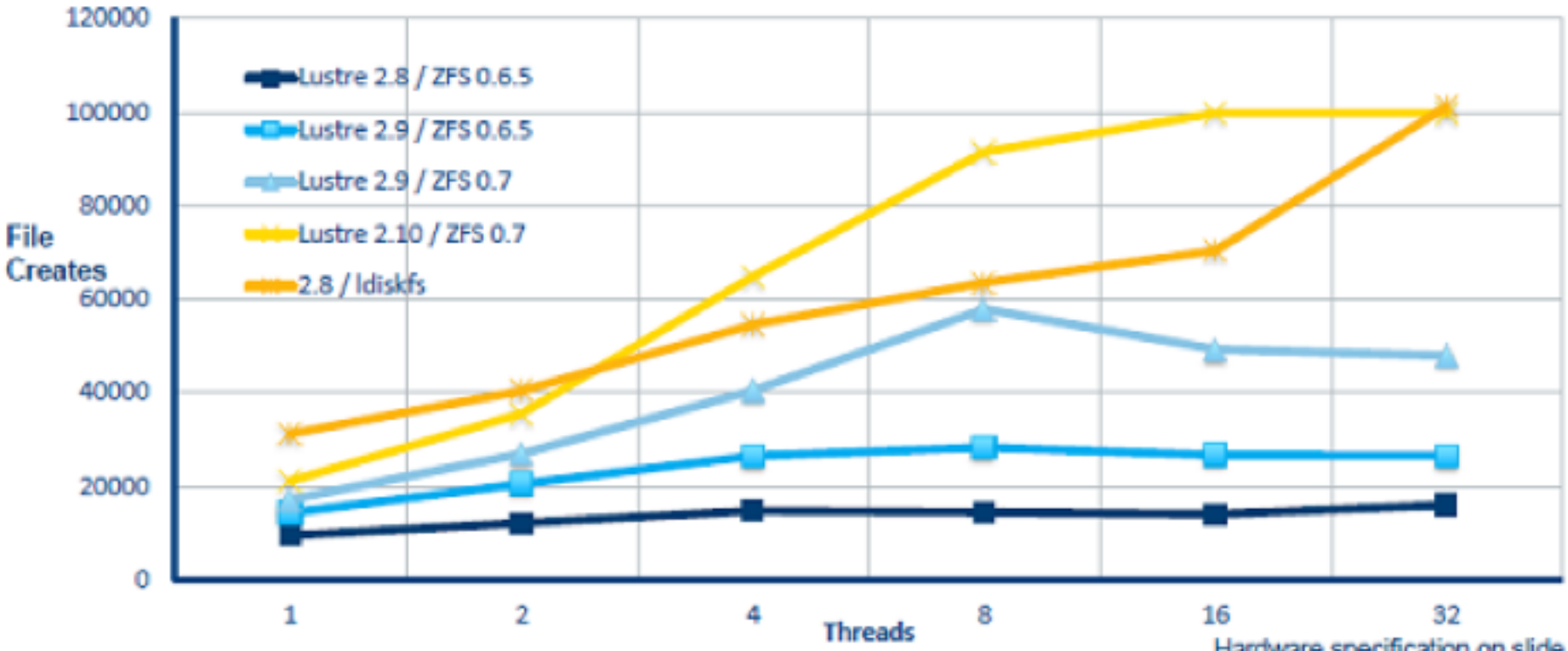
OSS02 came back

16:25:55 - 42.53 GB/sec

Lustre / ZFS Metadata
From Alexey Zhuravlev's 2016 LAD Talk

https://www.eofs.eu/_media/events/lad16/02_zfs_md_performance_improvements_zhuravlev.pdf

Lustre: step by step (mds-survey)



Hardware specification on slide 22

Metadata Apples and Oranges
Fastest ldiskfs MDS at IU versus Slate

DCRAM - Experimental all solid state Lustre cluster

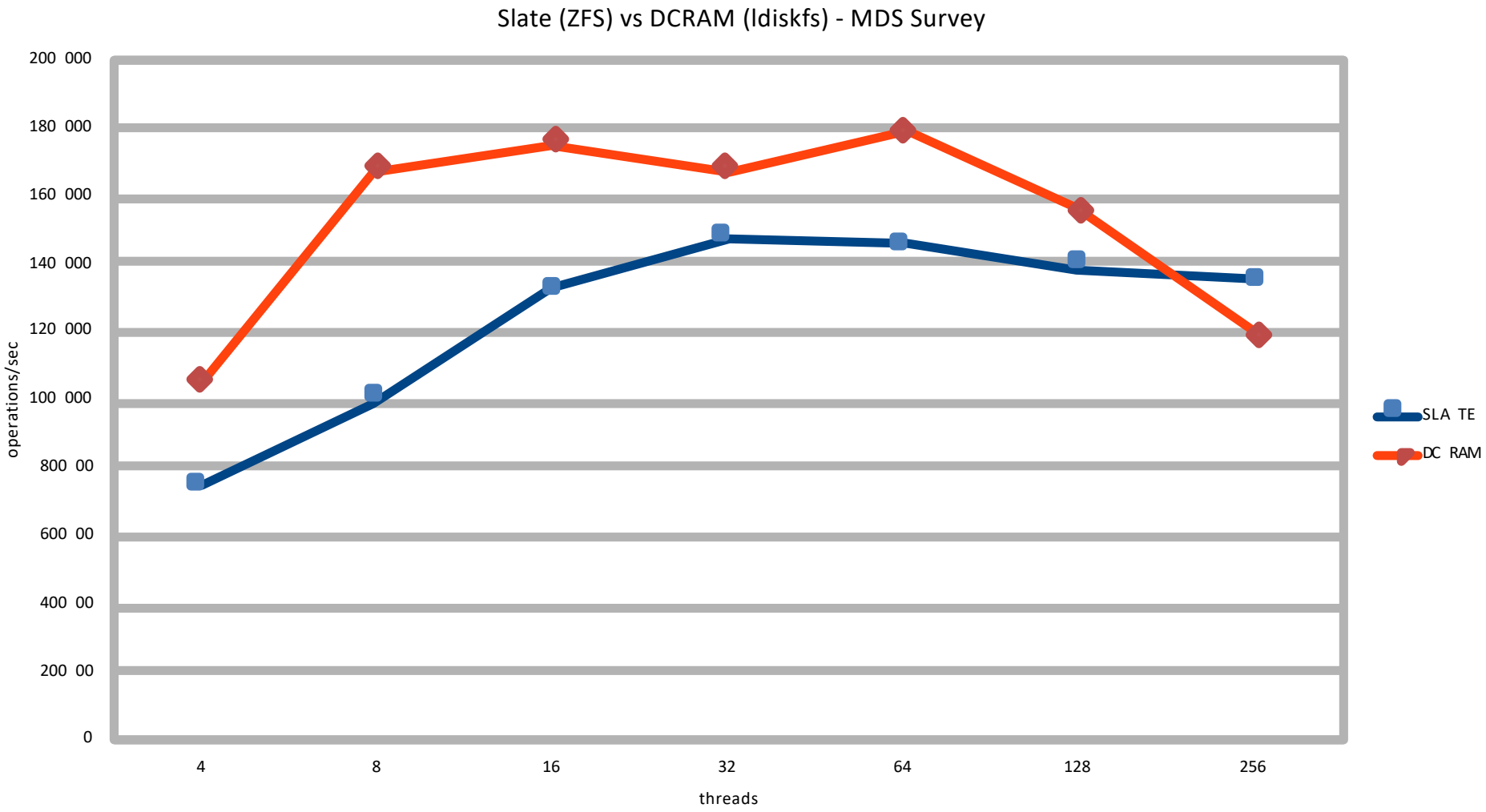
- MDS 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8
- MDT (ldiskfs)
  - SSDs in Hardware RAID controller
  - RAID 0

Slate

- MDS 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
- MDT (ZFS 0.7.5)
  - SSDs in JBOD
  - Striped RAID Mirrors (2+2+2+2+2)

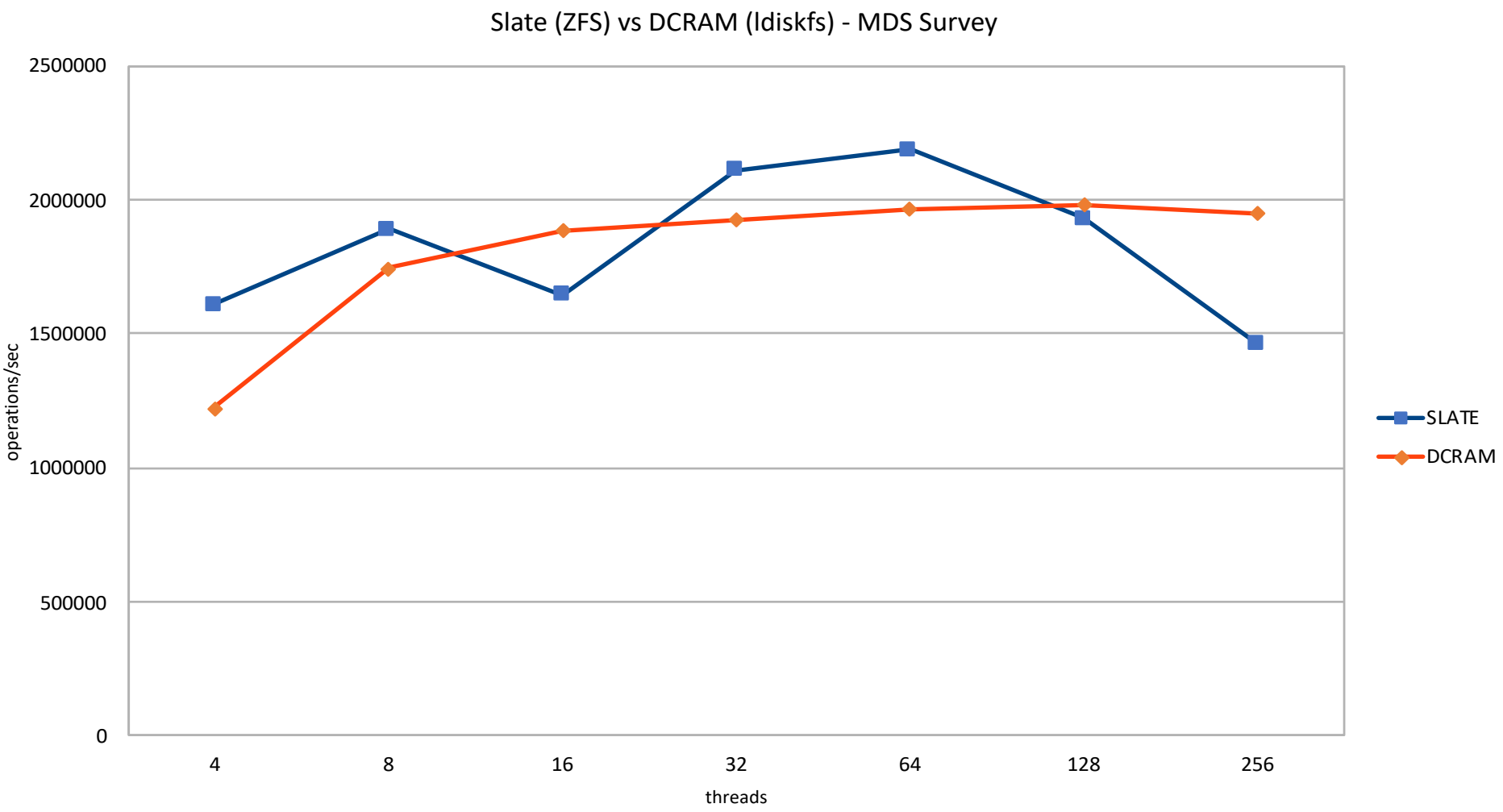
MDS Survey Creates

Slate (ZFS) 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
DCRAM (ldiskfs) 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8



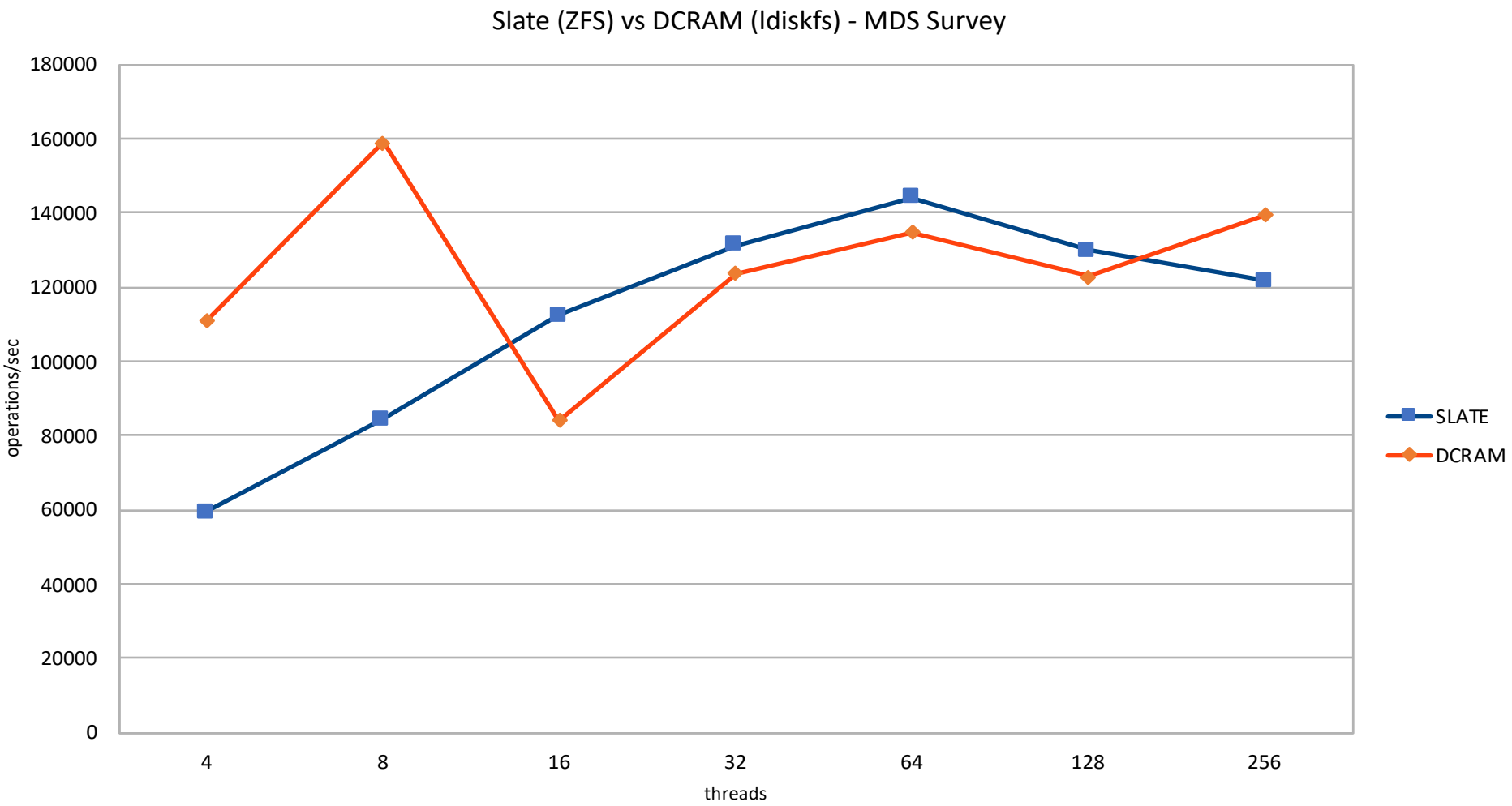
MDS Survey Lookups

Slate (ZFS) 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
DCRAM (ldiskfs) 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8



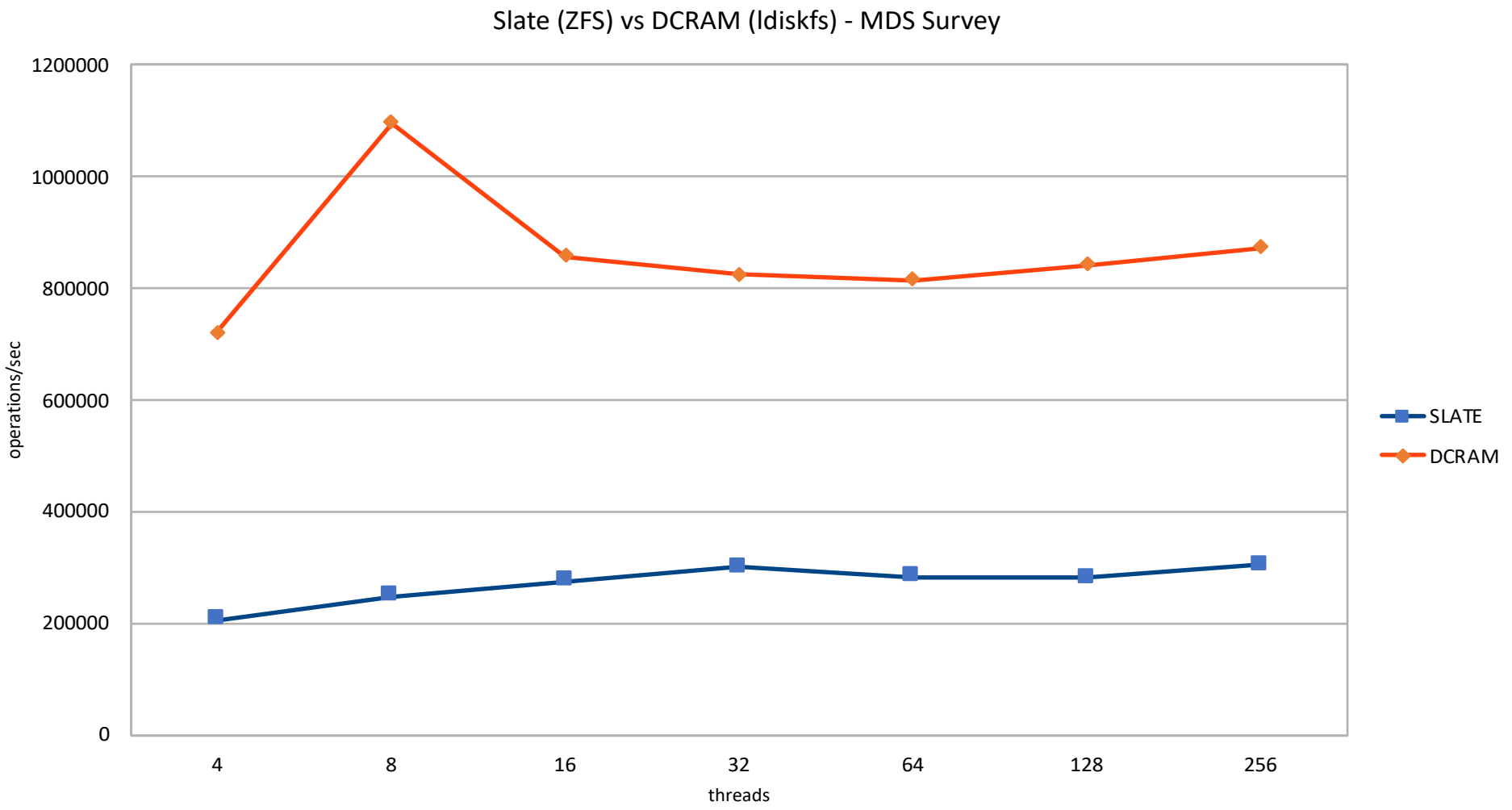
MDS Survey Destroys

Slate (ZFS) 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
DCRAM (ldiskfs) 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8



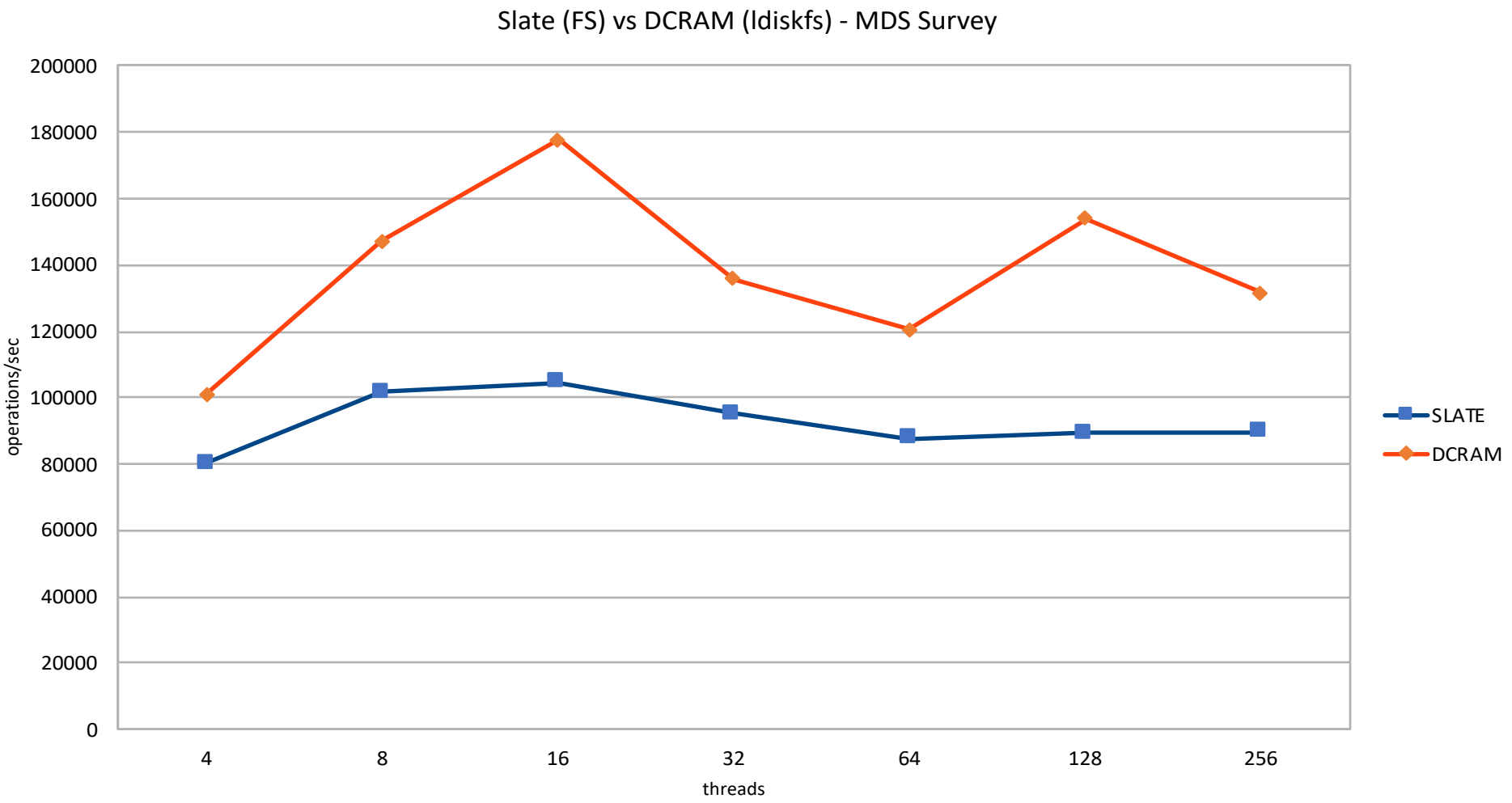
MDS Survey md_getattr

Slate (ZFS) 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
DCRAM (ldiskfs) 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8



MDS Survey setxattr

Slate (ZFS) 3.2 GHz – 8 core (E5-2667v4) – Lustre 2.10.3
DCRAM (ldiskfs) 3.3 GHz – 8 core (E5-2667v2) – Lustre 2.8



DDN Value Add

DDNs EXAScaler tools, like those for the SFA devices, ease management and administration of ZFS and the JBODs. I believe enhancements to these tools will continue.

DDN provided us with a set of Grafana based monitoring tools that (at the present time) are meeting our needs.

Most importantly, DDN has been very responsive to our needs and desires.

Acknowledgments

- DDN
  - Carlos Thomaz, Shuichi Ihara, Sebastien Buisson, Li Xi, Artur Novik, Frank Leers, Gregory Mason, Rich Arena, Bob Hawkins, James Coomer, and Robert Triendl
- The rest of IU's HPFS team
  - Chris Hanna, Nathan Heald, Nathan Lavender, Ken Rawlings, Shawn Slavin
- LLNL
  - Brian Behlendorf, Chris Morrone, Marc Stearman
- Whamcloud née Intel née Whamcloud née Oracle née Sun née Cluster File Systems
  - Special thanks to Alexey Zhuravlev

Thank You for Your Time!

Questions?