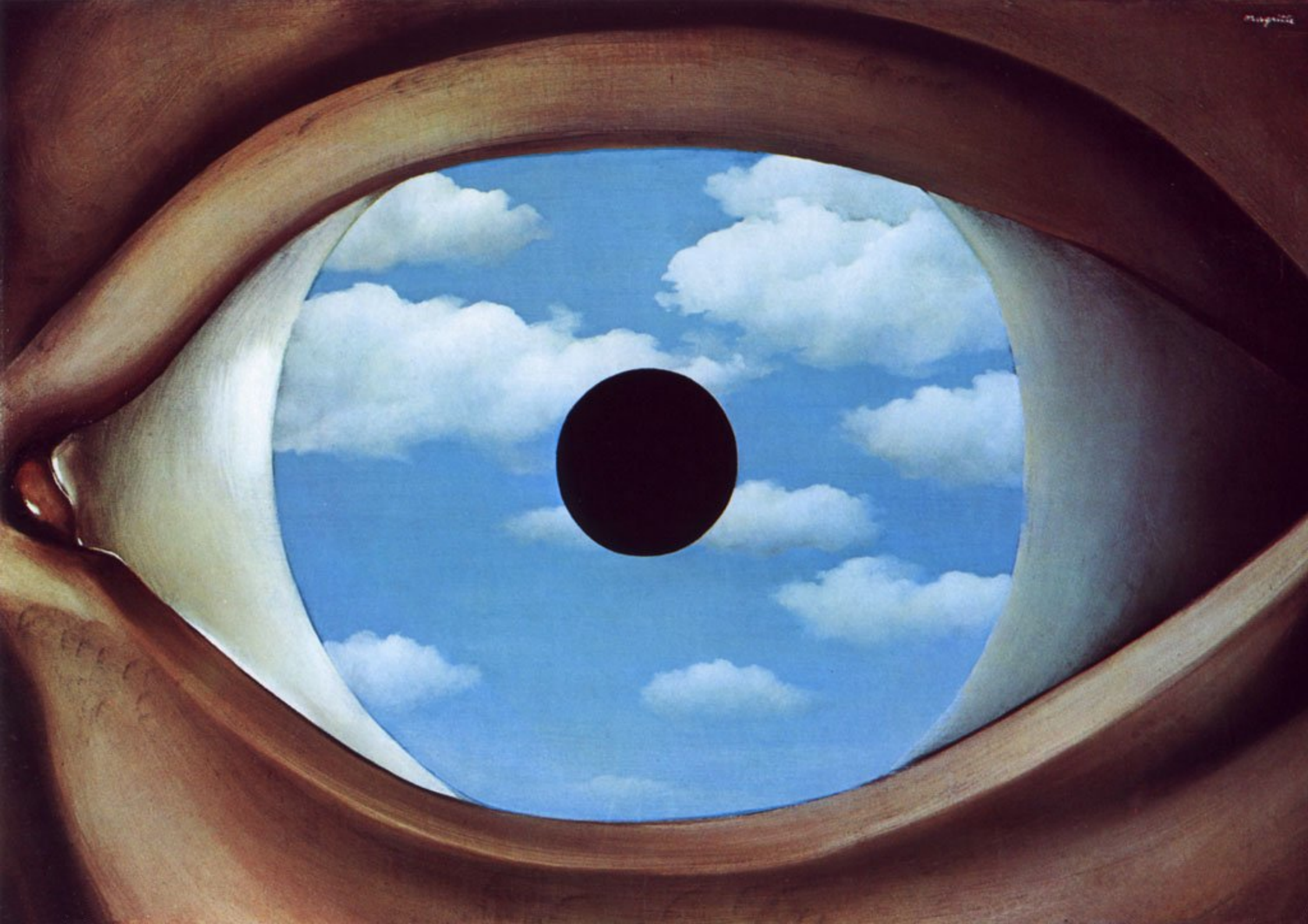
DDN's Vision for the Future of Lustre

LUG2015

Robert Triendl



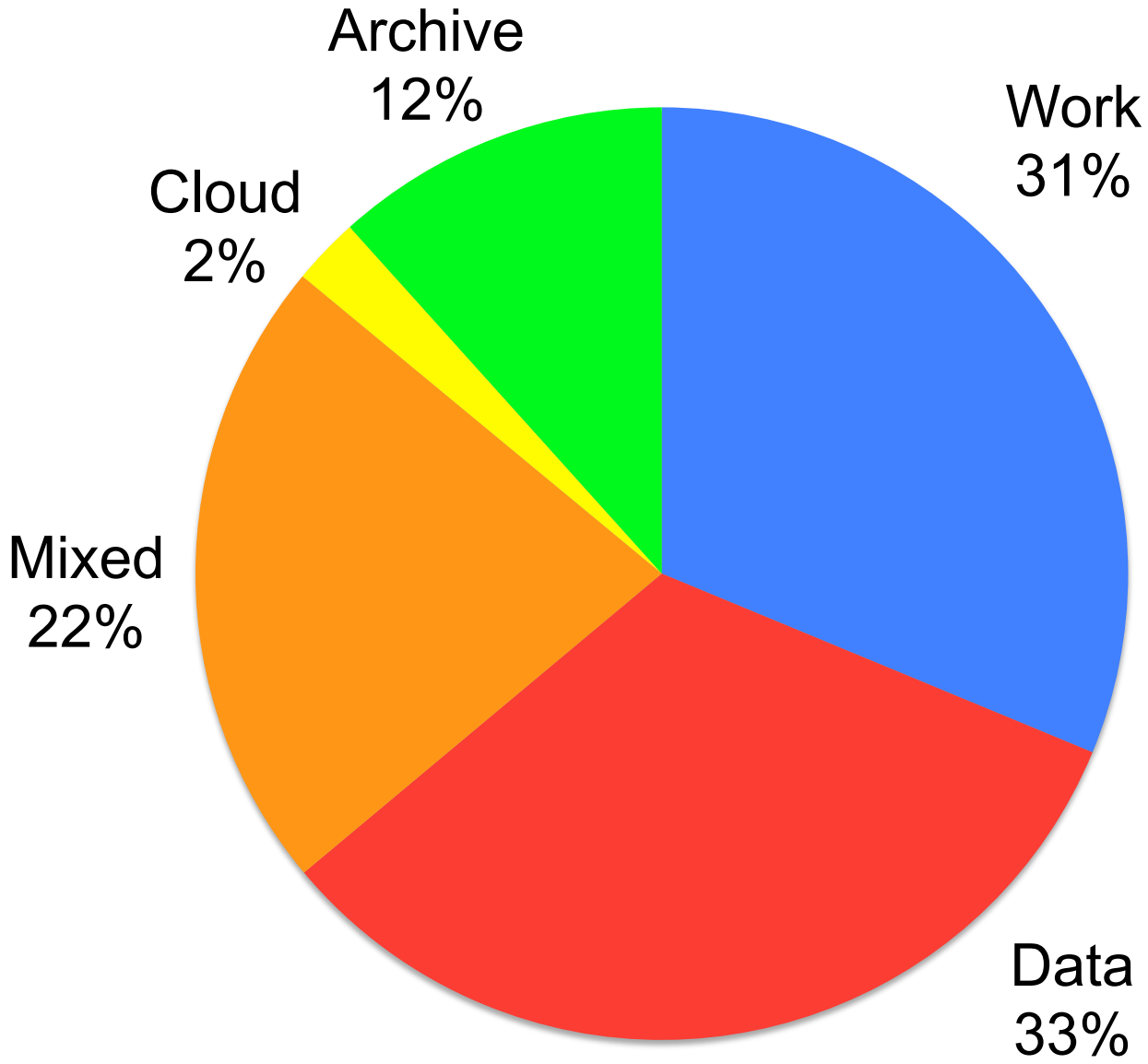
Topics

1. The Changing Markets for Lustre
2. A Vision for Lustre that isn't Exascale
3. Building Lustre for the Future
4. Peak vs. Operational Performance
5. Application Optimized Lustre
6. Why Conventional Storage Still Matters

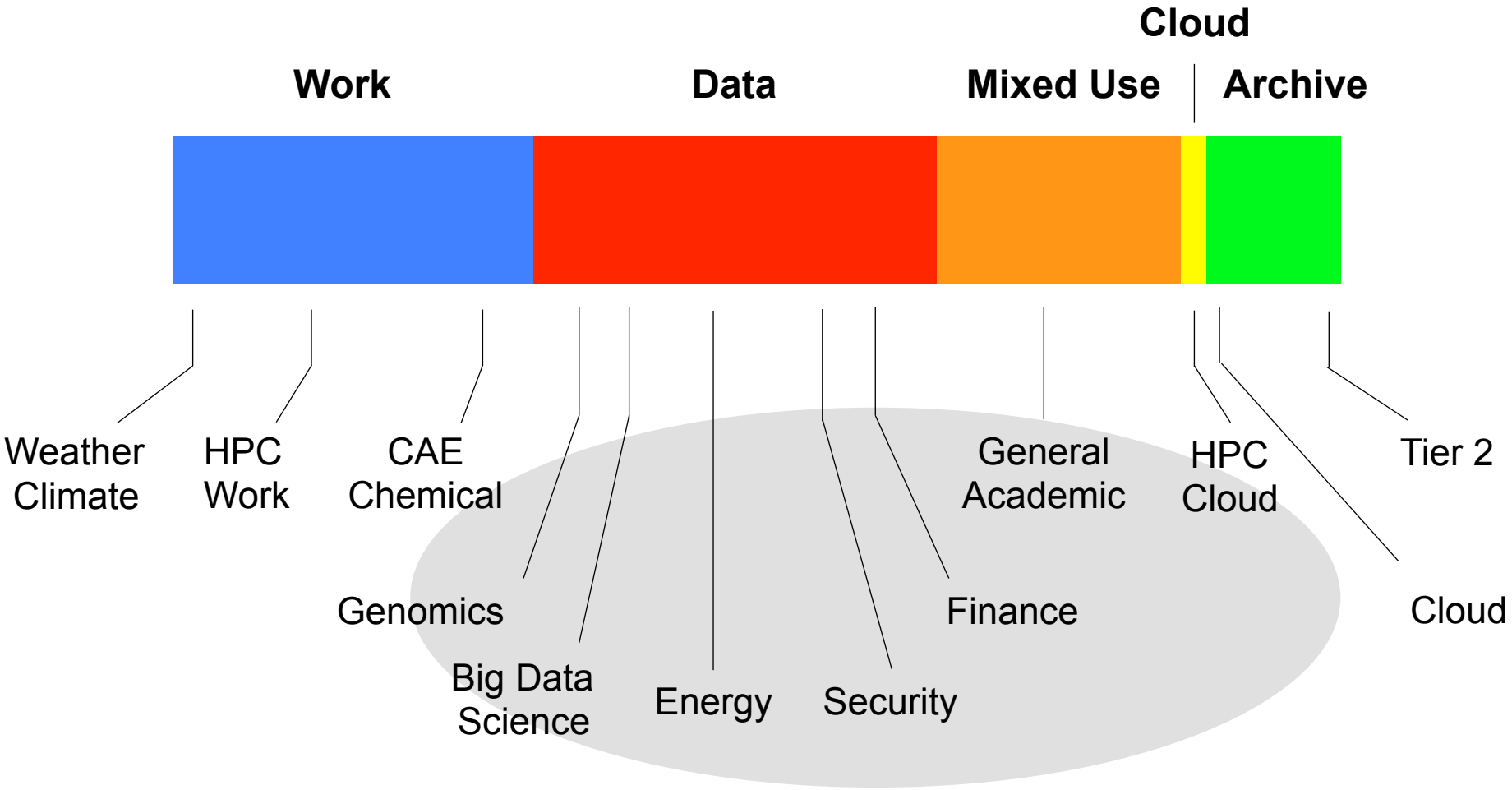
Hyperscale Storage Markets

HPC	Cloud
Scratch	WORM(N)
Petabytes	Billions of Files
Streaming Write	Random Read
Large Files	Small Files
Infiniband	Ethernet
Single Location	Distributed

Lustre Markets Today



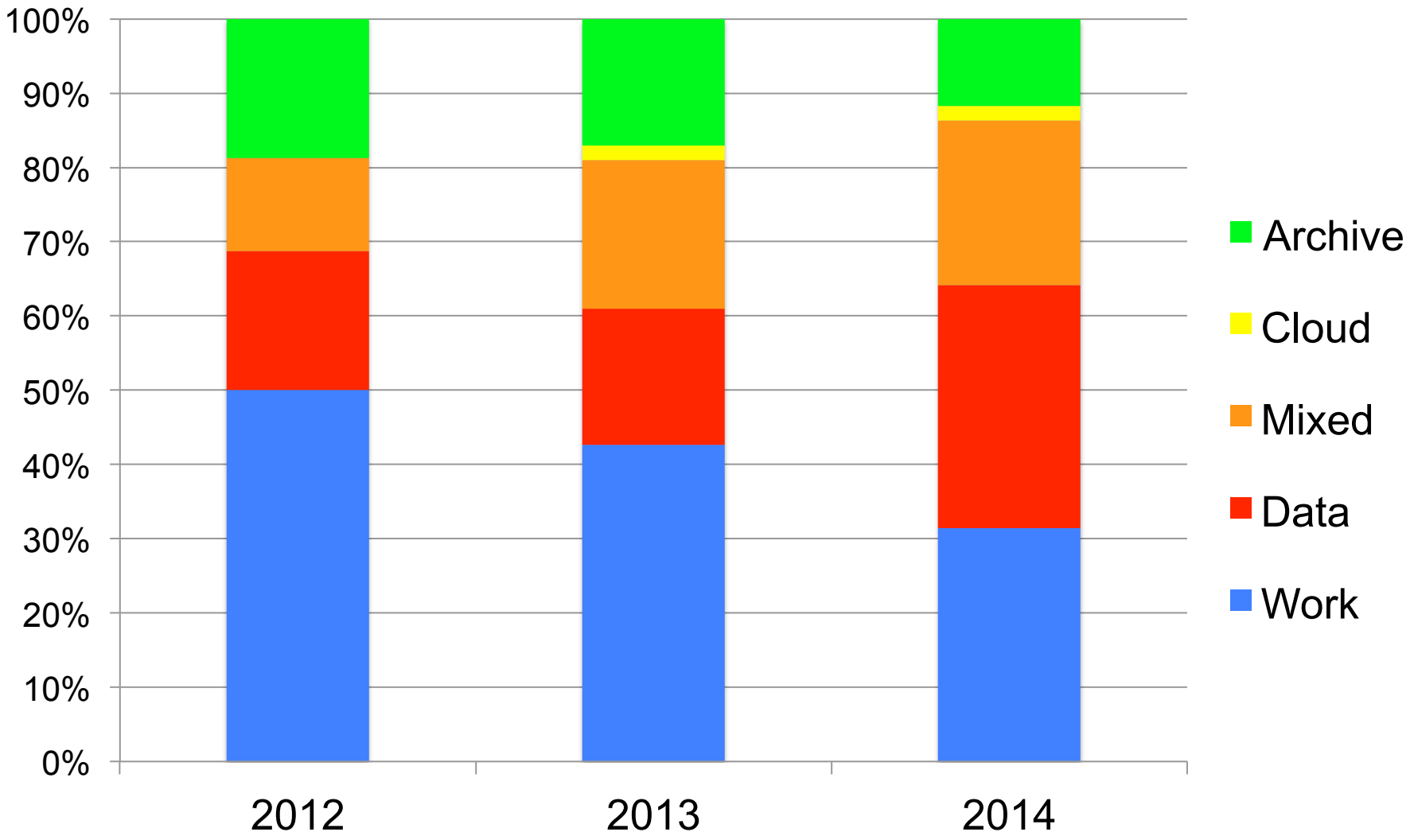
Market Diversification



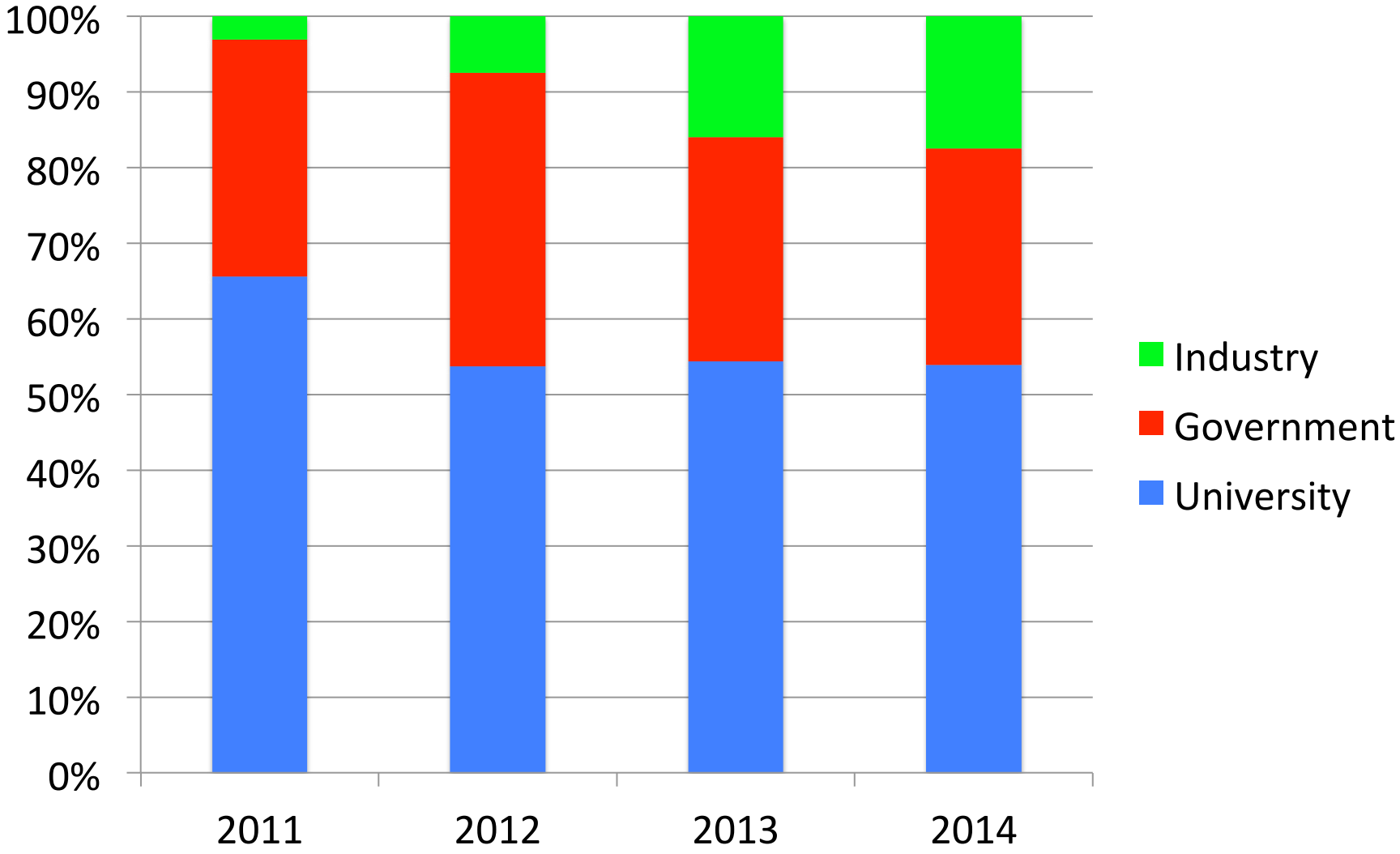
Lustre Futures Beyond Exascale

Manufacturing	CIFS/NFS Export, AD Integration, RAS Features, Snapshots, Data Management, etc.
Genomics	Random Performance, Small File & Metadata Performance, Data Management, Security, etc.
General Academic	Broad Application Support, Connectors, User Monitoring, User Access to Snapshot, etc.
Cloud	Virtualization, Snapshots, Small File Read Performance, Data Distribution, etc.
Archive	Data Management Features, SMR Drive Use, Data Scrubs, Data Distribution, etc.

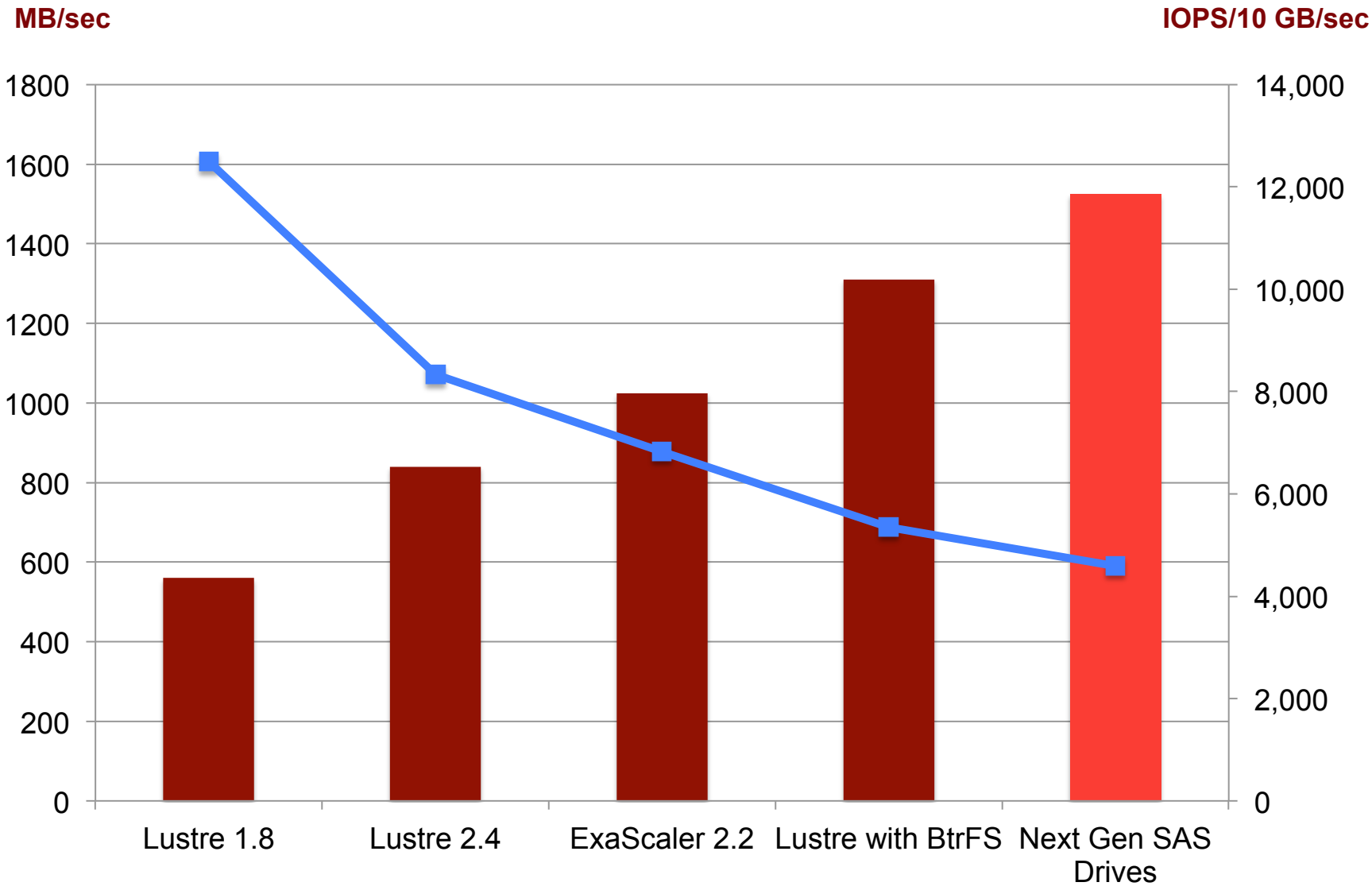
Market Evolution



Market Segments



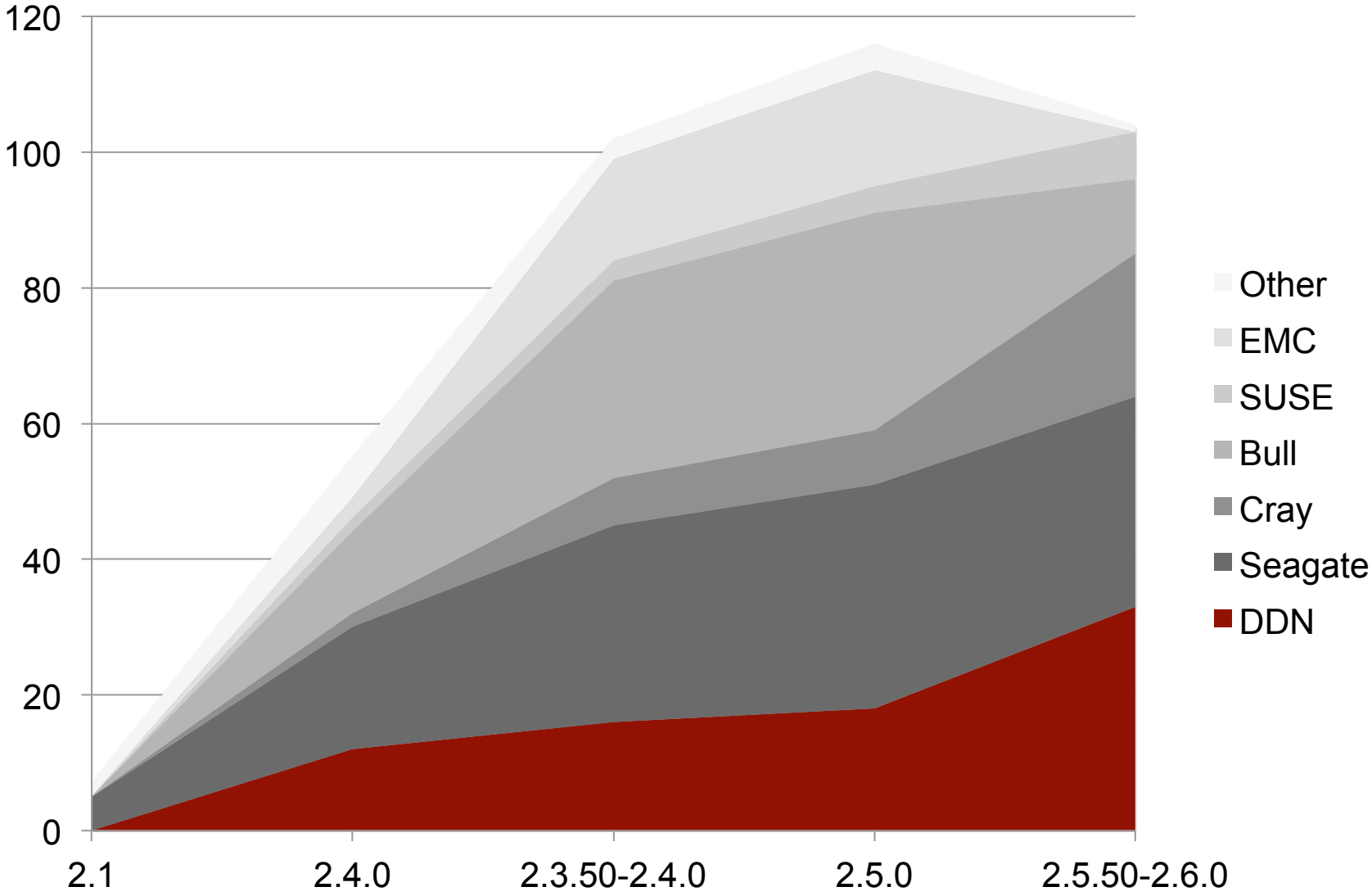
Disks: Throughput vs. IOPS



Lustre Development at DDN

- Lustre Usability Features
- Build-in Reliability and Availability
- Lustre Recovery
- Features for a Broader Market
- Performance for Broad Set of Applications
- Application-optimized Lustre

Lustre Code Contributions



DDN ExaScaler Software Stack

Data Management

Fast Data Copy | Tape | S3 Cloud | Object (WOS)

ExaScaler Data Management Framework

Monitoring & Management

DDN DirectMon | DDN ExaScaler Monitor | Intel IML

DDN ExaScaler

Core FS

DDN Clients | NFS/CIFS/S3 | DDN IME | Intel Hadoop

DDN Lustre Edition

ldiskfs | OpenZFS | btrfs

Intel DSS

Storage HW

DDN Block Storage with SFX Cache | Other HW

Why BtrFS?

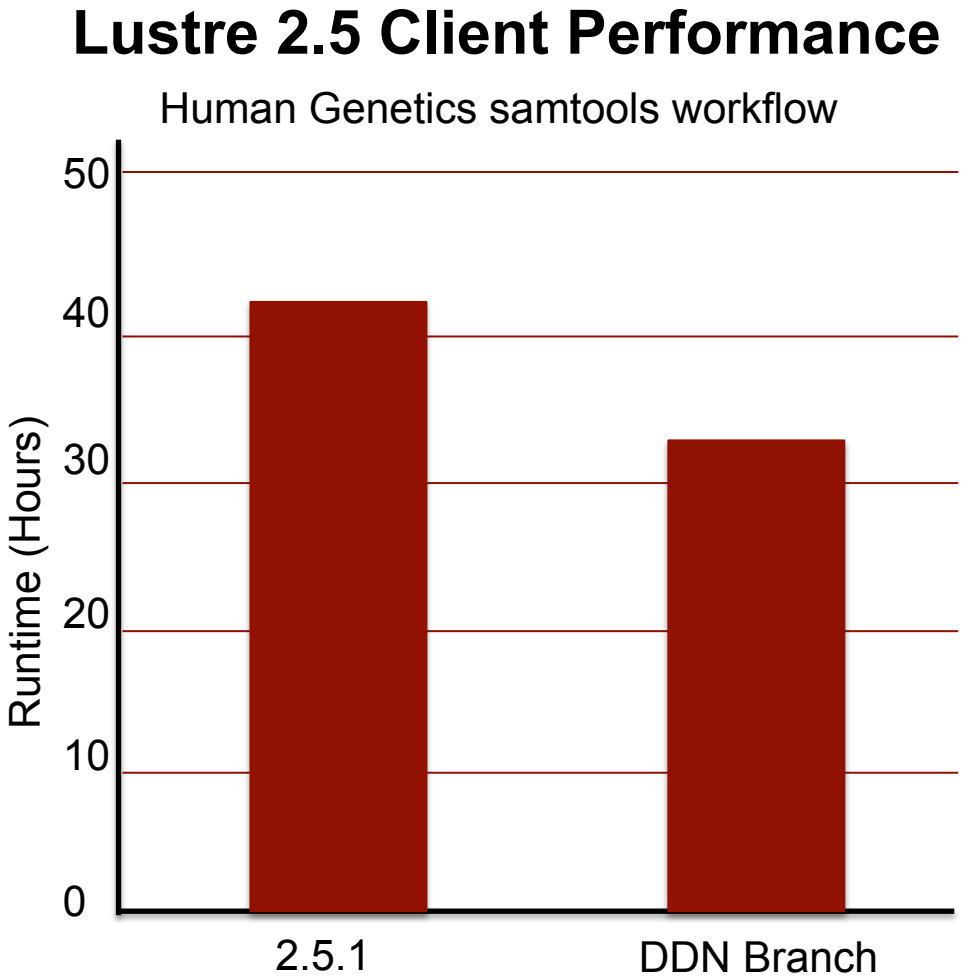
- Standard Local Filesystem in RHEL7
- Better Throughput Performance than ZFS
- Similar Feature Set, but all Linux
- No Possible Patent Infringement
- Simple Integration and Deployment

Application-Optimized Lustre

- Lustre for Specific Applications
- Workload Profiling
- Optimization Across I/O Calls
- Optimizing Application Runtime
- Working with Customers

Genome Pipeline Benchmarks

Samtools
20% faster with DDN
Lustre optimizations



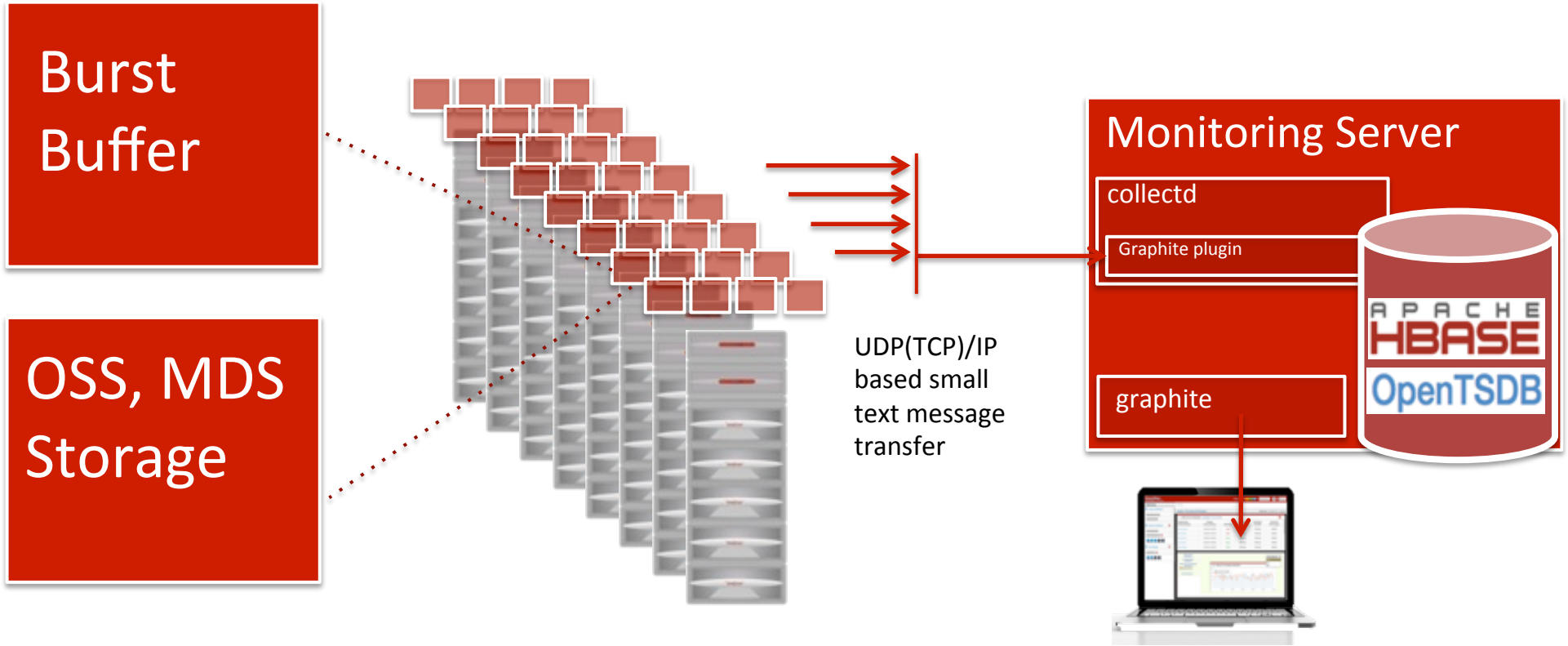
SSD Pools and Caching

- **DSS to Link File Layer to Block Layer**
- **Build into the File System**
- **Better use of SSDs for I/O Optimization**
- **Increased Small File Performance**
- **Increased Random Read to Large Files**
- **Additional specificity with** fadvice()

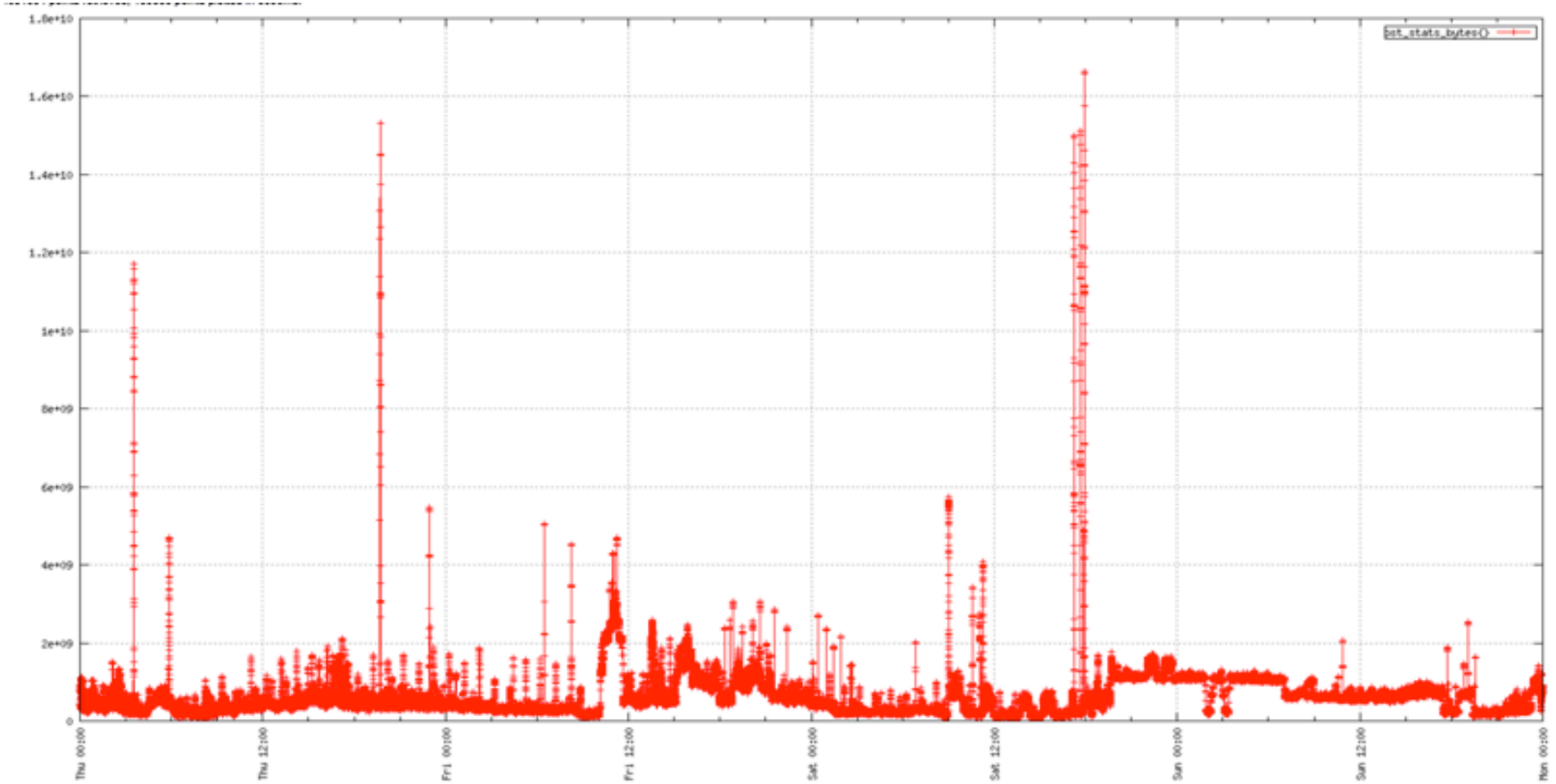
ExaScaler Monitoring

- Filesystem, OSS, MDS, OST, MDT, etc.
- JOB ID, UID/GID, application stats, etc.
- Archive of data by policy

- Lightweight
- Near real-time
- Massive scale



TITECH Examples



64 Billion of Lustre Stats in 15 days!

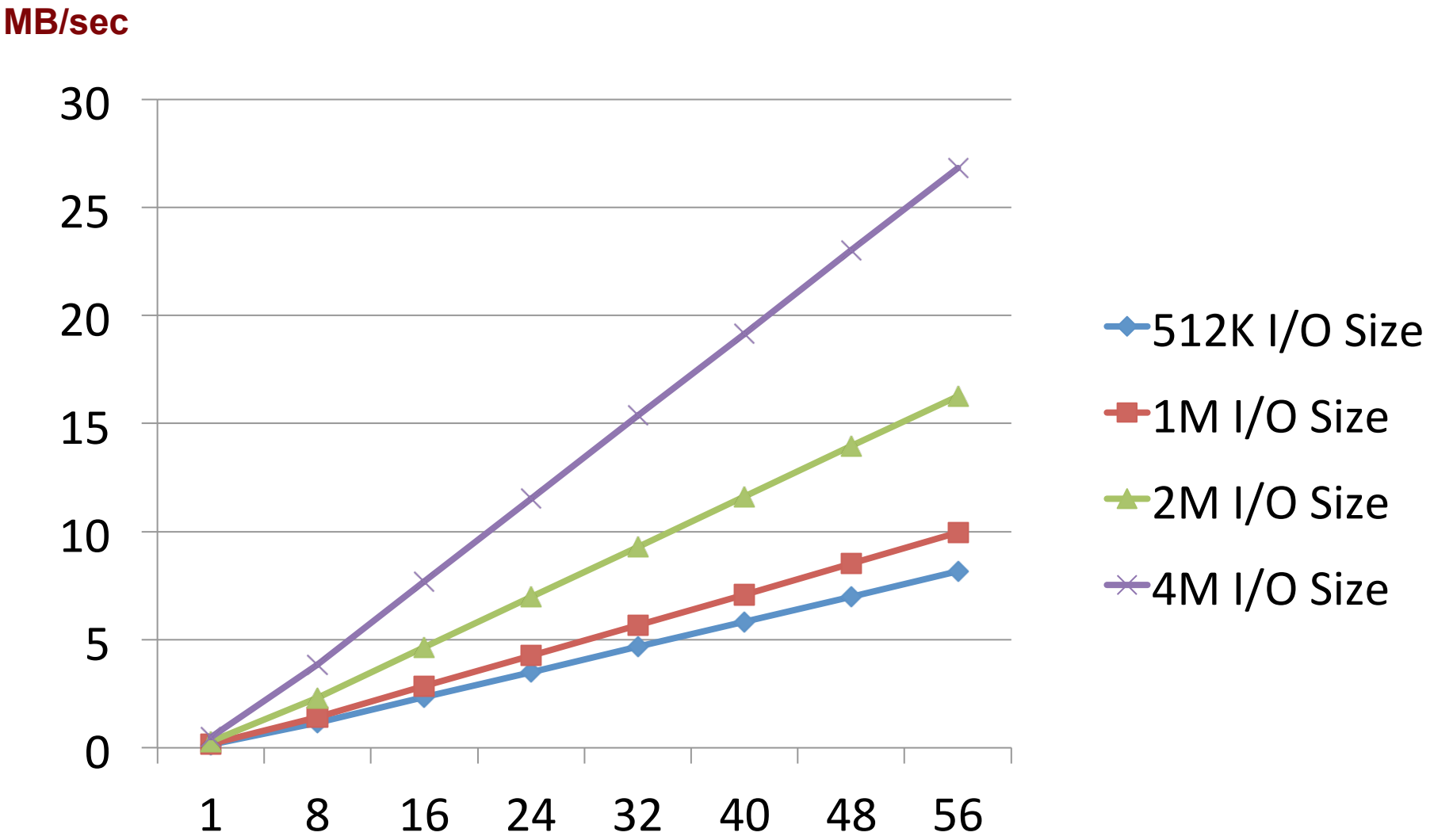
Why Block-Level Raid?

- Best Mixed I/O Performance
- Consistent Performance
- Hardware-optimized Performance
- Best Performance During Failure
- Integrated Storage Services

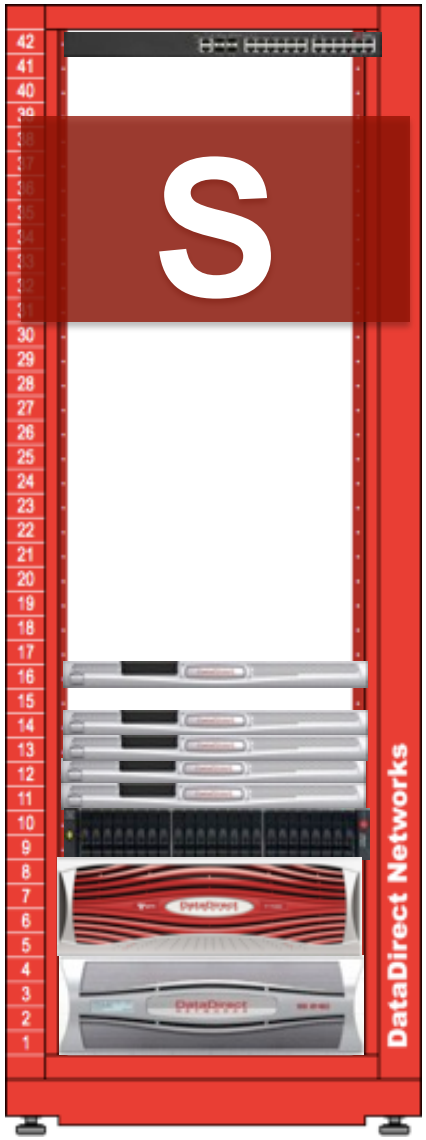
SFA RAID Stack Performance

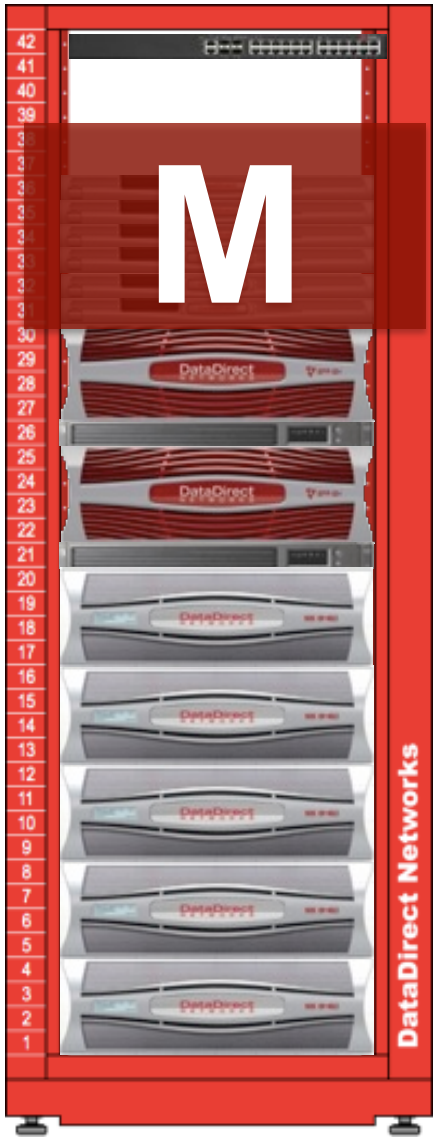
- Above 1 Million 4K IOPS per 8 CPU Cores
- Above 10 GB/sec per 8 CPU Cores
- 8 Cores Sufficient for PCI Infrastructure
- More Cores for File System Services
- Additional Cores for More Functionality

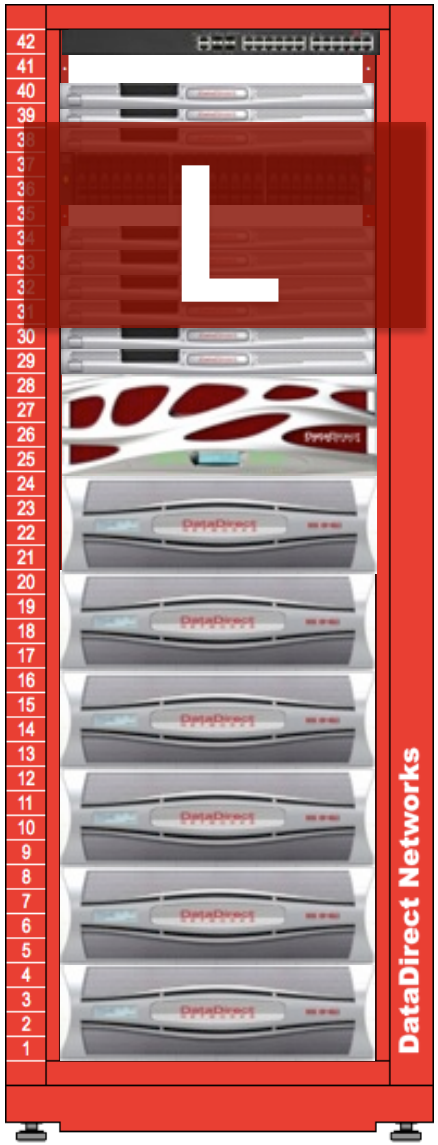
SFA Random Read

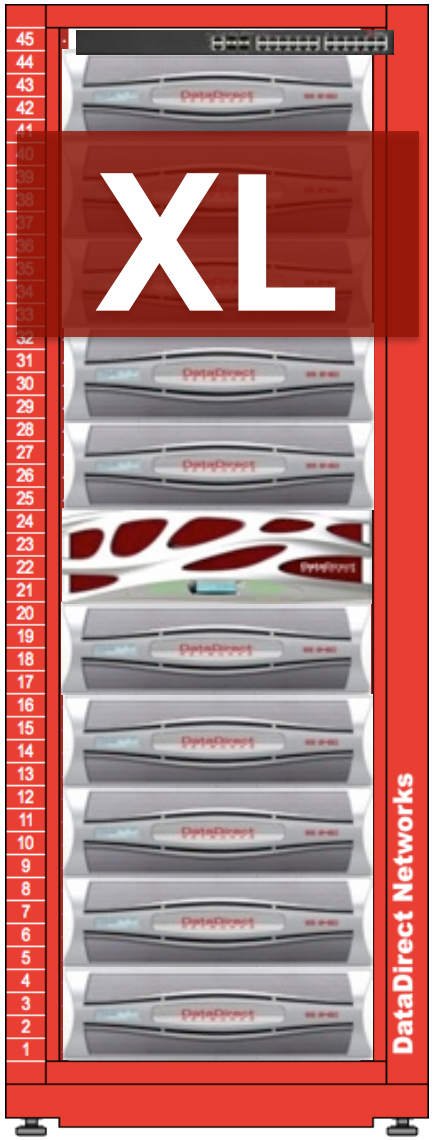


Flexible SSU Design









SFA14K Performance SSU

Up to 45 GB/sec
Up to 2950 TB
External MDT

Monitoring

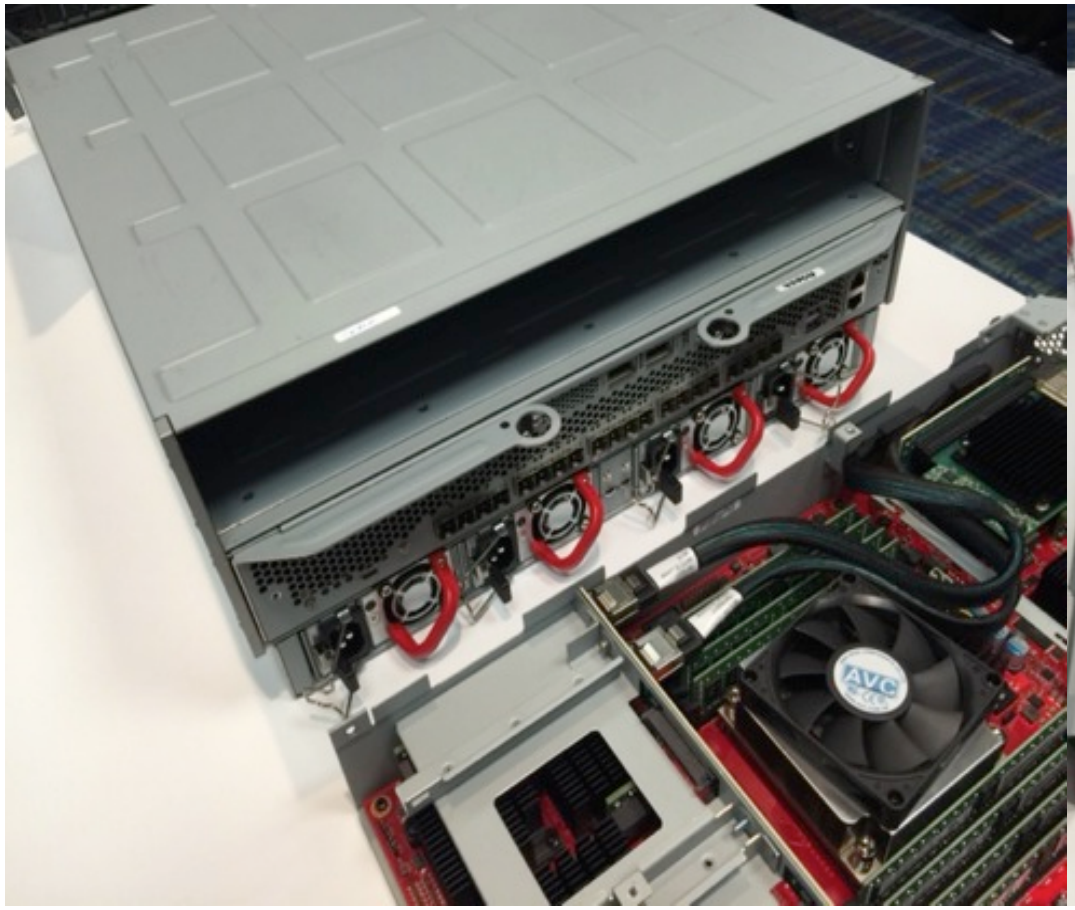
2-4 MDS

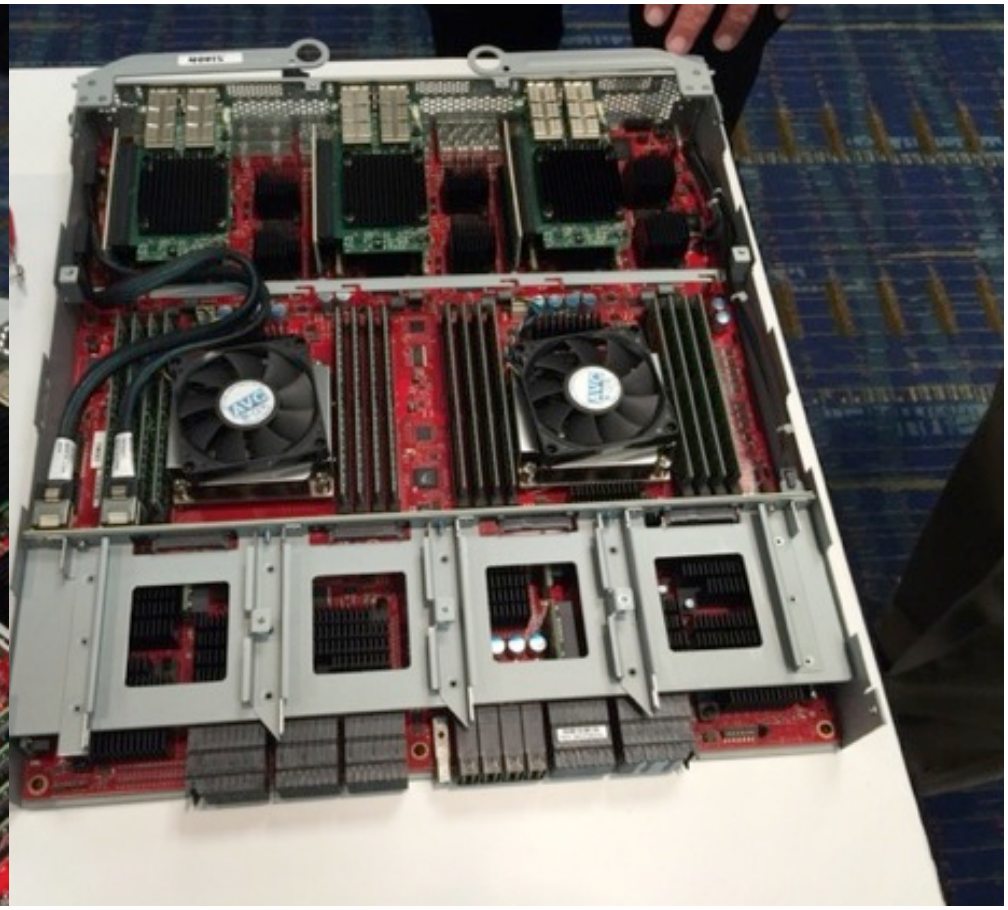
MDT Storage

4-6 OSS

OST Storage

“Wolfcreek” Hardware





“Wolfcreek” Hardware