ExaScaler

Lustre for HPC, Big Data, and AI

LUG 2018

DDN® STORAGE

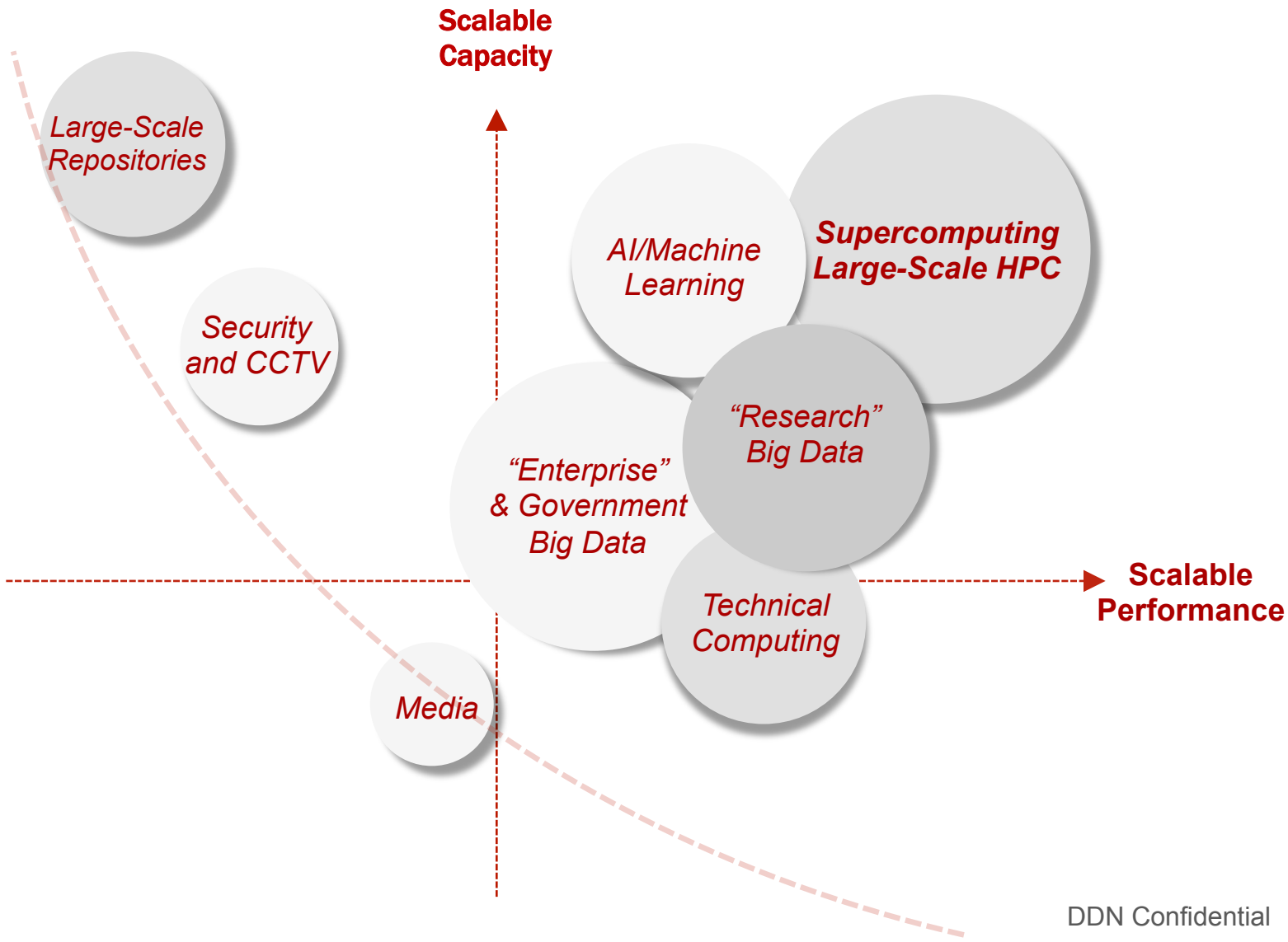
ExaScaler: Expanding the Market for Lustre

Scale Optimized Storage Platforms at Scale

Performance Bandwidth, IOPs, Latency, Small File I/O

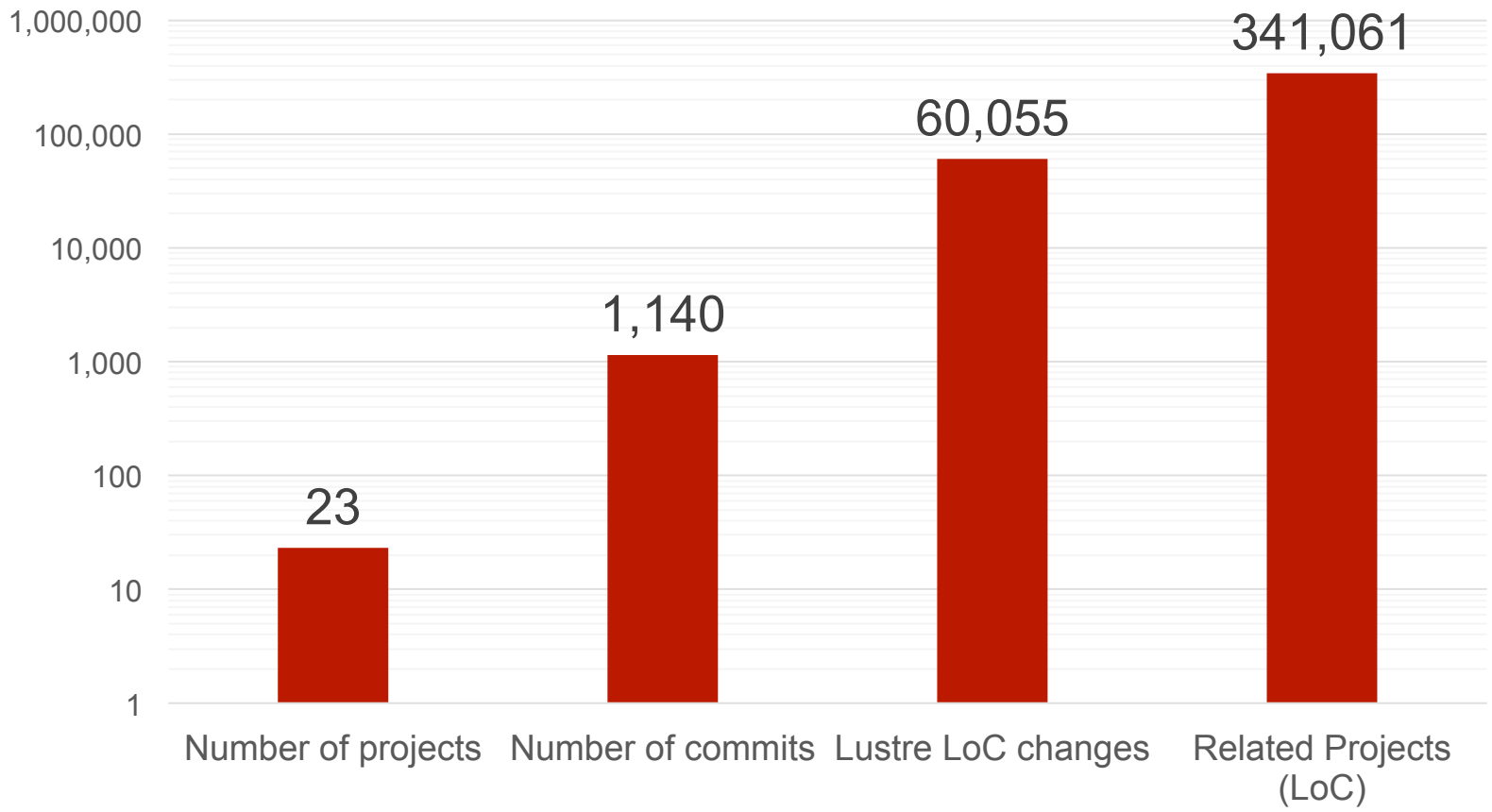
Flexible **On-Premise, Multi-tenancy, Cloud**

Reliability **Enterprise Lustre Building Blocks**

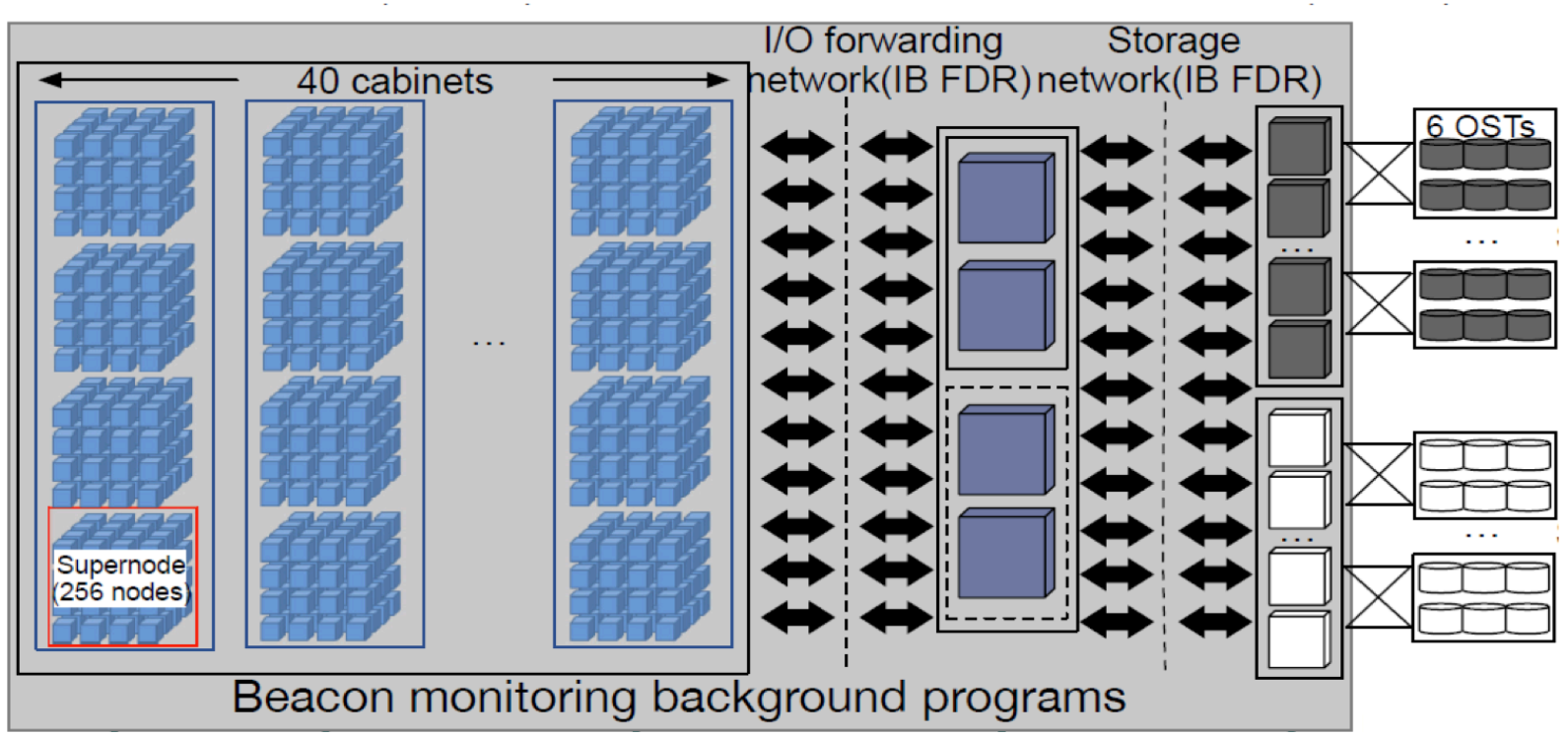


DDN Lustre Contributions

- **$20 Million DDN R&D & Support Investment in 2018-2020 for Lustre Enablement**
- Rapidly Growing Lustre Technical Team Today
- Powerful DDN Designed Build and Test Suite
- Forward-looking DDN Lustre Development Roadmap



ExaScale I/O Architecture for Sunway TaihuLight

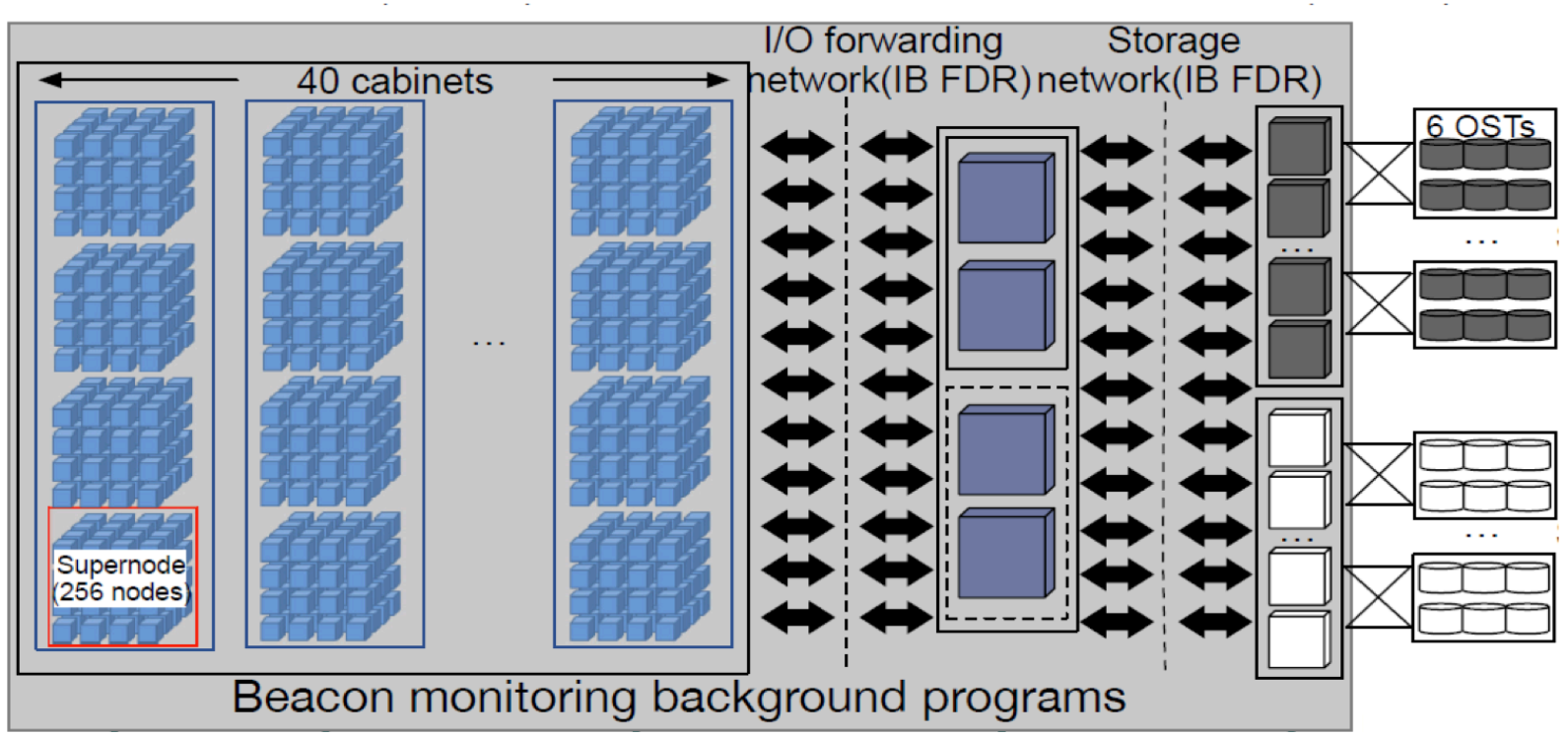


40,960 Compute Nodes

160 I/O Forwarding Nodes

288 Storage Nodes

ExaScale I/O Architecture for Sunway TaihuLight

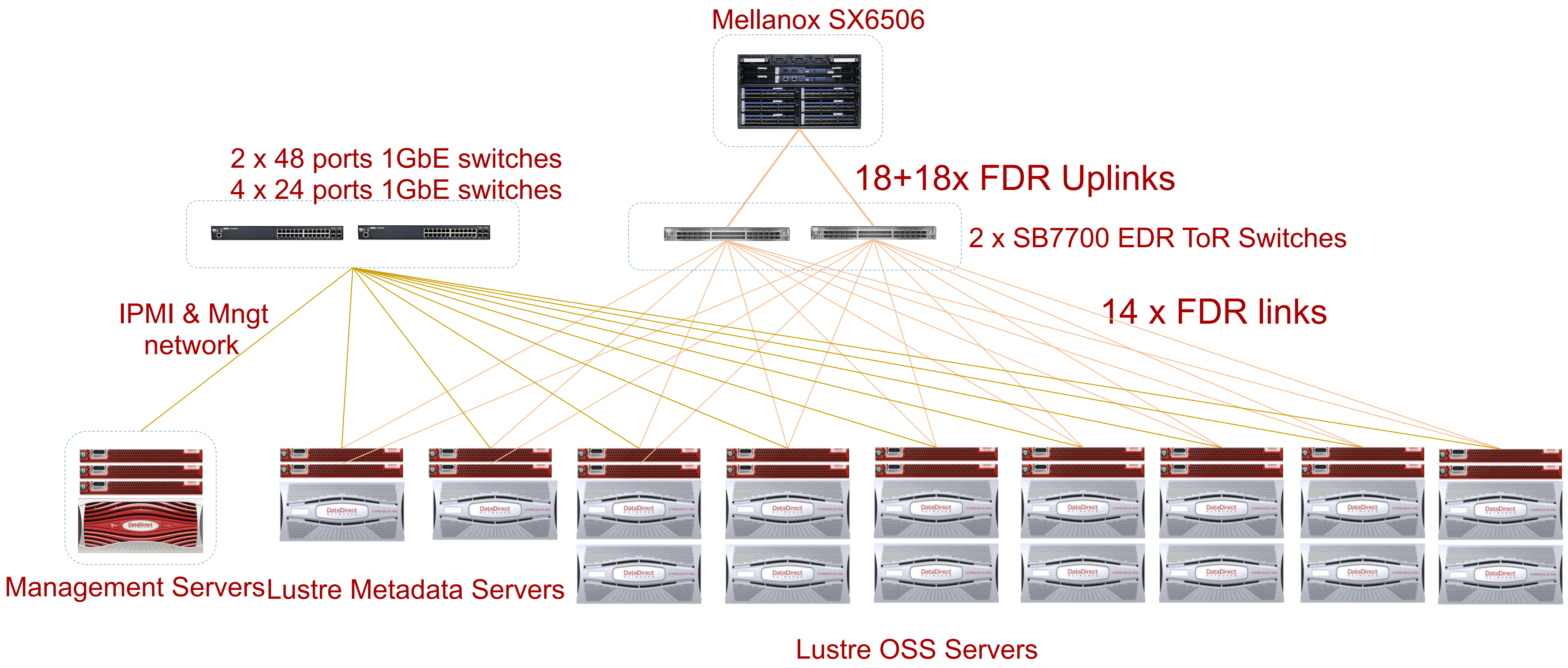


40,960 Compute Nodes

160 I/O Forwarding Nodes

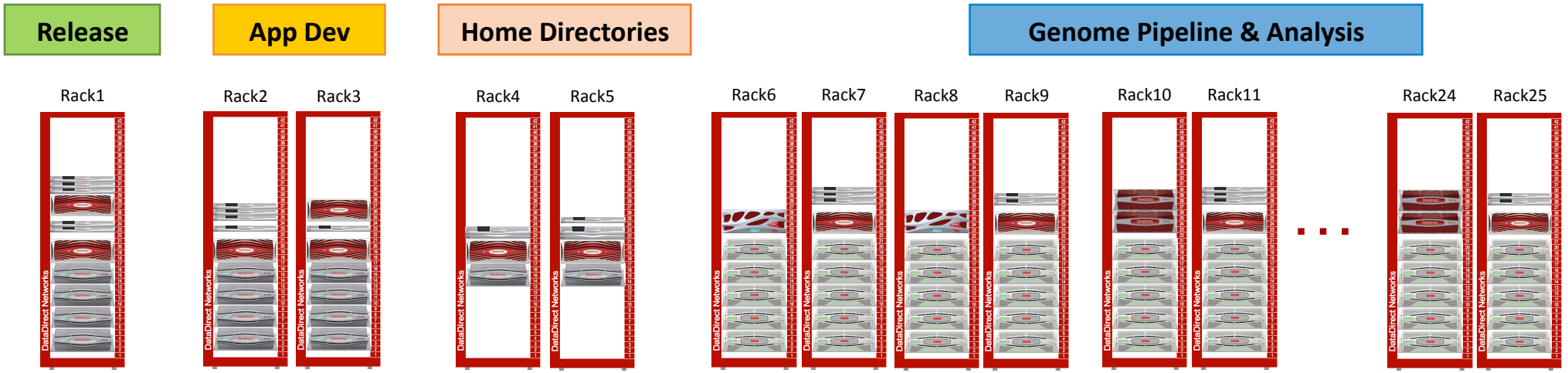
288 Storage Nodes

ExaScaler ZFS Mass Storage Deployment

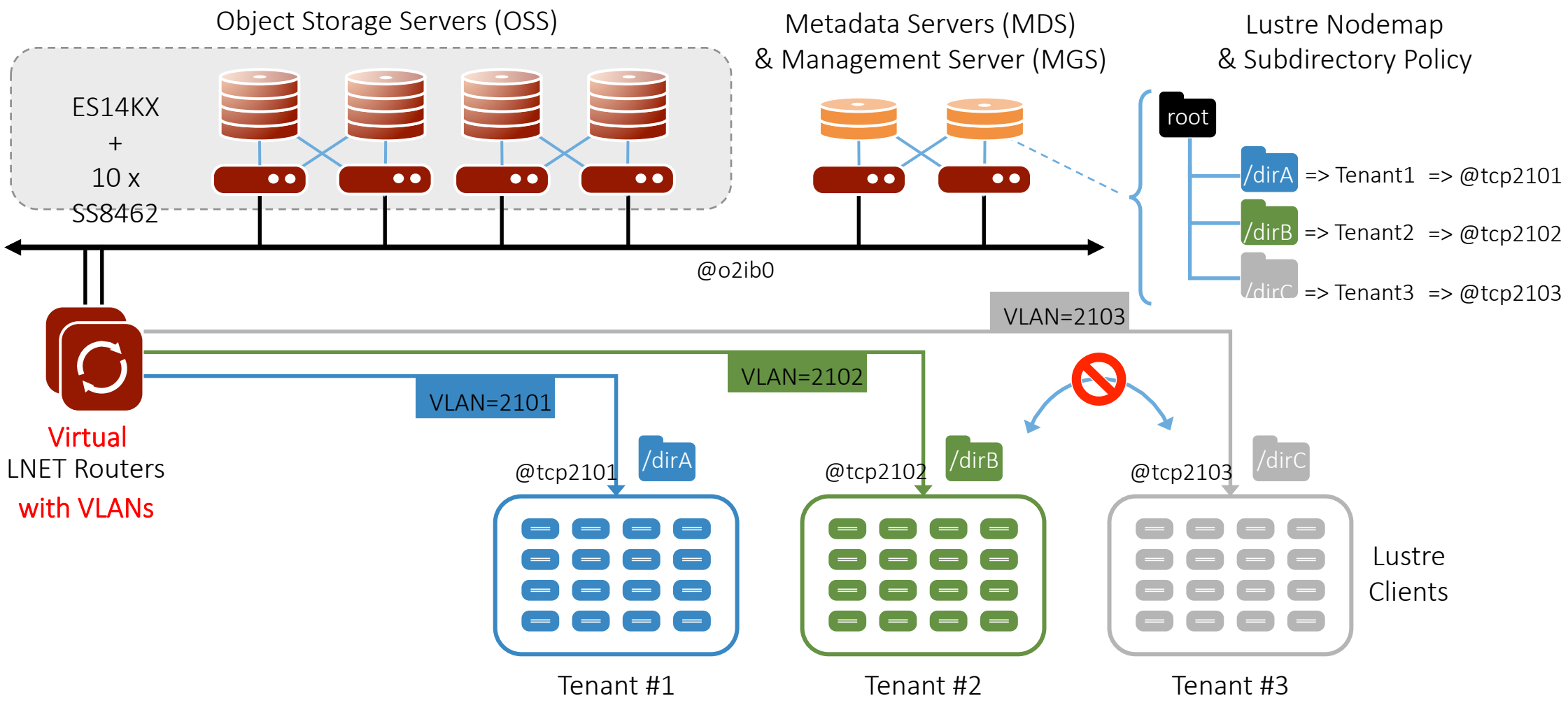


ExaScaler for Populations Genomics

- Large Lustre Environment for Populations Genomics
- Various File Systems for Different Workloads
- ExaScaler Provides both Performance and Capacity

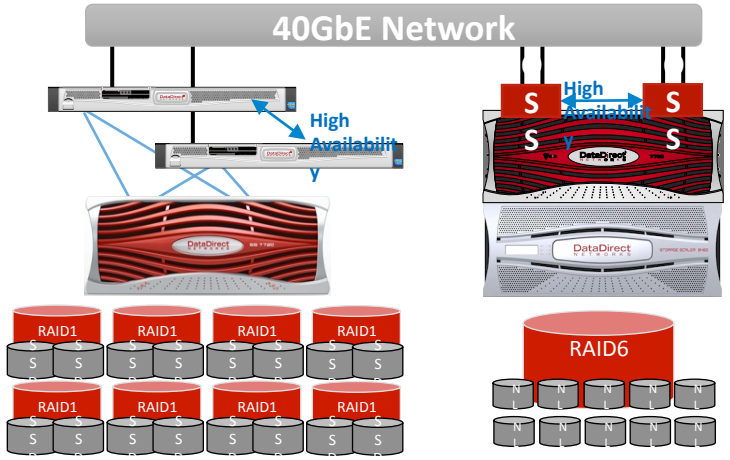


ExaScaler Isolation/Multi-tenancy for Clinical Genomics



ExaScaler Scalable Autonomous Driving Infrastructures

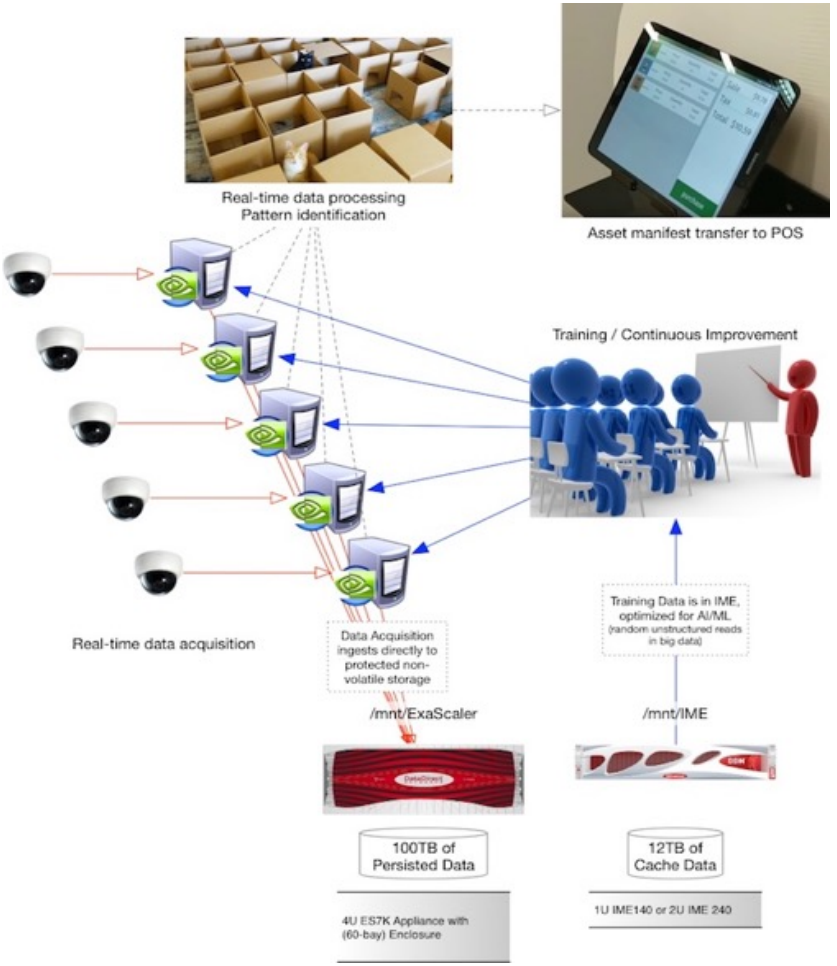
- Scalable Infrastructure for Autonomous Driving Applications
- 100s of OSS Servers, 100s of PBs of Storage
- Various IO requirements for Data Ingest, Data Curation, and Deep Learning Phases
- Many Industry Collaborations to Multi-tenancy and Isolation are Important



ExaScaler for Deep Learning

[Sc]
STANDARD COGNITION

- Machine Vision for Autonomous Checkout in Supermarkets
- ExaScaler for Data Ingest and Capacity Storage
- Low-latency IME Cache Tier for Training Phase
- Shared Workflow Acceleration with NVMe Caching

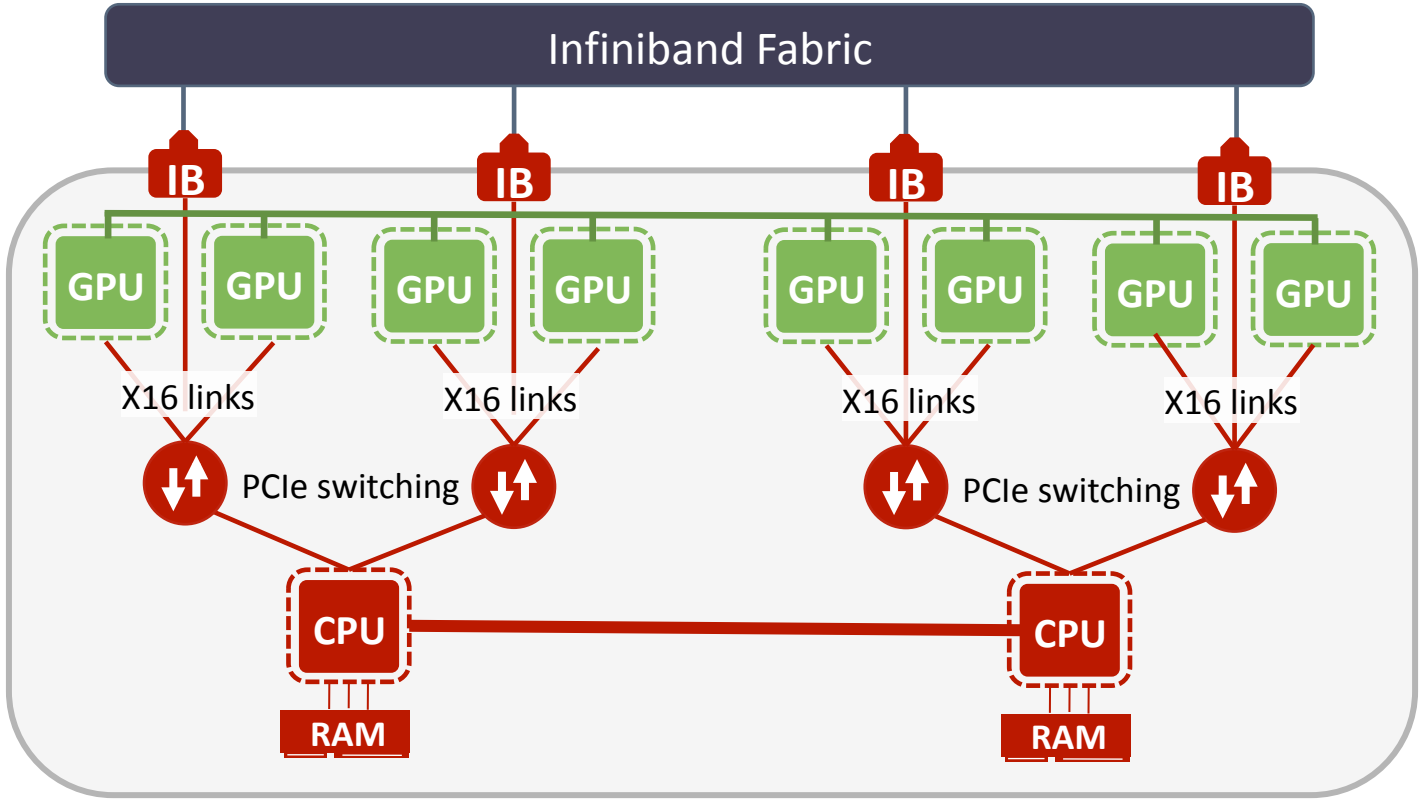


ExaScaler for AI Applications

Feature	Importance for AI	ExaScaler	Competing PFS
Support for high-performance mmap() I/O Calls	High - many AI applications use mmap() calls	✓ Strong	✘ Extremely poor
Container Support	High - most AI applications are containerized	✓ Available	✘ Poor (network complexity & root issues)
Data Isolation for Containers	Medium/High – important for shared environments	✓ Available	✘ Not available today
Unique Metadata Operations	Medium - depends on Installation Size and Application Workflow	✓ Highly scalable	✓ Highly scalable
Shared Metadata Operations	High - training data are usually curated into a single directory	✓ Up to 200K	✘ Lower than 10K (minimal improvements with v5)
Data-on-Metadata (small file support)	Medium/High – depends on data set	✓ DOM is highly tunable	✘ DOM only for files smaller than 3.4k

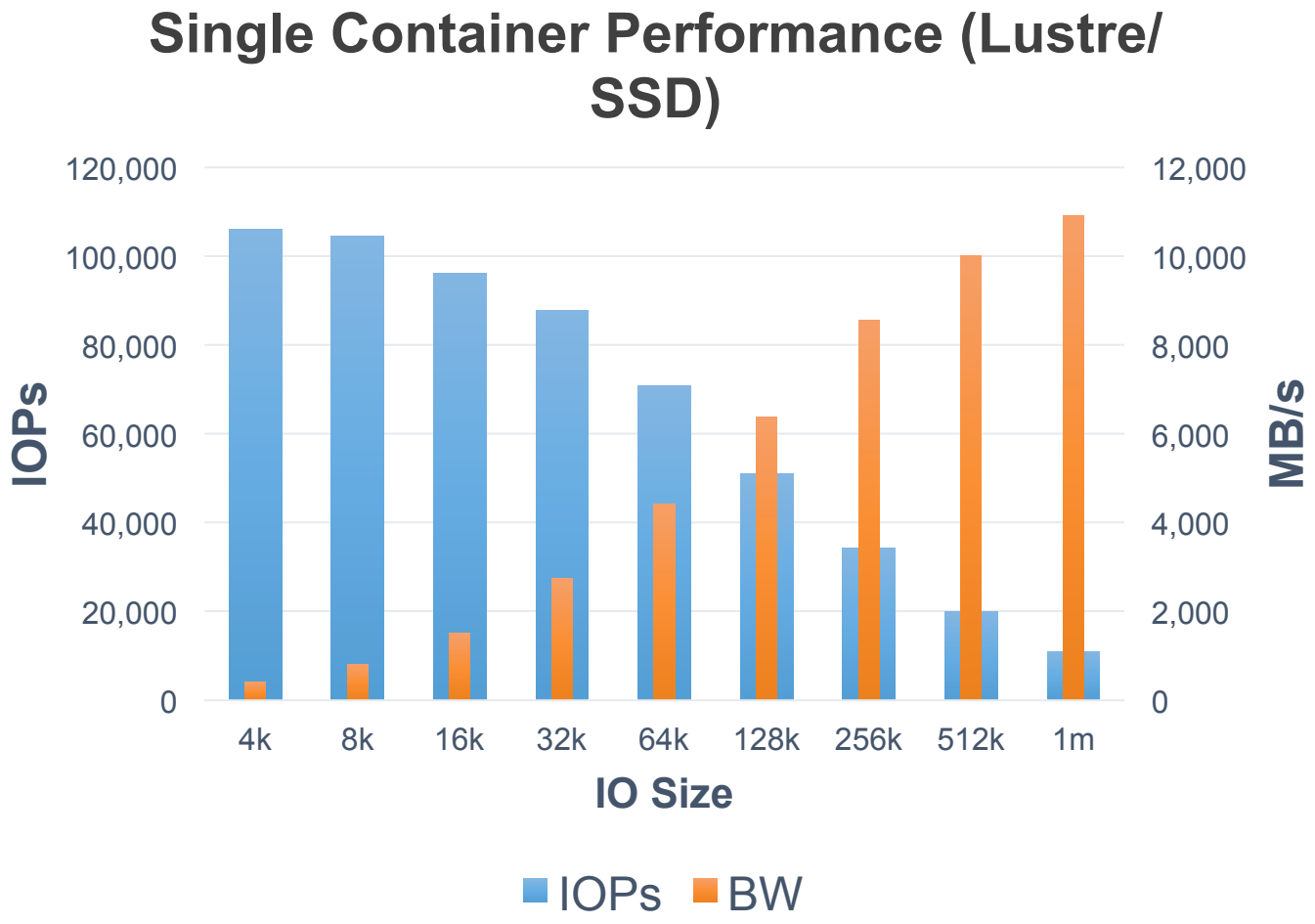
ExaScaler for DGX-1 Architecture

- Consists of Dual Intel Board with a total of 8 GPUs
- One Infiniband EDR HBA per two GPUs
- Containerized Software Stack
- Lustre Mount-point per Container
- Container-pinning for Streamlined I/O with Minimal Overhead



Single Container Performance

- Up to 11 GB/sec per Container
- Above 100K IOPS per Container
- Typical DL Workloads Read Data in 128K Chunks
- I/O Throughput of Around 6 GB/sec and 60K IOPS for Operational Workloads
- Scales with the Number of Containers



IME: Transparent Flash Acceleration for ExaScaler/ Lustre

Adaptive | Lean Data-Path | Write Anywhere

Full Scale-Out | Distributed | Declustered

- Reduces **Power** Up to 10x, Shrinks **Footprint** Up to 20x and Extends **SSD Life** Up to 5x
- Tough Workloads Become **NVMe Optimized**
- Wirespeed, RDMA Support, Linear Scaling
- Transparent I/O, Application, Filesystem Mgmt

IME® SCALE-OUT FLASH

Truly Software-Defined

Commodity Hardware

Highly Available

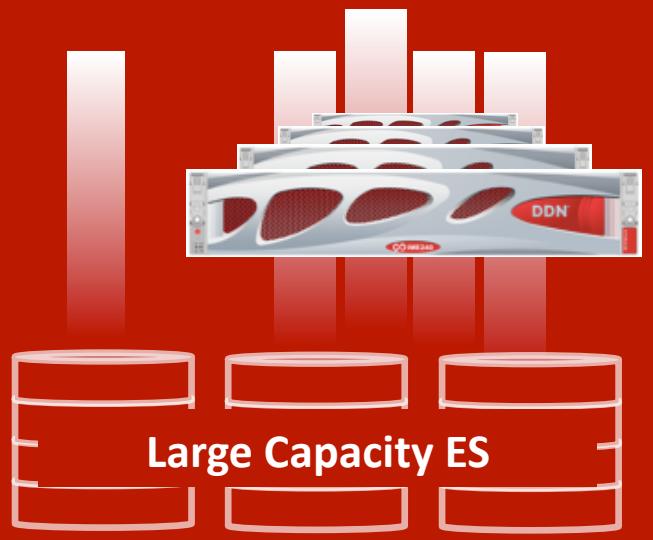
Low Power, High Density

Machine Learning

CPU/GPU Scale Out

600% Better IO Throughput

Massive Hot Data Cacheing

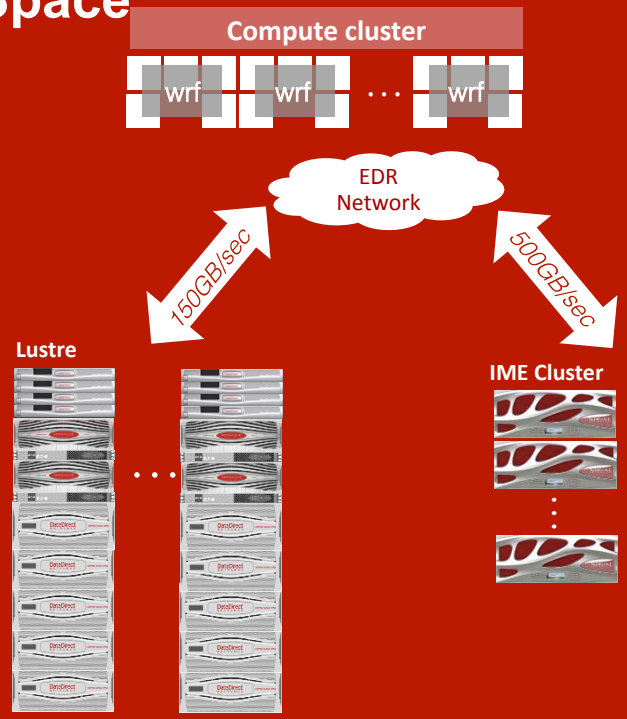


Weather

Concurrent Job Processing

400% Better IO Throughput

10X Less Power 20X Less Space

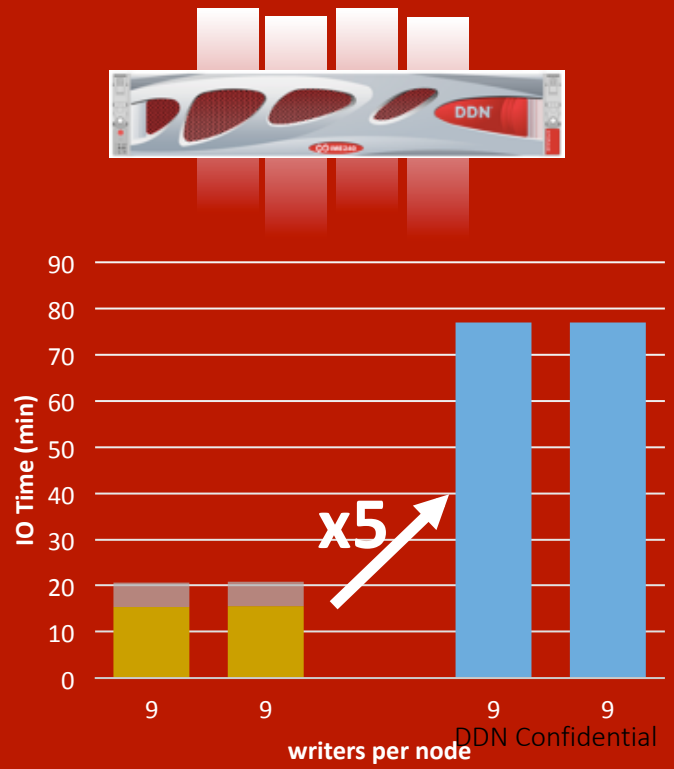


Oil & Gas

Concurrent Job Processing

5X More Writers Per Node

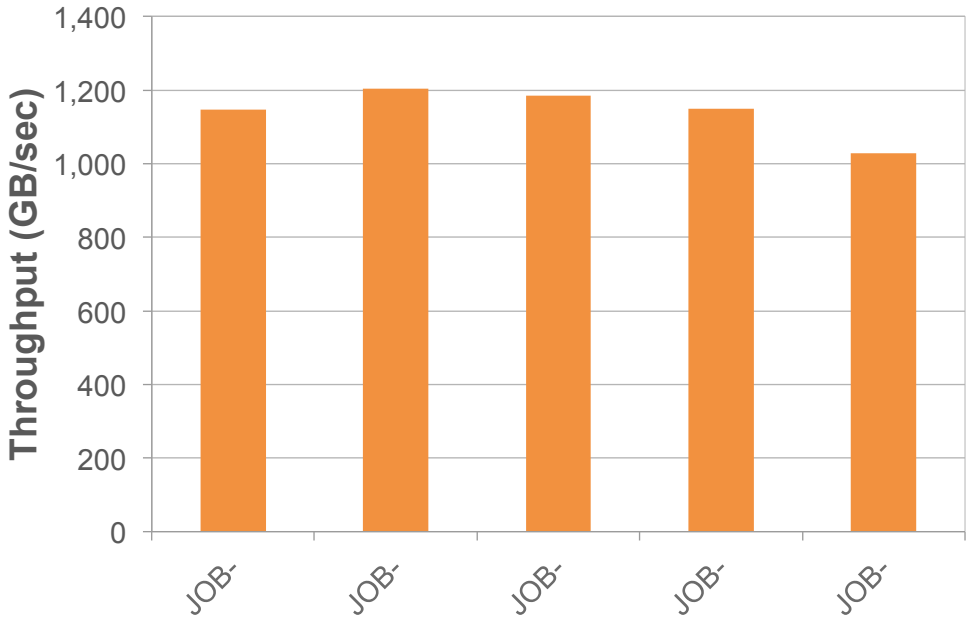
20X Faster Simulations



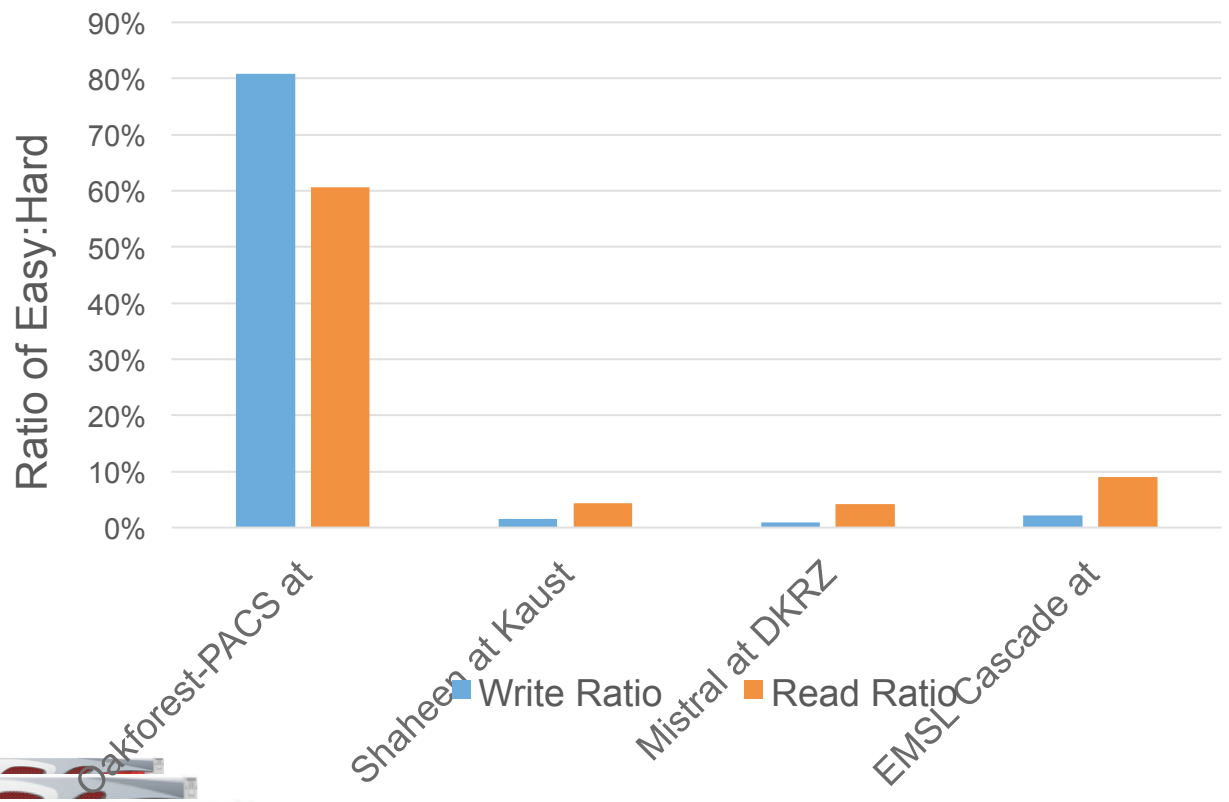
ExaScaler Transparent Caching with IME

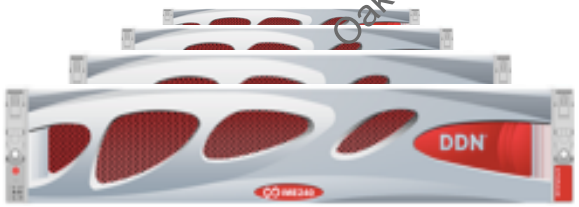
JCAHPC

Throughput Performance at OakForest PACS

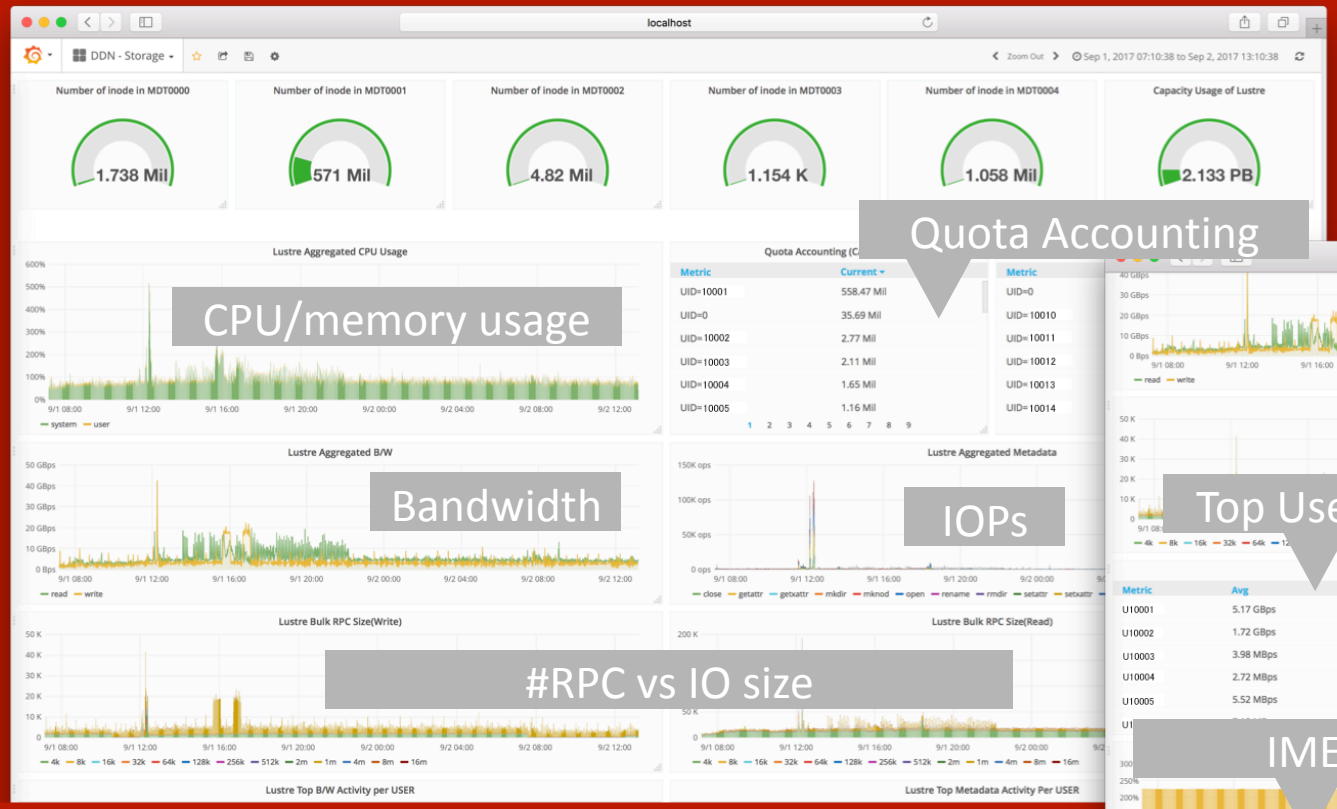


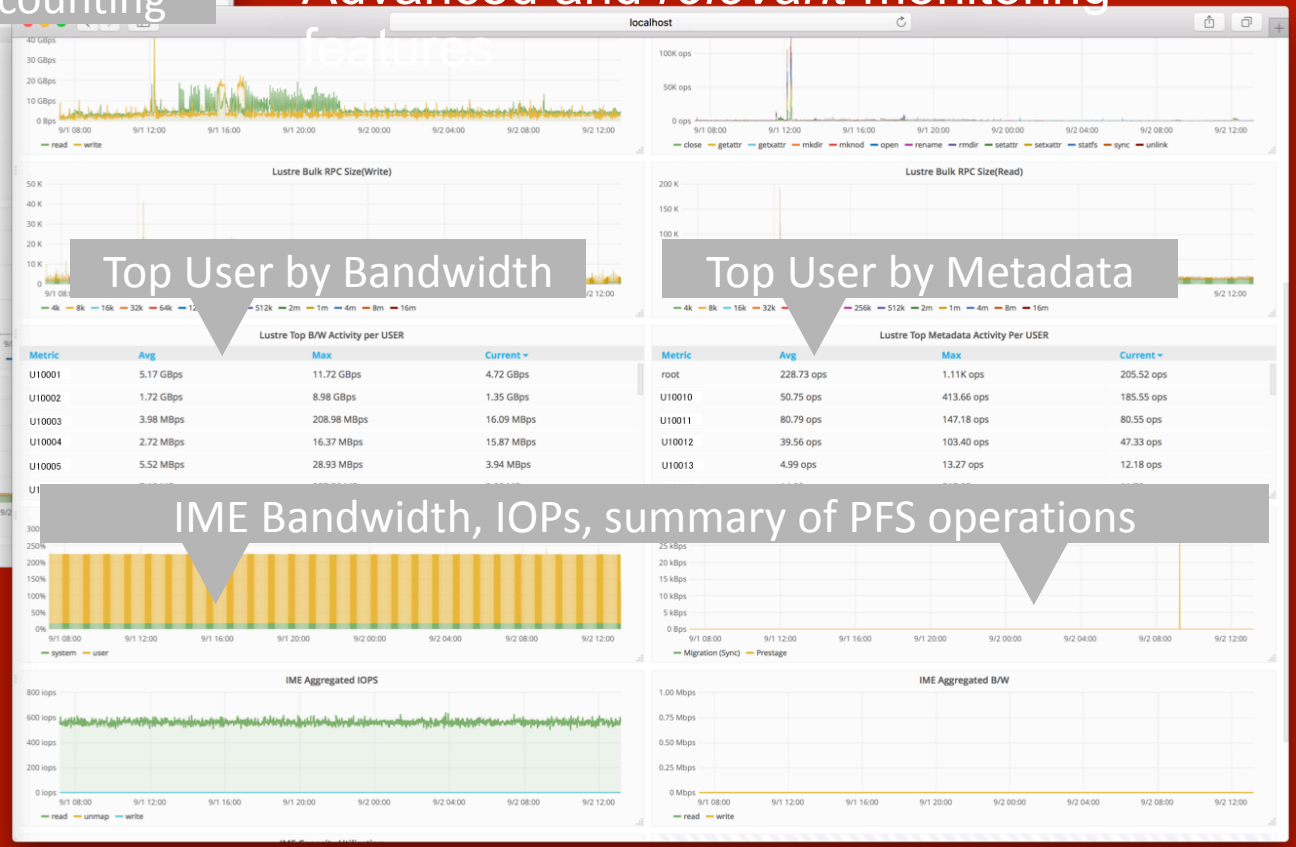
IO500 Results Ratio of "Easy" to "Hard"





ExaScaler Monitoring at Scale





Highly Scalable to 1000's of clients field-tested

Advanced and *relevant* monitoring features

Monitor Lustre File Systems

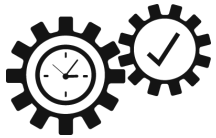
Integration with DDN Hardware

Extended features: JOBSTATS, IME Integration

DDN Lustre Development Focus

2016	2017	2018
Project Quota	Fast Metadata Scanning	LIPE Policy Engine
Online Upgrades	MPI File Utils	LIME Management
Monitoring	RAS Features	QoS Usability
Iadvice	QoS Framework	Metadata Performance
	Persistent Client Cache	Small File Performance
		ExaScaler for AI
Subdirectory Mounts	CoreOS Support	Secure Lustre FW
Lustre Isolation	Container Support	Scalable Archive
	ZFS Support	ExaScaler ZFS
		End-to-End T10-PI

ExaScaler Appliances



Maximum Usability

- Enterprise RAID Stack
- VirtIO SCSI Stack
- Integrated MDS

Flash Optimized

- Fast Metadata
- NoSQL, Lightweight
- MDS Integrated

Enhanced security

- MLS, Audit
- Isolation, Kerberos
- Encryption at Rest

OpenZFS

ZFS Enabled

- Reliable SES
- Lustre 2.10
- Exascaler Quality

DDN Product Offering

Storage Software

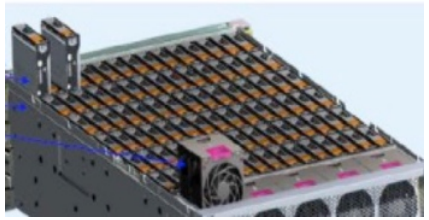
- WOS
- IME
- ExaScaler SW
- RED (End 2018)

SDS Appliances

- WOS
- IME
- ExaScaler ZFS
- RED (End 2018)

Enterprise Storage

- SFA Block
- SFA VIrtualization
- ES Appliances
- GS Appliances



Platforms

- JBODs
- Systems
- Special Development Projects

SCALE UP, SCALE OUT, OR BOTH

800 GB/s
10M IOPS

SCALE SPEED

~40 GB/s in 2RU

MAXMIUM FLEXIBILITY

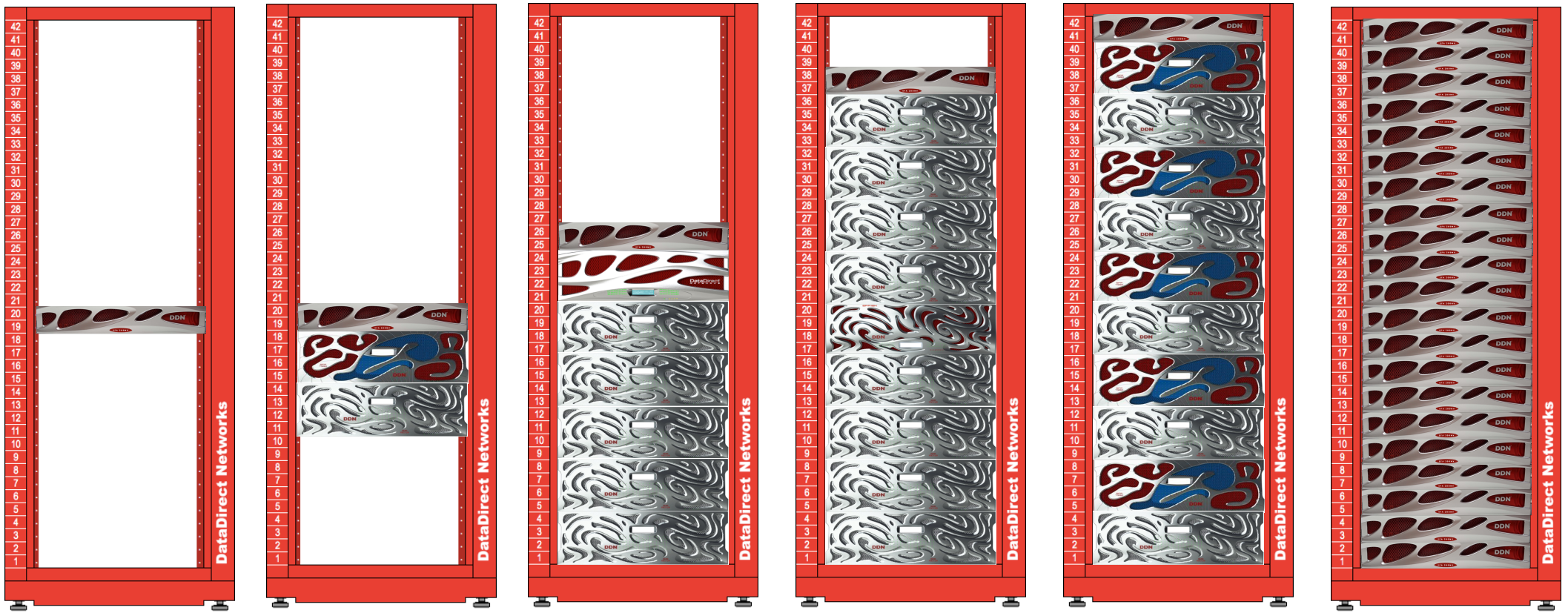
SCALE VOLUME

~1PB in 4RU

70 GB/s
8.9 PB

Up to 900 devices per Standard 42U Rack

ExaScaler Appliances – XS, S, M, L, XL



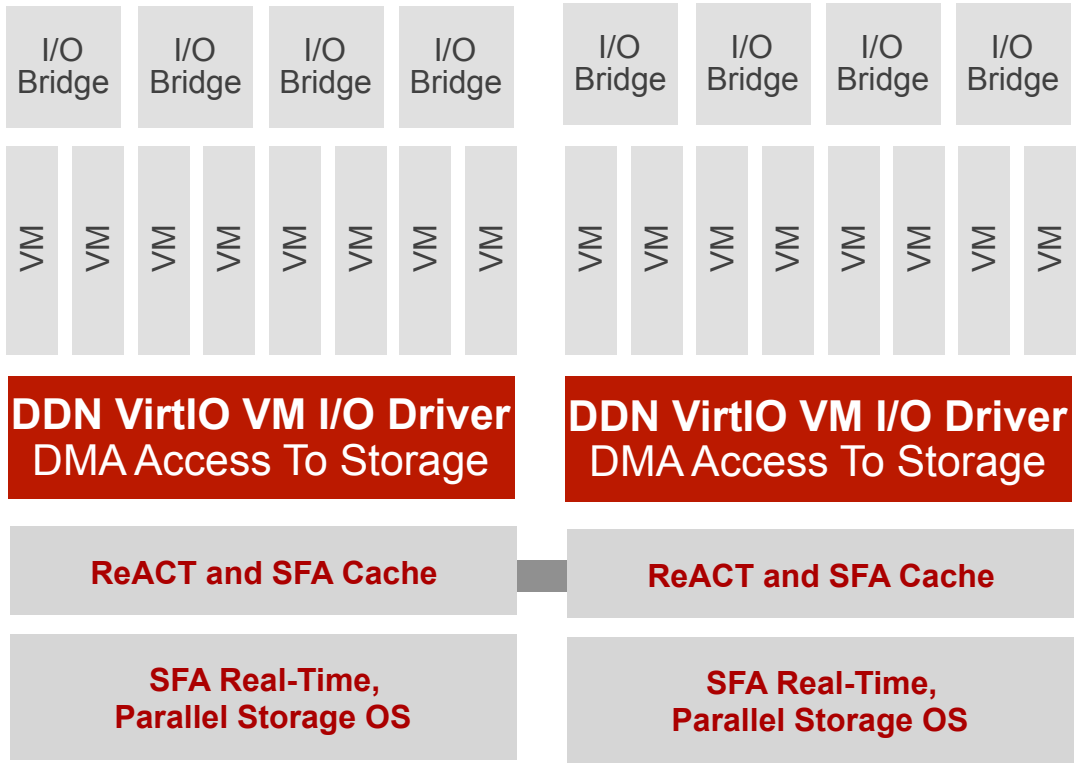
20-40 GB/sec 150 TB NVMe	20 GB/sec 2 PB HDD	40 GB/sec 5.3 PB HDD 550 TB SSD	70 GB/sec 8.6 PB HDD 550 TB SSD	100 GB/sec 9.6 PB HDD	800 GB/sec 3 PB NVMe

Enterprise Systems

ES14KX
40 GB/sec
1 Million IOPS
48 x NVMe/72 x SAS
450/900 HDDs
Redundant Enclosures
De-clustered RAID
VirtIO SCSI

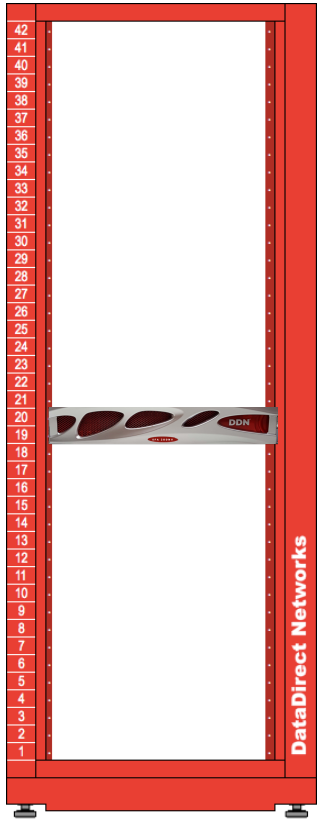
ES18K
70 GB/sec
1.5 Million IOPS
48 x NVMe/72 x SAS
720/1440 HDDs
Redundant Enclosures
De-clustered RAID
VirtIO SCSI

SFA Virtualization Stack



- Easy Management Through Industry Standard VM Stack
- Flexible Implementation for virtualized OSSs and MDSs
- RAID Stack and OSS Run in Isolated OS Instances
- Increase Performance for Throughput and IOPS

Modular Systems



ES200NV/
SFA or ZFS
2 or 4 CPU Skylake
20-40 GB/sec
Up to 24 x NVMe
MDS-MDT or OSS-OST
De-clustered RAID**
VirtIO SCSI**

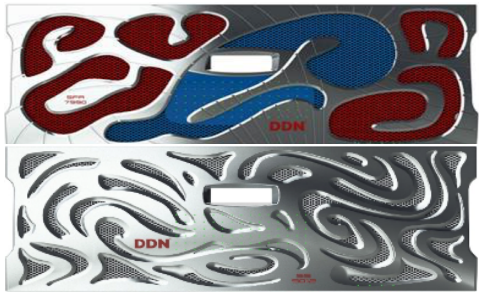
** SFA OS Only.

ES7990/ES7990-Z
SFA or ZFS
2 or 4* CPU Skylake
Up to 20 GB/sec
80/160/260 x HDD
Mainly OSS-OST
De-clustered RAID**
VirtIO SCSI**

* ZFS Only.

ldiskfs with SFA vs. ZFS

Metadata SSU

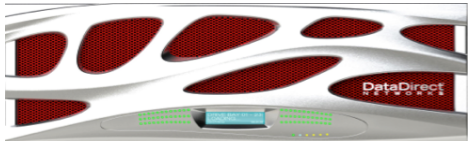


Data SSU

	ldiskfs with SFA	ZFS with RAID Z
Throughput	Higher	Lower
Metadata	Higher	Lower
IOPS	Higher	Lower
Boot & Failover	Faster	Slower
Usable Capacity	~79%	~73%
De-clustered RAID	Available	Future
Virtualization	VirtIO SCSI	Not Available
System Management	Highly Mature	Early Stage
Devices per CPU	Up to 400	80-100

All Flash Systems

ES200NV/
NVMe
1.5 Million IOPS
20-40 GB/sec
SFA or ZFS



ES14KX
NVMe or SAS
1 Million IOPS
40 GB/sec
SFA

ES18KX
NVMe or SAS
1.5 Million IOPS
70 GB/sec
SFA

Future

1. Expanding the Market for Lustre	2. New Features for New Markets	3. Robustness, Availability, Usability
AI, Deep Learning, and Enterprise Analytics	DOM and Scalable Metadata	Continuing Investments in Test Automation
Multi-tenancy for Cloud and Data Isolation	Improved Support for Flash and NVMe	Lustre Health Monitoring
Lustre for Scalable Archive	Enhanced Protocol Export: NFS, CIFS, S3	Improved Logging and Failure Reporting