Whamcloud

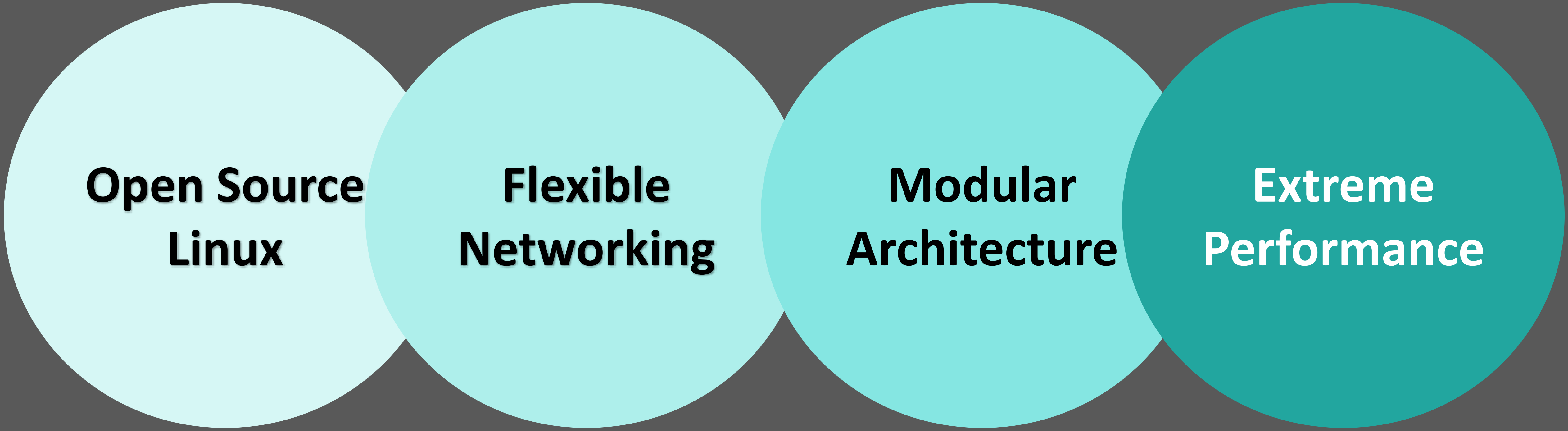
Lustre for HPC and AI

DDN® STORAGE

SC18

Dallas, TX | hpc inspires.

CONFIDENTIALITY NOTICE



Open Source Linux

Flexible Networking

Modular Architecture

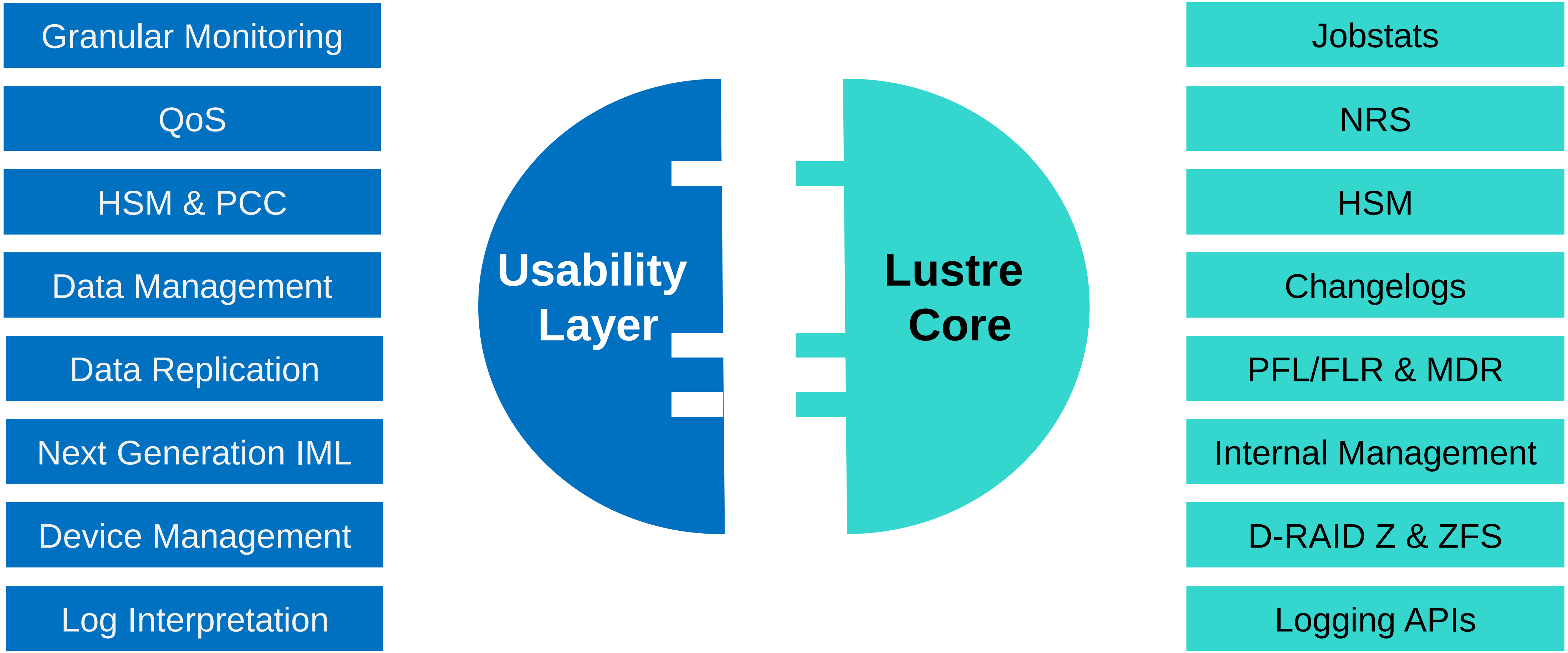
Extreme Performance

Whamcloud & DDN

Whamcloud

- Open Source Lustre
- Advanced Management Features
- Integration with DDN Hardware
- Performance Optimization
- New Application Areas
- Workload Optimization
- Packaged Solutions
- End-to-End Support

DDN/Whamcloud Open Source Model



Granular Monitoring

QoS

HSM & PCC

Data Management

Data Replication

Next Generation IML

Device Management

Log Interpretation

Usability Layer

Lustre Core

Jobstats

NRS

HSM

Changelogs

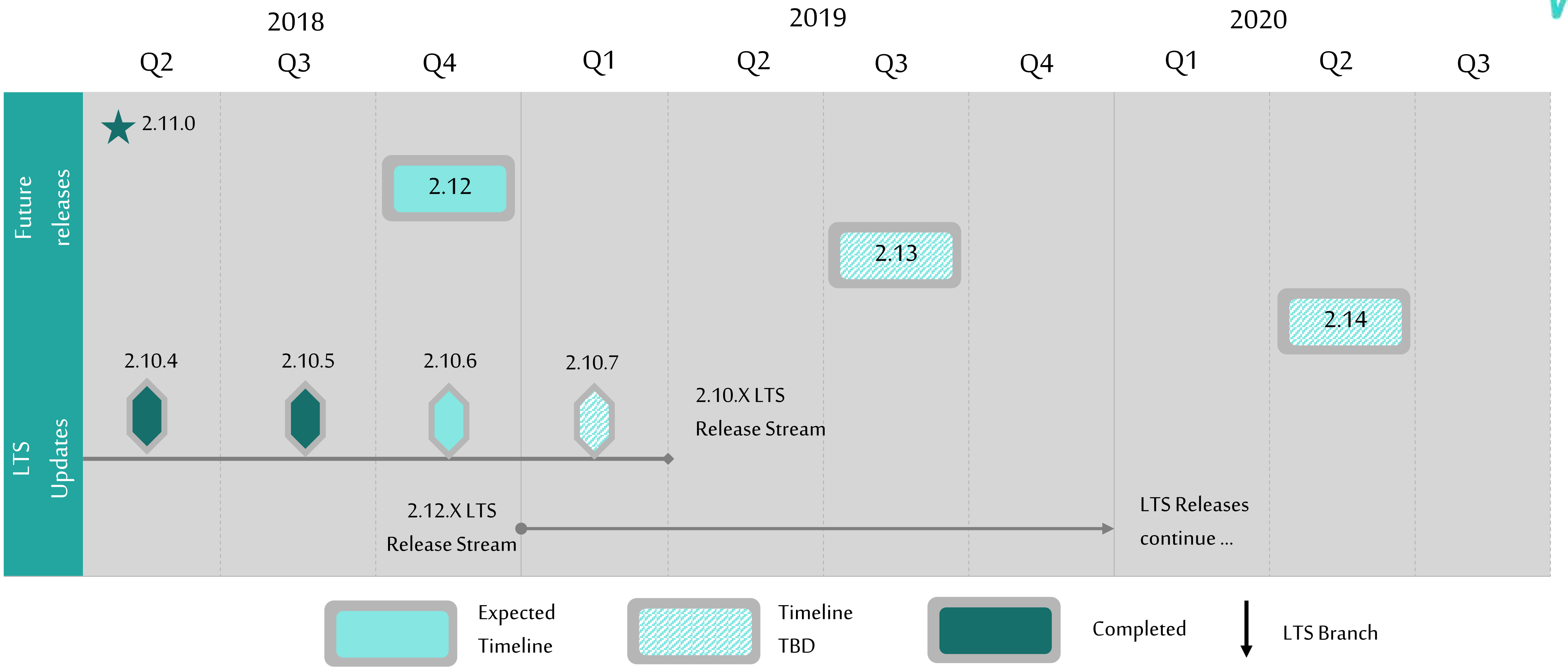
PFL/FLR & MDR

Internal Management

D-RAID Z & ZFS

Logging APIs

Official Lustre Community Roadmap



2.11	2.12	2.13	2.14
- Data on Metadata	- Lazy size on MDT	- Persistent Client Cache	- FLR Erasure Coding
- FLR Delayed resync	- Lnet Health	- LNet Selection Policy	- Health Monitoring
- Lock Ahead	- DNE Dir restriping	- Self Extending Layouts	- DNE Auto Restriping

Estimates are not commitments and are provided for informational purposes only
Fuller details of features in development are available at http://wiki.lustre.org/Projects

Lustre Core – Key Features Roadmap

					Future Versions	
	2.10.x	*2.11*	***2.12 LTS***	*2.13*	*Possibly 2.14*	*Possibly 2.15*
Lustre Core Key Features	✓ Progressive File Layouts ✓ NRS Delay Policy ✓ Project Quotas ✓ Multi-rail LNET ✓ Simplified User Space ✓ Snapshot (ZFS)	✓ Data on Metadata ✓ FLR Delayed Sync ✓ Lock Ahead	❑ LNET Multi-rail health ❑ DNE Directory restriping ❑ Lazy Size on Metadata ❑ T10 DIF support ❑ Support for Open ZFS 0.8.x	❑ Persistent Client Cache ❑ Lnet Selection Policy ❑ Self Extending Layouts	❑ FLR Erasure code striping ❑ DNE Auto Remote Dir striping ❑ Health Monitoring	❑ Client metadata write back cache (CWBC) ❑ Client Container Image (CCI) ❑ Metadata Replication
	Features landed or about to be landed			Features under development	Features being designed and/or prototyped Not committed yet	

Dates and version are subjected to change without prior notification

Lustre-on-Demand: Ephemeral Namespaces

Dynamic File Systems

Temporary fast Lustre filesystem across the compute nodes

LOD creates Lustre on computes nodes dynamically

Scheduler Integration

User turn LOD on/off per job at job submission

Currently integrated into SLURM’s Burst Buffer option, but other job scheduler also could work

Automated Data Staging

User can define file/directory list on stage-in/out to LOD at Job submission

LOD automatically sync/migrate data from persistent Lustre to created temporary Lustre filesystem

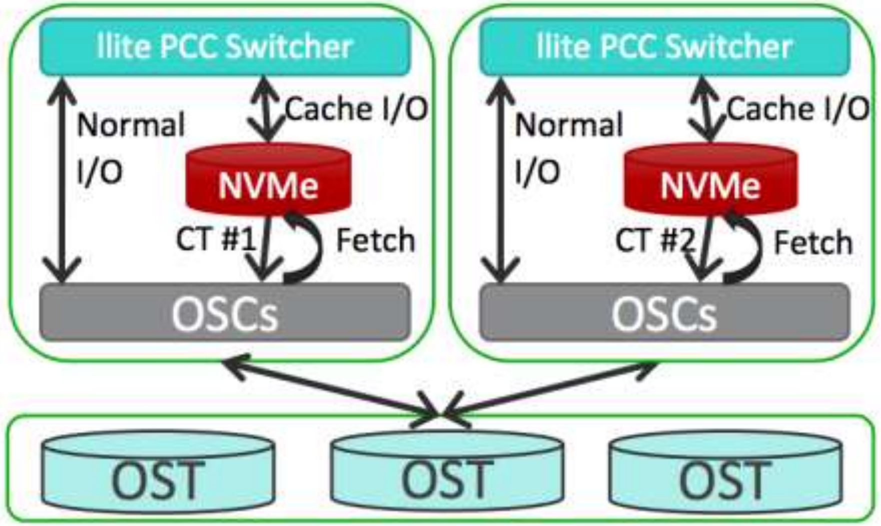
Flexibility

Flexible MDT/OST configuration for advanced users

Persistent Client Cache

Lustre 2.13

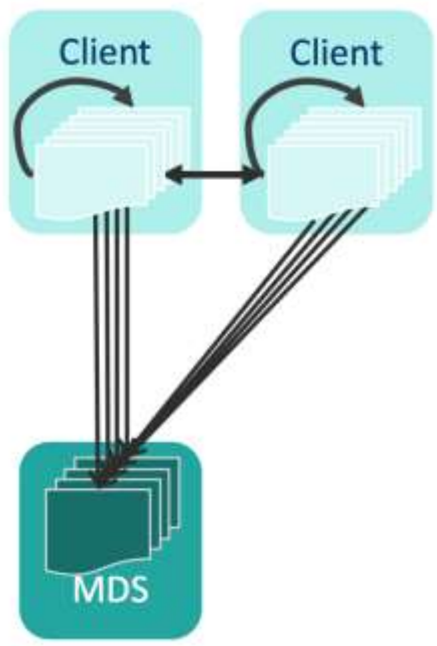
- ✓ Reduce latency, improve small IOPS, reduce network traffic
- ✓ PCC integrates Lustre with persistent per-client local cache devices
  - Each client has own cache (SSD/NVMe/NVRAM) as a local filesystem (e.g. ext4/ldiskfs)
  - No global/visible namespace is provided by PCC, data is local to client only
  - Existing files pulled into PCC by HSM copytool per user directive, job script, or policy
  - New files created in PCC is also created on Lustre MDS
- ✓ Kernel uses local file if in cache or normal Lustre IO
  - Further file read/write access "directly" to local data
  - No data/IOPS/attributes leave client while file in PCC
  - File migrated out of PCC via HSM upon remote access
- ✓ Separate functionality read vs. write file cache
- ✓ Could later integrate with DAX for NVRAM storage



Client Metadata Write Back Cache

Lustre 2.15 or Later

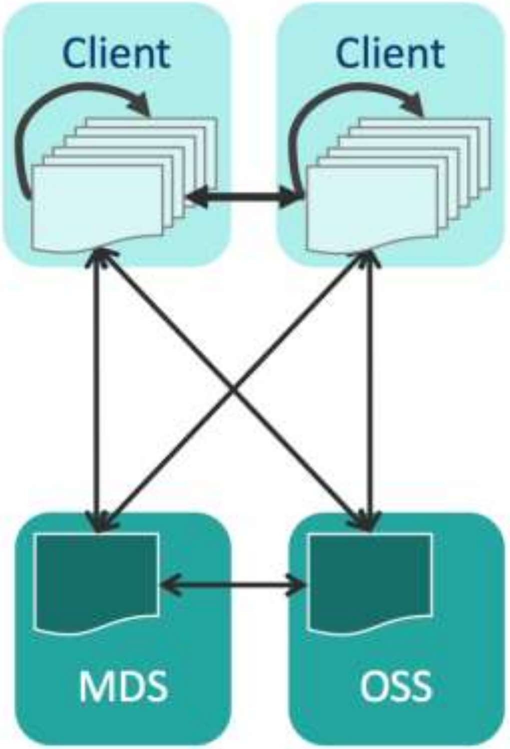
- ✓ Metadata WBC creates new files in RAM in new directory
  - Avoid RPC round-trips for each open/create/close
  - Lock directory exclusively, avoid other DLM locking
  - Cache file data only in pagecache until flush
  - Flush tree incrementally to MDT/OST in background batches
- ✓ Could prefetch directory contents for existing directory
- ✓ Can integrate with PCC to avoid initial MDS create
- ✓ Early WBC prototype in progress
  - Discussions underway for how to productize it
  - Early results show 10-20x improvement for some workloads



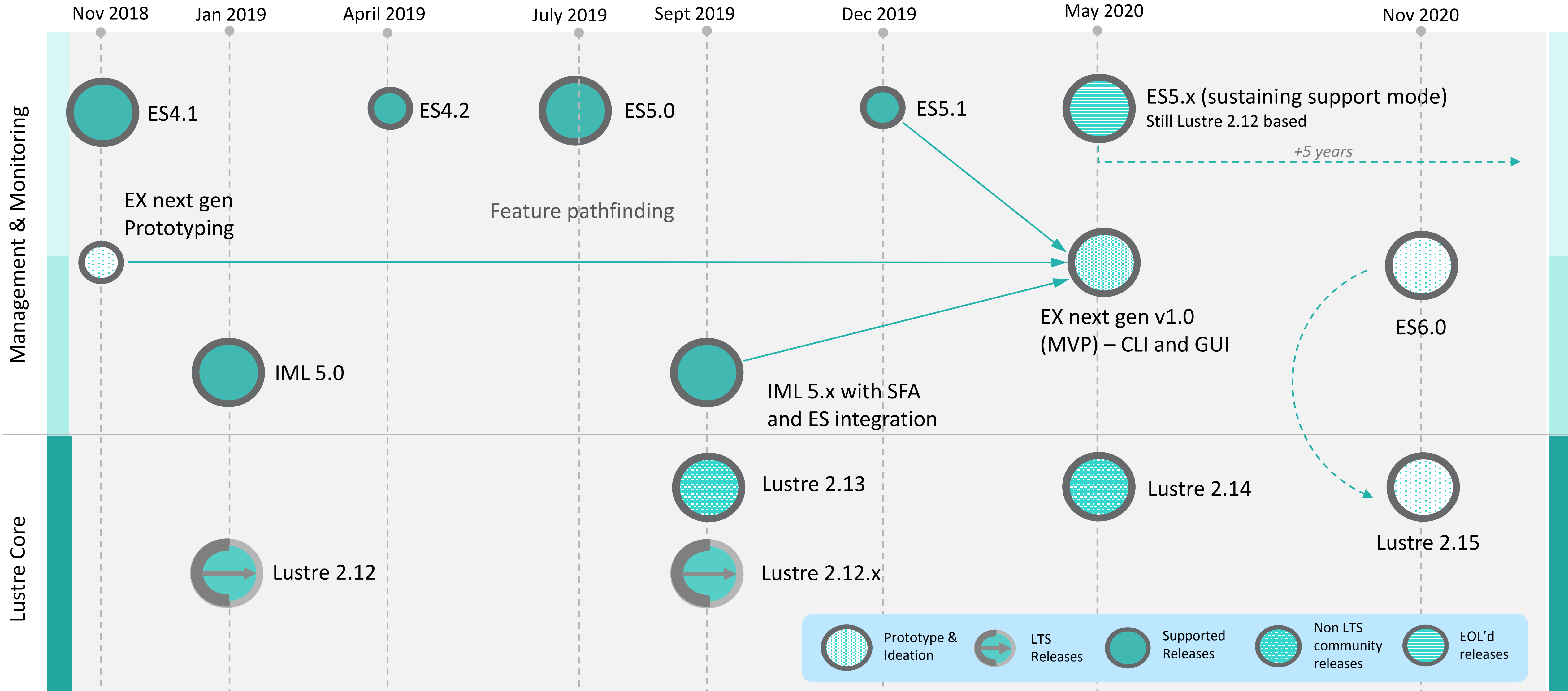
Client Container Image (CCI)

Lustre 2.15 or later

- ✓ Filesystem images were used *ad hoc* with Lustre in the past
  - Read-only cache of many small files manually mounted on clients
  - Root filesystem images for diskless clients
- ✓ Container Image is local ldiskfs image mounted on client
  - Holds a whole directory tree stored as a single Lustre file
- ✓ CCI integrates container handling with Lustre
  - Mountpoint is registered with Lustre for automatic mount
  - Automatic local loopback mount of image at client upon access
  - Image file read on demand from OST(s) and/or cached in PCC
  - Low I/O overhead, few file extent lock(s), high IOPS per client
  - Access, migrate, replicate image with large read/write to OST(s)
- ✓ MDS can mount and re-export image files for shared use
- ✓ CCI can hold archive of whole directory tree for HSM
- ✓ Unregister/delete image and all files therein with a few ops



Whamcloud Unified Roadmap Overview



Nov 2018

Jan 2019

April 2019

July 2019

Sept 2019

Dec 2019

May 2020

Nov 2020

Management & Monitoring

ES4.1

ES4.2

ES5.0

ES5.1

ES5.x (sustaining support mode)
Still Lustre 2.12 based

+5 years

EX next gen
Prototyping

Feature pathfinding

EX next gen v1.0
(MVP) – CLI and GUI

ES6.0

IML 5.0

IML 5.x with SFA
and ES integration

Lustre Core

Lustre 2.13

Lustre 2.14

Lustre 2.15

Lustre 2.12

Lustre 2.12.x

Prototype & Ideation

LTS Releases

Supported Releases

Non LTS community releases

EOL'd releases

Dates and version are subjected to change without prior notification

ExaScaler – Key Features Roadmap

Whamcloud

Exascaler Features

ES4.1

- ✓ A^{3}I – Support for AI200 (factory deployment)
- ✓ A^{3}I – Support for AI200 (ES_Wizard)

ES4.2

New Platform Support

- □ AI7990 and AI400 support
- □ ES7990, NV400 and ES18K support
- □ Exascaler Docker Container for A^{3}I deployments
- □ IBM Power 9 client support
- □ ARM client support
- □ Large IO & Read Performance
- □ DDN Insight and DDN Performance Monitoring convergence
- □ Lustre on GCP
- □ LORIS v1
- □ *Tech Preview: NFS & NFS Gateway Reference Configuration*
- □ *Tech Preview: Lustre-on-Demand*

ES5.0

SFA Performance

- □ Lustre 2.12 LTS based with Support for all 2.12 Features
- □ IML Management for ExaScaler
- □ Call Home v1
- □ IOPS Improvements
- □ Small IO Performance v1
- □ Metadata Performance for “E”
- □ DNE2 Scaling
- □ Lustre-on-Demand
- □ T10-PI End-to-End Data Integrity
- □ Whamcloud Protocol Gateway Cluster v1: NFS & CIFS
- □ DDN Data Flow Support v1
- □ *Tech Preview: LiPE FS Accounting*
- □ *Tech Preview: DCS Platform*

ES5.1

NVMe & Data Management

- □ LiPE GA v1
- □ Transparent SSD Pools v1
- □ SFA FStrim Support
- □ Optimized DDN Hardware support for Data on Metadata
- □ Lustre Persistent Client Cache (LPCC)
- □ Dsync Integration v1
- □ Whamcloud Protocol Gateway Cluster v2: Object Support
- □ *Tech Preview of New Management and Monitoring Framework*
- □ *Tech Preview: Object Support for Whamcloud Gateway Cluster*
- □ *Tech Preview: LWBC*

ES6.0

Lustre Next Gen

- □ Integrated Management and monitoring
- □ New Lustre Management framework for High Availability, deployment and scalability
- □ Lustre Write Back Cache (LWBC)
- □ Lustre Client Container (LCC)
- □ Lustre Metadata Replication
- □ Small IO and Metadata Performance Improvements

Features Ready

Features under development

Features being designed and/or prototyped

Dates and version are subjected to change without prior notification

Exascaler – Feature Development

	ES4.1	*ES4.2*	*ES5.0*	*ES5.1*	*ES6.0*
A³I	• Ai200	• AI400, AI7990 • A3I Client Docker containers	• Small IO Optimizations • Metadata Optimizations	• Lustre PCC	• Lustre WBC (LWBC) • Lustre Client Container (LCC)
DDN SFA	• NV200 • ES14KX • ES7700	• NV400 • SFA7990 • SFA18K	• T10-PI Support • *DCS Hardware Tech Preview*	• SFA 18KX Support • SFA7990X, SFA 200/400NV X	• Next Gen SFA (PCIe Gen4)
DDN Performance		• Large IO & SFA Read Performance Increases	• IOPS Optimizations • Small IO Optimizations • SFA “E” Platform Metadata Performance Optimizations • DNE Performance Scaling	• Local Caching with LPCC • *LWBC Tech Preview*	• Metadata Performance with CWBC • Small File Performance with CCI • Metadata Replication
Monitoring & management	• LustrePerfMon Framework Open Source	• IML Monitoring for SFA • DDN Insight and DDN Perf Mon convergence • LORIS v1	• Whamcloud Call Home v0 • IML Management for SFA • Enhanced Lustre monitoring with Insight for Lustre	• *New Management and Monitoring framework Tech Preview*	• Whamcloud Lustre Manager: New Integrated Management and Monitoring for Lustre
Cloud & Enterprise	• Lustre on AWS	• Lustre on GCP • CIFS and NFS Gateway Preview • *Lustre-on-Demand Tech Preview*	• CIFS and NFS Gateway v1 • Lustre-on-Demand	• CIFS and NFS Gateway v2: Object Support	

Dates and version are subjected to change without prior notification

DDN T10-PI End to End Data Integrity

Fully transparent End-to-End Data integrity from Lustre client to disk

Relies on open standard format T10PI/DIX

Any T10PI/DIX supported hardware is usable

Minimum performance impacts

Keep compatibility for old Lustre version or non-T10PI supported hardware

DDN SFA Ready!

DDN CIFS and NFS Gateway

CIFS and NFS (Object in Future)

Consistency & Cross-Protocol Locking

UNIX passwords, LDAP and Microsoft AD

High Availability

Horizontal scalability

Phase 1

- Blue Print – Architectural reference
- Ready to run external servers
- 2-node HA configuration

Phase 2

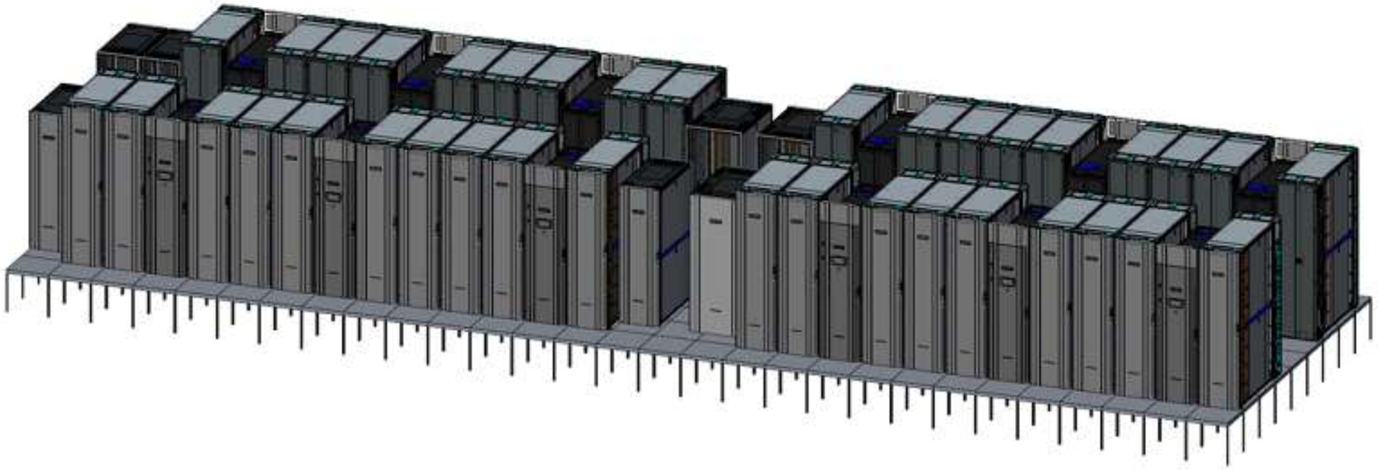
- Integrated appliance model(s)
- No need of external servers
- Integration with Lustre to provide scalability & Performance
- Multi-node HA configuration

Broad CPU Support

- Server & Client
  - x86 Intel
  - x86 AMD
  - ARM (Fujitsu, Cavium)
- GPUs
  - NVIDIA, AMD
- Client Only
  - Power/PowerPC
  - Other

Sandia National Laboratories

Astra



Lustre and Cloud

- ✓ **Today**: Lustre on Public Cloud
  - Support IO-Intensive Applications (e.g. SAS Analytics)
  - Easy Set-up Process
  - DDN/Whamcloud Expert Support
  - Import/Export Data via S3
- ✓ **Future**: Hybrid Cloud for HPC
  - Support data tiering between *on-premise* and *public* cloud instances
  - Migrate temporary workloads to cloud quickly
  - Stage-out cold data to public cloud storage

DDN DATAFLOW | ACCELERATE DATA WORKFLOWS AT SCALE

Protect, vault, move and synchronize

Integrated with Lustre

Access and store data on any storage platform

Safeguard critical information and ensure availability

Enrich data archives with advanced metadata collection

Accelerate data workflows with scalable distributed architecture

Structure data operations with management, monitoring and reporting

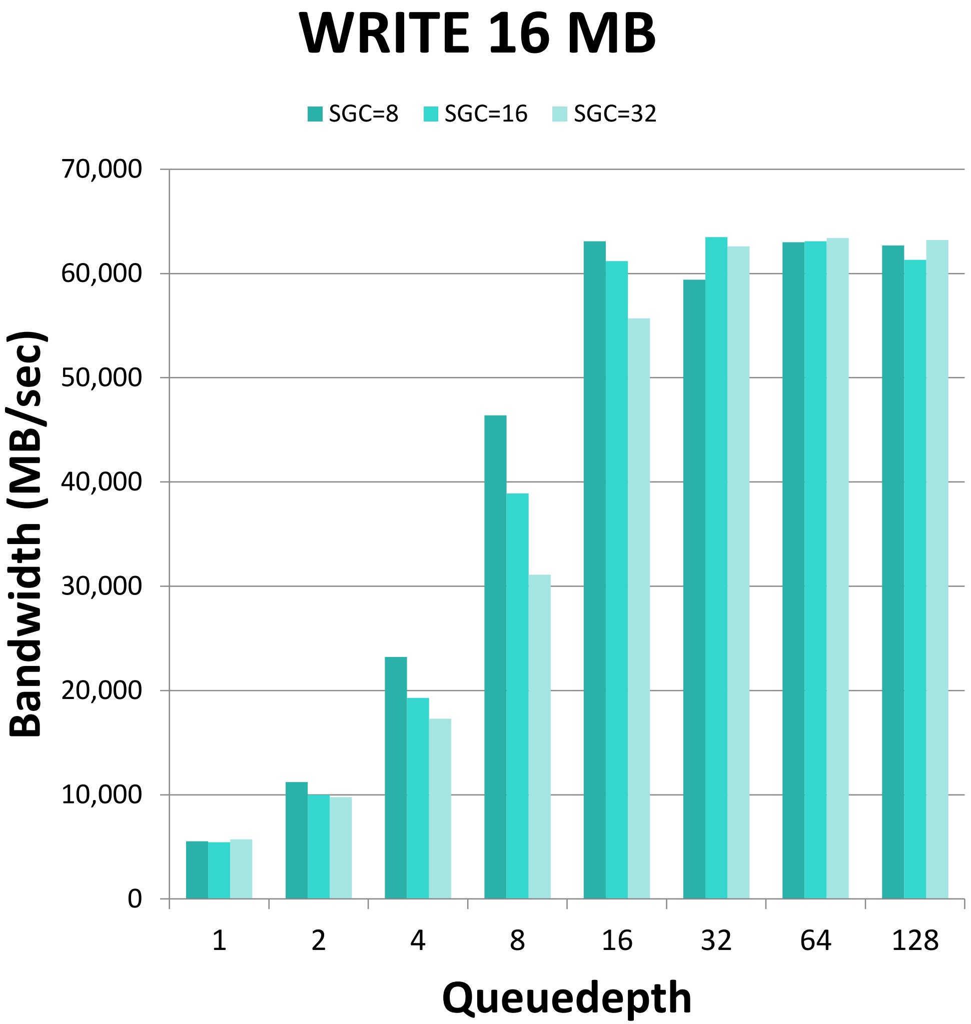
Intuitive interfaces make data management simple and easy

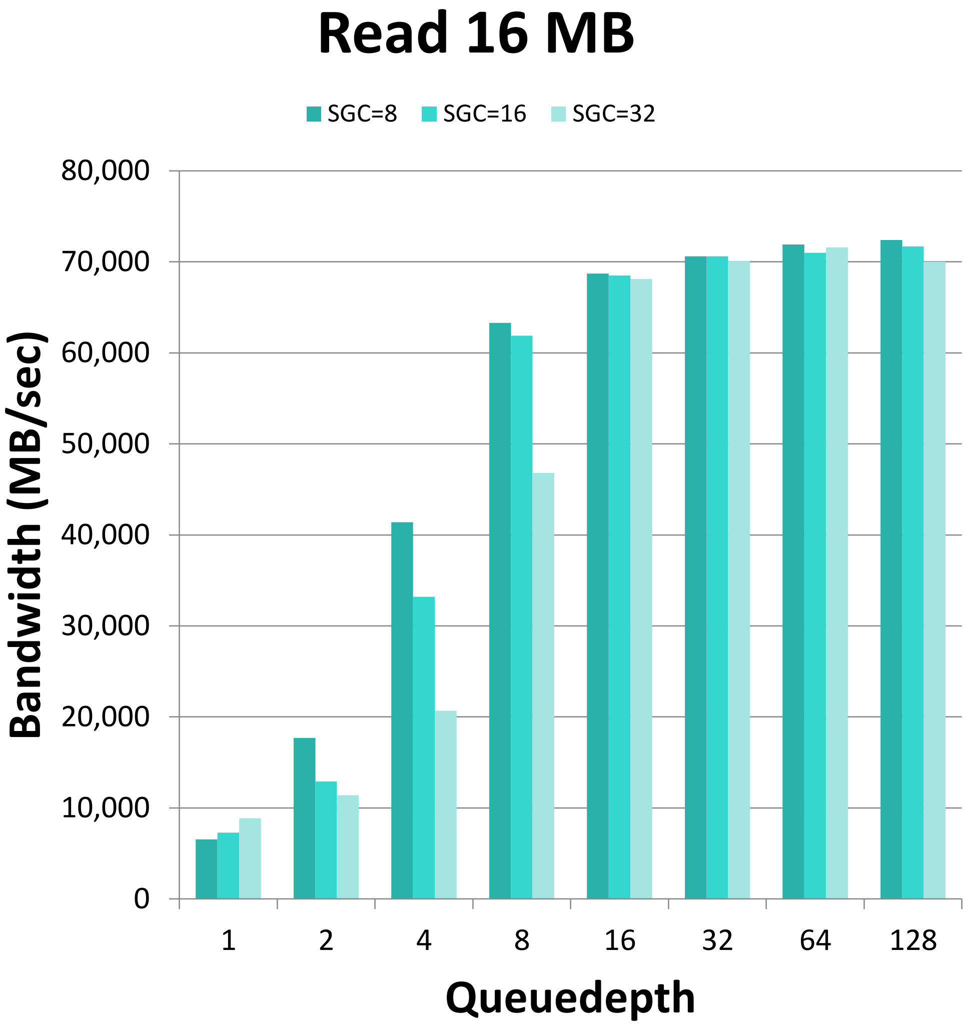
BACKUP ARCHIVE MOVE SYNC

WITH

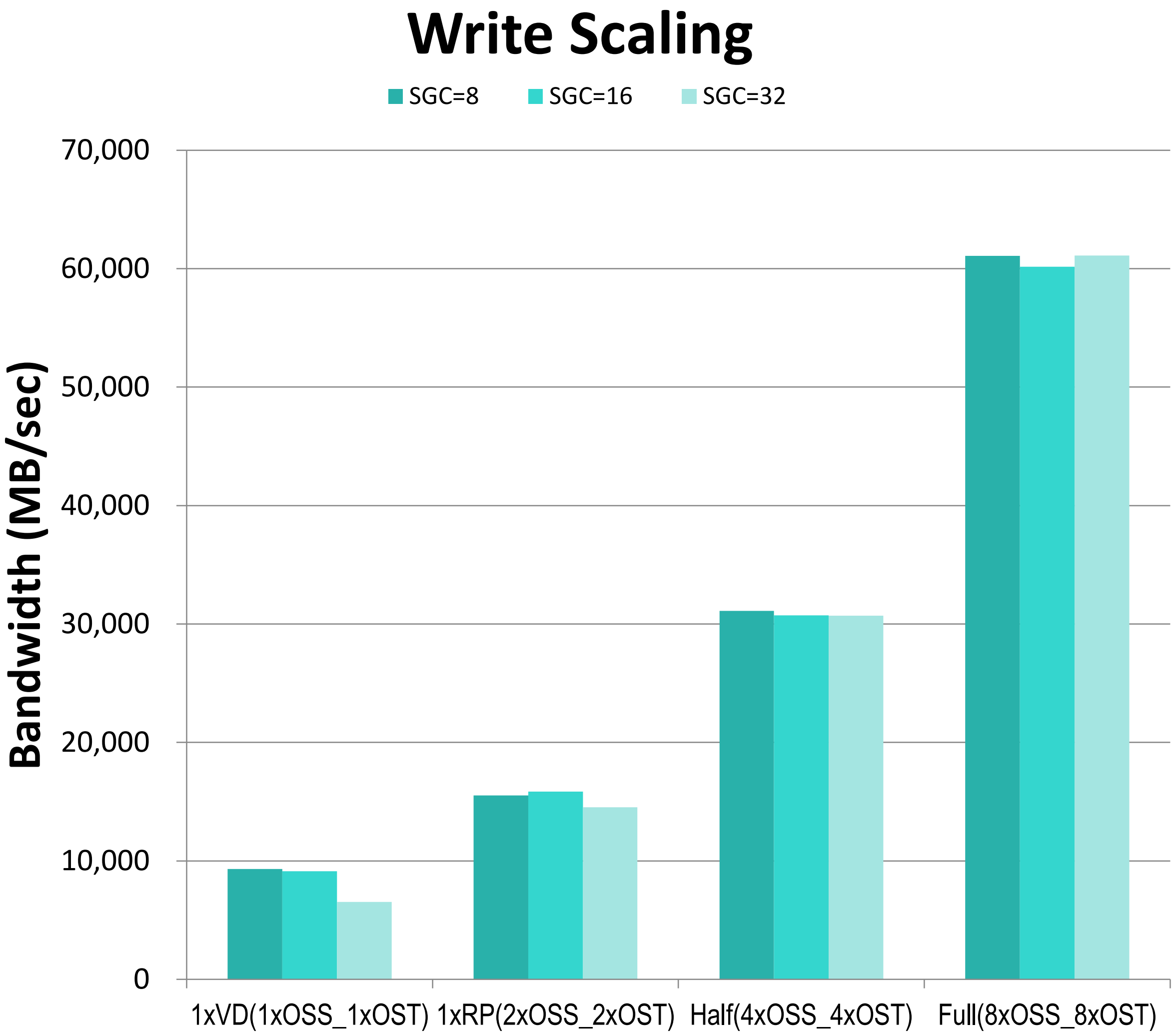
FILE OBJECT CLOUD TAPE

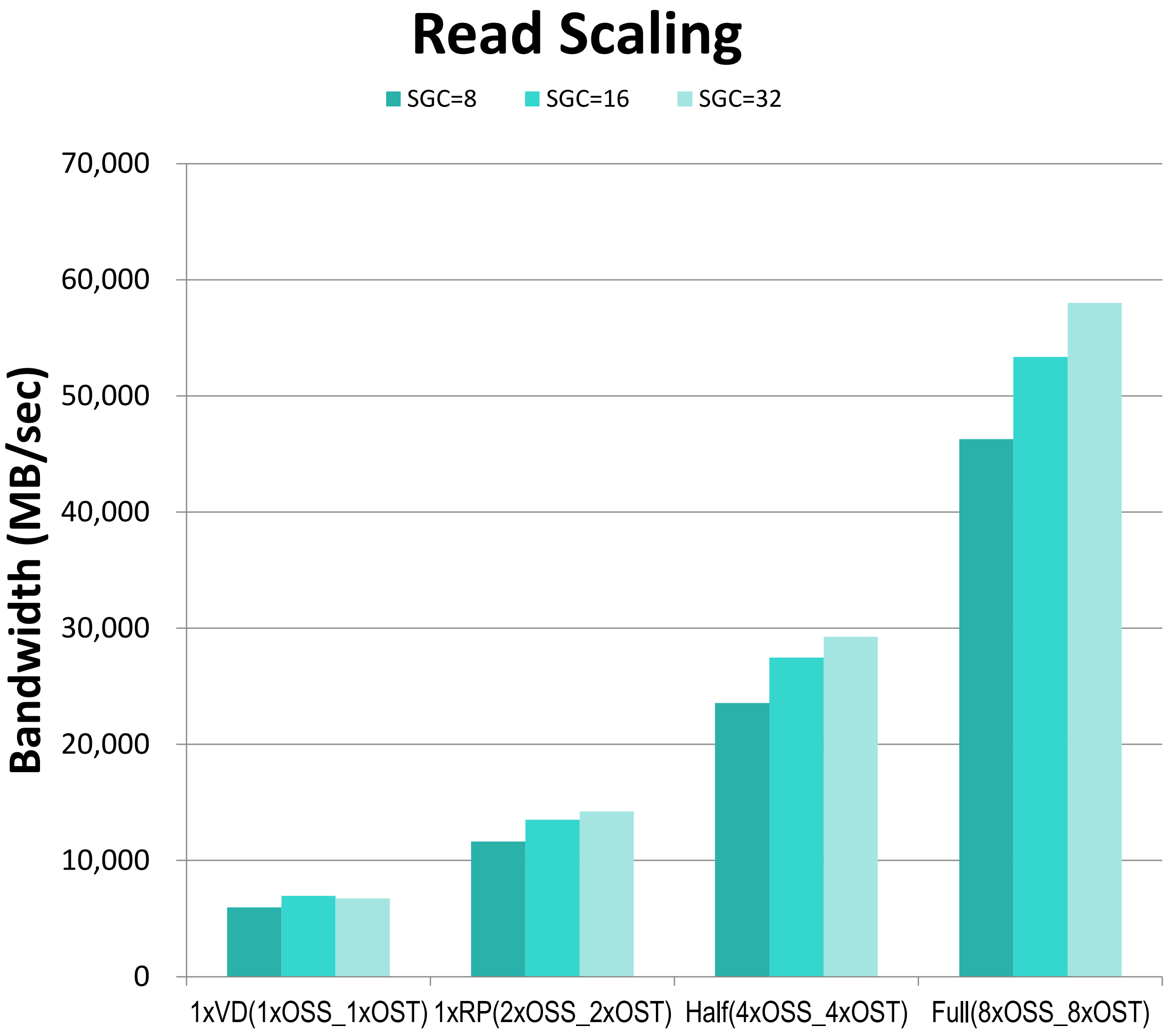
DDN ES18K – Block Performance (400 x HDD)





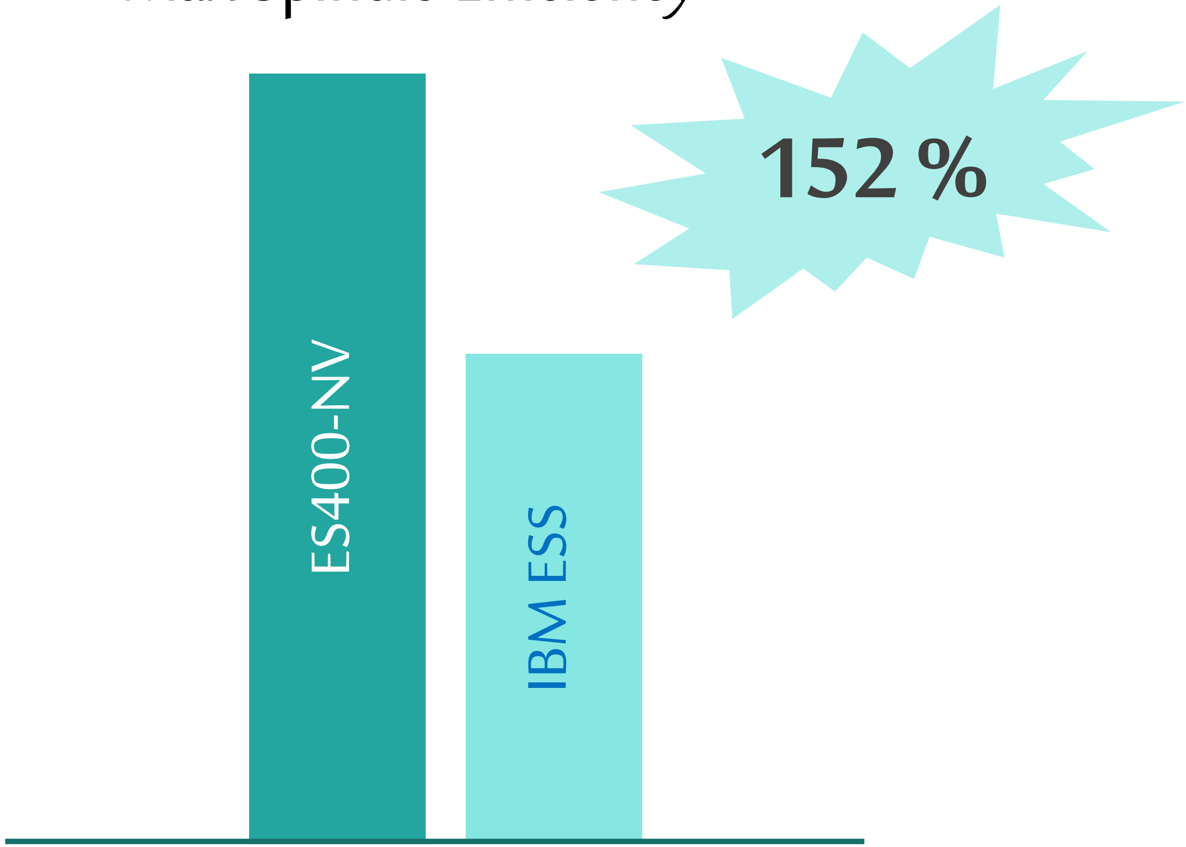
DDN ES18K – IOR 1MB FPP (400 HDD)



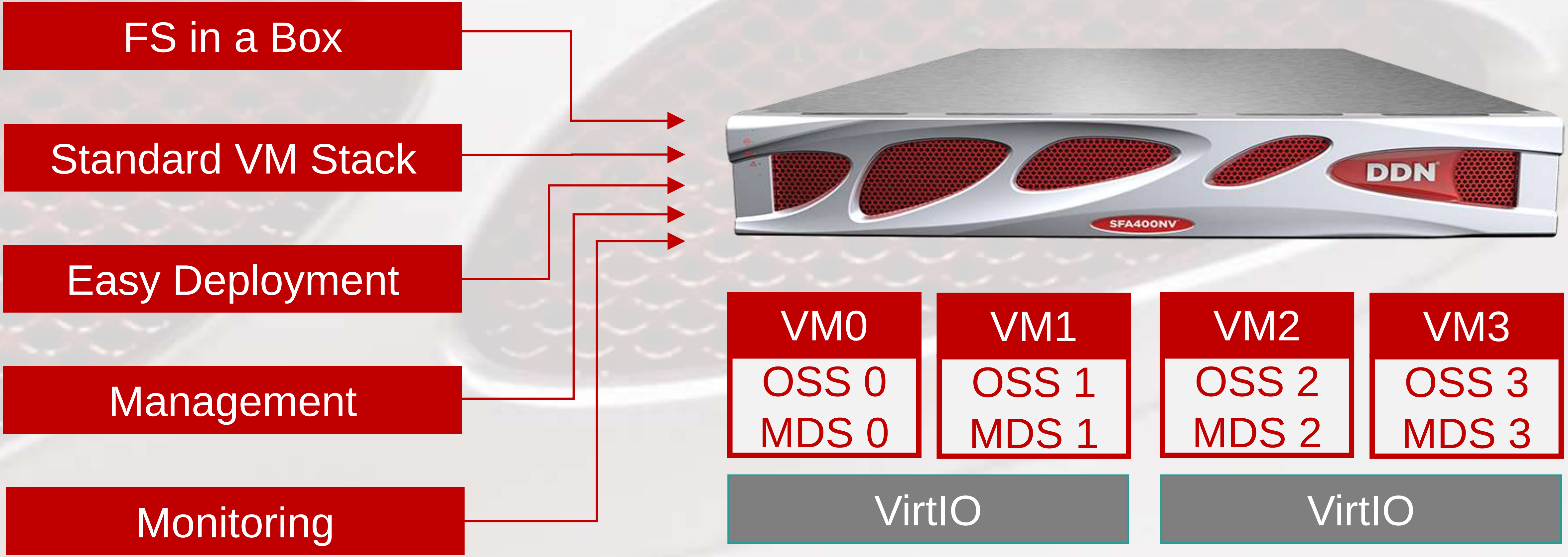


DDN ExaScaler vs. IBM ESS

Max Spindle Efficiency



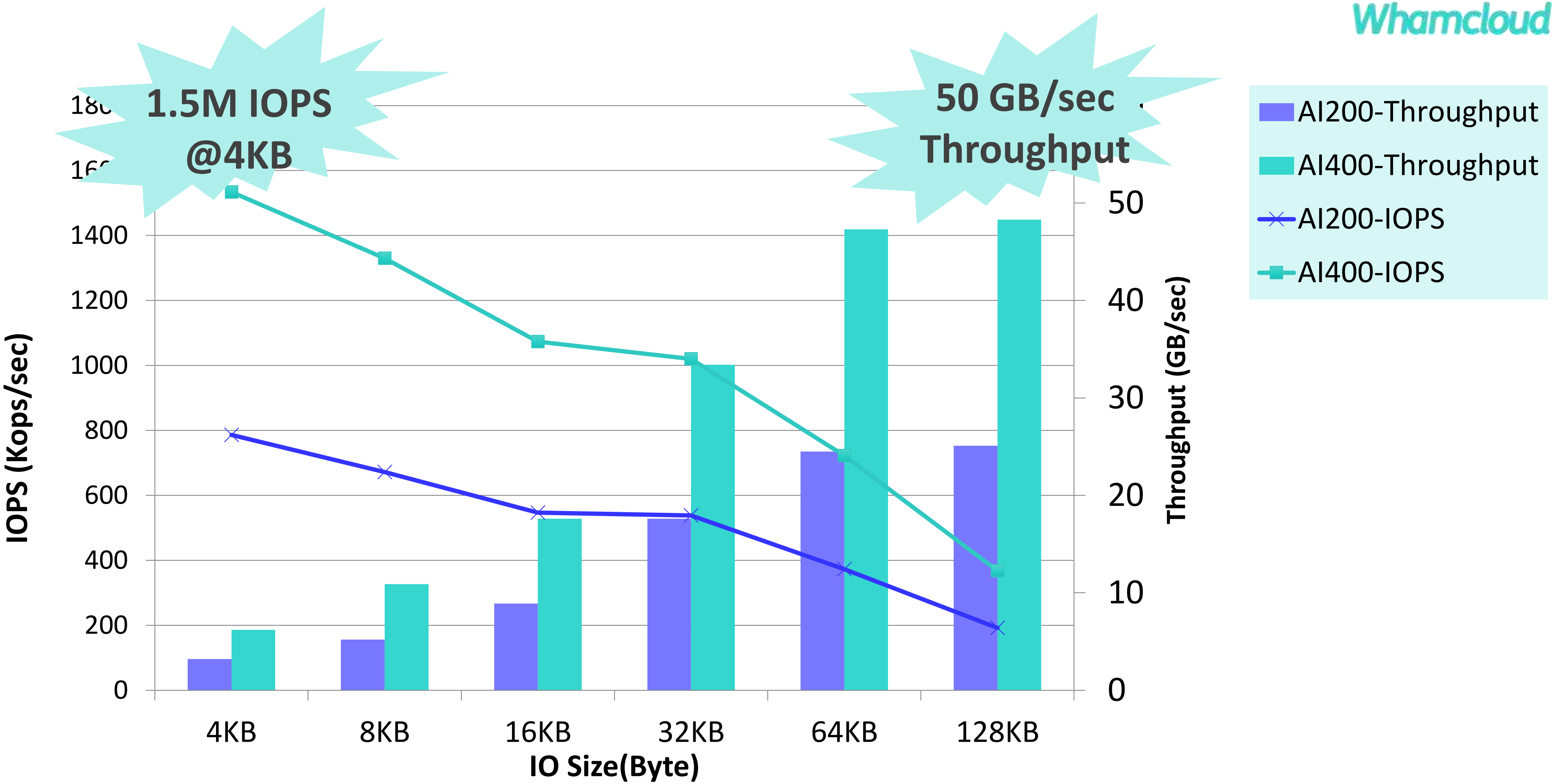
AI/ES 200/400



- FS in a Box
- Standard VM Stack
- Easy Deployment
- Management
- Monitoring

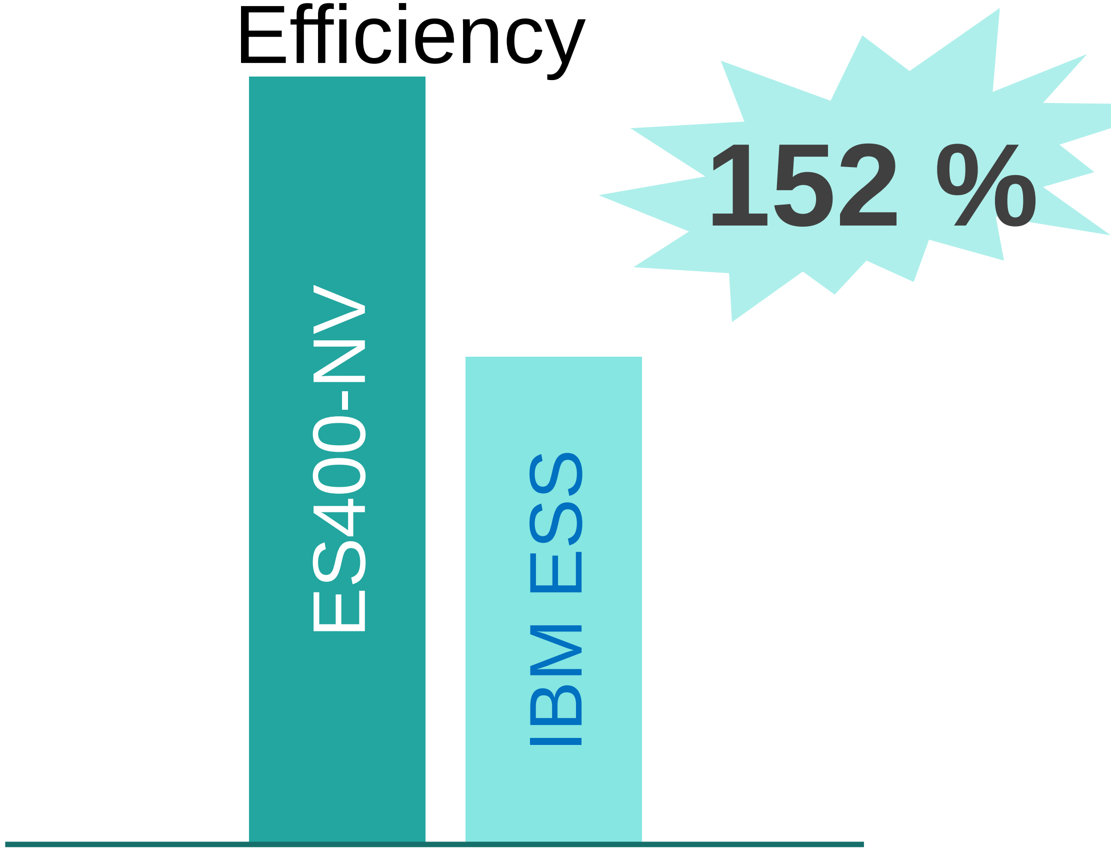
VM0	VM1	VM2	VM3
OSS 0 MDS 0	OSS 1 MDS 1	OSS 2 MDS 2	OSS 3 MDS 3
VirtIO		VirtIO	

AI-200/400 Random Read IOPS and B/W

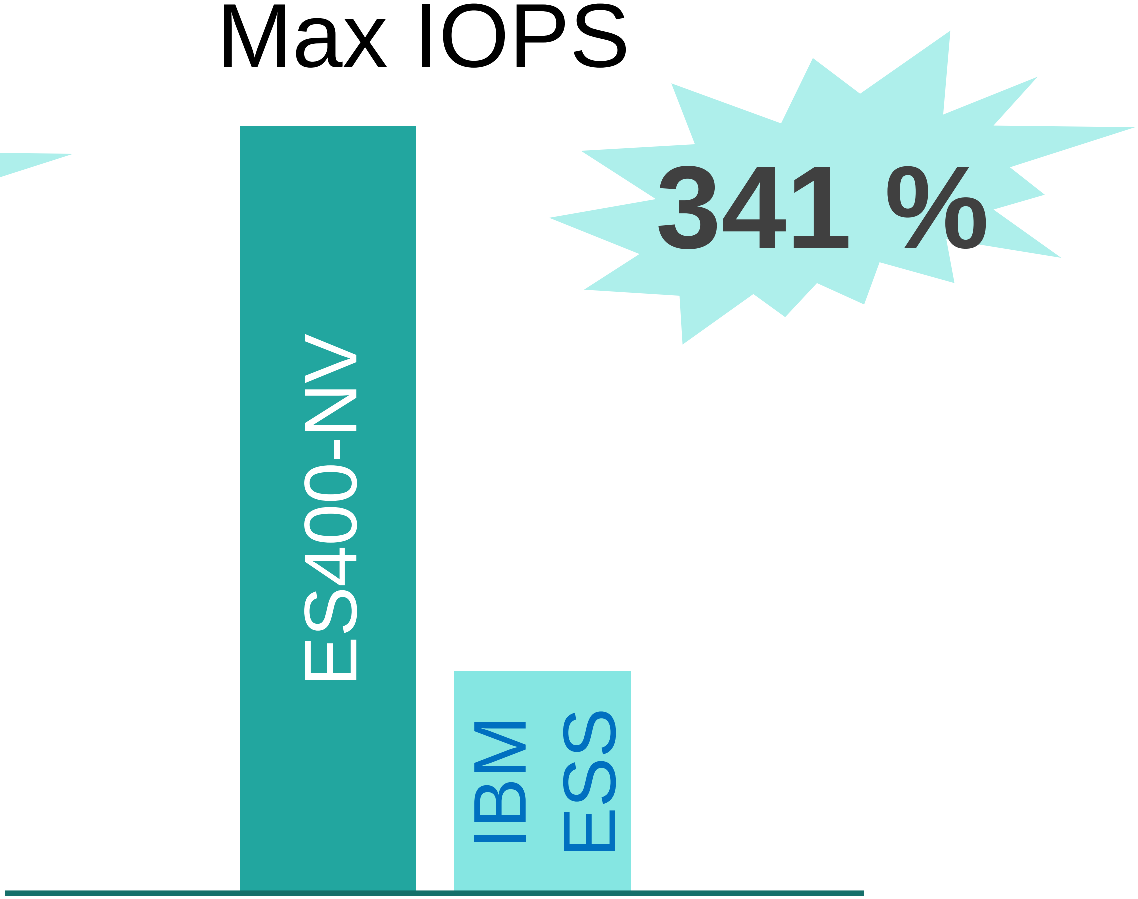


DDN ExaScaler vs. IBM ESS

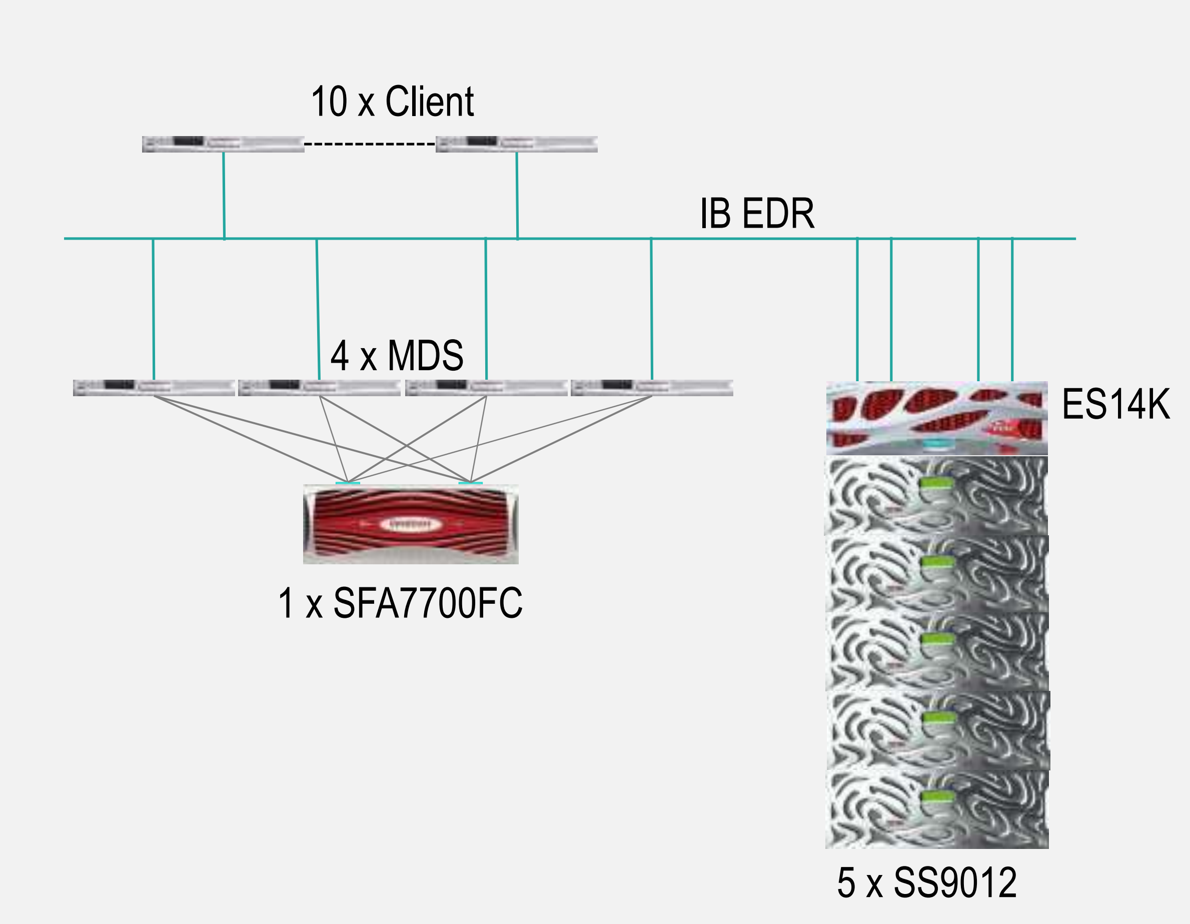
Max Spindle Efficiency



Max IOPS



IO-500 on Pre-ES5.0 Test Configuration



MDS

- 4 x MDS
- 1x Platinum 8160, 96GB RAM, 1 x IB EDR
- 1 x SFA7700 FC
- 2 x MDT(2 x RAID1 SSD) per MDS

OSS

- 1 x ES14K + 5 x SS9012
- 408 x NL-SAS 10TB
- 8 x DCR POOL (51/MR=1)
- 4 x vOSS(8 CPU Core, 90GB RAM, 1x IB EDR

CLIENT

- 10 x Intel Server
- 2 x E5-2650v4, 128GB RAM, 1x IB EDR
- CentOS7.4

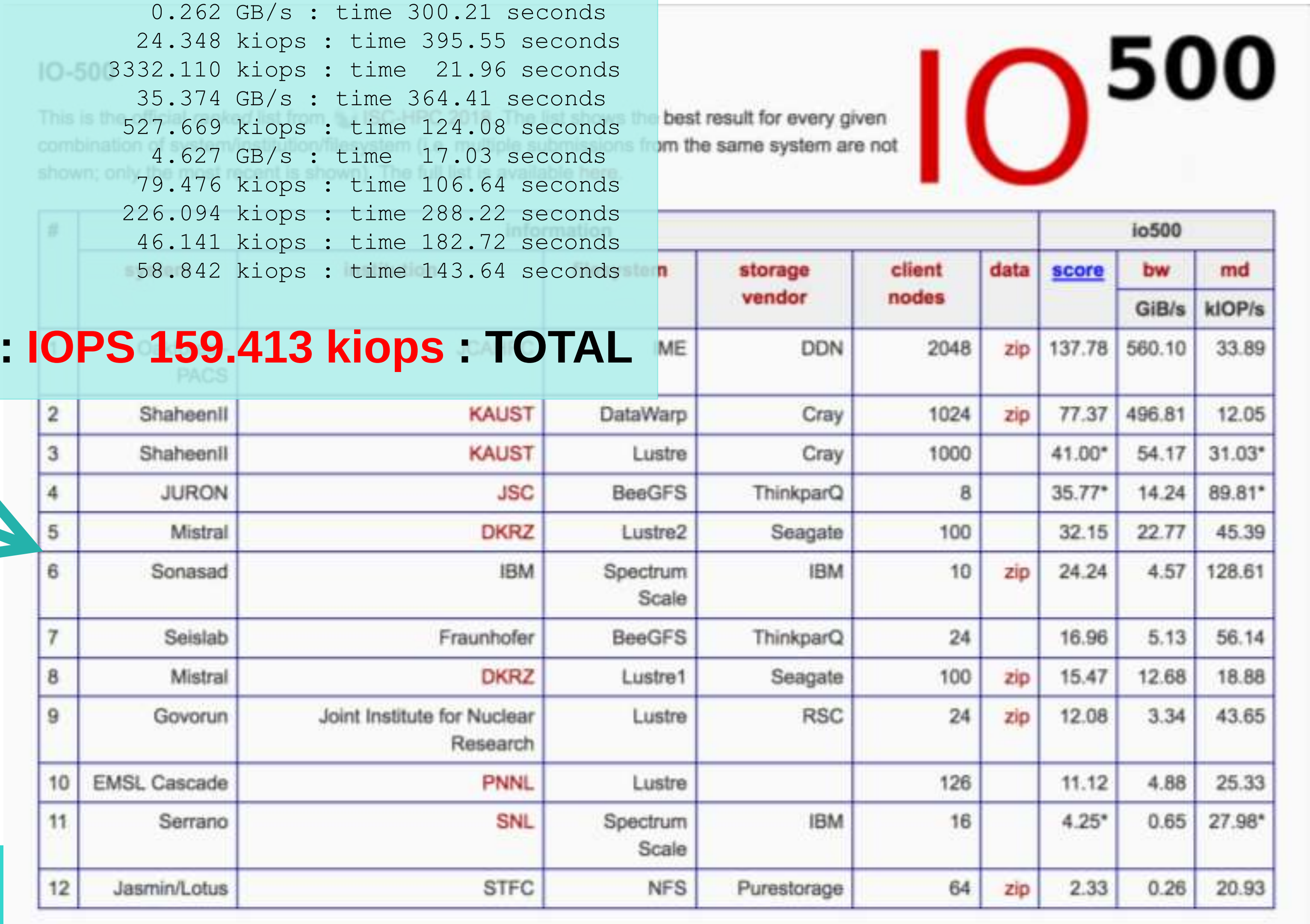
SW

- Pre-ES5.0
- master branch for Lustre-2.12

IO-500 10 Client Node Challenge (Pre-ES5.0 Results)

```
[RESULT] BW   phase 1            ior_easy_write               37.540 GB/s : time 343.38 seconds
[RESULT] IOPS phase 1         mdtest_easy_write              199.685 kiops : time 325.87 seconds
[RESULT] BW   phase 2            ior_hard_write                0.262 GB/s : time 300.21 seconds
[RESULT] IOPS phase 2         mdtest_hard_write               24.348 kiops : time 395.55 seconds
[RESULT] IOPS phase 3                      find             3332.110 kiops : time  21.96 seconds
[RESULT] BW   phase 3             ior_easy_read               35.374 GB/s : time 364.41 seconds
[RESULT] IOPS phase 4          mdtest_easy_stat              527.669 kiops : time 124.08 seconds
[RESULT] BW   phase 4             ior_hard_read                4.627 GB/s : time  17.03 seconds
[RESULT] IOPS phase 5          mdtest_hard_stat               79.476 kiops : time 106.64 seconds
[RESULT] IOPS phase 6        mdtest_easy_delete              226.094 kiops : time 288.22 seconds
[RESULT] IOPS phase 7          mdtest_hard_read               46.141 kiops : time 182.72 seconds
[RESULT] IOPS phase 8        mdtest_hard_delete               58.842 kiops : time 143.64 seconds
```

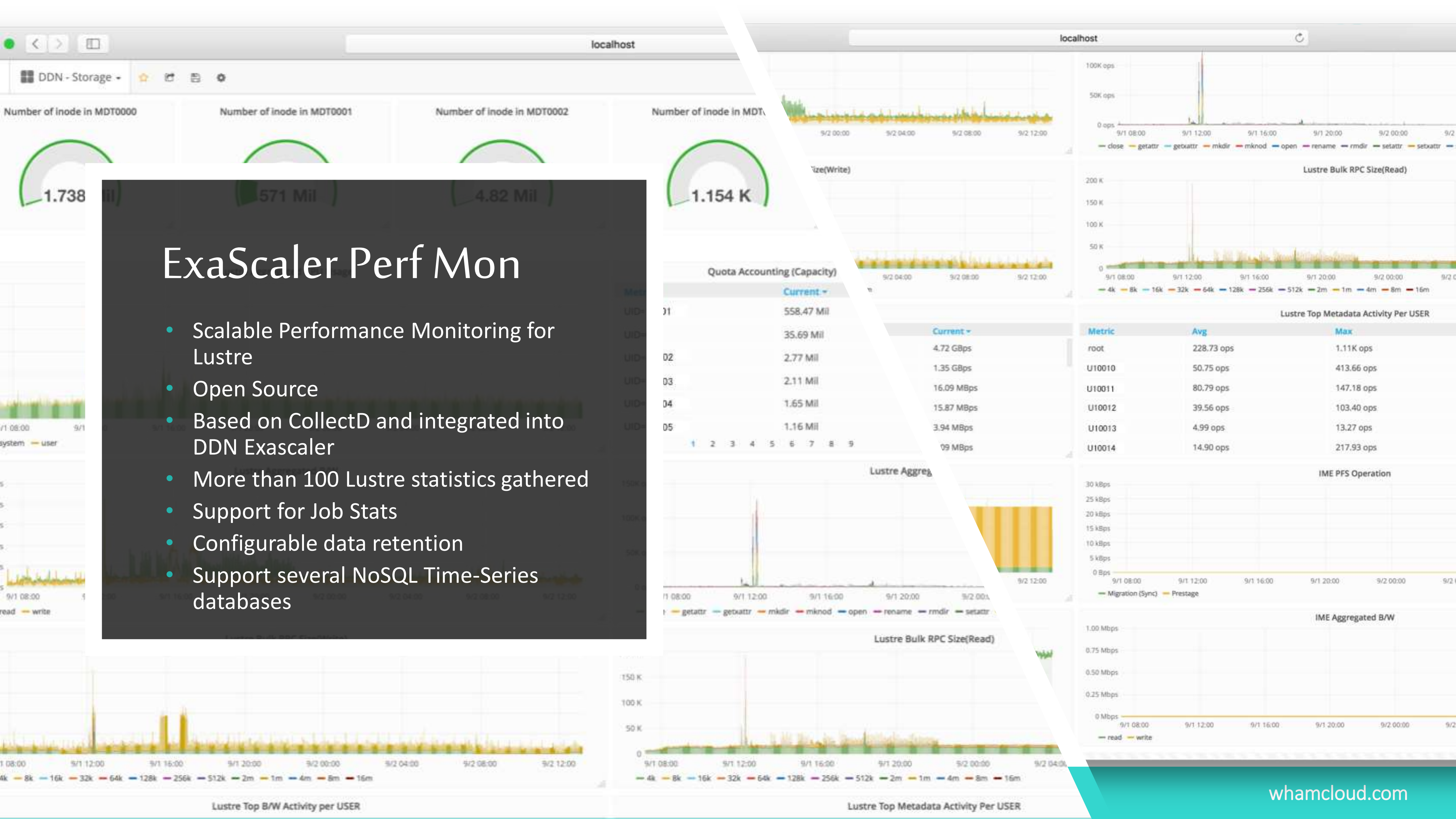
[SCORE] Bandwidth 6.33725 GB/s : IOPS 159.413 kiops : TOTAL 31.7843



IO-500

This is the official ranked list from ISC-HPC 2018. The list shows the best result for every given combination of system/institution/filesystem (i.e. multiple submissions from the same system are not shown; only the most recent is shown). The full list is available here.

#	information						io500		
	system	institution	filesystem	storage vendor	client nodes	data	score	bw	md
								GiB/s	kIOP/s
	Oakforest-PACS	JCAHPC	IME	DDN	2048	zip	137.78	560.10	33.89
2	ShaheenII	KAUST	DataWarp	Cray	1024	zip	77.37	496.81	12.05
3	ShaheenII	KAUST	Lustre	Cray	1000		41.00*	54.17	31.03*
4	JURON	JSC	BeeGFS	ThinkparQ	8		35.77*	14.24	89.81*
5	Mistral	DKRZ	Lustre2	Seagate	100		32.15	22.77	45.39
6	Sonasad	IBM	Spectrum Scale	IBM	10	zip	24.24	4.57	128.61
7	Seislab	Fraunhofer	BeeGFS	ThinkparQ	24		16.96	5.13	56.14
8	Mistral	DKRZ	Lustre1	Seagate	100	zip	15.47	12.68	18.88
9	Govorun	Joint Institute for Nuclear Research	Lustre	RSC	24	zip	12.08	3.34	43.65
10	EMSL Cascade	PNNL	Lustre		126		11.12	4.88	25.33
11	Serrano	SNL	Spectrum Scale	IBM	16		4.25*	0.65	27.98*
12	Jasmin/Lotus	STFC	NFS	Purestorage	64	zip	2.33	0.26	20.93


localhost
DDN - Storage
Number of inode in MDT0000
Number of inode in MDT0001
Number of inode in MDT0002
1.738
571 Mil
4.82 Mil
1.154 K
Lustre Bulk RPC Size(Read)
IME PFS Operation
IME Aggregated B/W
Lustre Top B/W Activity per USER
Lustre Top Metadata Activity Per USER
Migration (Sync)
Prestage
read
write

Quota Accounting (Capacity)

	Current
01	558.47 Mil
	35.69 Mil
02	2.77 Mil
03	2.11 Mil
04	1.65 Mil
05	1.16 Mil

Current
4.72 GBps
1.35 GBps
16.09 MBps
15.87 MBps
3.94 MBps
09 MBps

Lustre Top Metadata Activity Per USER

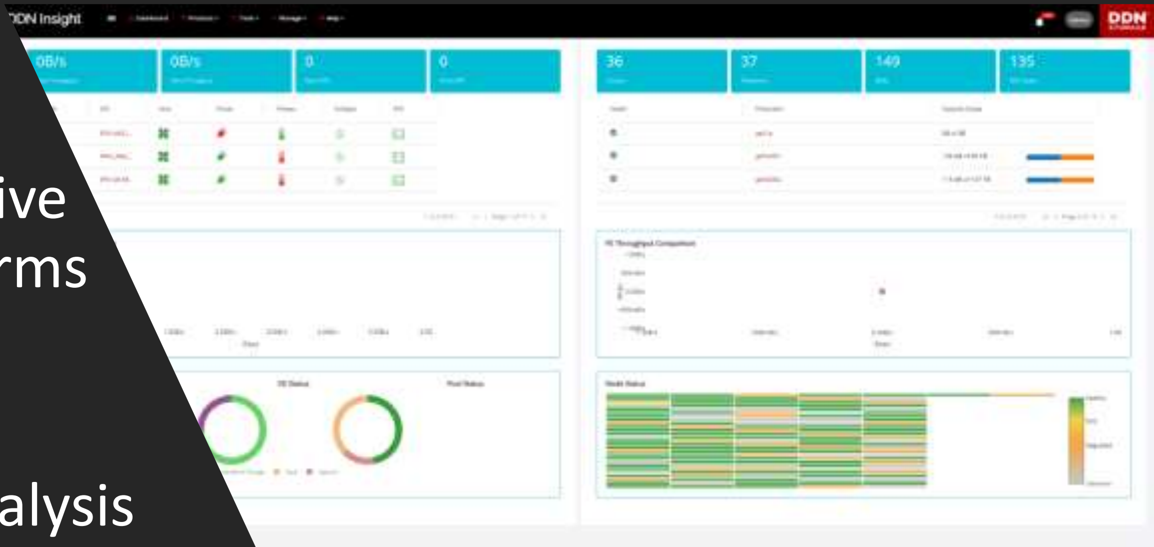
Metric	Avg	Max
root	228.73 ops	1.11K ops
U10010	50.75 ops	413.66 ops
U10011	80.79 ops	147.18 ops
U10012	39.56 ops	103.40 ops
U10013	4.99 ops	13.27 ops
U10014	14.90 ops	217.93 ops

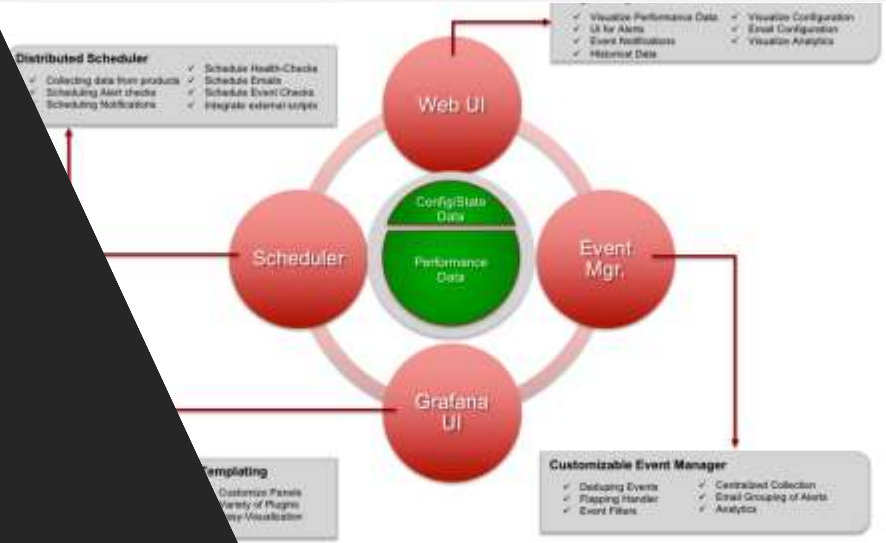
ExaScaler Perf Mon

- Scalable Performance Monitoring for Lustre
- Open Source
- Based on CollectD and integrated into DDN Exascaler
- More than 100 Lustre statistics gathered
- Support for Job Stats
- Configurable data retention
- Support several NoSQL Time-Series databases

DDN Insight for ExaScaler

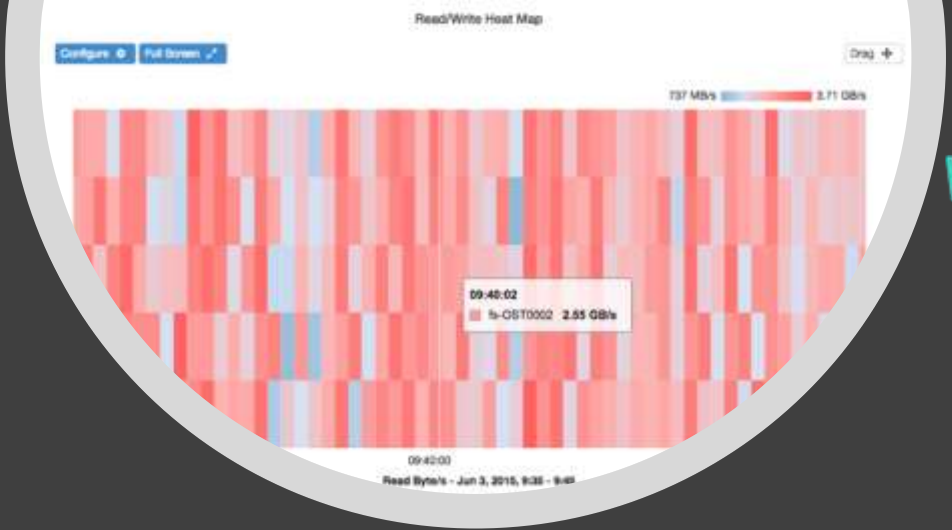
- Multi-solutions Monitoring
  - Centralized Web interface for comprehensive monitoring of all DDN solutions and platforms
- Unprecedented Vision
  - Derive the most value of your storage infrastructure through a deep real-time analysis from file system to hardware platform
- Open and Extensible
  - Extensively customizable DDN Insight open back end database and integration with Grafana allows you to correlate storage performance with metrics beyond the DDN system

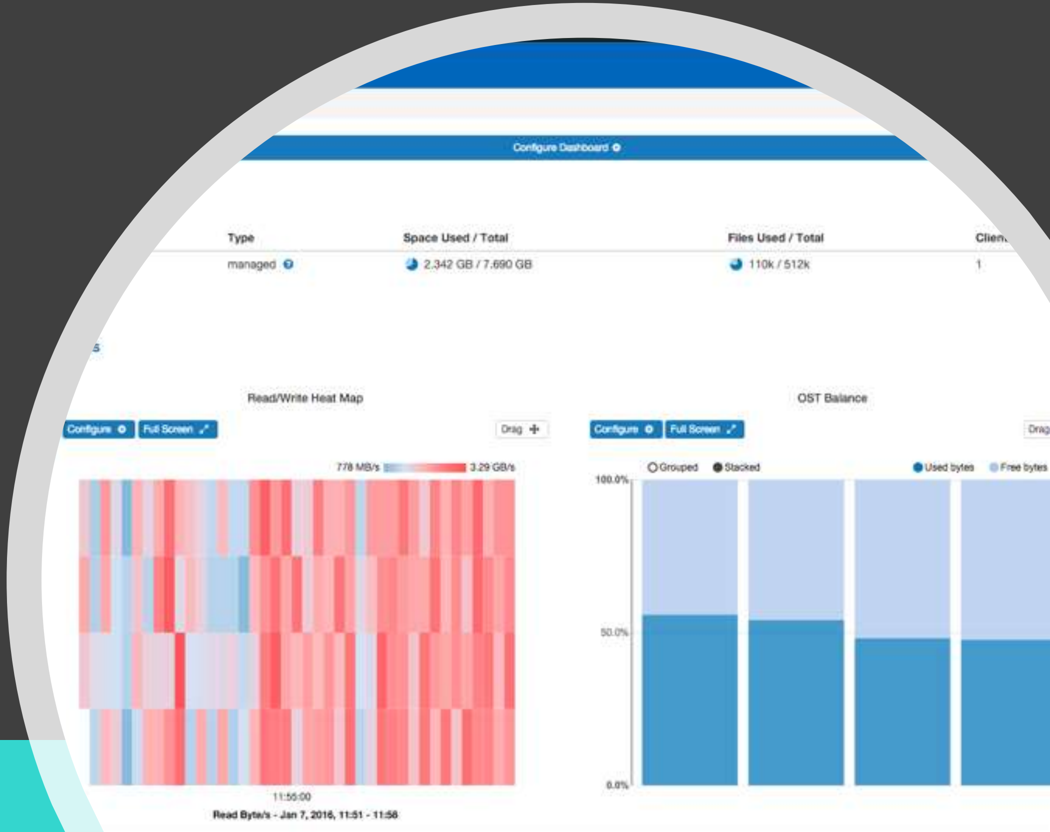




IML - Integrated Management for Lustre

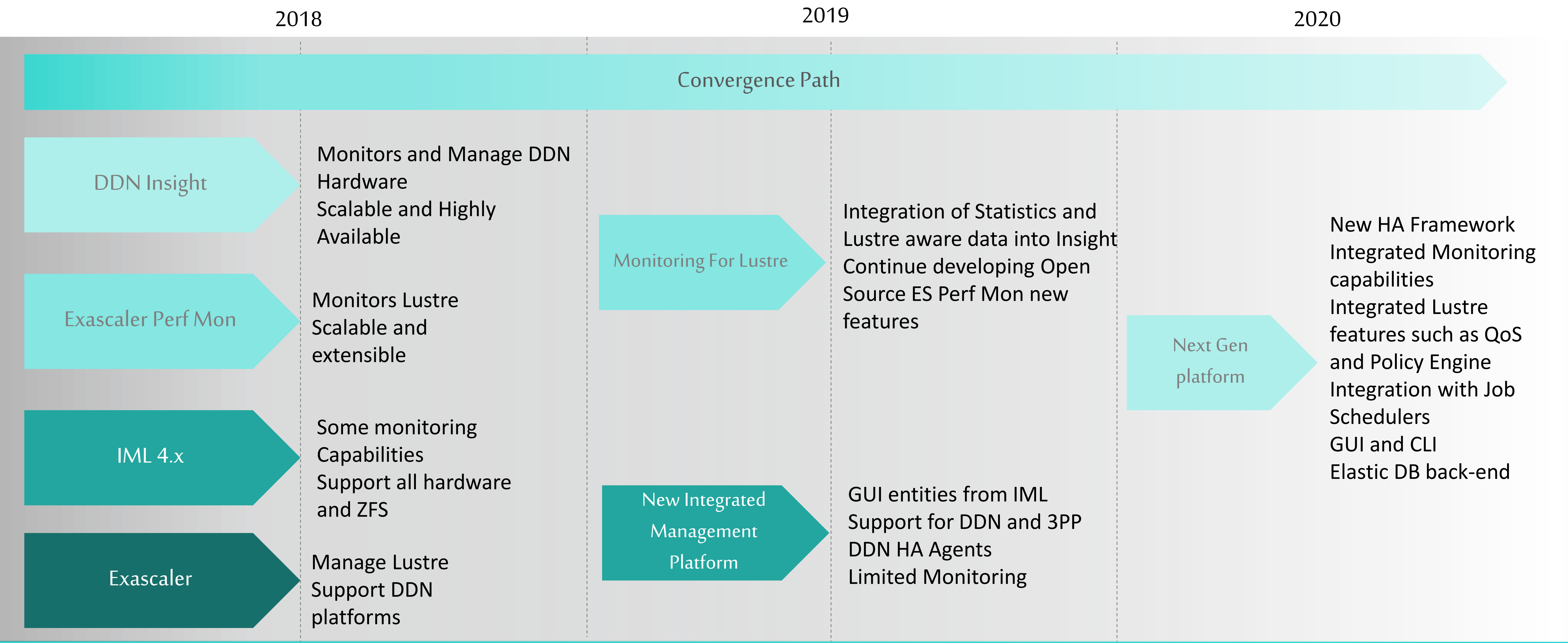
- Open source suite of tools for deploying, managing and monitoring Lustre
- IML simplifies Lustre administration with intuitive interfaces and near real -time feedback
- Works with new and existing Lustre installations
- Monitors performance and system health





Type	Space Used / Total	Files Used / Total	Clien
managed	2,342 GB / 7,690 GB	110k / 512k	1

Manage and Monitoring Strategy for Lustre



2018

2019

2020

Convergence Path

DDN Insight

Monitors and Manage DDN Hardware
Scalable and Highly Available

Exascaler Perf Mon

Monitors Lustre
Scalable and extensible

IML 4.x

Some monitoring Capabilities
Support all hardware and ZFS

Exascaler

Manage Lustre
Support DDN platforms

Monitoring For Lustre

Integration of Statistics and Lustre aware data into Insight
Continue developing Open Source ES Perf Mon new features

New Integrated Management Platform

GUI entities from IML
Support for DDN and 3PP
DDN HA Agents
Limited Monitoring

Next Gen platform

New HA Framework
Integrated Monitoring capabilities
Integrated Lustre features such as QoS and Policy Engine
Integration with Job Schedulers
GUI and CLI
Elastic DB back-end

Future Lustre Events

SC18	**Tuesday, November 13th, 12:15-1:15pm – Room C140/142**
May 14-17	Lustre User Group 2019 (OpenSFS)
June 16-20	ISC Lustre User Group/BOF
Mar 11-14	APAC Lustre User Event Sept at SCASIA19 (TBD)
Sept	Lustre Admins & Dev Workshop (EOFS)
Oct	Japan Lustre User Group (Whamcloud)
Oct	China Lustre User Group (Whamcloud)
Nov	SC19 Lustre User Group/BOF

Whamcloud

Thank you!

DDN® STORAGE

SC18 Dallas, TX | hpc inspires.