



LUSTRE USER GROUP 2023

Bringing Lustre to the masses through a fully-managed cloud service

Darryl Osborne

Principal Solutions Architect
Amazon Web Services

Agenda

Demo

Presentation

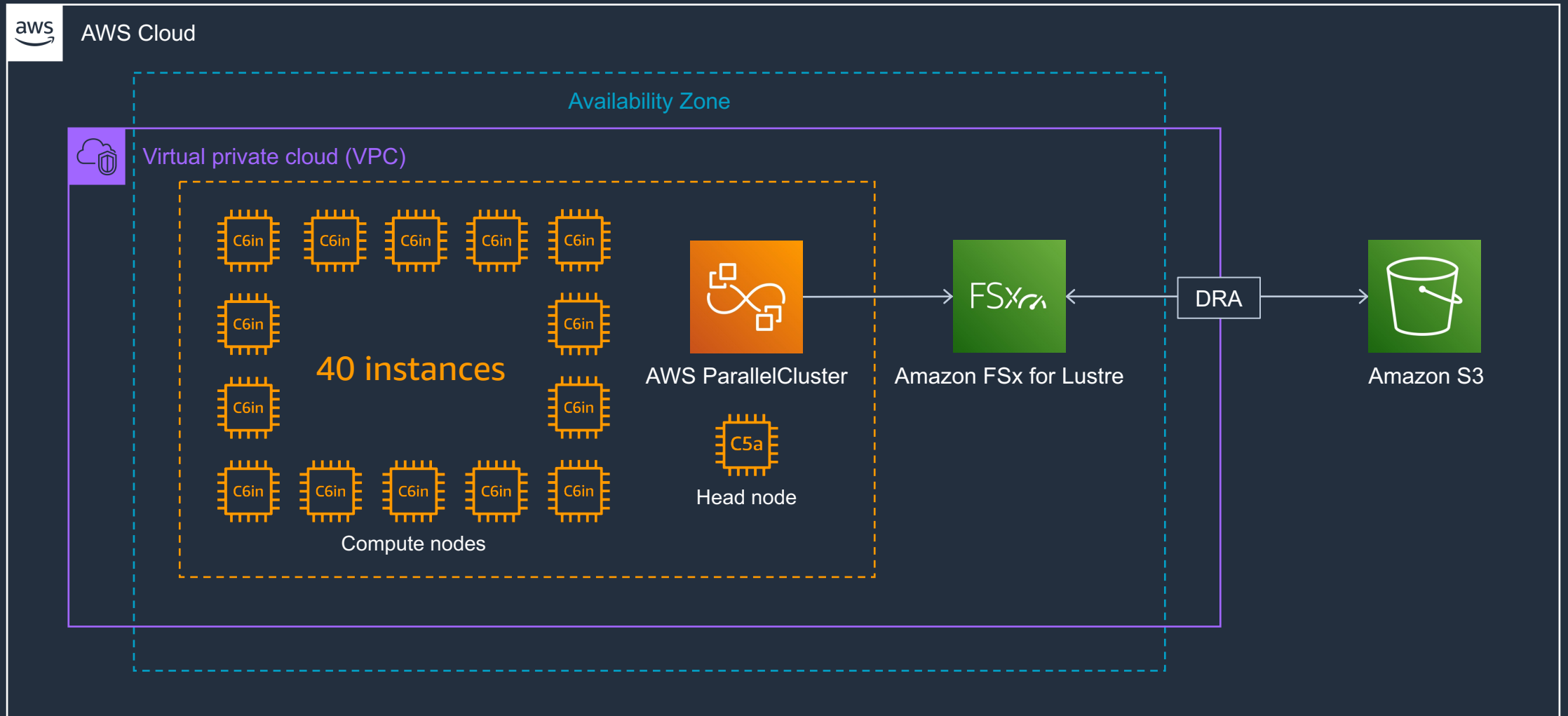
Juggle



Demo



From nothing to 200+ GB/s in 30 minutes or less



Agenda

Demo – From nothing to 200+ GB/s in 30 minutes or less

Use cases

Architecture

HSM solution using Amazon S3

Performance

Q&A



Use cases





Rivian used **Amazon FSx for Lustre and Amazon EC2** to support new concepts, crash and vibration testing, and simulations and **achieved a up to a 56% workload acceleration.**

"This is accelerating adoption across the board."

Madhavi Isanaka
Chief Information Officer, Rivian



Learn More



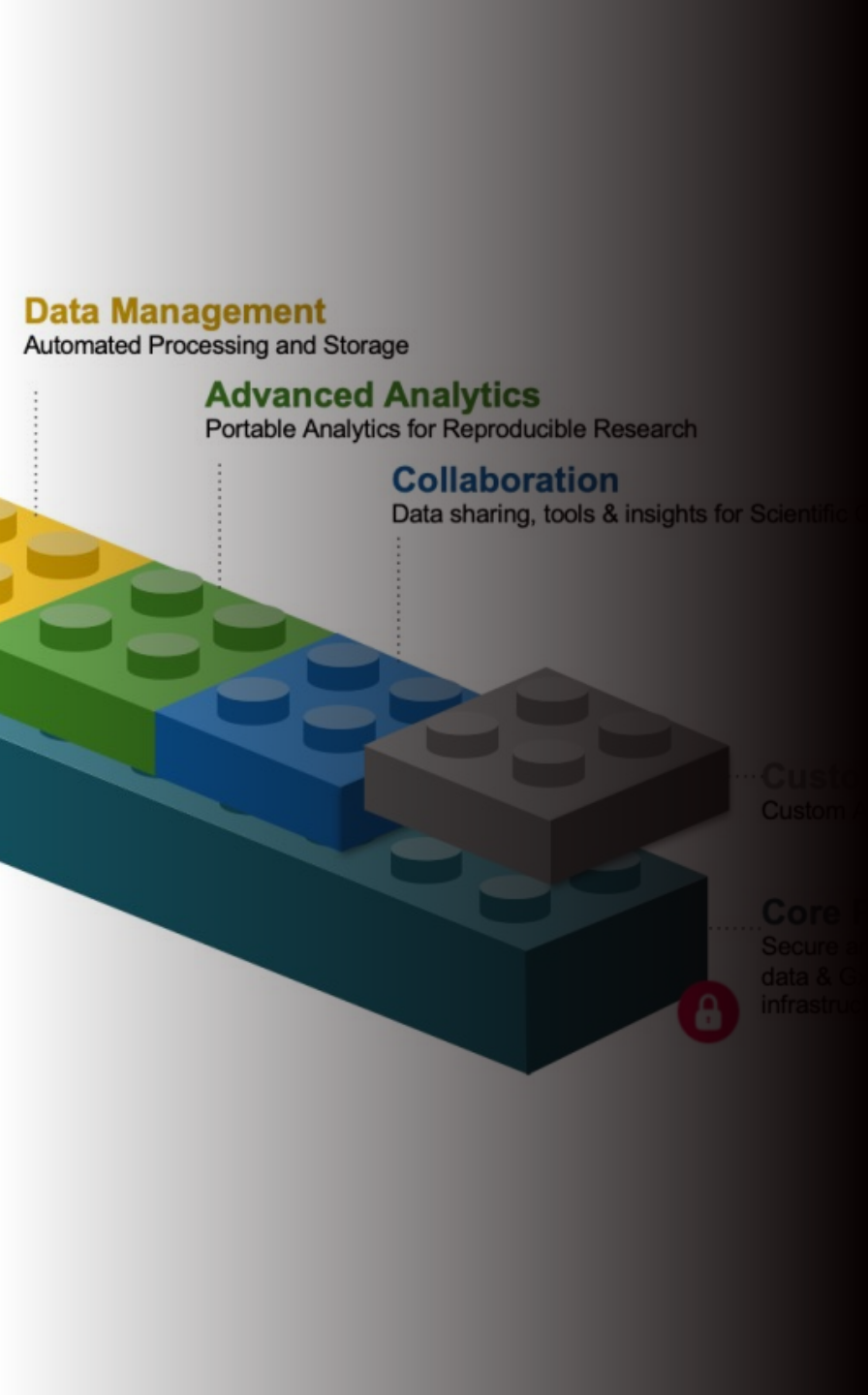


Image files that previously took 2–3 days for processing are now ready in hours, and modular electronic health record datasets get processed within minutes.

“Roche is taking steps closer towards its mission to **provide every patient with the best treatment possible in the fastest time.**”

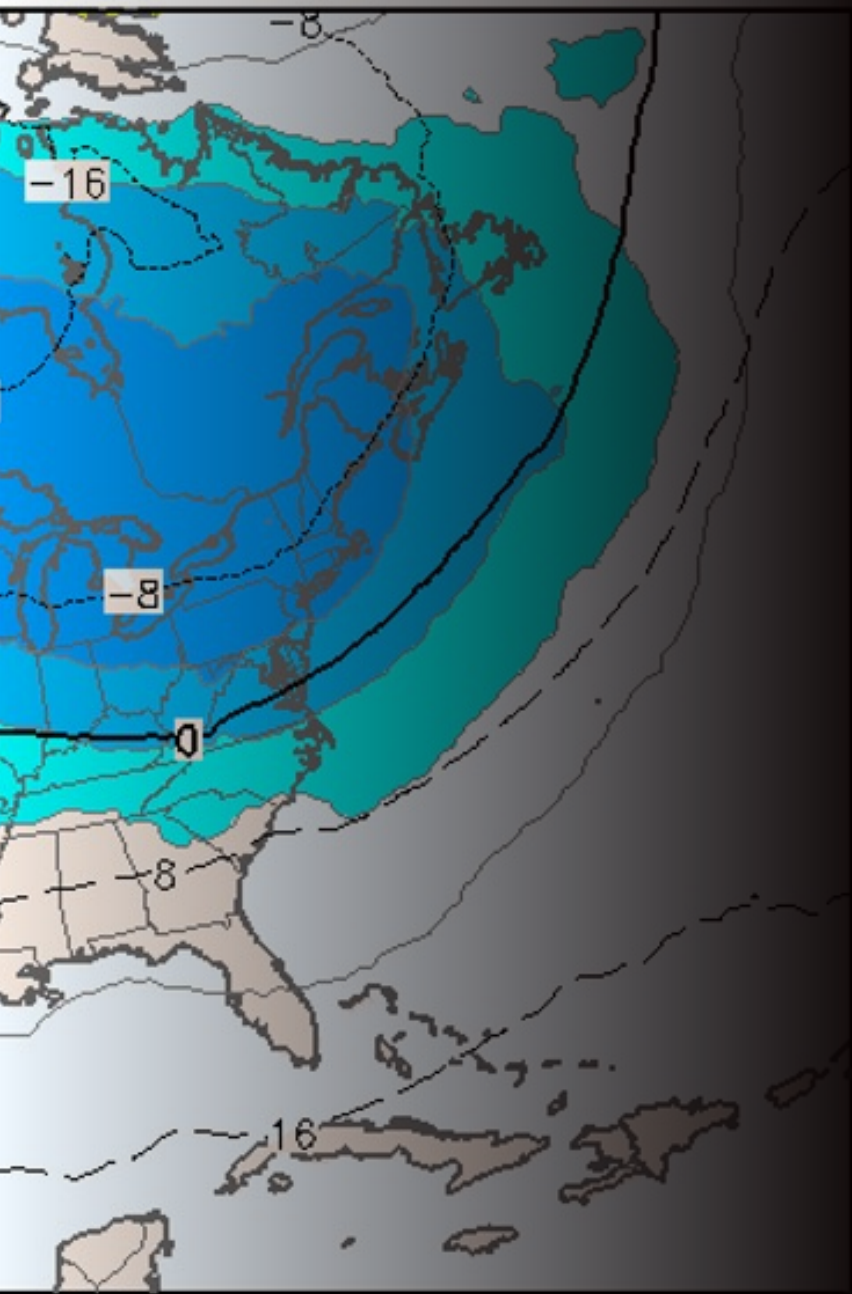
Mustaqhusain Kazi, Head of Personalized Healthcare,
Pharma Informatics at Roche



Learn More



Temp - 12Z Nov07-22



MAXAR

Maxar uses AWS to deliver forecasts 58% faster than weather supercomputer

“Maxar used Amazon FSx for Lustre in our AWS HPC solution for running NOAA’s numerical weather forecasting model. This allowed us to reduce compute time by 58%, generating the forecast in about 45 minutes for a much more cost-effective price point. Maximizing our AWS compute resources was an incredible performance boost for us.”

Stefan Cecelski,
PhD Senior Data Scientist & Engineer, Maxar Technologies



[Learn More](#)



Architecture

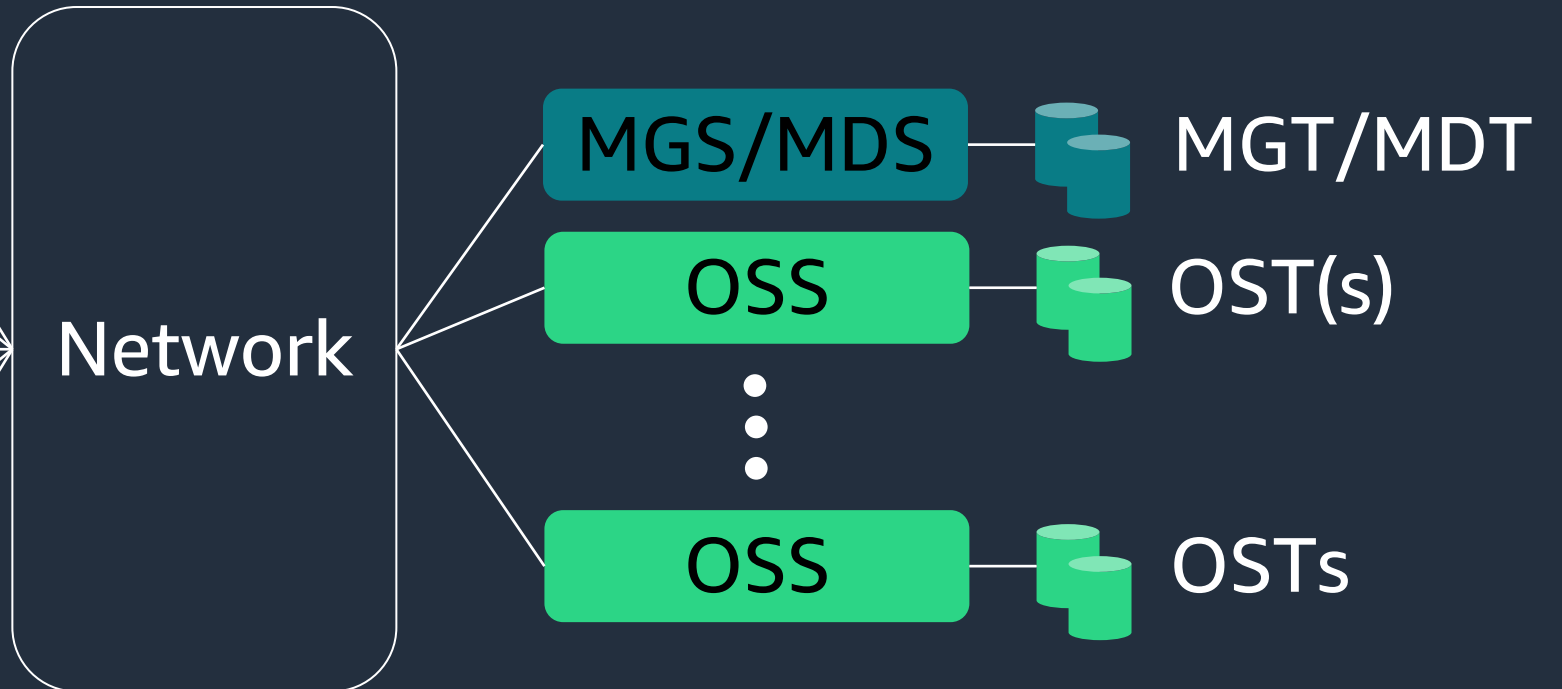


Lustre architecture

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)



Storage and deployment types

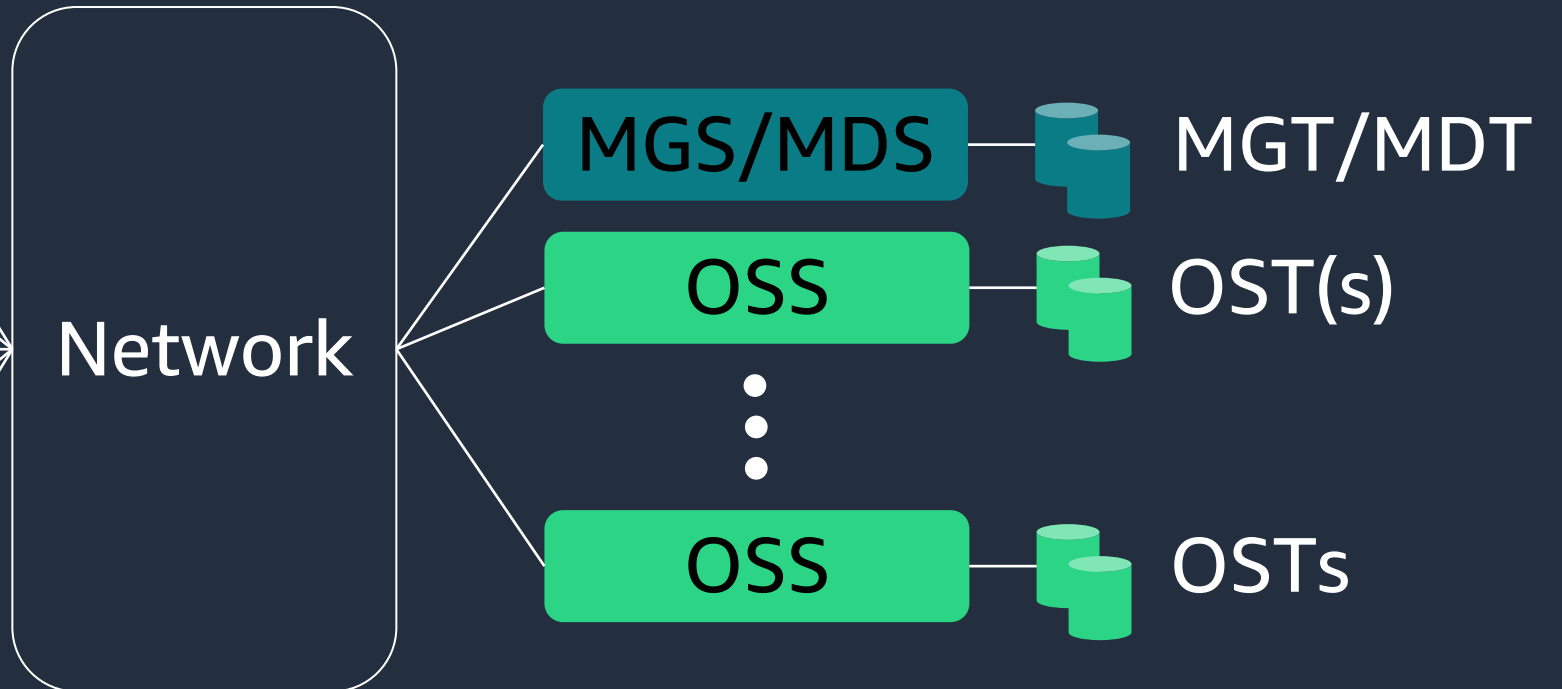
Storage type	Deployment type
HDD	Persistent
SSD	Scratch
	Persistent

Lustre architecture

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)

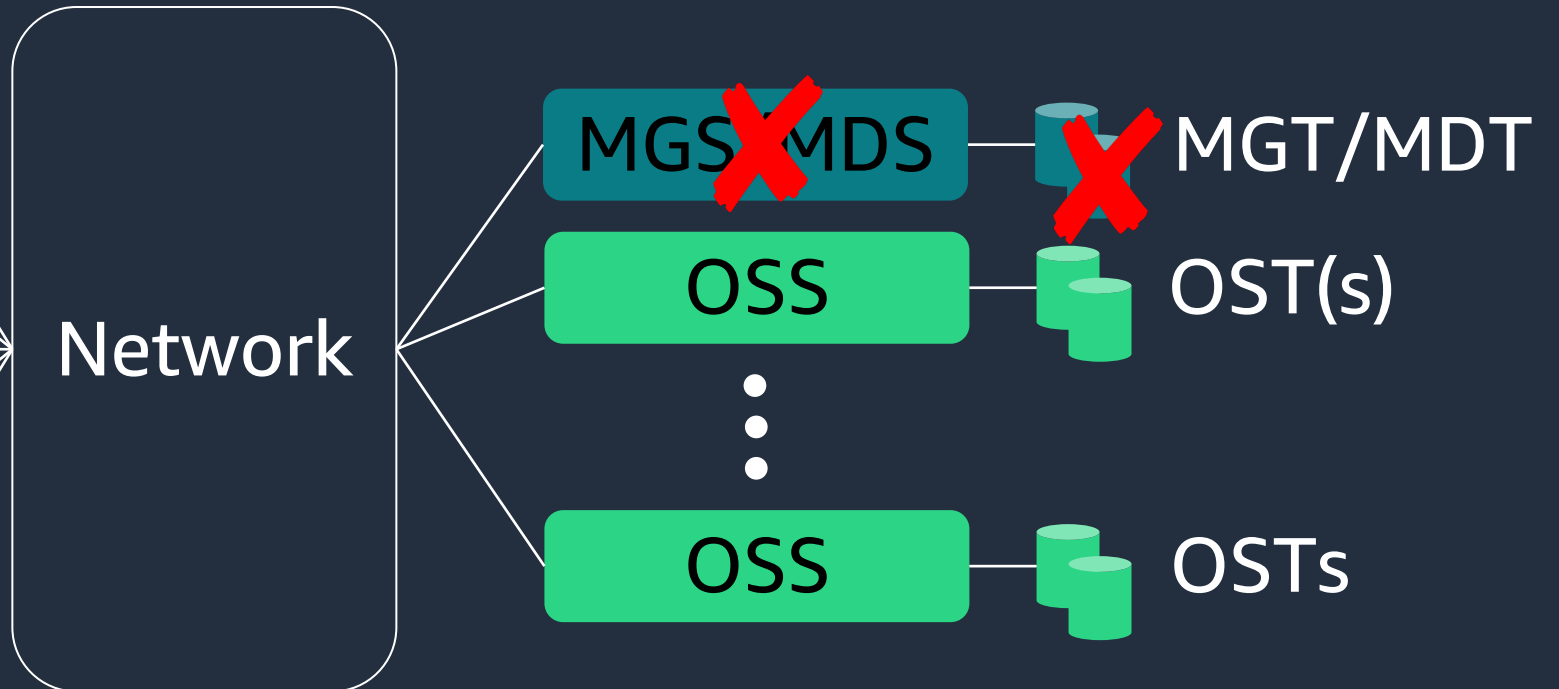


Lustre architecture and persistent file systems

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)

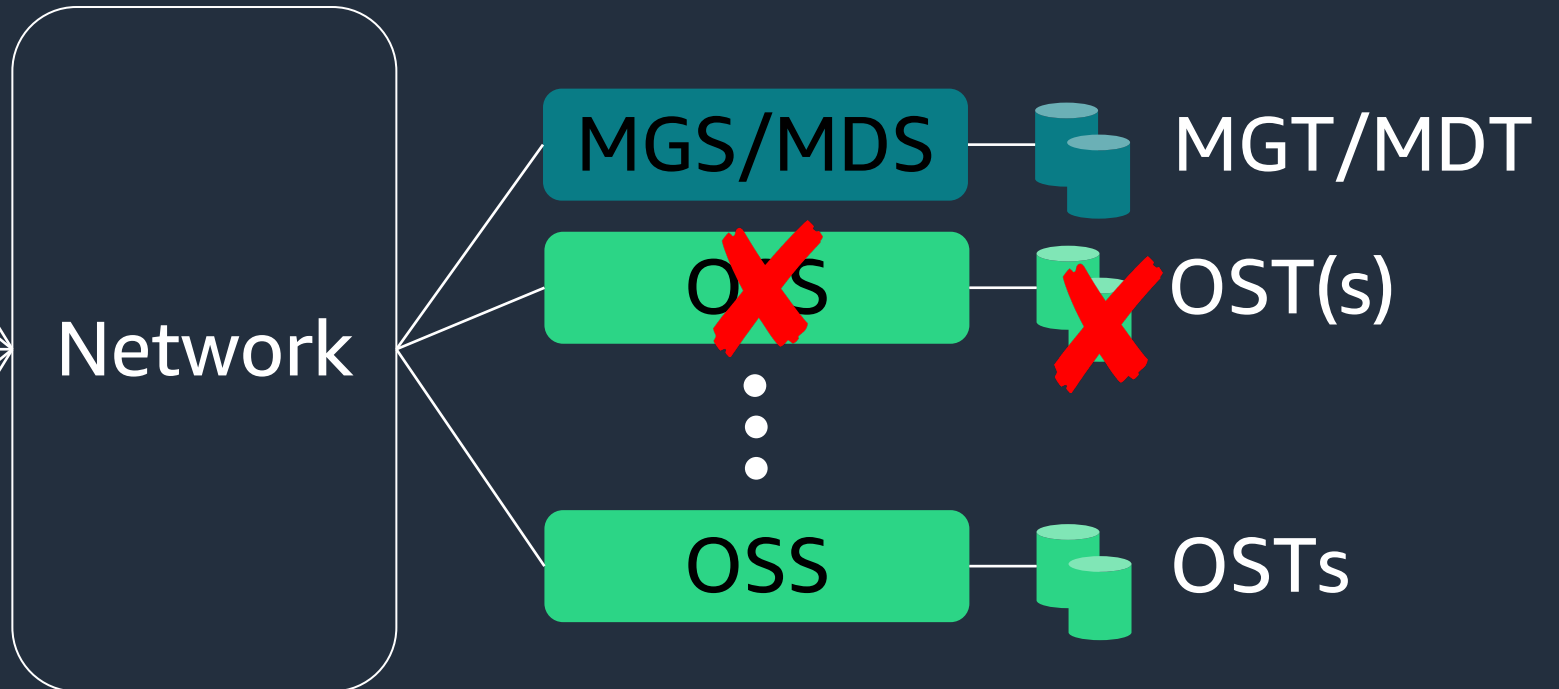


Lustre architecture and persistent file systems

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)

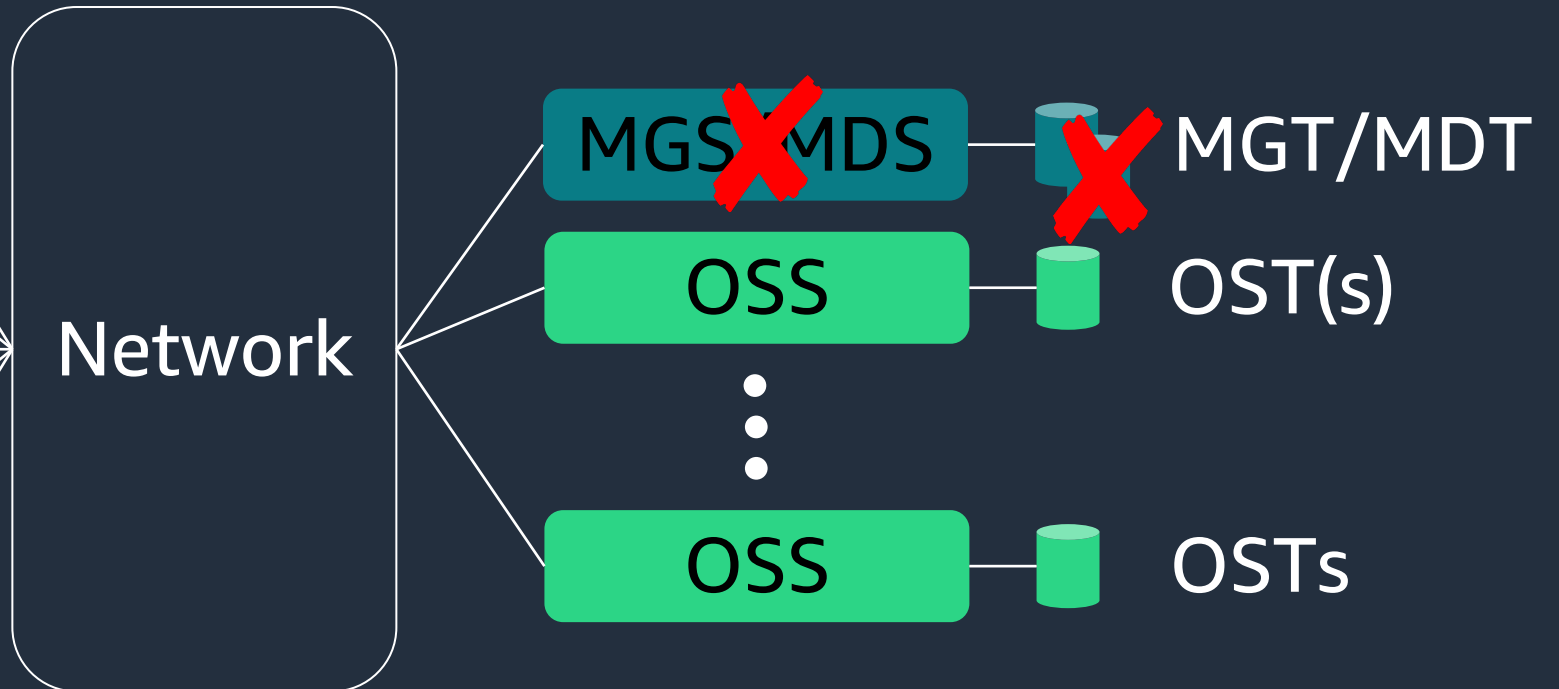


Lustre architecture and scratch file systems

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)

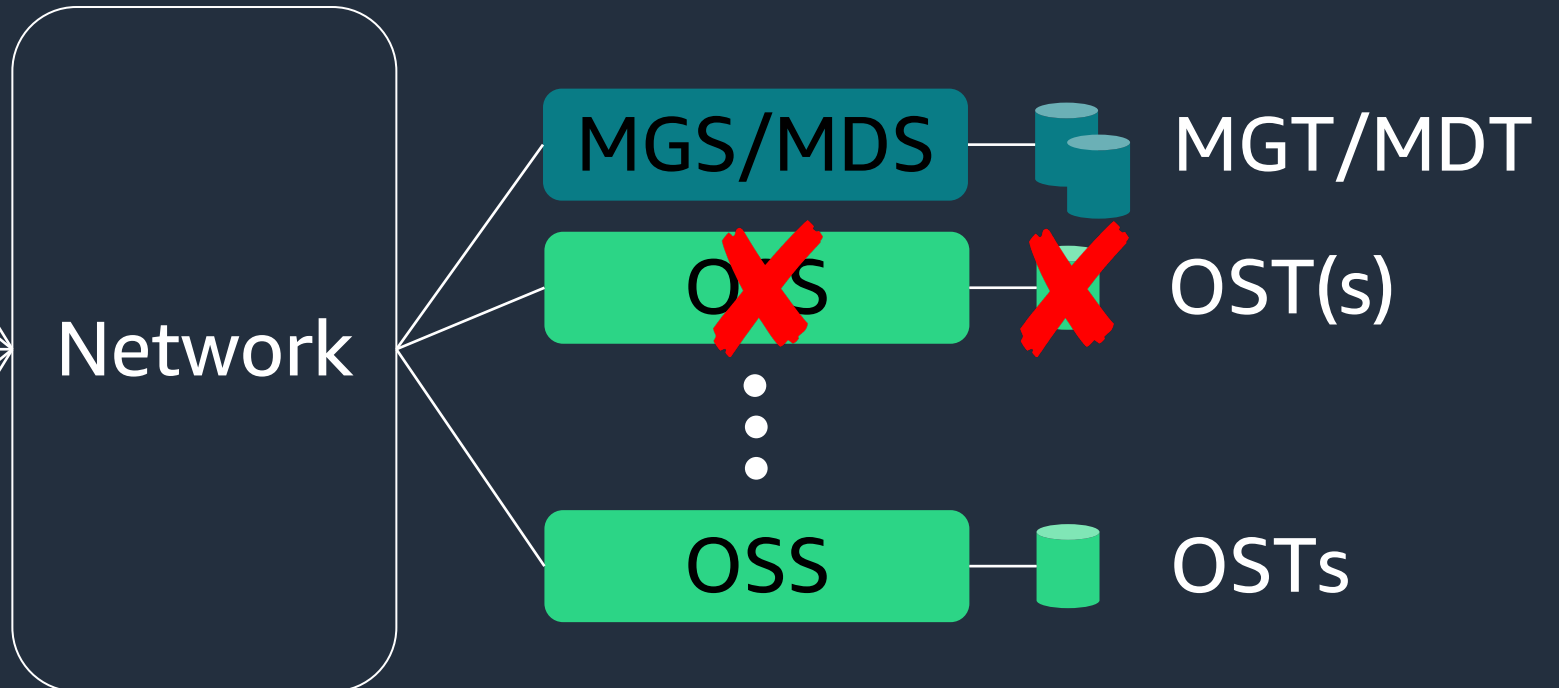


Lustre architecture and scratch file systems

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)



Storage and deployment types

Storage type	Deployment type	Disk storage throughput (MB/s per TiB of storage)	SSD read cache throughput (MB/s per TiB of cache*)	Price per GB-month**
HDD	Persistent	12	-	\$0.025
			200	\$0.041
		40	-	\$0.083
			200	\$0.099
SSD	Scratch	200	-	\$0.140
	Persistent	125	-	\$0.145
		250	-	\$0.210
		500	-	\$0.340
		1000	-	\$0.600

* Read cache sized at 20% of HDD storage capacity

** US East (N. Virginia) pricing

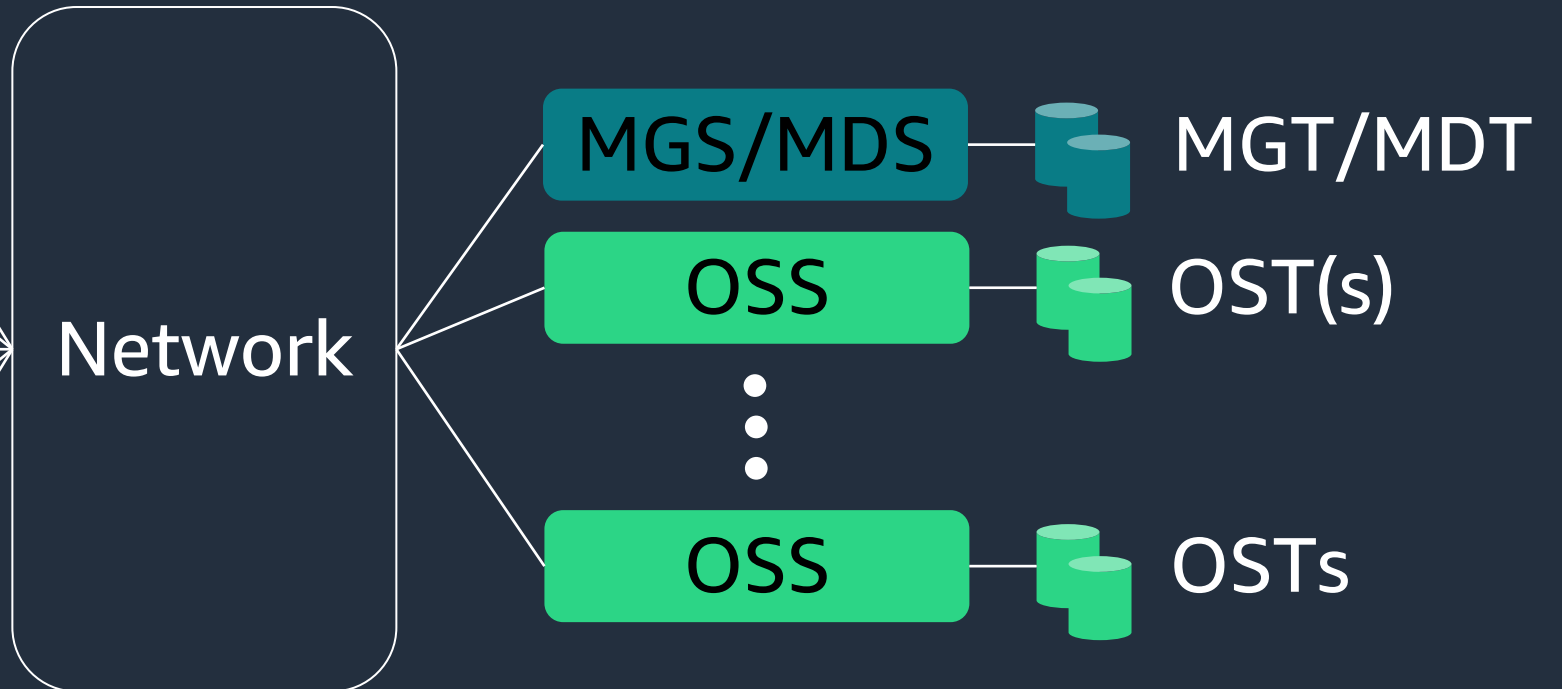


Lustre architecture

Client instances



Lustre file system



Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)



Storage and deployment types

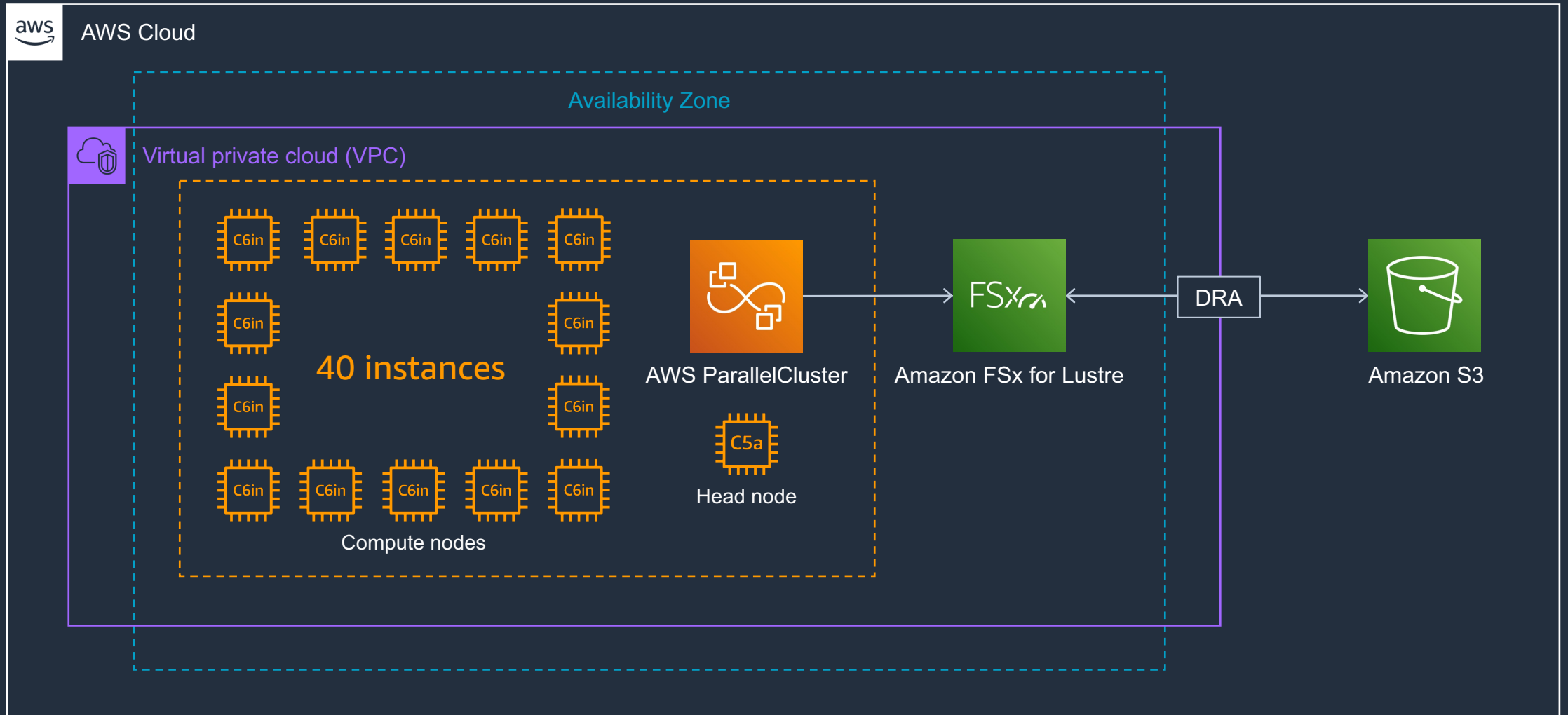
Storage type	Deployment type		Minimum size (TiB)	Incremental size (TiB)	OSS count (1 per x TiB)	OST size (TiB)	OST count per OSS	MDS and MDT count
HDD	Persistent	12 MB/s	6.0	6.0	6.0	1.5	4	1 and 1
		40 MB/s	1.8	1.8	1.8	1.8	1	
SSD	Scratch							
	Persistent		1.2	2.4	2.4	1.2	2	

Example:

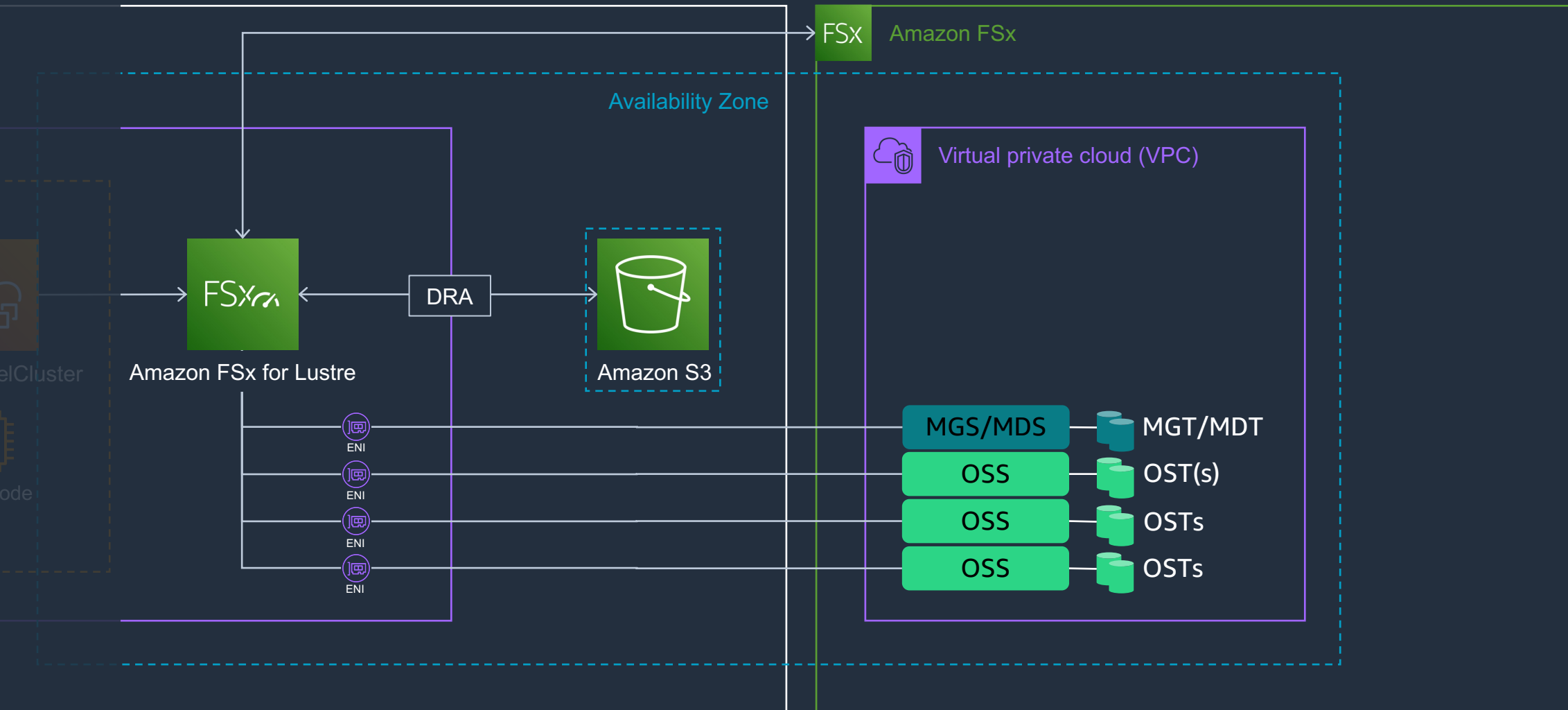
SSD Persistent 100.8 TiB: 1 MDS/MDT = 1 Elastic Network Interface (ENI)
 42 OSSs ($100.8 \div 2.4$) = 42 Elastic Network Interfaces (ENIs)
 84 OSTs ($100.8 \div 1.2$)



Demo environment



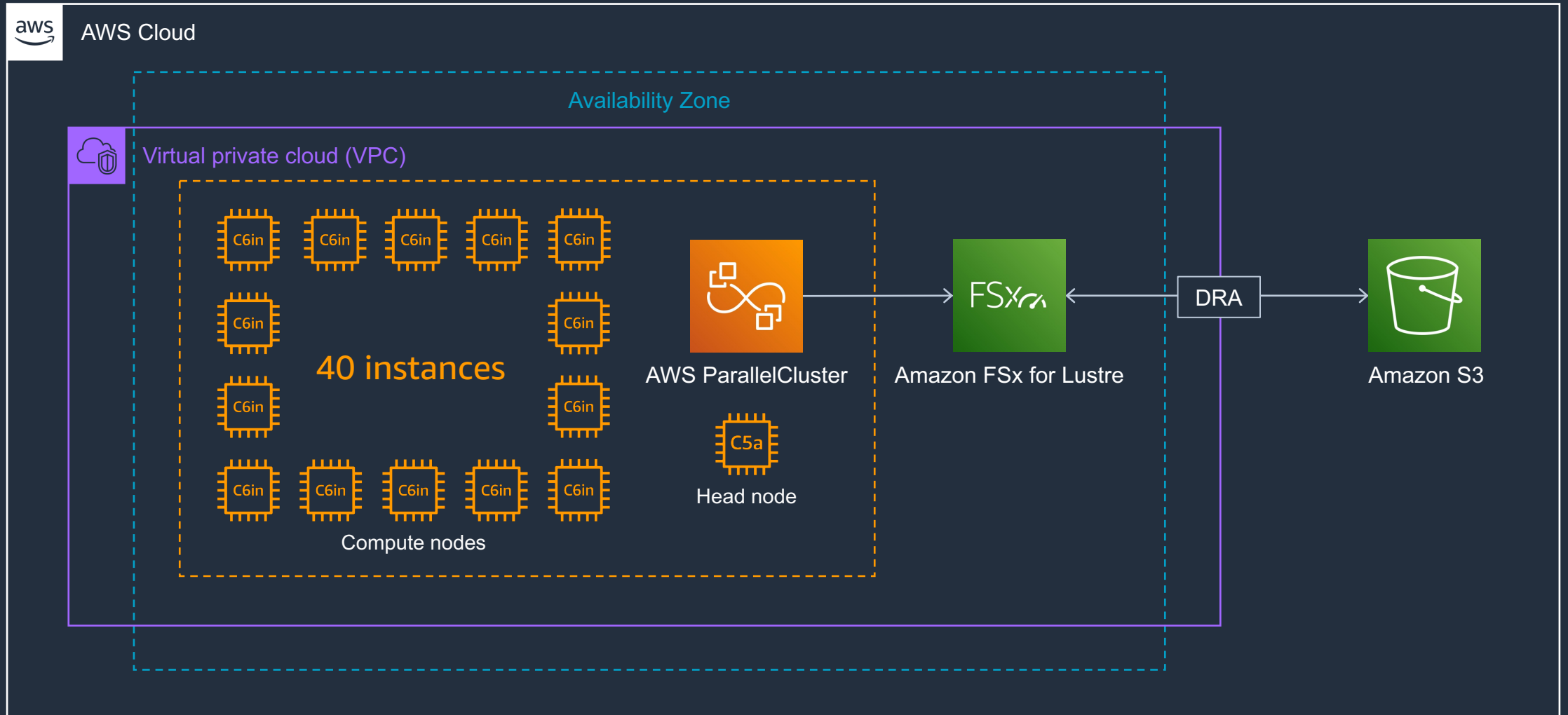
Demo environment



HSM using Amazon S3



Demo environment



Hierarchical Storage Management (HSM) using Amazon S3

Data Repository Association (DRA)

Up to eight (8) per file system

DRA path is an S3 bucket or prefix

Links file system path to a DRA path

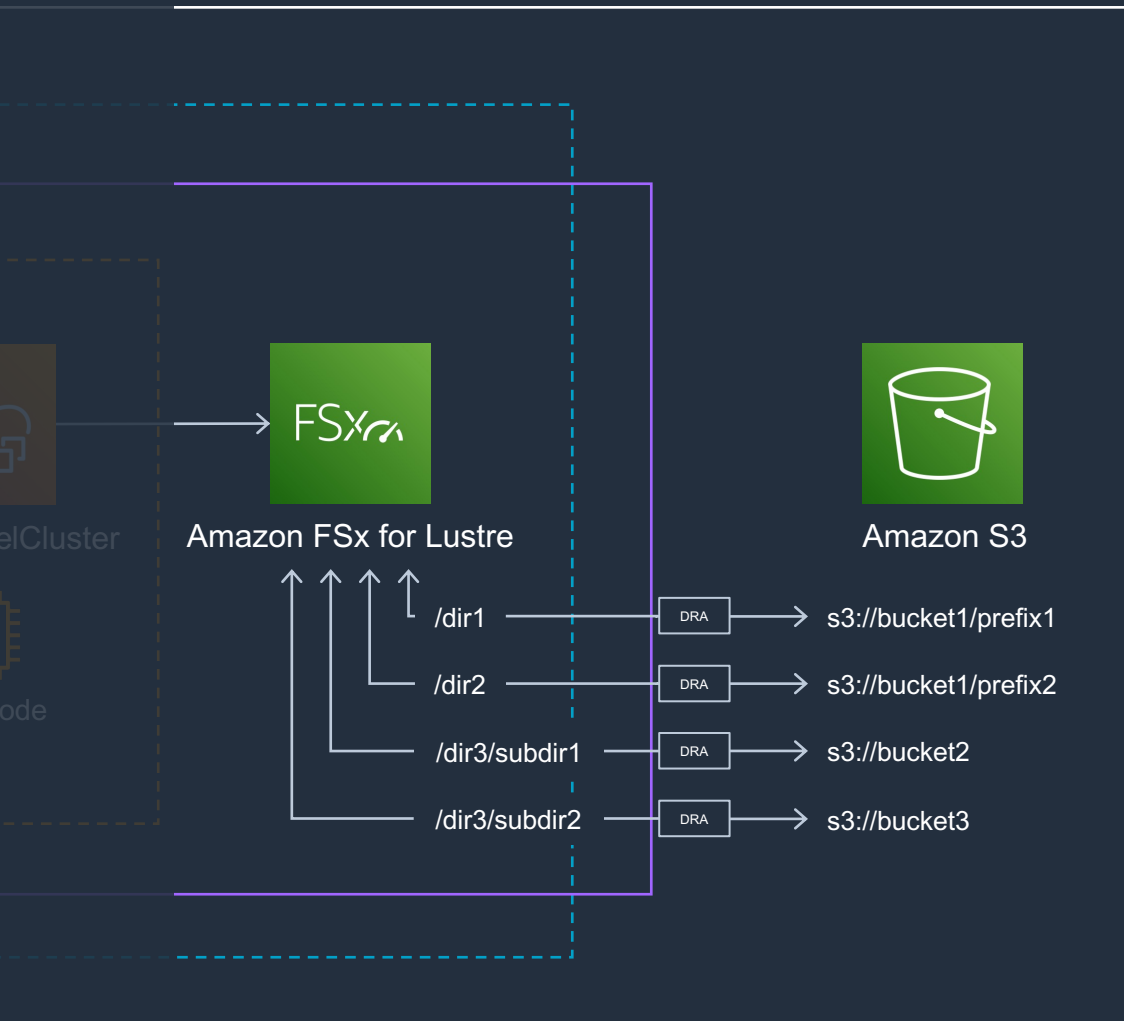
Cannot overlap file system paths

Cannot overlap DRA paths

Import policy – DRA path updates propagated to file system path

Export policy – File system path updates propagated to DRA path

1:1 mapping between file system path and object keys



Demo – HSM solution on Amazon S3



Demo – hsm_restore

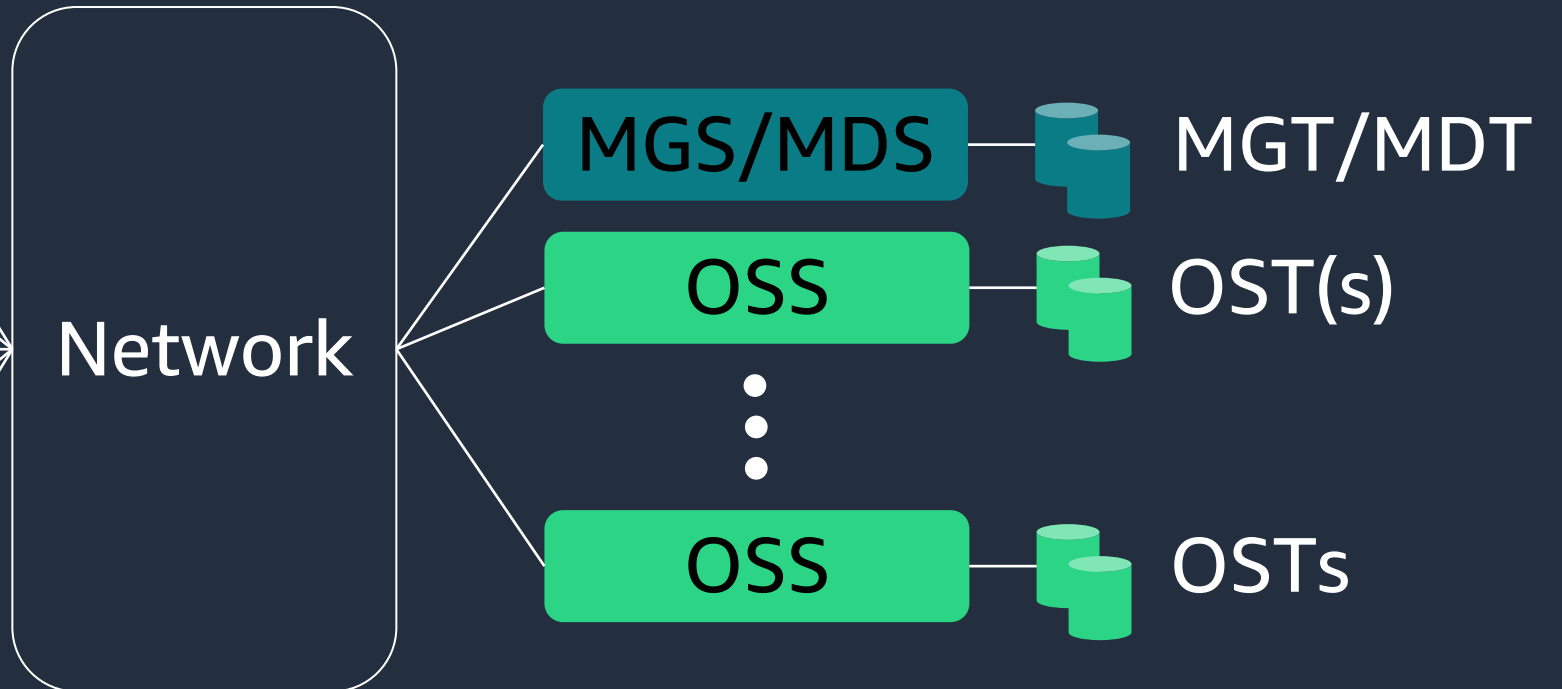


Lustre architecture

Client instances



Lustre file system



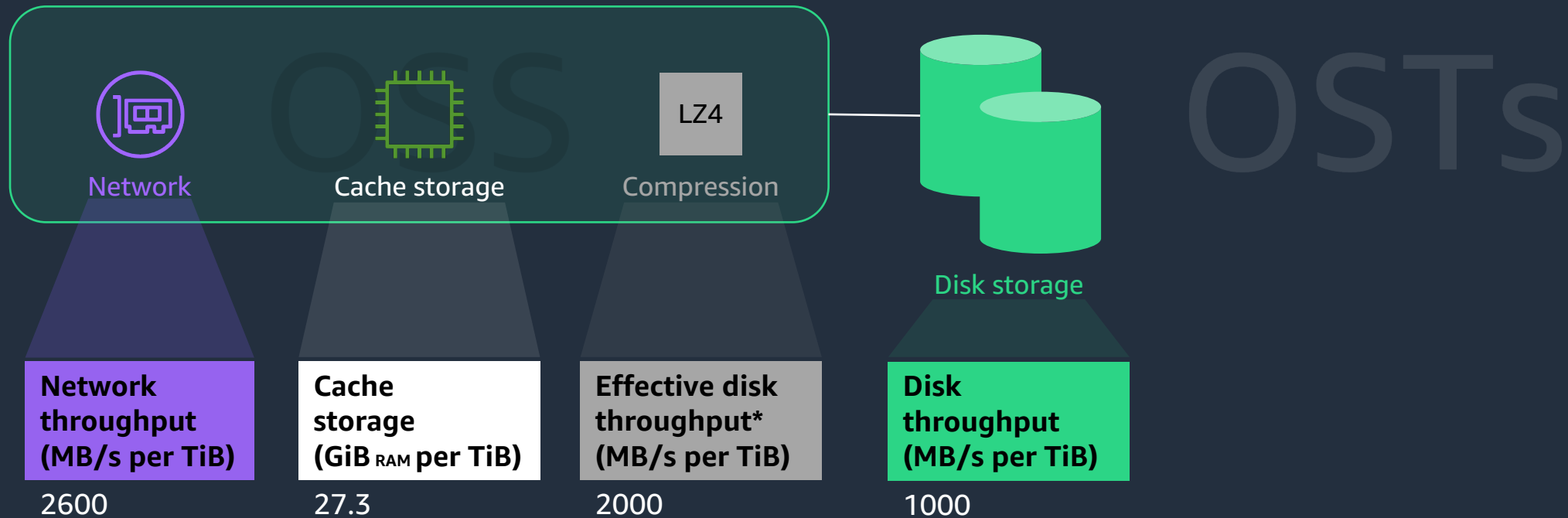
Object Storage Client (OSC) | Metadata Client (MDC) | Management Server (MGS) | Management Target | Metadata Server (MDS) | Metadata Target (MDT) | Object Storage Server (OSS) | Object Storage Target (OST)



Lustre architecture



Lustre architecture and performance



SSD Persistent 2 1000 MB/s per TiB

* 2:1 compression ratio



Demo – Parallel cluster read & auto export



Feature summary

HDD and SSD storage types

Persistent and scratch deployment types

LZ4 compression

Configurable file striping (PFLs)

Online storage capacity increases

AWS service integrations

Storage quotas

Root squash

Encryption at rest and in transit

HSM solution using Amazon S3

Automatic backups

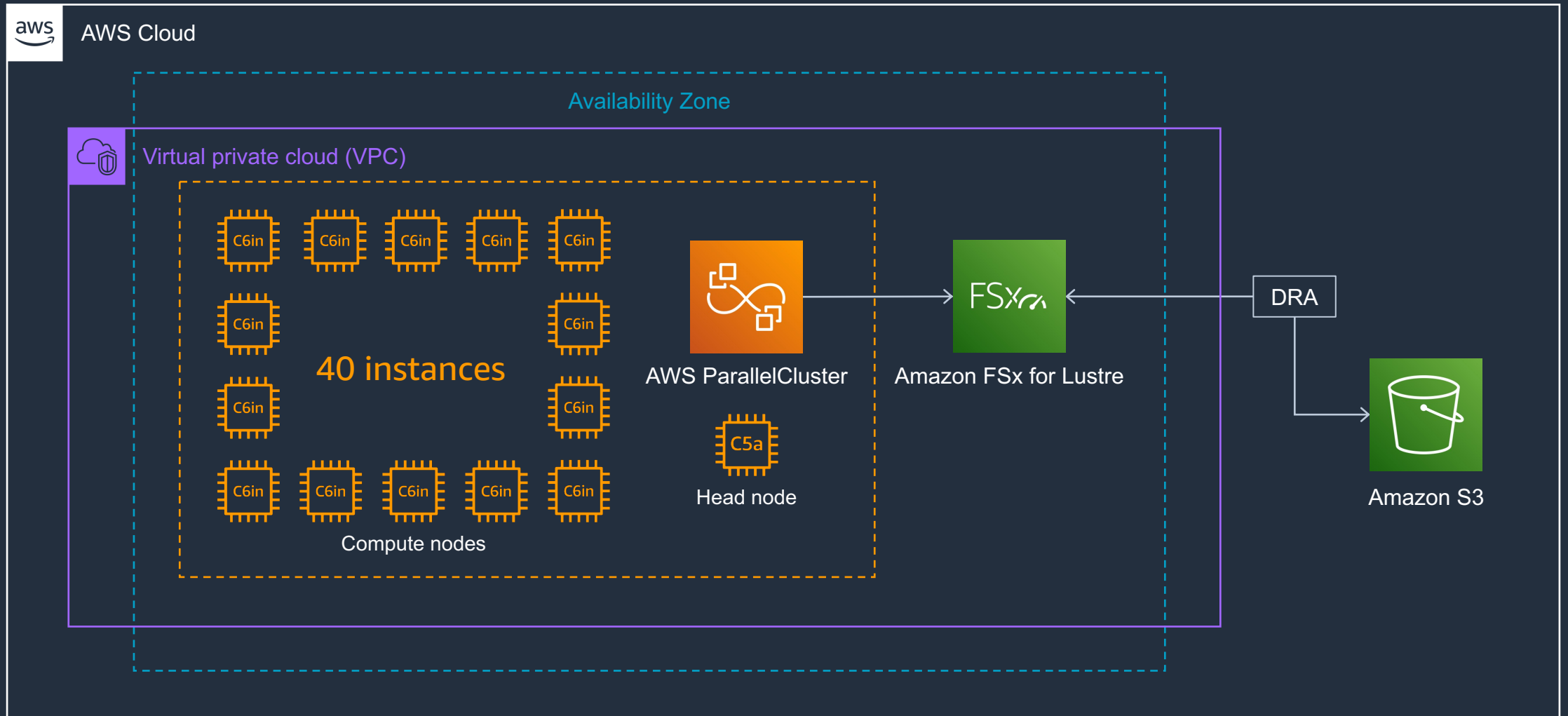
Weekly maintenance window



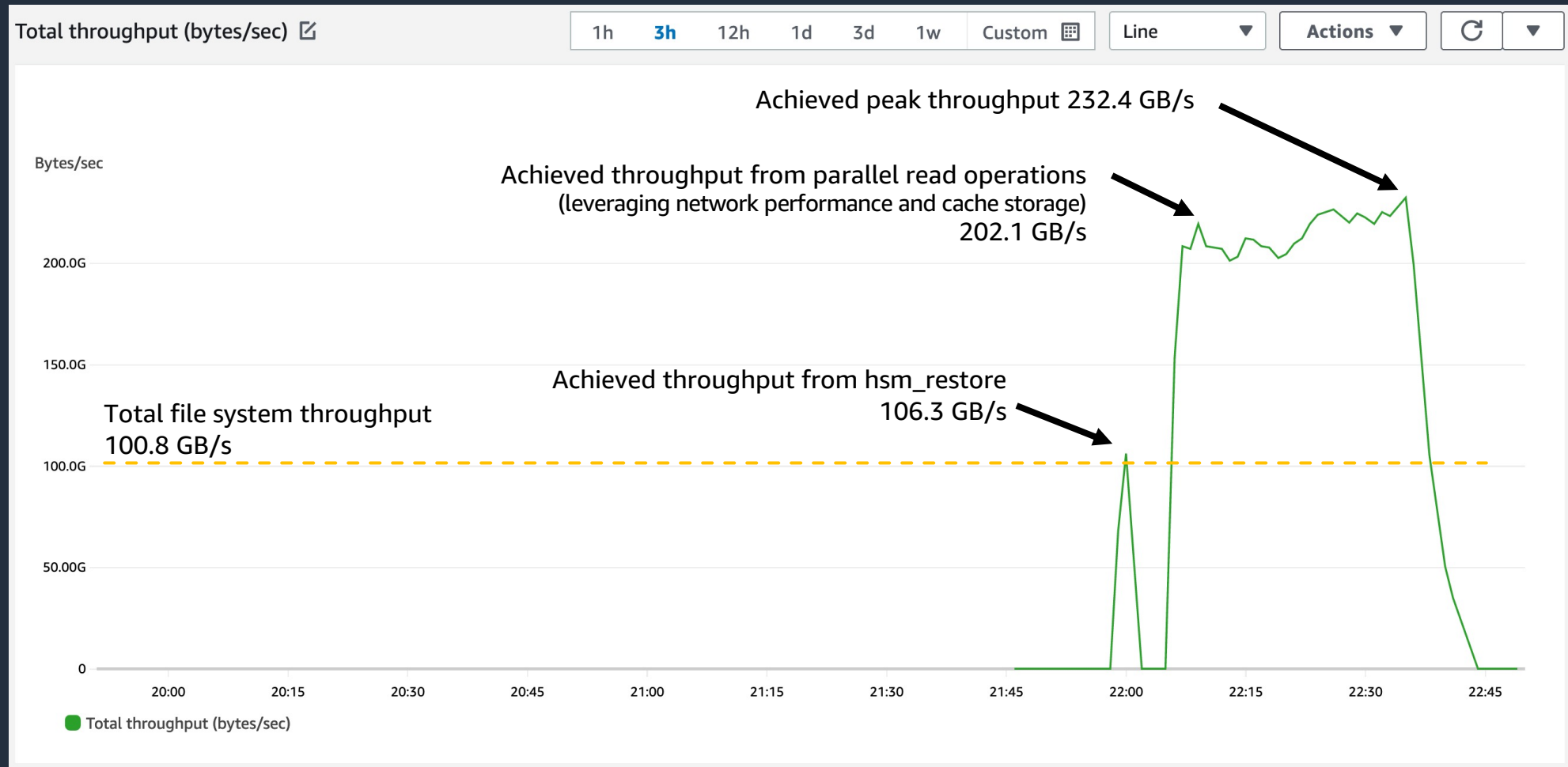
Q&A



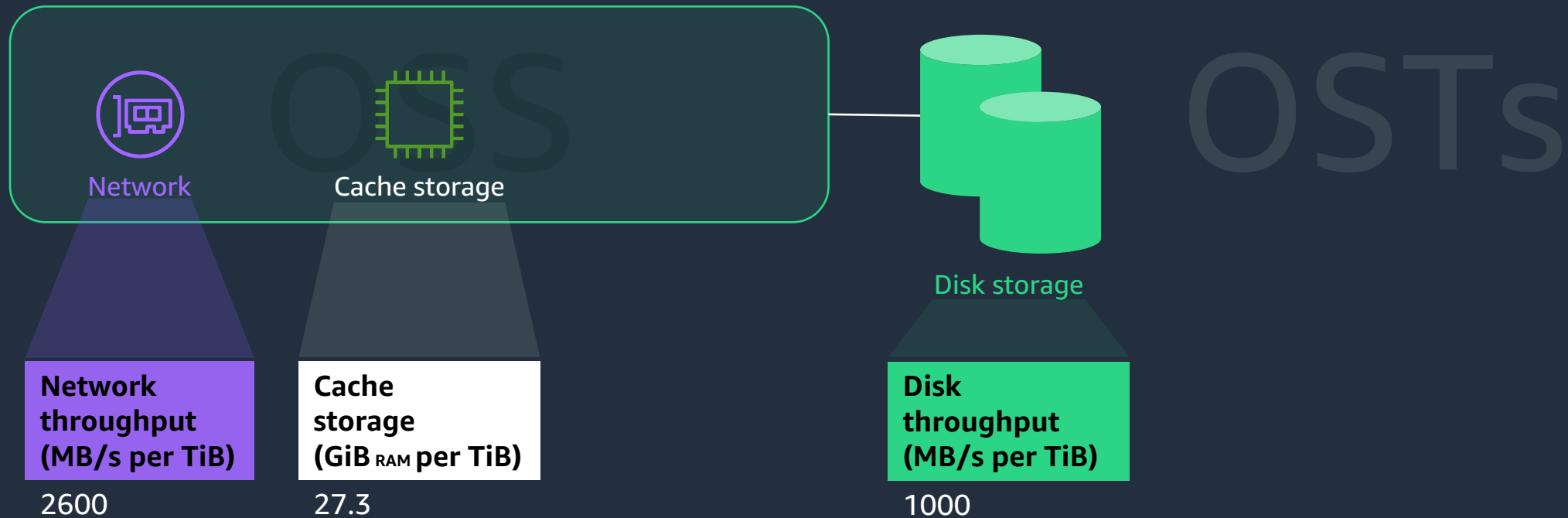
From nothing to 200+ GB/s in 30 minutes or less



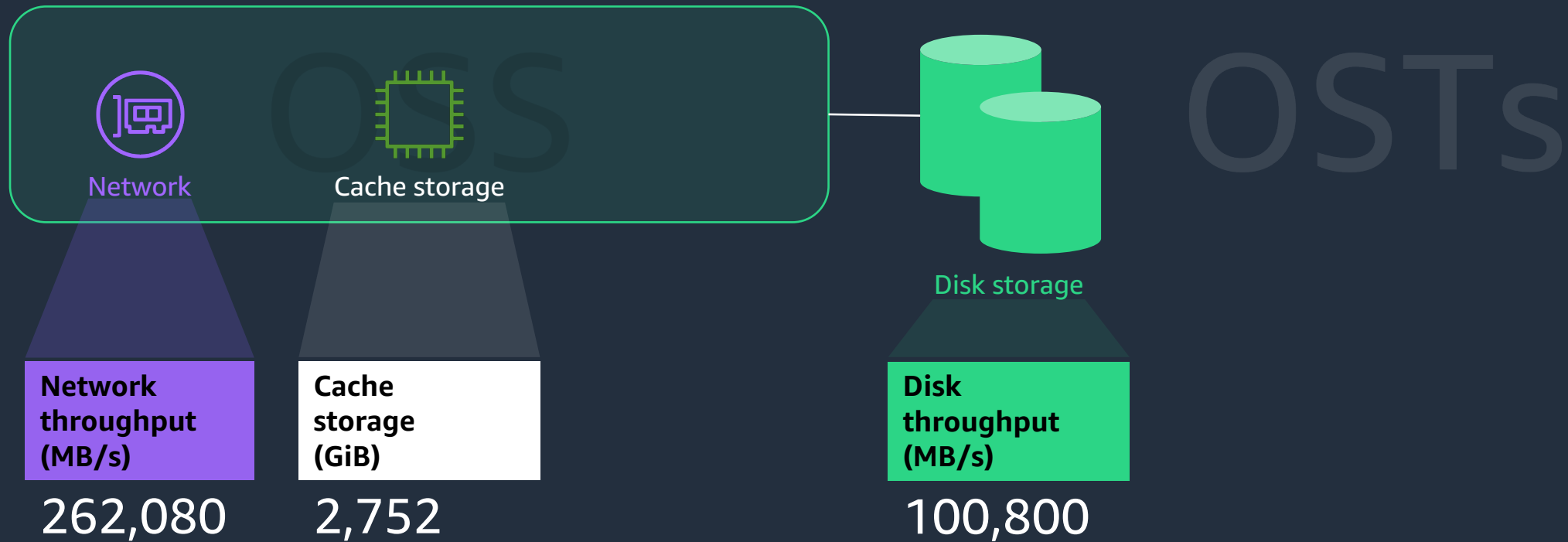
Demo results – Persistent 2 SSD 100.8 TiB at 1000 MB/s per TiB



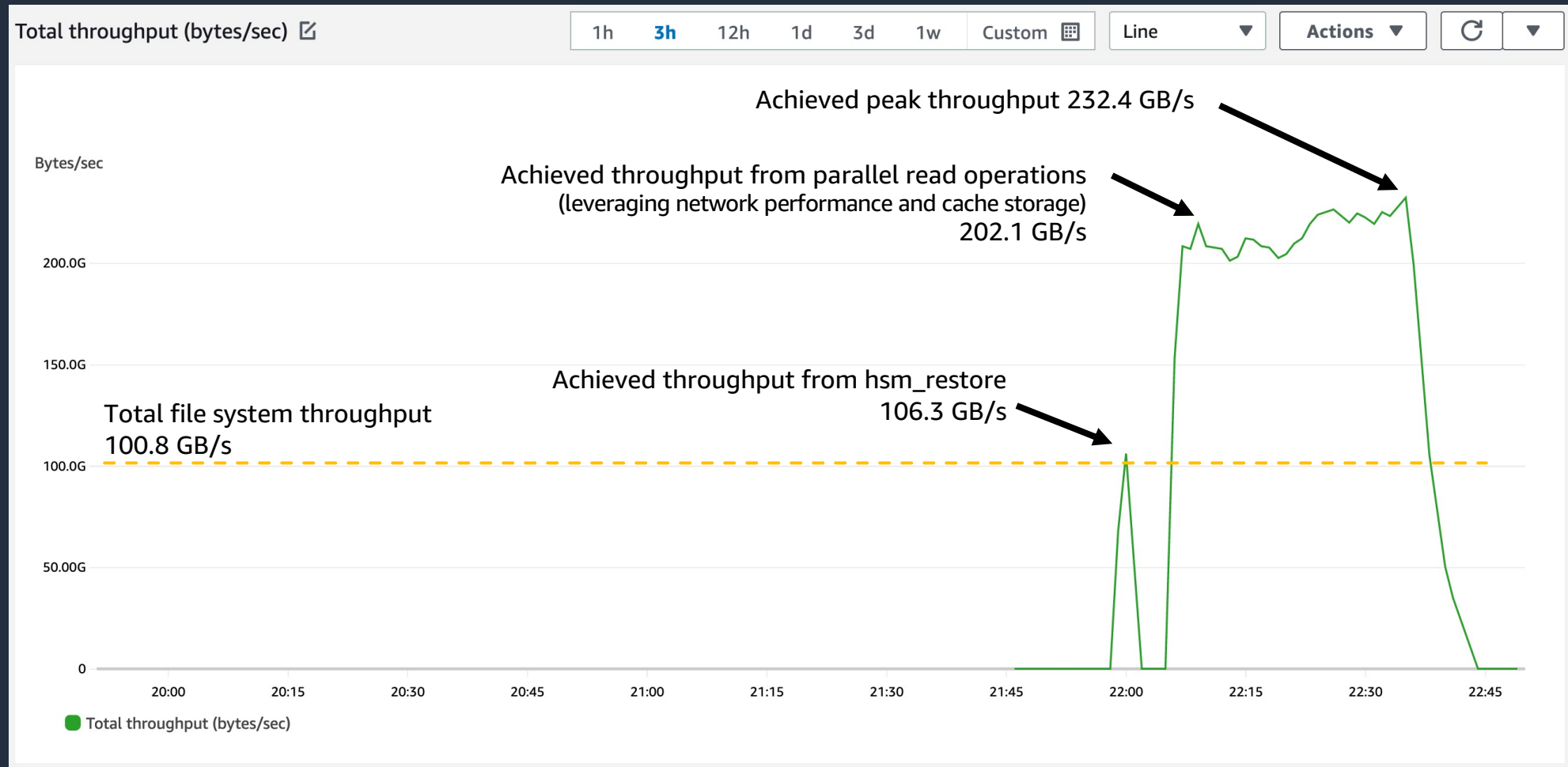
Demo results – Persistent 2 SSD 100.8 TiB at 1000 MB/s per TiB



Demo results – Persistent 2 SSD 100.8 TiB at 1000 MB/s per TiB



Demo results – Persistent 2 SSD 100.8 TiB at 1000 MB/s per TiB



Lustre architecture and performance

The screenshot shows the AWS documentation page for 'Aggregate file system performance' in the Lustre User Guide. The page includes a navigation sidebar on the left with a search bar at the top. The main content area contains a section titled 'Aggregate file system performance' with two paragraphs of text and a table titled 'File system performance for SSD storage options'. The table has five columns: Deployment Type, Network throughput (MB/s/TiB of storage provisioned), Network IOPS (IOPS/TiB of storage provisioned), Cache storage (GiB of RAM/TiB of storage provisioned), and Disk IOPS per file system (millions per P50). There are also social media icons and a QR code on the right side of the page.

Aggregate file system performance

The throughput that an FSx for Lustre file system supports is proportional to its storage capacity. Amazon FSx for Lustre file systems scale to hundreds of GBps of throughput and millions of IOPS. Amazon FSx for Lustre also supports concurrent access to the same file or directory from thousands of compute instances. This access enables rapid data checkpointing from application memory to storage, which is a common technique in high performance computing (HPC). You can increase the amount of storage and throughput capacity as needed at any time after you create the file system. For more information, see [Managing storage and throughput capacity](#).

FSx for Lustre file systems provide burst read throughput using a network I/O credit mechanism to allocate network bandwidth based on average bandwidth utilization. The file systems accrue credits when their network bandwidth usage is below their baseline limits, and can use these credits when they perform network data transfers.

The following tables show performance that the FSx for Lustre deployment options are designed for.

File system performance for SSD storage options				
Deployment Type	Network throughput (MB/s/TiB of storage provisioned)	Network IOPS (IOPS/TiB of storage provisioned)	Cache storage (GiB of RAM/TiB of storage provisioned)	Disk IOPS per file system (millions per P50)

<https://docs.aws.amazon.com/fsx/latest/LustreGuide/performance.html>





Thank you!

Darryl Osborne

 darrylsosborne

 darrylo@amazon.com