BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

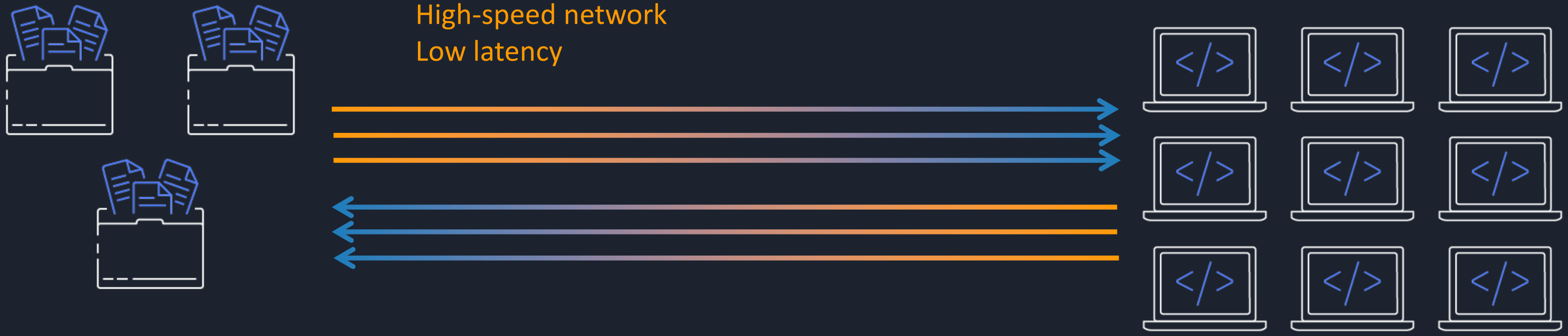
Amazon FSx For Lustre Deep Dive and its importance in Machine Learning

Suman Debnath
Amazon Web Services

What is a high performance workload?

Any workload that processes data at a rapid pace with lots of compute power

Typically involves:



Vast data sets

Ability to serve data quickly:

Fast storage

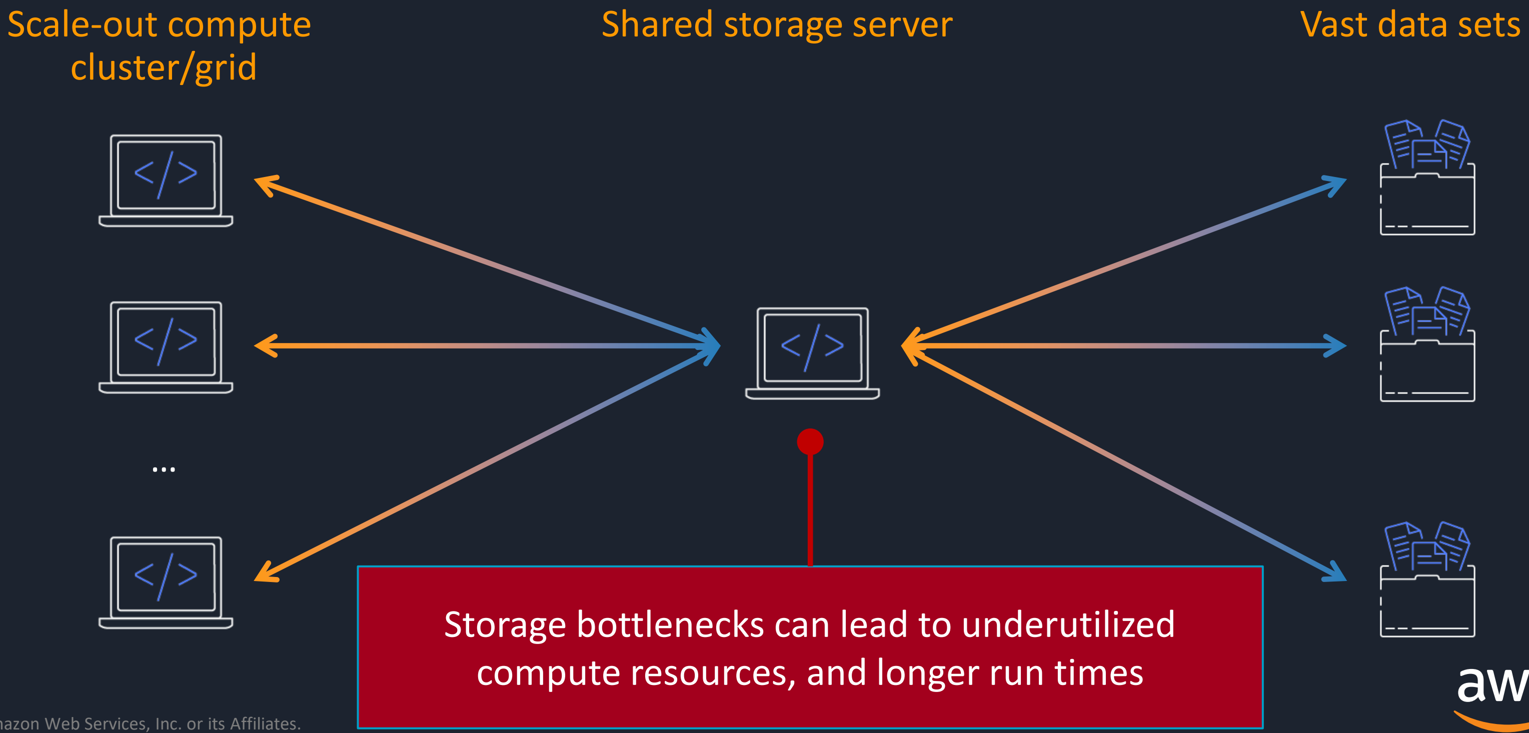
High-speed network

Low latency

Scale-out compute capacity

(hundreds–millions of cores)

For large workloads, compute is distributed across a compute cluster/grid and data is accessed through shared storage



Storage bottlenecks can lead to underutilized compute resources, and longer run times

Why do we need fast parallel file systems?

Amazon FSx for Lustre

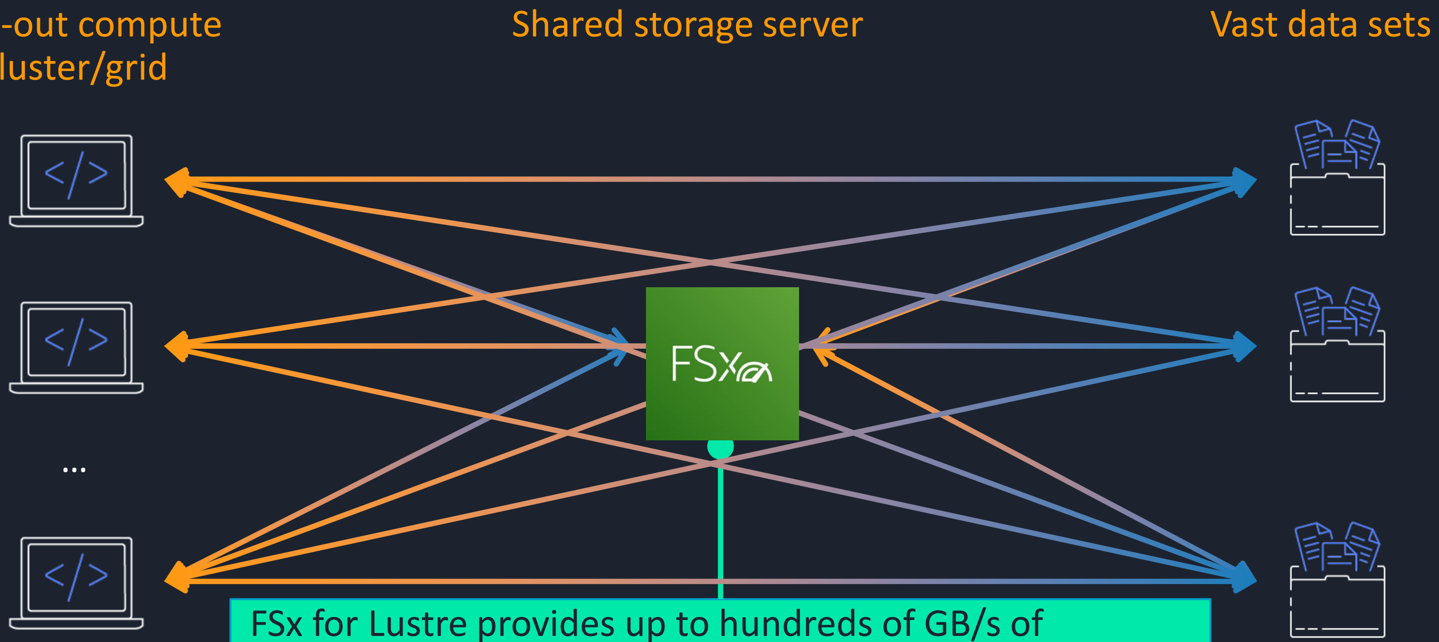
For every $1 spent on high performance computing, businesses see $463 in incremental revenues and $44 in incremental profit[1]

To efficiently utilize high performance processors, memory and networking, these workloads depend on high performance file systems to avoid storage bottlenecks

High performance storage reduces workload runtimes, accelerate business insights, and save costs by keeping compute resources fully utilized

[1] High Performance Computing on AWS Redefines What is Possible

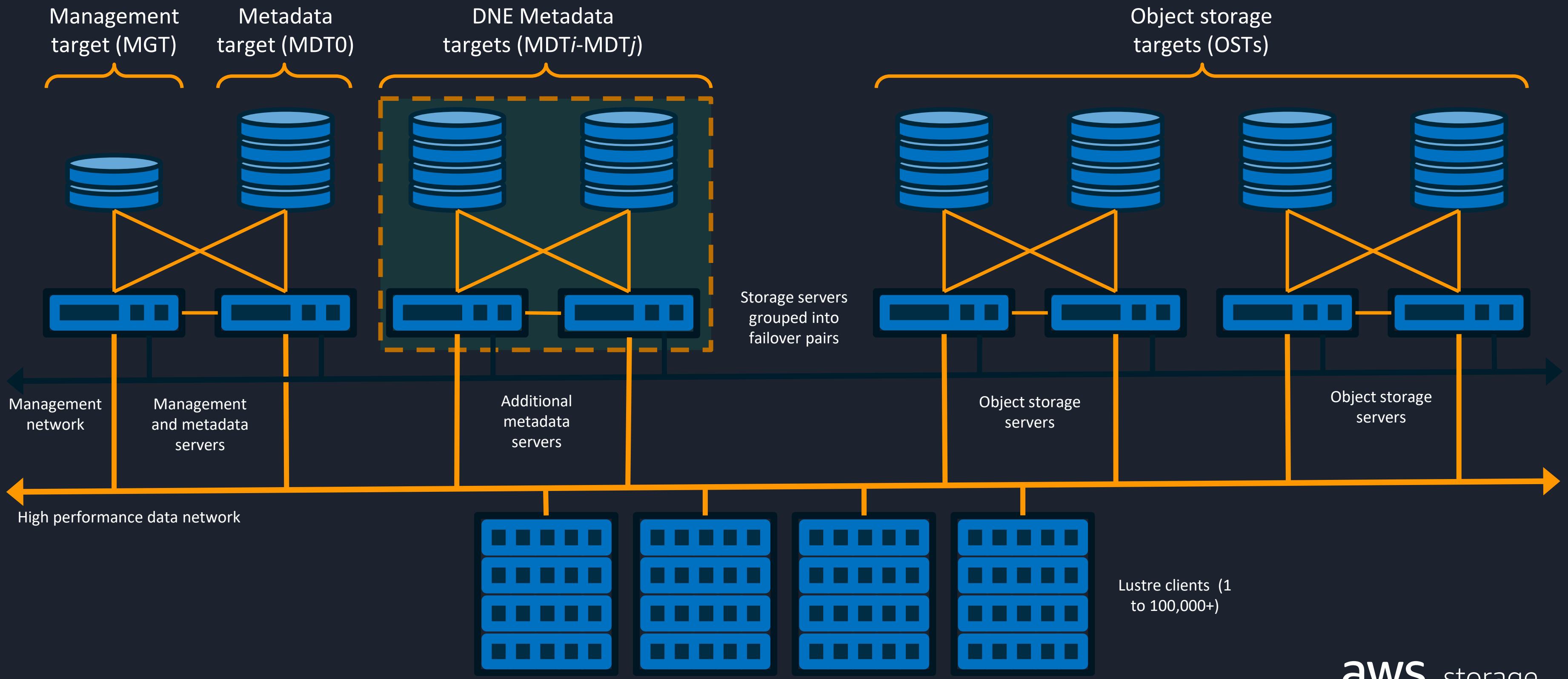
FSx for Lustre provides a scale-out shared file system that avoids storage bottlenecks when running large workloads



FSx for Lustre provides up to hundreds of GB/s of throughput, sub-ms latencies, and millions of IOPS.

aws storage

How Lustre works in FSx for Lustre


Management target (MGT)
Metadata target (MDT0)
DNE Metadata targets (MDT*i*-MDT*j*)
Object storage targets (OSTs)
Storage servers grouped into failover pairs
Management network
Management and metadata servers
Additional metadata servers
Object storage servers
Object storage servers
High performance data network
Lustre clients (1 to 100,000+)

aws storage

Customers continue to increase the size of their workloads on AWS across industry verticals and application areas

Industries and example use cases

Financial services:
Modeling and analytics

Life Sciences:
Genome analysis

Media and Entertainment:
Rendering and transcoding

Automotive:
ECU simulations and object detection

Semiconductor:
Electronic design automation

Oil and gas:
Seismic data processing

Application areas

Big data analytics

Machine learning

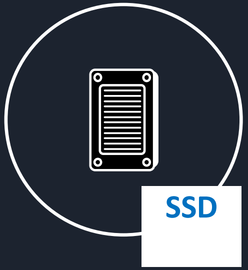
High-performance computing

For every $1 spent on high performance computing, businesses see $463 in incremental revenues and $44 in incremental profit[1]

[1] High Performance Computing on AWS Redefines What is Possible

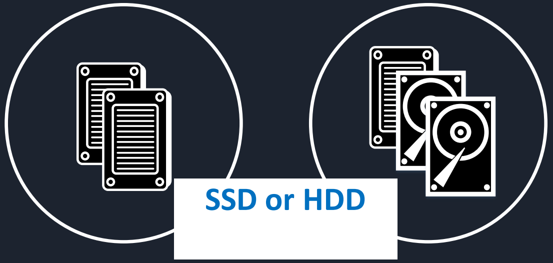
High and scalable performance

High and
scalable
performance



Scratch

Short-term processing
Spin up > process > spin down
Single copy of data



Persistent

Longer-term processing
HA file servers
Replicated copies of data

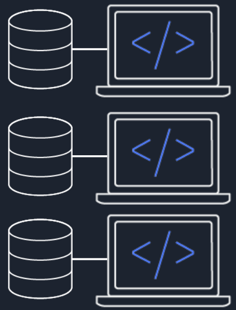
Amazon FSx for Lustre Control Plane (API, management layer, file system control) designed to be highly available (HA) for both options

Amazon FSx for Lustre

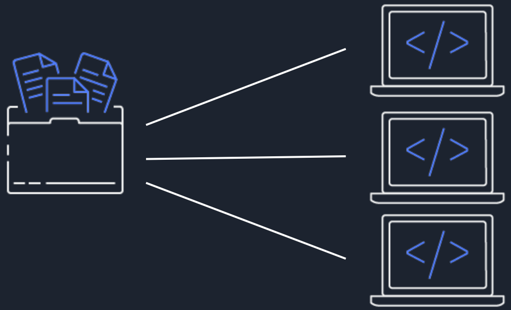
S3 connectivity

For many customers, running large workloads requires transferring data to and from an S3 data lake

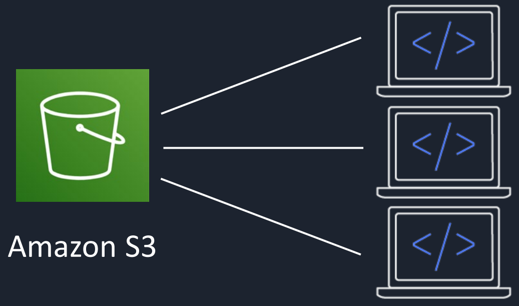
To process your data sets in Amazon S3, you either move them to temporary storage or process them directly on S3



On EBS or instance storage



Self-managed file systems

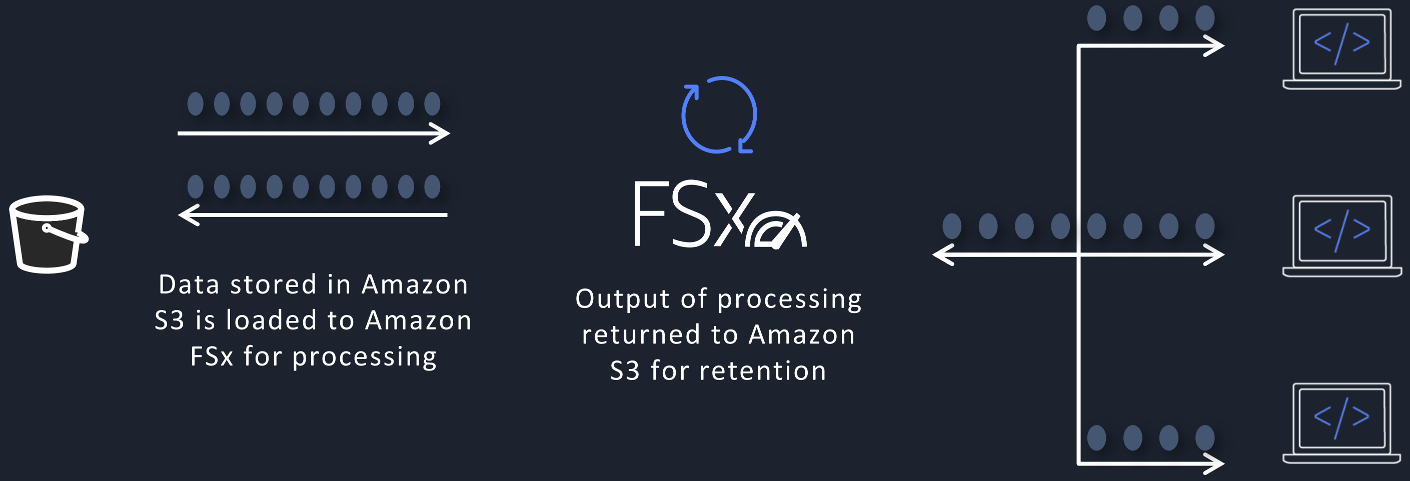


Directly on S3

Objects stored in S3 can be accessed as files on FSx for Lustre with sub-millisecond latencies

Amazon FSx for Lustre

Link your Amazon S3 data set to your Amazon FSx for Lustre file system to see S3 objects represented as files, then…

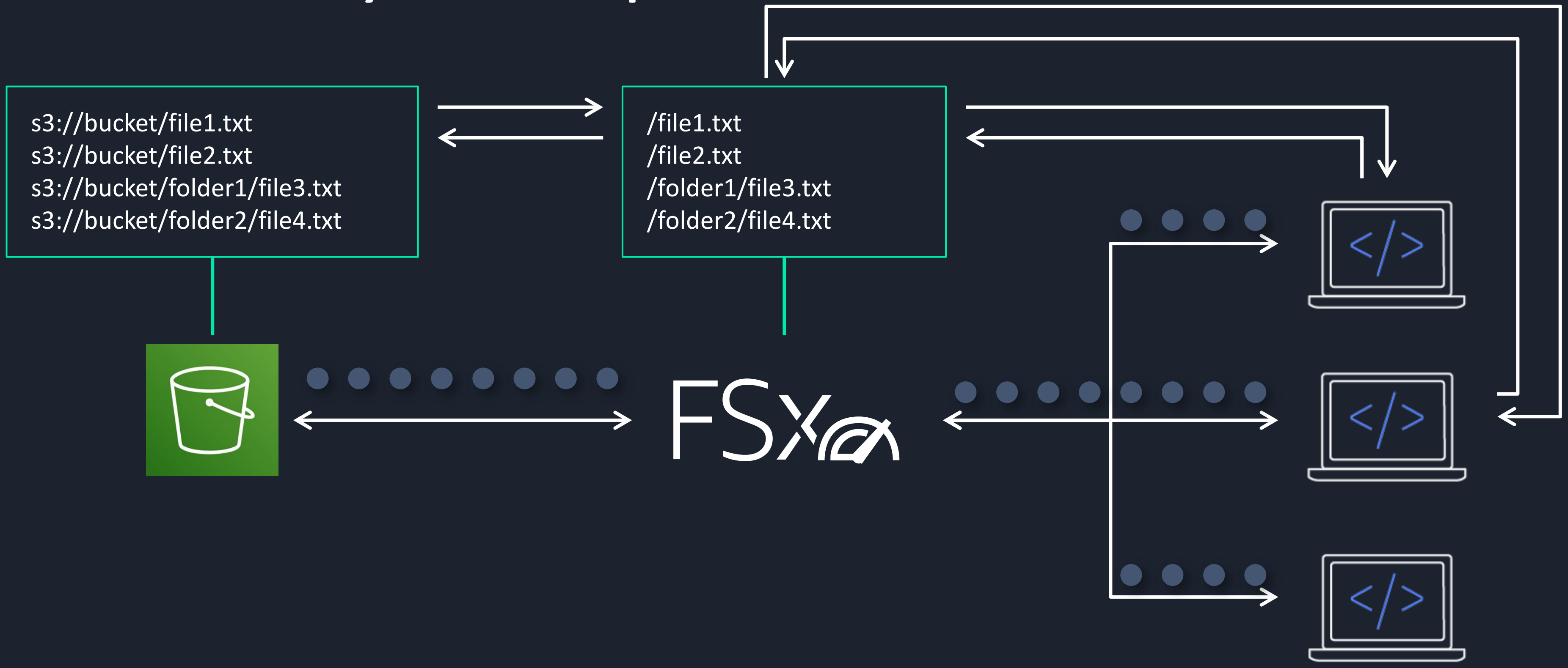


Data stored in Amazon S3 is loaded to Amazon FSx for processing

Output of processing returned to Amazon S3 for retention

… use Amazon FSx as a shared high performance file system to keep up with the storage needs of thousands of compute instances

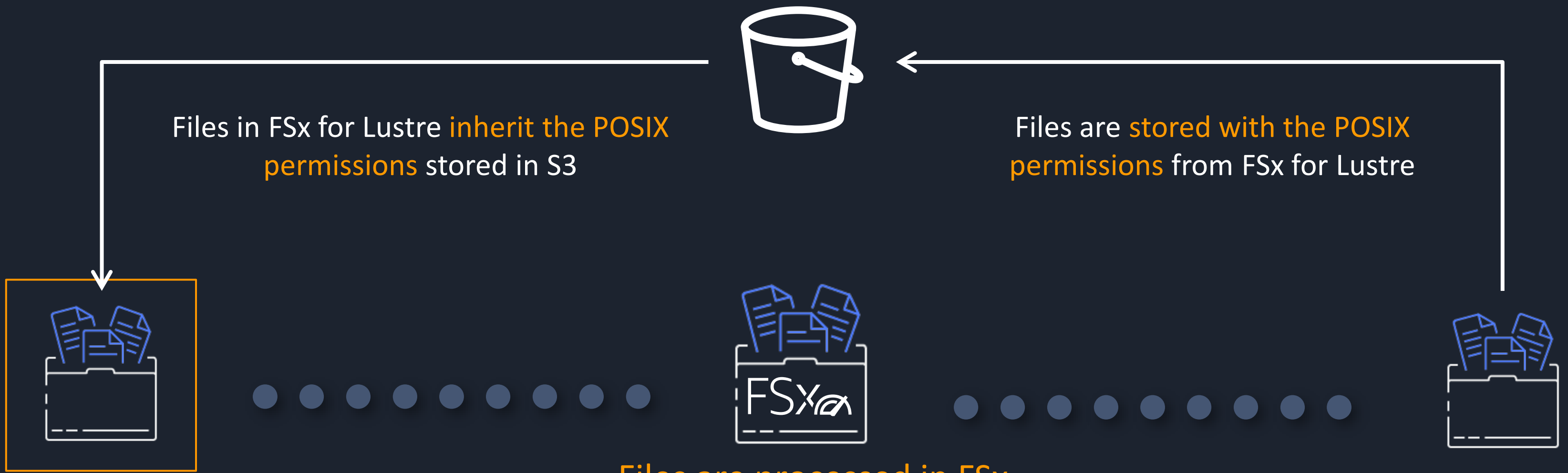
Amazon S3 lazy load example



s3://bucket/file1.txt
s3://bucket/file2.txt
s3://bucket/folder1/file3.txt
s3://bucket/folder2/file4.txt

/file1.txt
/file2.txt
/folder1/file3.txt
/folder2/file4.txt

Preserve POSIX metadata across Amazon FSx and S3



Files in FSx for Lustre inherit the POSIX permissions stored in S3

Files are stored with the POSIX permissions from FSx for Lustre

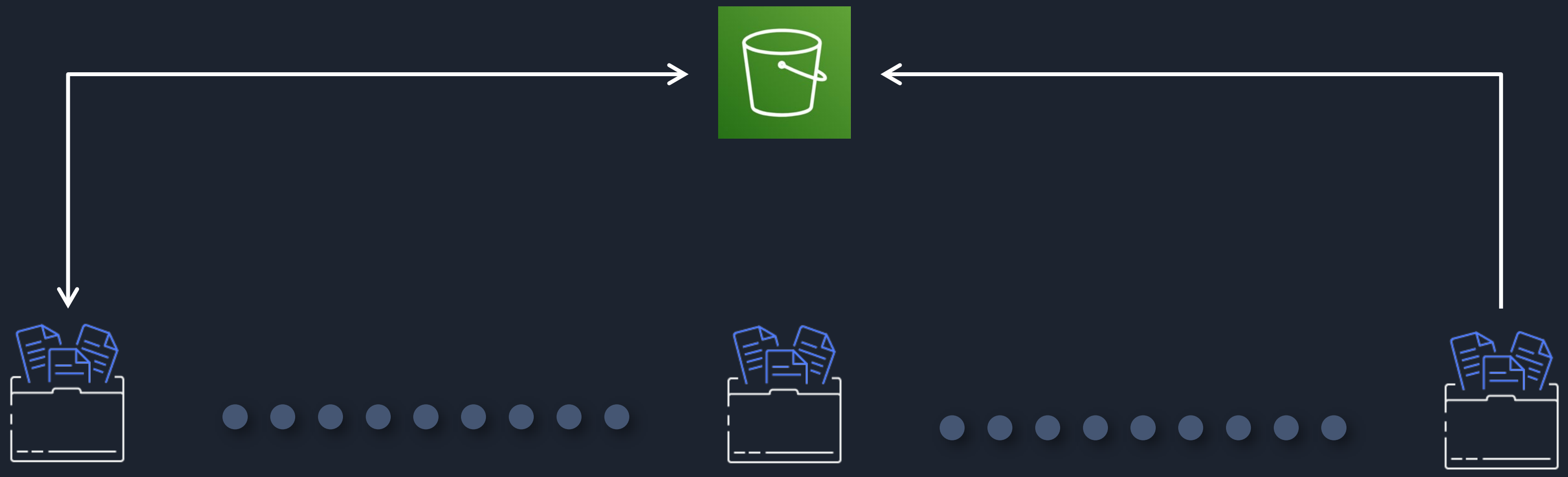
When you create a file system linked to S3 (*CreateFileSystem* API)

Files are processed in FSx for Lustre

Export changes to S3 (*DataRepositoryTask* API)

aws storage

Release inactive data sets to S3 to free up space



Create myfs1 linked to S3 bucket s3://mybucket and mount myfs1

Files are processed in FSx for Lustre

hsm_release files from myfs1 to S3://mybucket

aws storage

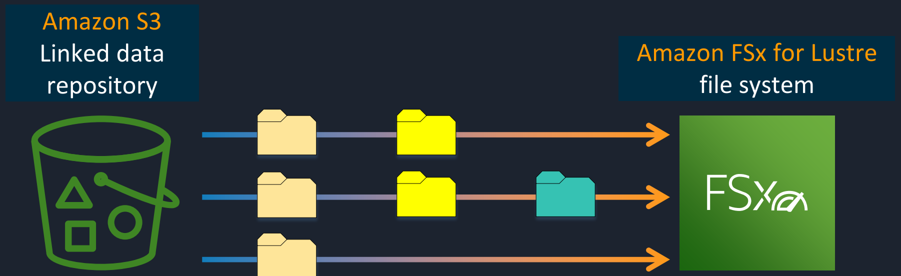
Hierarchical Storage Management (HSM) commands for data movement

hsm_archive – Copy files to Amazon S3 from FSx for Lustre

hsm_release – Free disk space associated with files, once archived

hsm_restore – Bring back file data to FSx for Lustre from Amazon S3 *(also done automatically when accessing a file for the first time)*

Auto-Import with S3 and FSx for Lustre



Original objects and directories initially loaded during file creation

New objects and directories added to S3 buckets after initial load

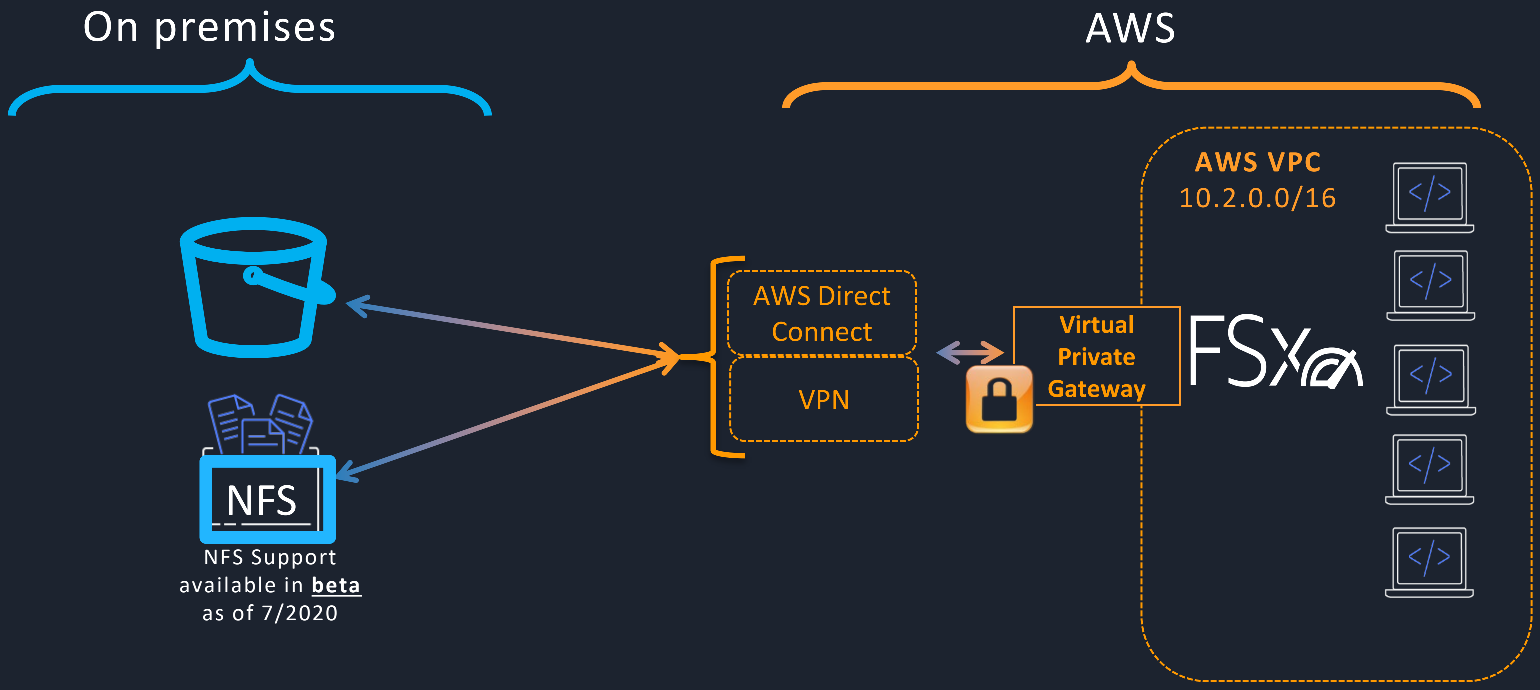
Modified objects and directories changed after initial load

Three ways to manage S3 Auto-Import

1. Update my file and directory as objects are added to my bucket
2. Update my file and directory listing as objects are added to or changed in my bucket
3. Do not update my file and directly listing when objects are added to or changed in my bucket

FSx for Lustre supports cloud bursting from on premises, also supports NFS repository

Amazon FSx for Lustre



NFS Support available in **beta** as of 7/2020

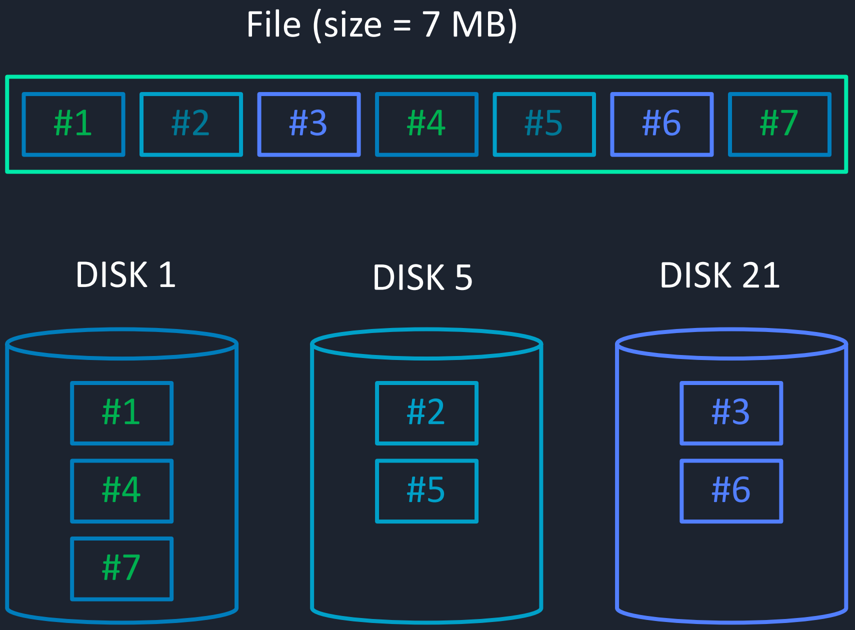
Amazon FSx for Lustre

Performance Tuning

What is striping, why use it?

- Striping refers to sharding large files in to fragments and storing them across disks in multiple servers
- It allows you to parallelize access to individual files, driving higher aggregate throughput
- By default each file is stored in one disk
- Striping can be set per directory or per file
- All files in a directory inherit it's striping parameters

How striping works in FSx for Lustre



stripe_count = 3

stripe_size = 1 MB

Specify stripe_count and stripe_size (lfs setstripe)

Striping can be set per directory or per file, all files in a directory inherit it’s striping parameters

Stripe files across disks based on CloudWatch Max metric

Set ImportedFileChunkSize = (dominant file size / # of disks)

Optimizing I/O performance on FSx for Lustre

Best practices for striping file system data

- Stripe files to optimize I/O performance when concurrent access is common

Average I/O size

- Throughput increases with higher average I/O size

Client selection

- Choose EC2 instance type with enough memory, CPU, and bandwidth

Best Practices to optimize Performance

- Parallelize your workload

  Use multiple threads per client. If a client are fully utilized, add additional clients.

- Balance workload across OSTs

  Stripe files to optimize I/O performance when concurrent access is common

  Set ImportedFileChunkSize = (dominant file size / # of disks)

- Average I/O size

  Throughput increases with higher average I/O size

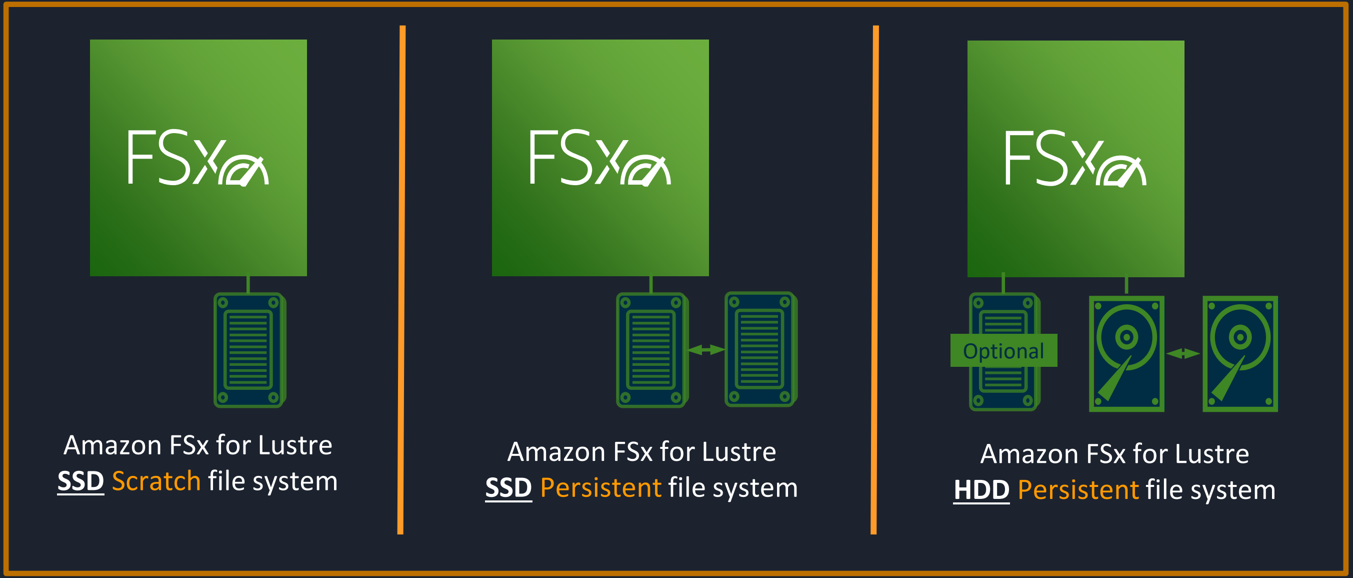
- Client selection

  Choose EC2 instance type with enough memory, CPU, and bandwidth

Tiers and Performance Options

FSx for Lustre deployment options

High and scalable performance



Amazon FSx for Lustre **SSD** Scratch file system

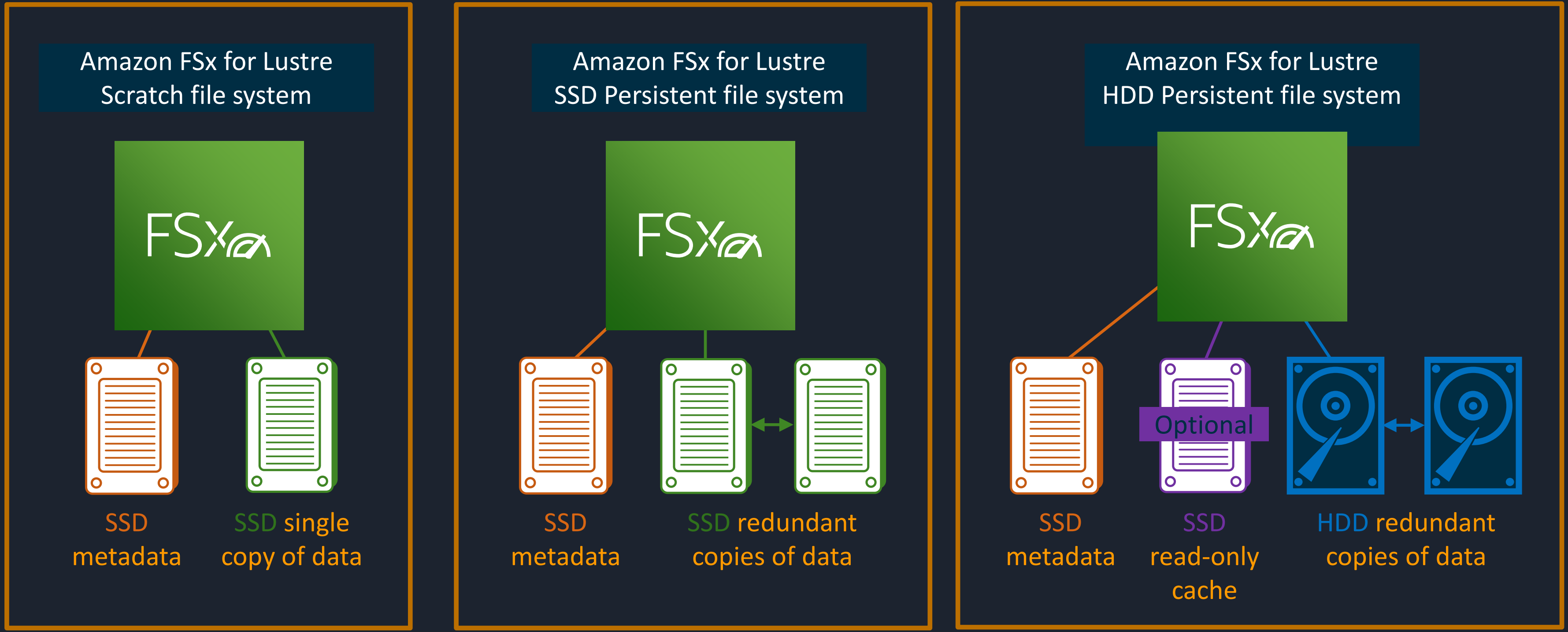
Amazon FSx for Lustre **SSD** Persistent file system

Amazon FSx for Lustre **HDD** Persistent file system

In all options, we support encryption at-rest and in-transit*

aws storage

FSx for Lustre SSD & HDD Tiers



Amazon FSx for Lustre
Scratch file system

SSD metadata

SSD single copy of data

Amazon FSx for Lustre
SSD Persistent file system

SSD metadata

SSD redundant copies of data

Amazon FSx for Lustre
HDD Persistent file system

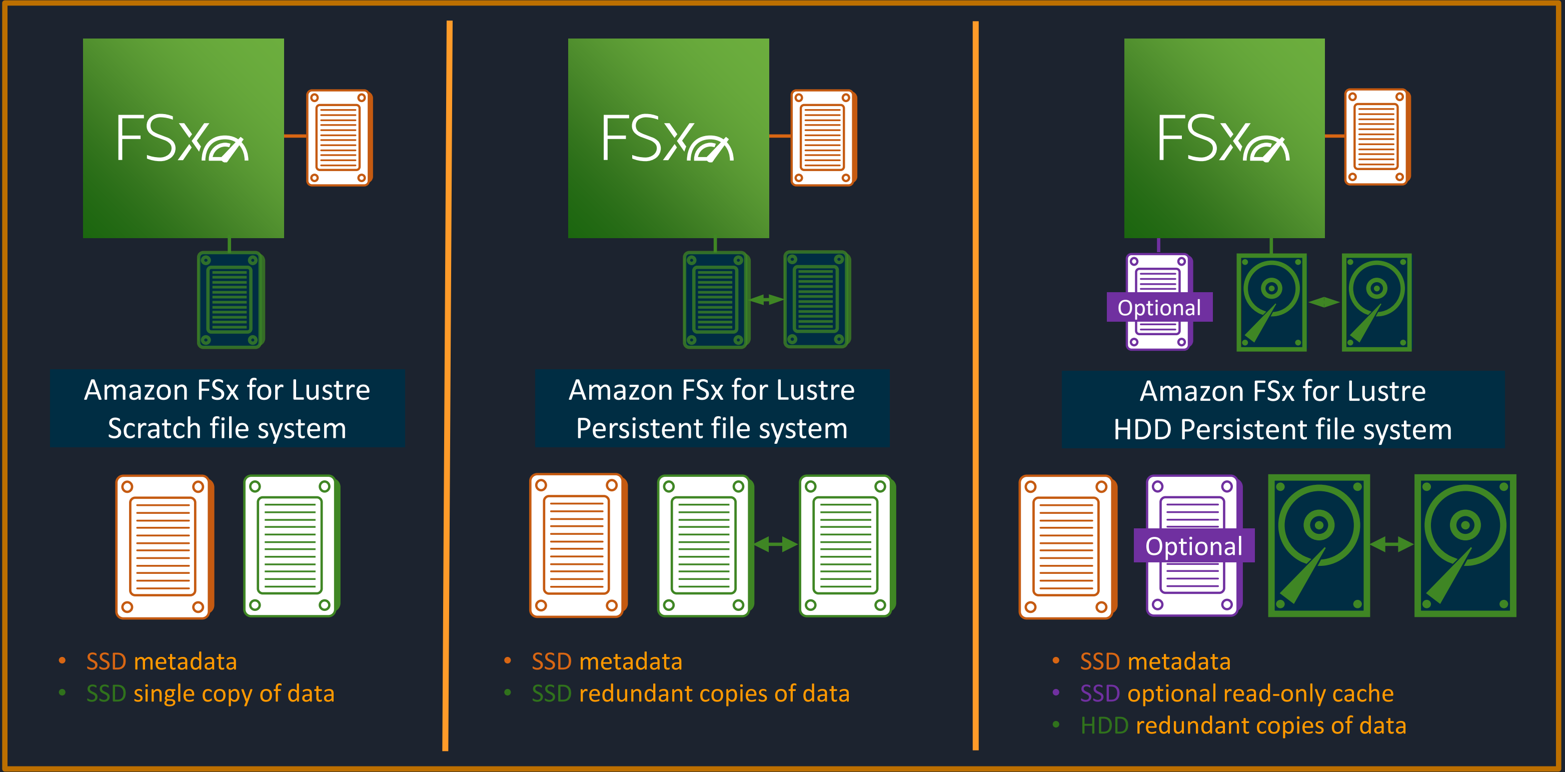
SSD metadata

Optional

SSD read-only cache

HDD redundant copies of data

FSx for Lustre SSD & HDD Tiers with Optional Cache



Amazon FSx for Lustre
Scratch file system

- SSD metadata
- SSD single copy of data

Amazon FSx for Lustre
Persistent file system

- SSD metadata
- SSD redundant copies of data

Amazon FSx for Lustre
HDD Persistent file system

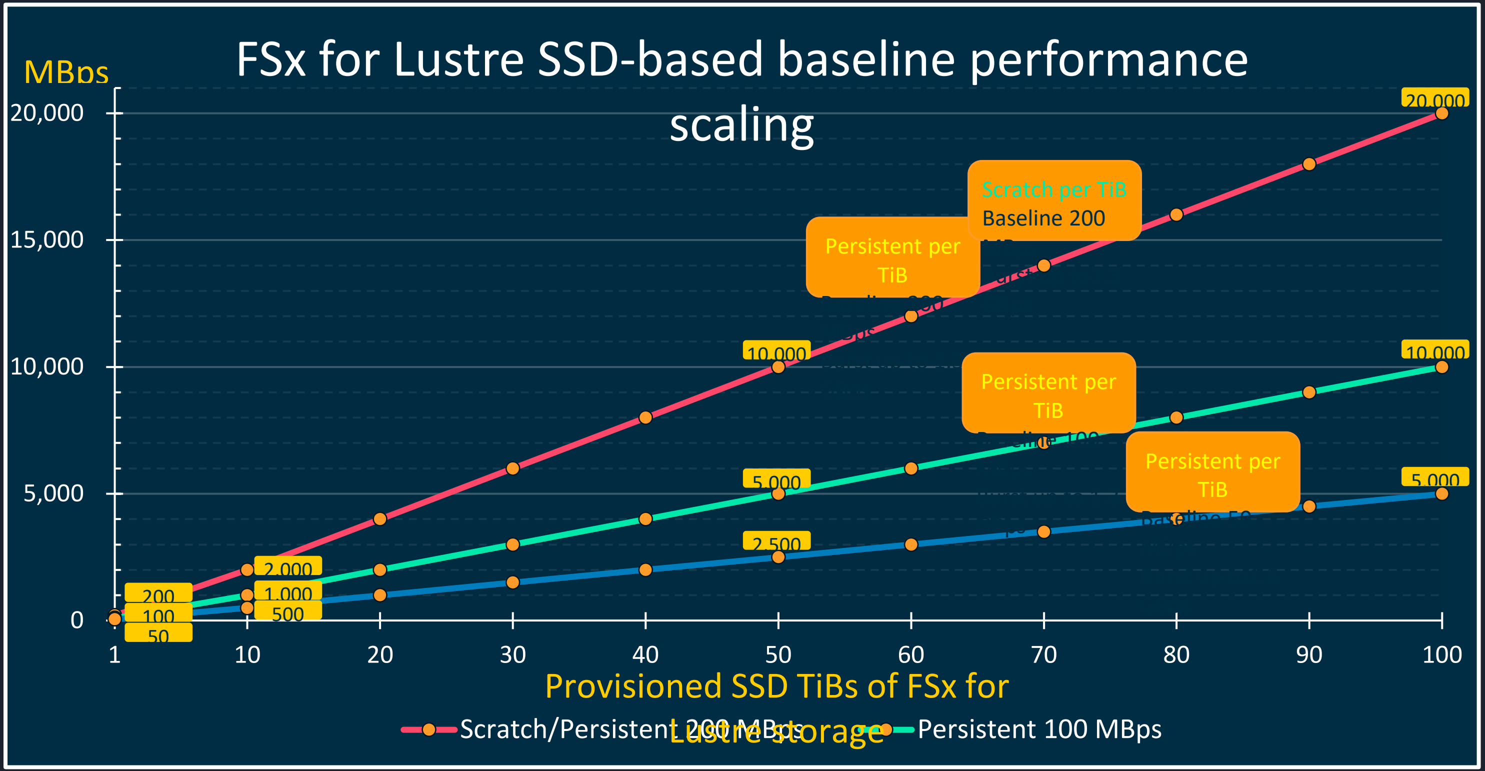
- SSD metadata
- SSD optional read-only cache
- HDD redundant copies of data

aws storage

FSx for Lustre SDD Performance Scaling

Provisioned storage (TiBs)	Scratch 200 MBps baseline	Persistent 200 MBps baseline	Persistent 100 MBps baseline	Persistent 50 MBps baseline	Burst up to 1.3 GBps
1	200	200	100	50	1,300
10	2,000	2,000	1,000	500	13,000
50	10,000	10,000	5,000	2,500	65,000
100	20,000	20,000	10,000	5,000	130,000
1,000	200,000	200,000	100,000	50,000	1,300,000

FSx for Lustre SSD Performance Scaling



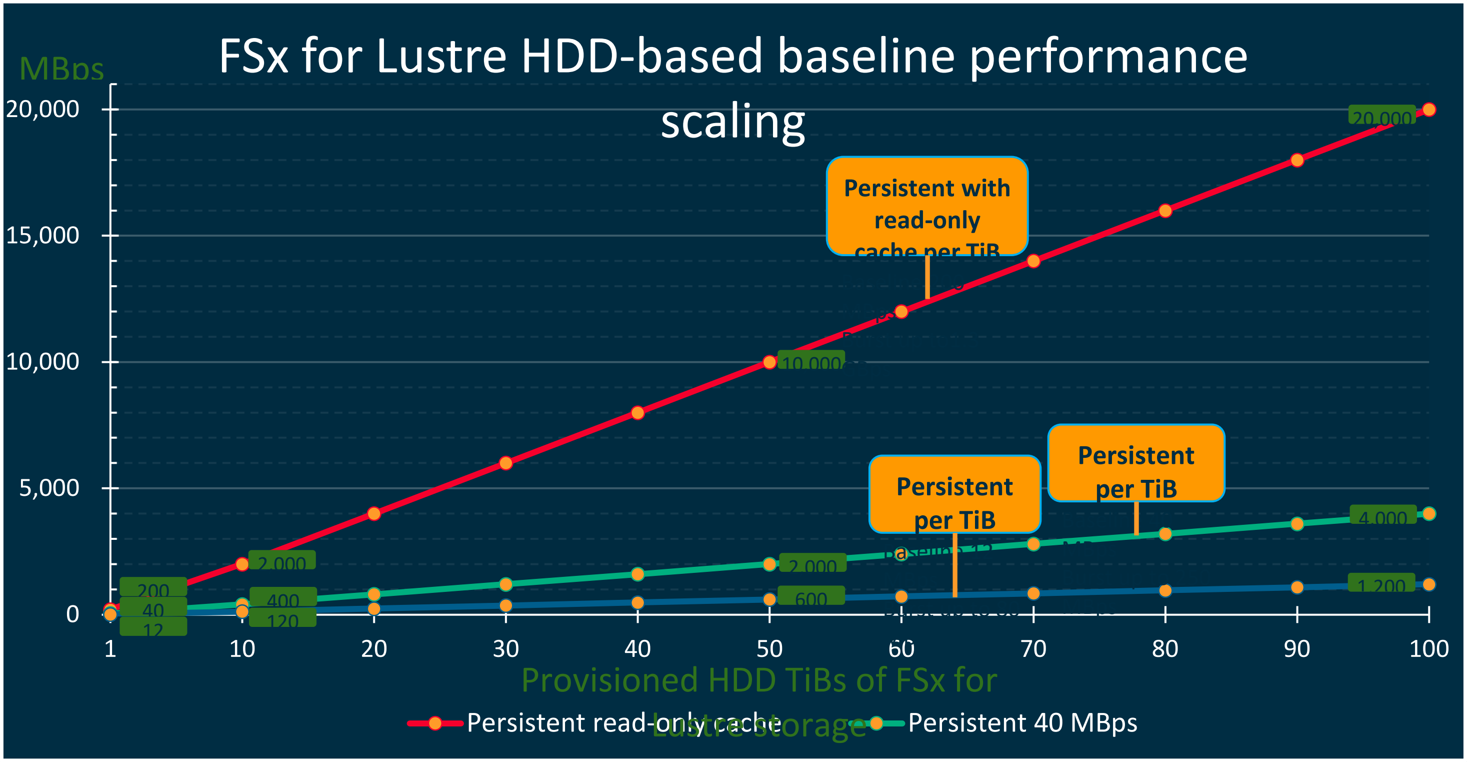
FSx for Lustre SSD-based baseline performance scaling

aws storage

FSx for Lustre HDD Performance Scaling

Provisioned storage (TiBs)	Read-only cache 200 MBps baseline / burst *	Persistent 40 MBps baseline / burst	Persistent 12 MBps baseline / burst
1	200 baseline 1,300 burst	40 baseline 250 burst	12 baseline 80 burst
10	2,000 baseline 13,000 burst	400 baseline 2,500 burst	120 baseline 800 burst
50	10,000 baseline 65,000 burst	2,000 baseline 12,500 burst	600 baseline 4,000 burst
100	20,000 baseline 130,000 burst	4,000 baseline 25,000 burst	1,200 baseline 8,000 burst
1,000	200,000 baseline 1,300,000 burst	40,000 baseline 250,000 burst	12,000 baseline 80,000 burst

FSx for Lustre HDD Performance Scaling



Multiple FSx for Lustre throughput options and deployment types allow customers to optimize storage cost and performance

Storage type	Baseline throughput	Price per GB-month (in IAD) [1]	
		Persistent storage	Scratch Storage
HDD (New!)	12 MB/s per TiB	$0.025 $0.041 (with SSD cache)	-
	40 MB/s per TiB	$0.083 $0.099 (with SSD cache)	-
SSD	50 MB/s per TiB	$0.140	-
	100 MB/s per TiB	$0.190	-
	200 MB/s per TiB	$0.290	$0.14

- Scratch file systems are ideal for temporary storage and shorter-term processing of data.
- Data is not replicated and does not persist if a file server fails.

- File systems with SSD storage can burst up to 1.3 GB/s per TiB

Sample pricing for AID - US East (N. Virginia)

[1] Prices are as of August 14, 2020 and subject to change without notice. Pricing varies by AWS Region. For current pricing information, see the Amazon FSx for Lustre Pricing page on the AWS website.

aws storage

AWS integrations | Amazon FSx for Lustre

FSx

Amazon FSx for Lustre

Integrations

- Amazon VPC
- Amazon IAM
- Amazon KMS
- Amazon CloudWatch
- Amazon CloudTrail
- Amazon CloudFormation
- Amazon S3
- AWS Mgt. Console
- AWS CLI
- Amazon SageMaker
- Amazon Elastic Kubernetes
- AWS ParallelCluster
- AWS Batch
- AWS Direct Connect
- AWS VPN
- Amazon ECS
- Amazon EC2

aws storage

Amazon FSx for Lustre availability *

US West (Oregon)
US West (N. California)
US East (N. Virginia)
US East (Ohio)
Canada (Montreal)
Europe (Ireland)
Europe (Frankfurt)
Europe (London)
Europe (Stockholm)
Europe (Paris)
Asia Pacific (Sydney)
Asia Pacific (Singapore)
Asia Pacific (Tokyo)
Asia Pacific (Hong Kong)
Asia Pacific (Seoul)
Asia Pacific (Mumbai)

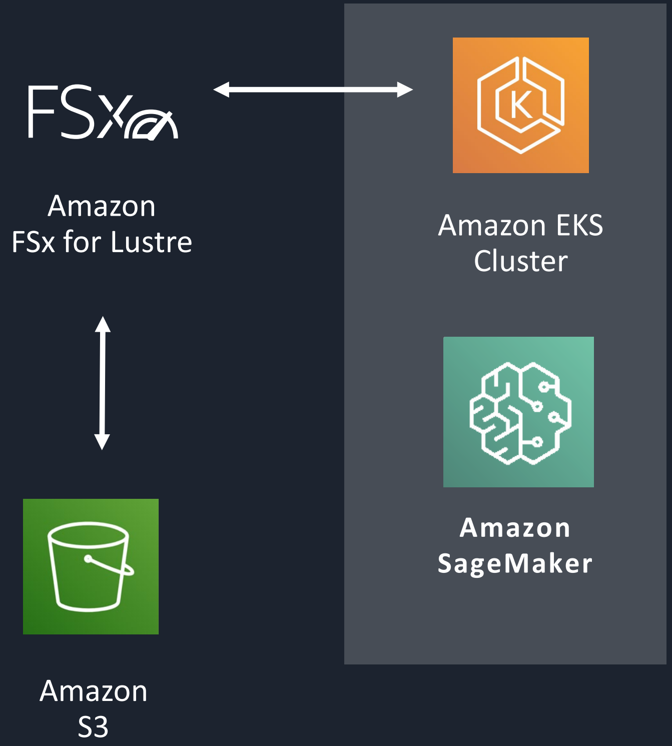
Additional AWS Regions coming soon

** Availability as of August 12, 2020*

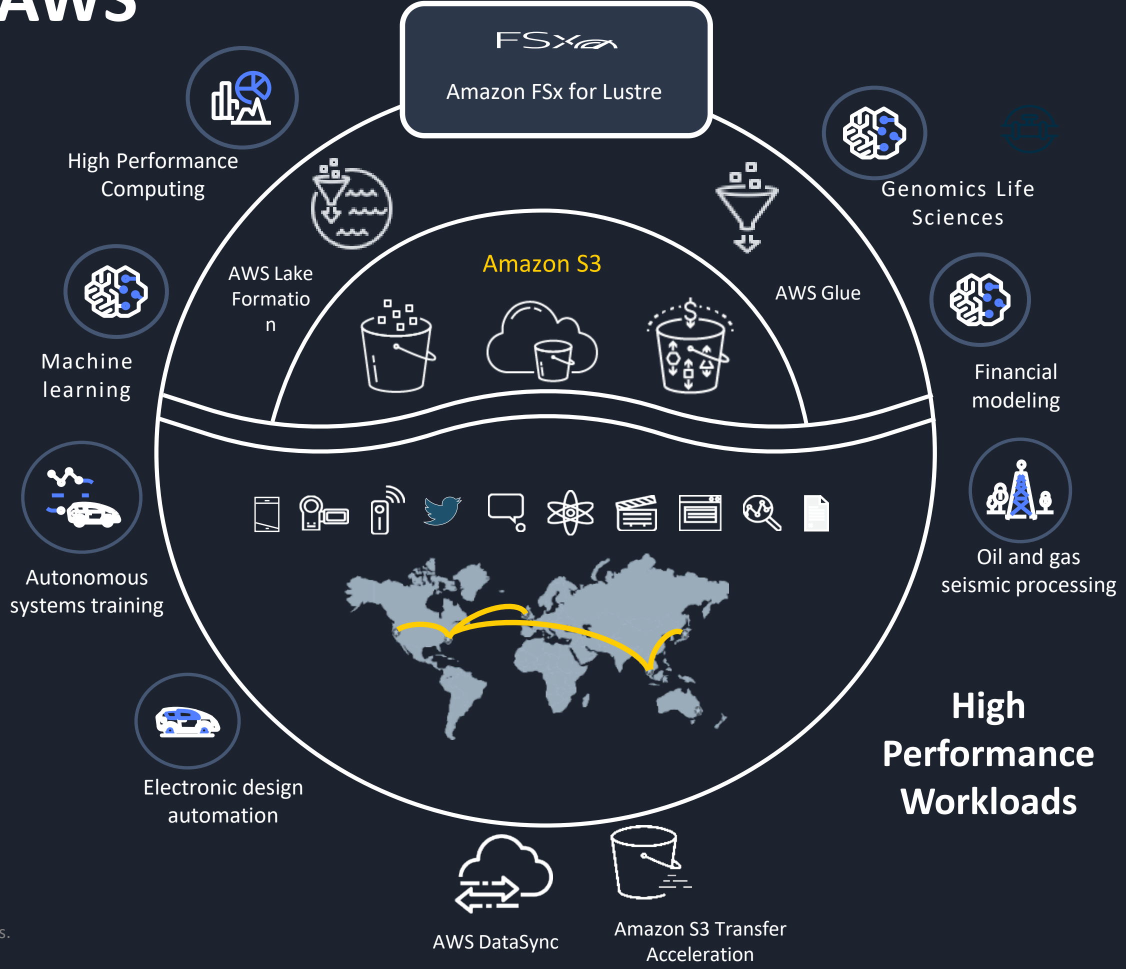
Use Cases

Amazon EKS & SM integration

- FSx for Lustre can be used as persistent volume (PVC) for self-managed Kubernetes or Amazon EKS cluster.
- Allows data to persistent beyond the lifecycle of a Kubernetes pod.
- Can be used as input data source for machine learning jobs on EKS using SageMaker Operators for Kubernetes.



Data Lake on AWS


Amazon FSx for Lustre
High Performance Computing
Machine learning
Autonomous systems training
Electronic design automation
AWS Lake Formatio n
Amazon S3
AWS Glue
Genomics Life Sciences
Financial modeling
Oil and gas seismic processing
AWS DataSync
Amazon S3 Transfer Acceleration

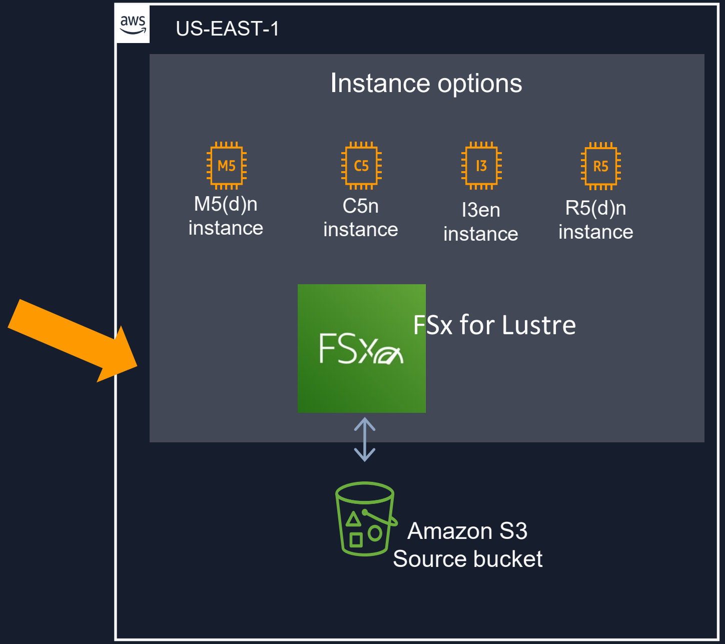
High Performance Workloads

aws storage

Autonomous Vehicles - Architectures in AWS

<Streaming versus Batch architectural consideration>

Architecture with FSx Lustre



US-EAST-1

Instance options

M5(d)n instance

C5n instance

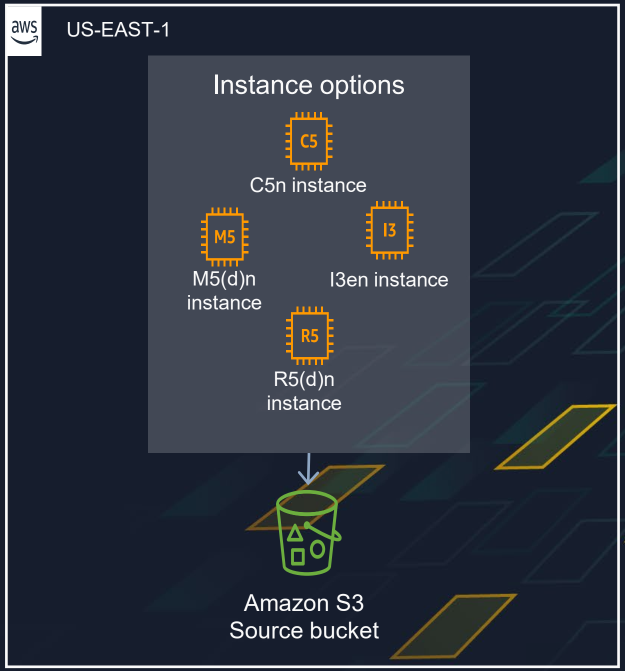
I3en instance

R5(d)n instance

FSx for Lustre

Amazon S3 Source bucket

Architecture with S3



US-EAST-1

Instance options

C5n instance

M5(d)n instance

I3en instance

R5(d)n instance

Amazon S3 Source bucket

Thank you!