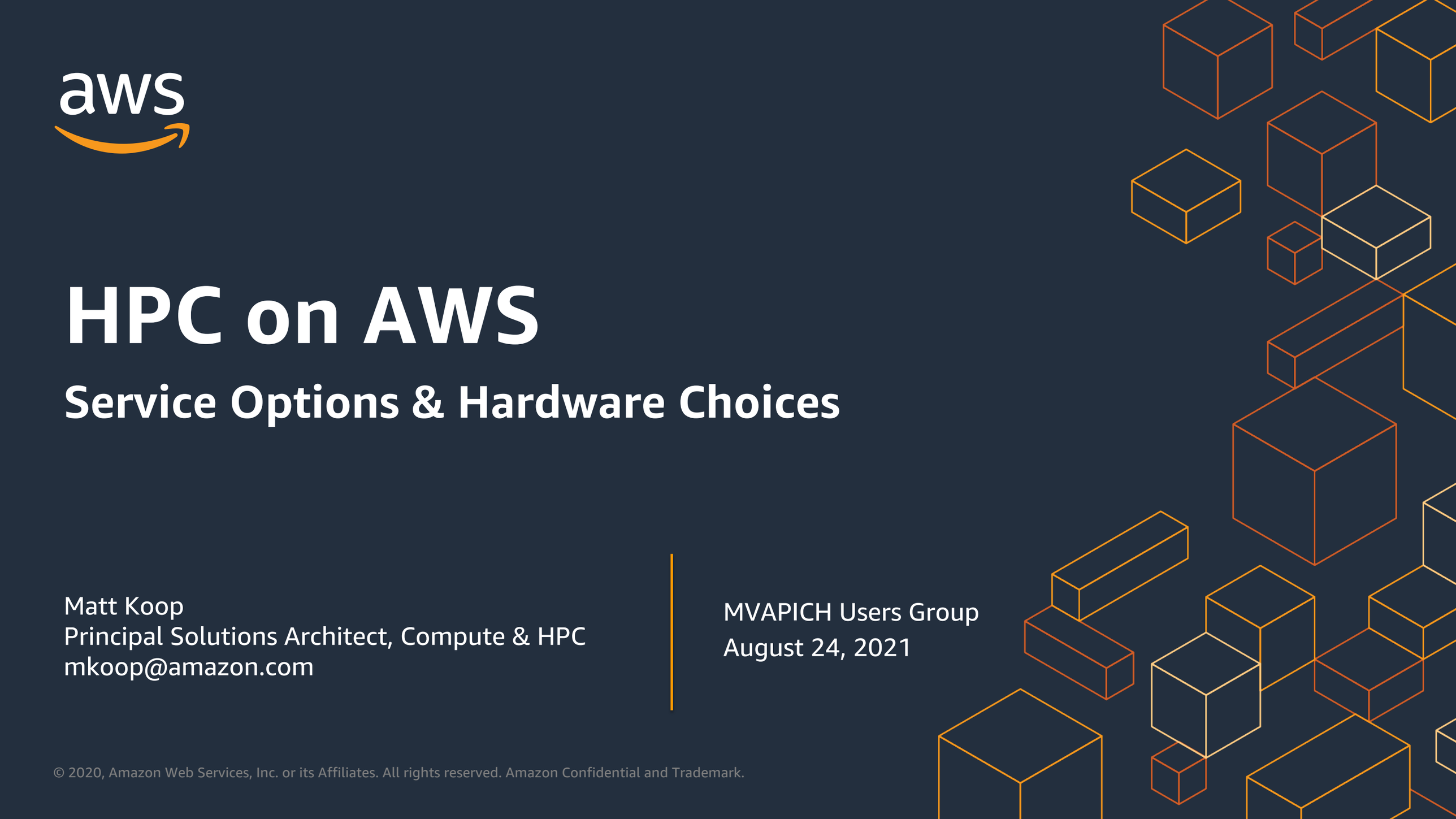


HPC on AWS

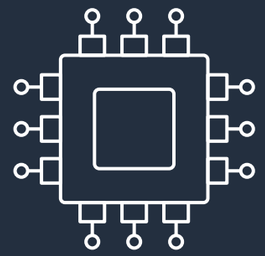
Service Options & Hardware Choices

Matt Koop
Principal Solutions Architect, Compute & HPC
mkoop@amazon.com

MVAPICH Users Group
August 24, 2021

Key services and hardware that enable HPC on AWS

Compute



Amazon EC2

Bare metal performance
AWS Graviton 2
EC2 Ultra Clusters

Networking

Elastic Fabric Adapter (EFA)

MVAPICH2-X-AWS is designed for EFA

Storage

Amazon FSx for Lustre

Workflow

AWS ParallelCluster

AWS Batch

EC2: Broad and Deep Instance Choice

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage
(high I/O, dense)

GPU compute

Graphics intensive

CAPABILITIES

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.5 GHz)

High memory footprint
(up to 24 TiB)

Instance storage
(HDD and SSD)

Accelerated computing
(GPU, FPGA, and ASIC)

Networking
(up to 400 Gbps)

Bare metal

Size
(Nano to 32xlarge)

OPTIONS

Linux, Unix, Windows, macOS

Amazon EBS

Amazon Elastic Inference

Elastic Fabric Adapter

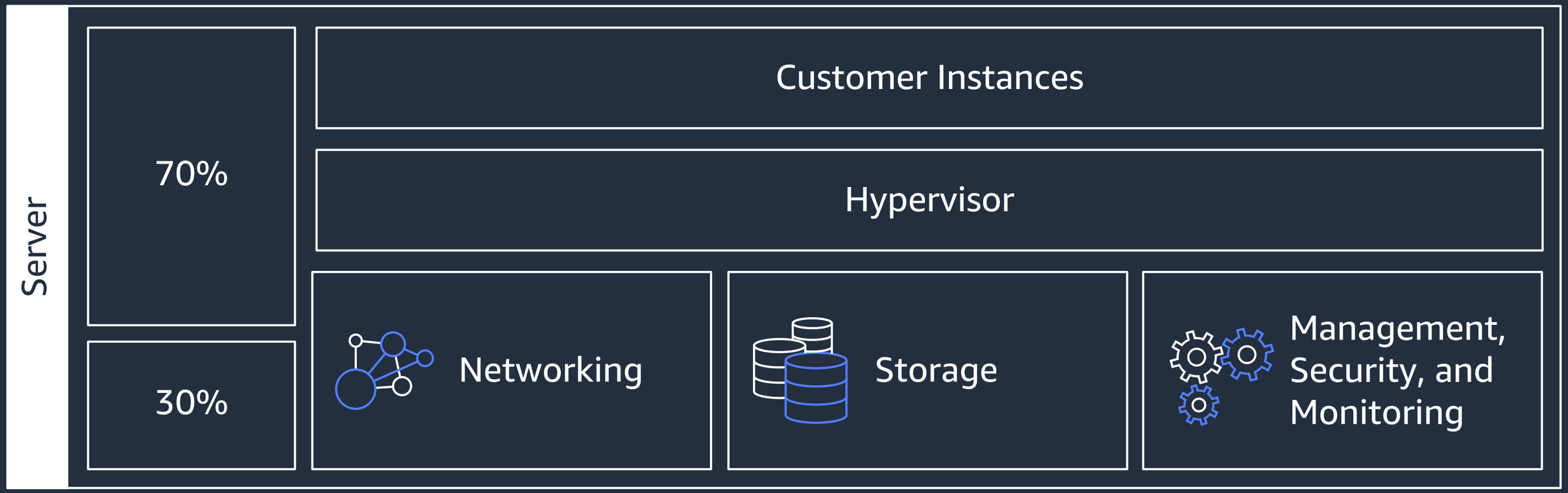
400+

INSTANCE TYPES

for virtually every workload and business need

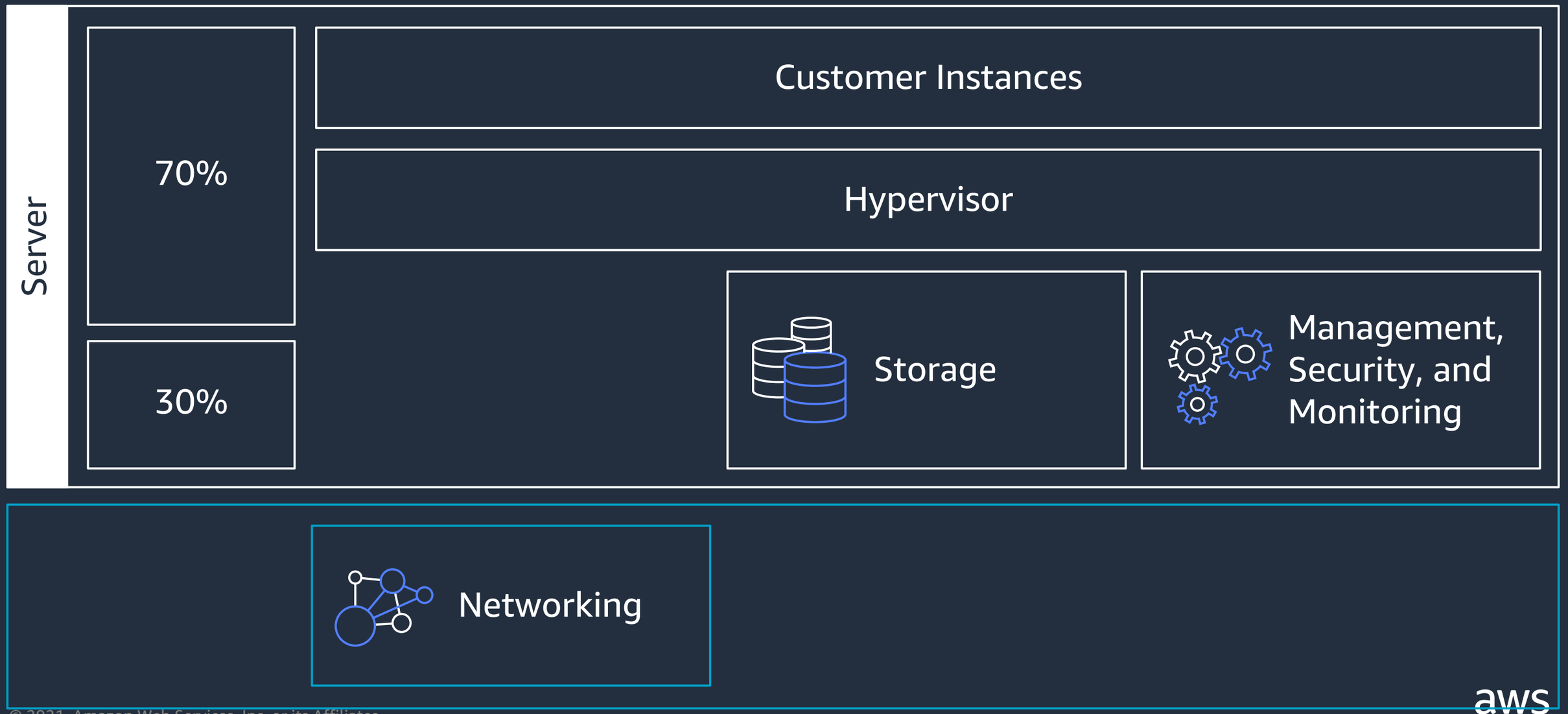
aws

EC2 “Instance” host architecture



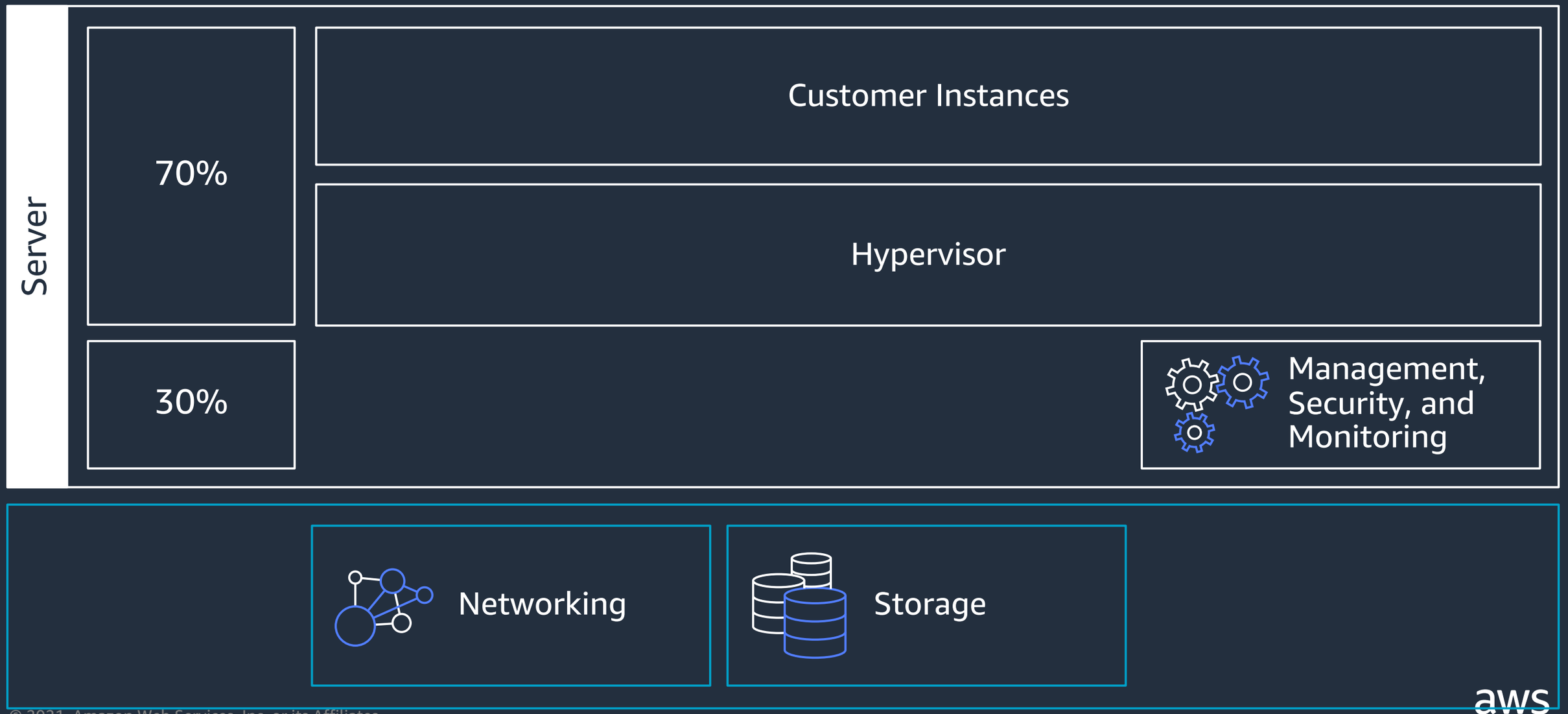
aws

2012 EC2 “Instance” host architecture



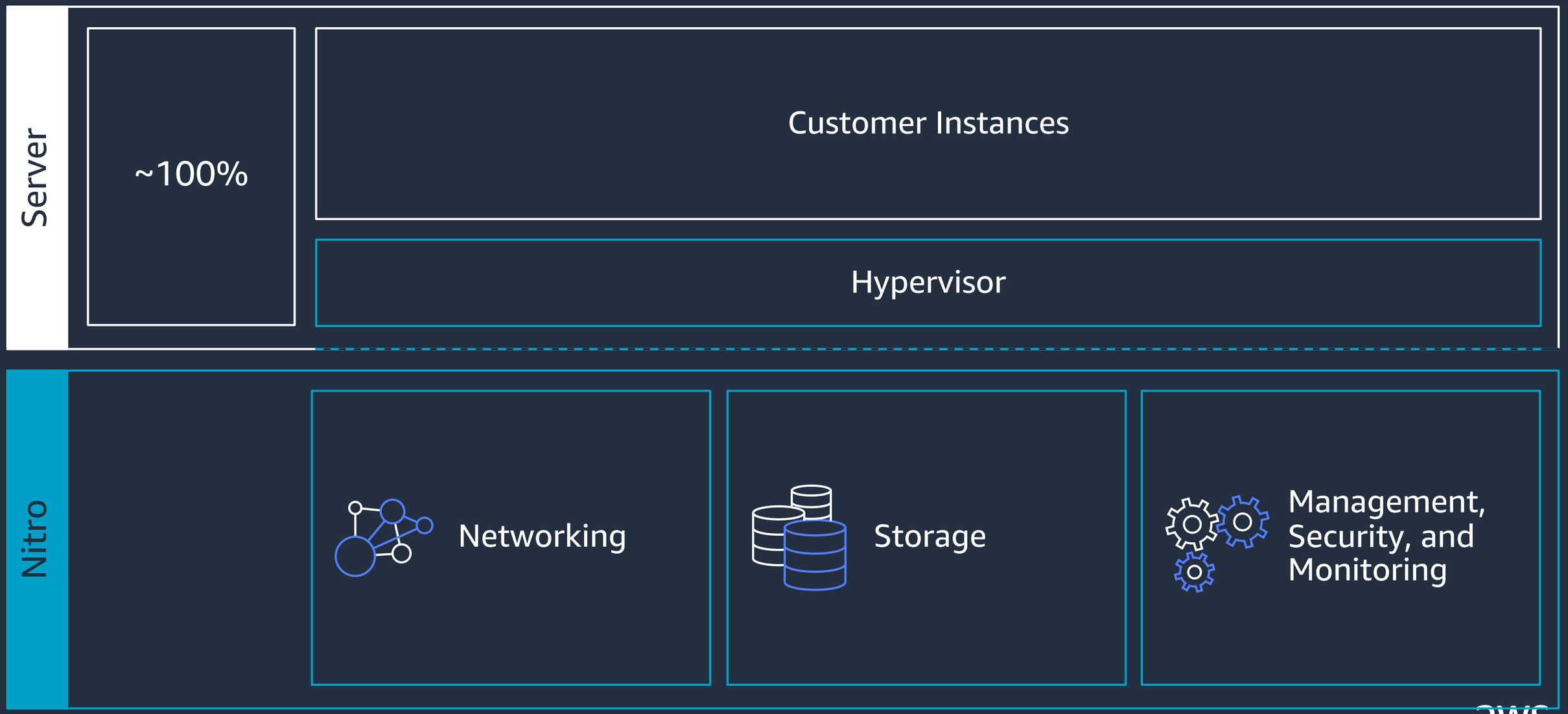
aws

2013 EC2 “Instance” host architecture



aws

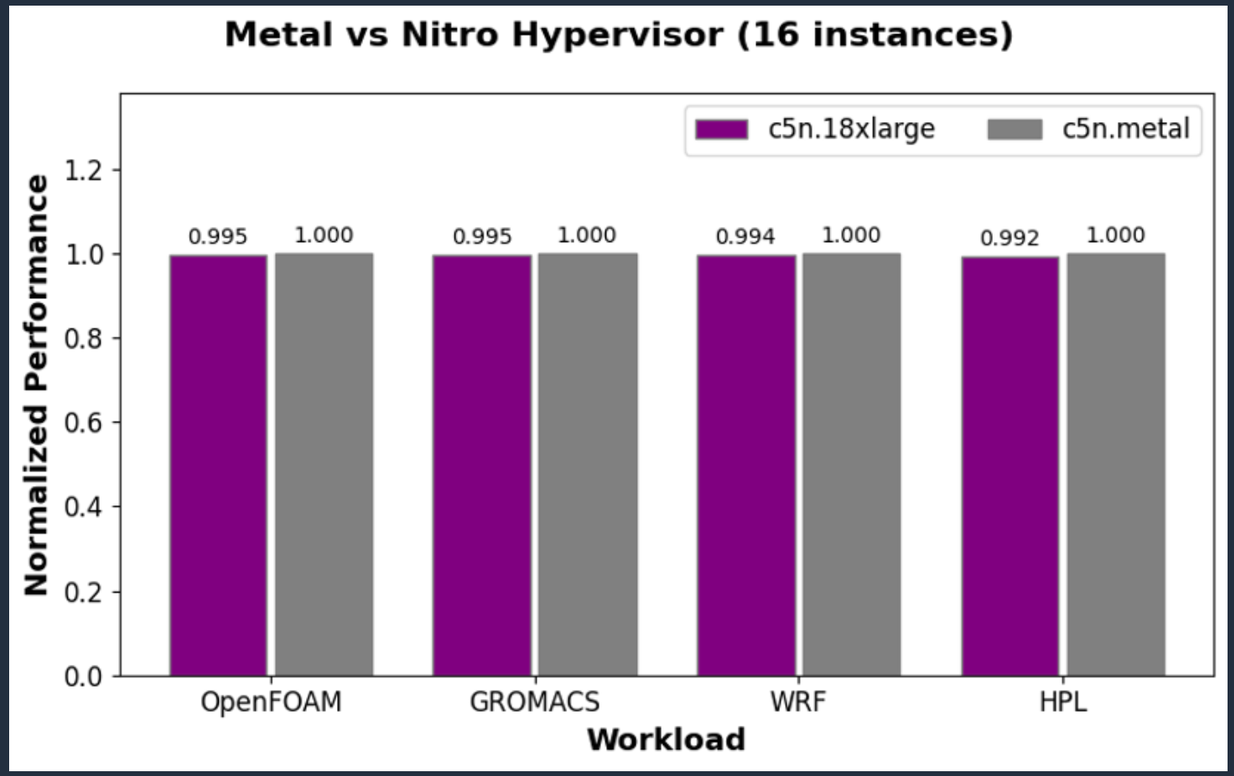
The Nitro Architecture



aws

Metal vs. Nitro Hypervisor Instances on AWS

- AWS offers ".metal" instances, which remove the hypervisor entirely
- The Nitro Hypervisor has minimal overhead in the evaluated HPC applications/benchmarks



https://aws.amazon.com/blogs/hpc/bare-metal-performance-with-the-aws-nitro-system/

AWS Graviton processors

Custom AWS silicon with 64-bit Arm Neoverse cores

Targeted optimizations for cloud-native workloads

Rapidly innovate, build, and iterate on behalf of customers

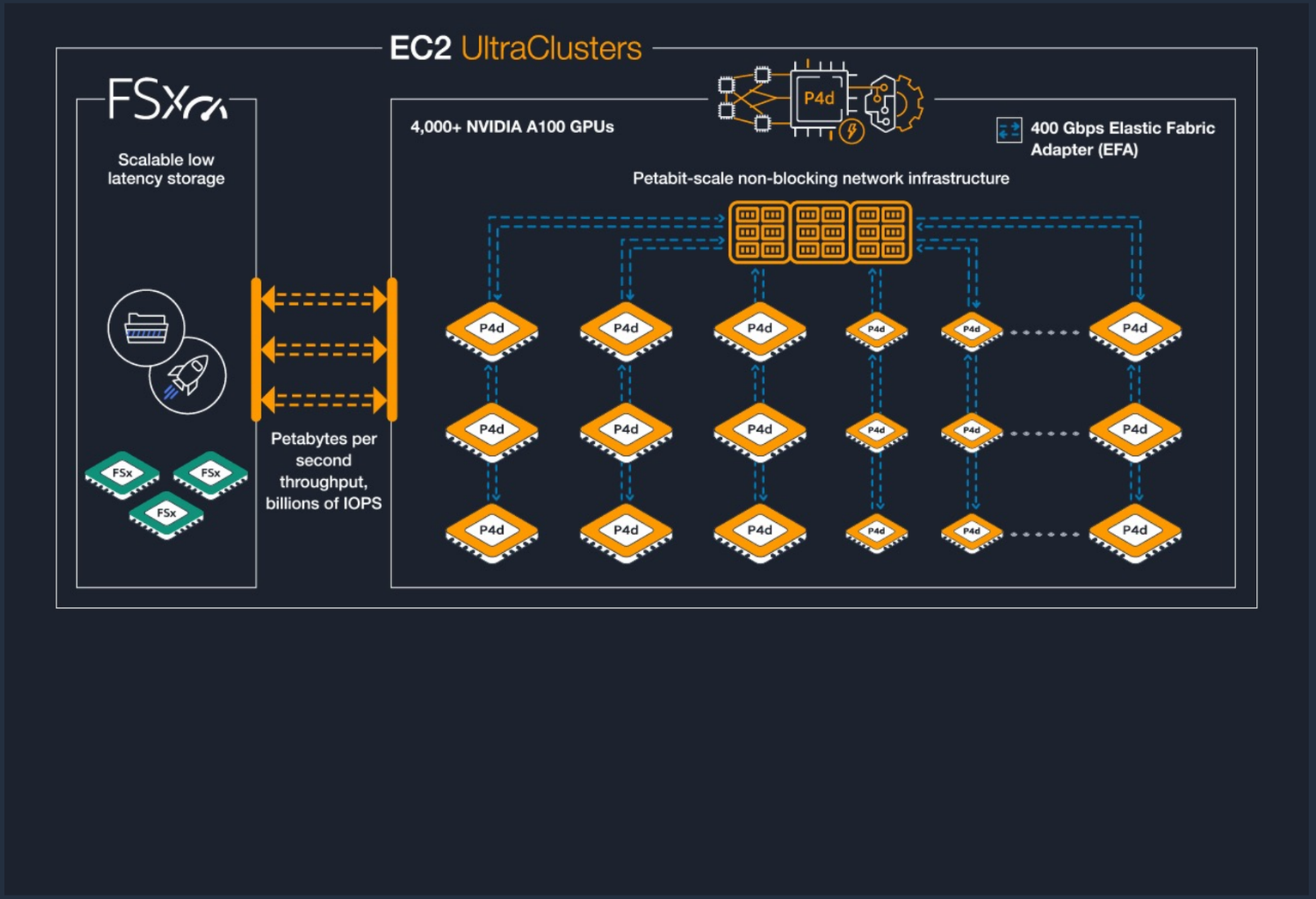
AWS Graviton 2

- 64 Arm® Neoverse™ N1 cores
- Arm v8.2 compliant
- Worked closely with Arm on creation of N1
  - Large 64KB L1 caches and 1MB L2 cache per vCPU
  - Coherent Instruction cache
  - Lower overheads of interrupts, virtualization, and context switching
  - 4-wide front-end with 8-wide dispatch per issue
  - Dual-SIMD units
  - Data types to accelerate ML inference: int8 and fp16
- Every vCPU is a physical core
  - No simultaneous multithreading (SMT)
- No NUMA concerns

aws

EC2 UltraClusters of P4d instances

On-demand access to a world-class supercomputer



Over 4,000 A100 GPUs

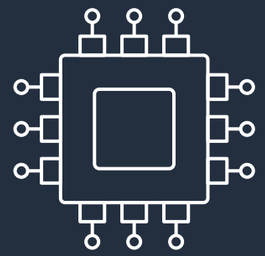
Fully non-blocking petabit-scale network infrastructure

High-throughput, low-latency storage from Amazon FSx for Lustre

aws

Key services and hardware that enable HPC on AWS

Compute



Amazon EC2

Bare metal performance
AWS Graviton 2
EC2 Ultra Clusters

Networking

Elastic Fabric Adapter (EFA)

MVAPICH2-X-AWS is designed for EFA

Storage

Amazon FSx for Lustre

Workflow

AWS ParallelCluster

AWS Batch

Elastic Fabric Adapter (EFA)

Elastic Fabric Adapter

OS bypass

GPUdirect and RDMA

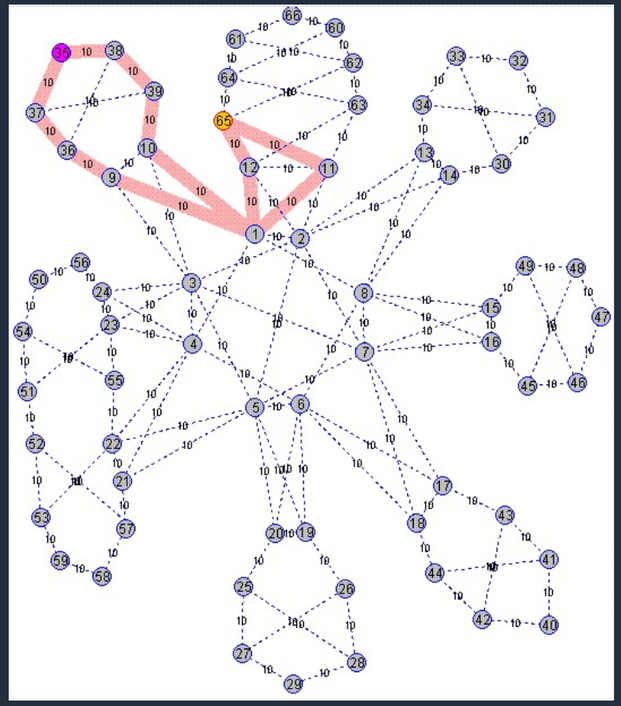
Libfabric core supports wide array of MPIs and NCCL

Scalable Reliable Datagram*

ECMP-enabled packet spraying

Cloud-scale congestion control

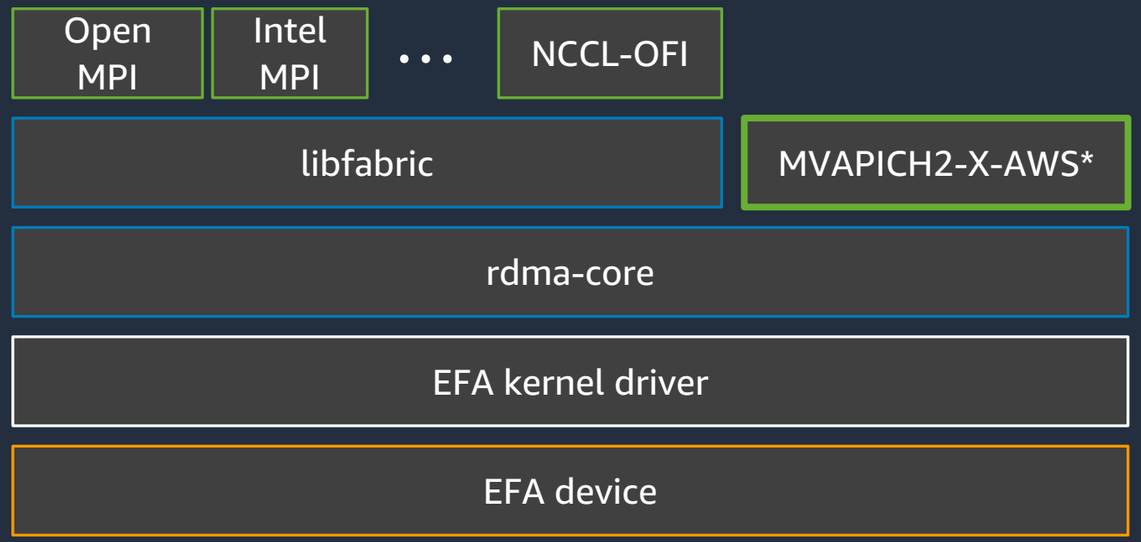
Fast recovery from packet loss or link failure



*L. Shalev, H. Ayoub, N. Bshara and E. Sabbag, "Supercomputing on Nitro in AWS Cloud," in IEEE Micro 2020

Thanks to Wikipedia and Peter Ashwood-Smith for the snappy animated GIF explaining ECMP

EFA Software Stack



- **MVAPICH2-X-AWS** is implemented on rdma-core (verbs)
  - Re-ordering with copy-out
  - Immediate data for sequence IDs
  - Packetization for large messages
- Open MPI, Intel MPI, and NCCL-OFI are implemented using libfabric.
- The EFA libfabric provider implements ordering, packetization, and additional semantic options

*S. Chakraborty, S. Xu, H. Subramoni and D. K. Panda, Designing Scalable and High-Performance MPI Libraries on Amazon Elastic Adapter, Hot Interconnect, 2019

Wide Variety of Instance Types Available with EFA

Graviton2

C6gn.16xlarge
1S, 64c, 2GB/core
100Gb

Storage-Dense

i3en.{24xlarge,metal}
60 TB NVMe
100Gb

x86

"Ice Lake"
m6i.32xlarge
2Sx 32c, 8GB/core
50Gb

"Cascade Lake"
m5n.{24xlarge,metal}
2Sx 24c, 8GB/core
100Gb

"Cascade Lake"
r5n.{24xlarge,metal}
2Sx 24c, 8GB/core
100Gb

"Skylake"
c5n.{18xl,metal}
2Sx 18c, ~4GB/core
100Gb

"Cascade Lake"
m5zn.24xlarge (4.5GHz)
2Sx 12c, 8GB/core
100Gb

+ 'd' variants with NVMe (r5dn, m5dn)

Accelerator

p4d.24xlarge
8x A100 GPUs
400Gb

p3dn.24xlarge
8x V100 GPUs
100Gb

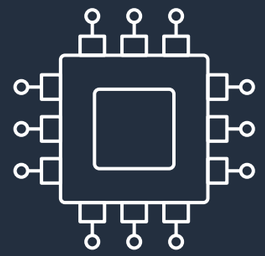
inf1.24xlarge
16 Inferentia chips
100Gb

G4dn.{16xlarge, metal}
8x T4 GPUs
100Gb

aws

Key services and hardware that enable HPC on AWS

Compute



Amazon EC2

Bare metal performance
AWS Graviton 2
EC2 Ultra Clusters

Networking

Elastic Fabric Adapter (EFA)

MVAPICH2-X-AWS is designed for EFA

Storage

Amazon FSx for Lustre

Workflow

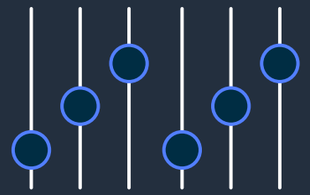
AWS ParallelCluster

AWS Batch

Amazon FSx for Lustre

High and scalable performance

Parallel file system

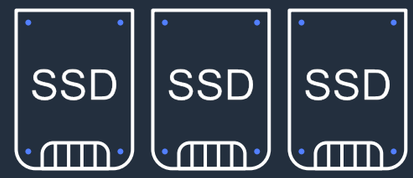


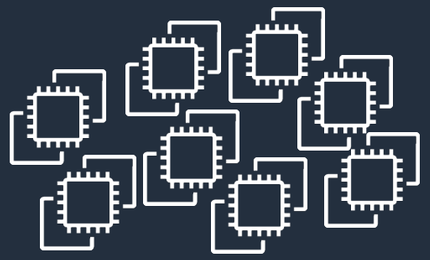
+100 GiB/s throughput

Millions of IOPS

Consistent submillisecond latencies

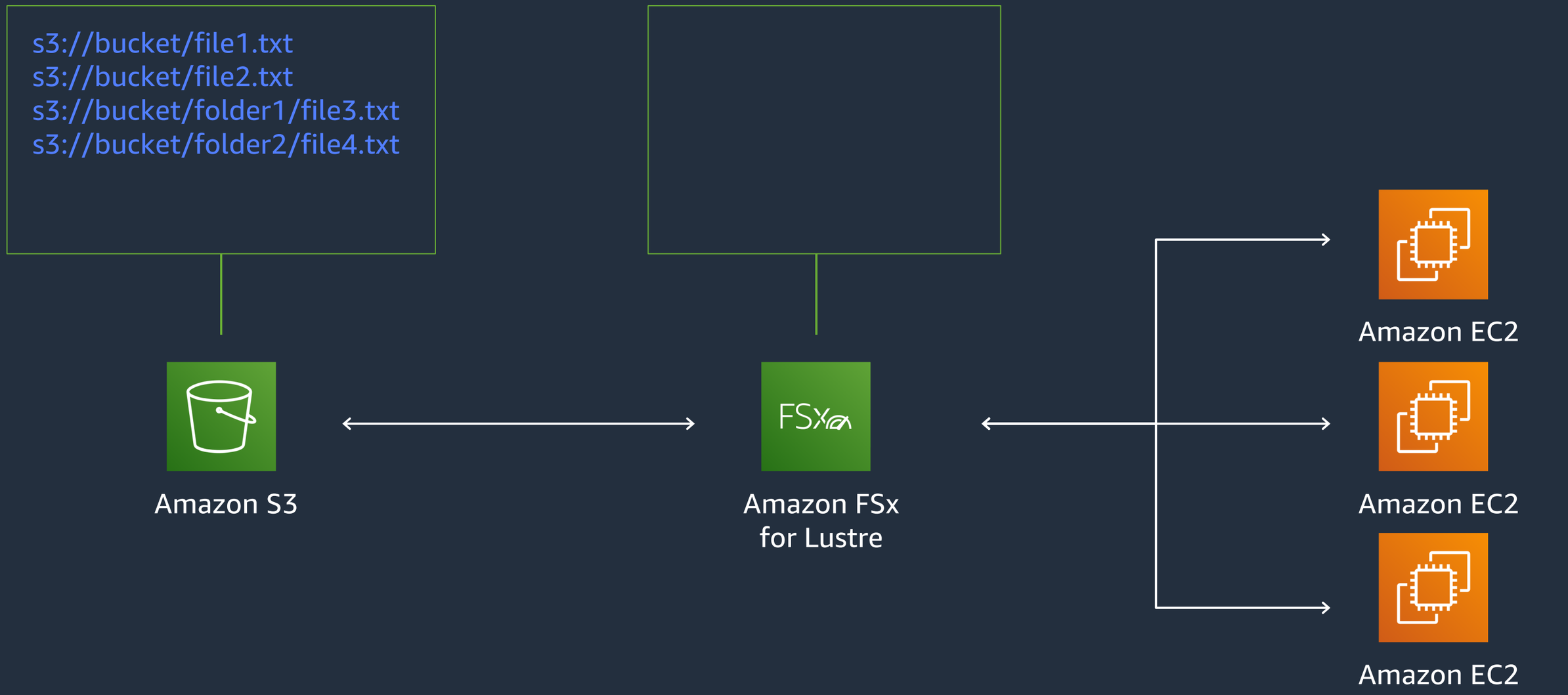
SSD-based



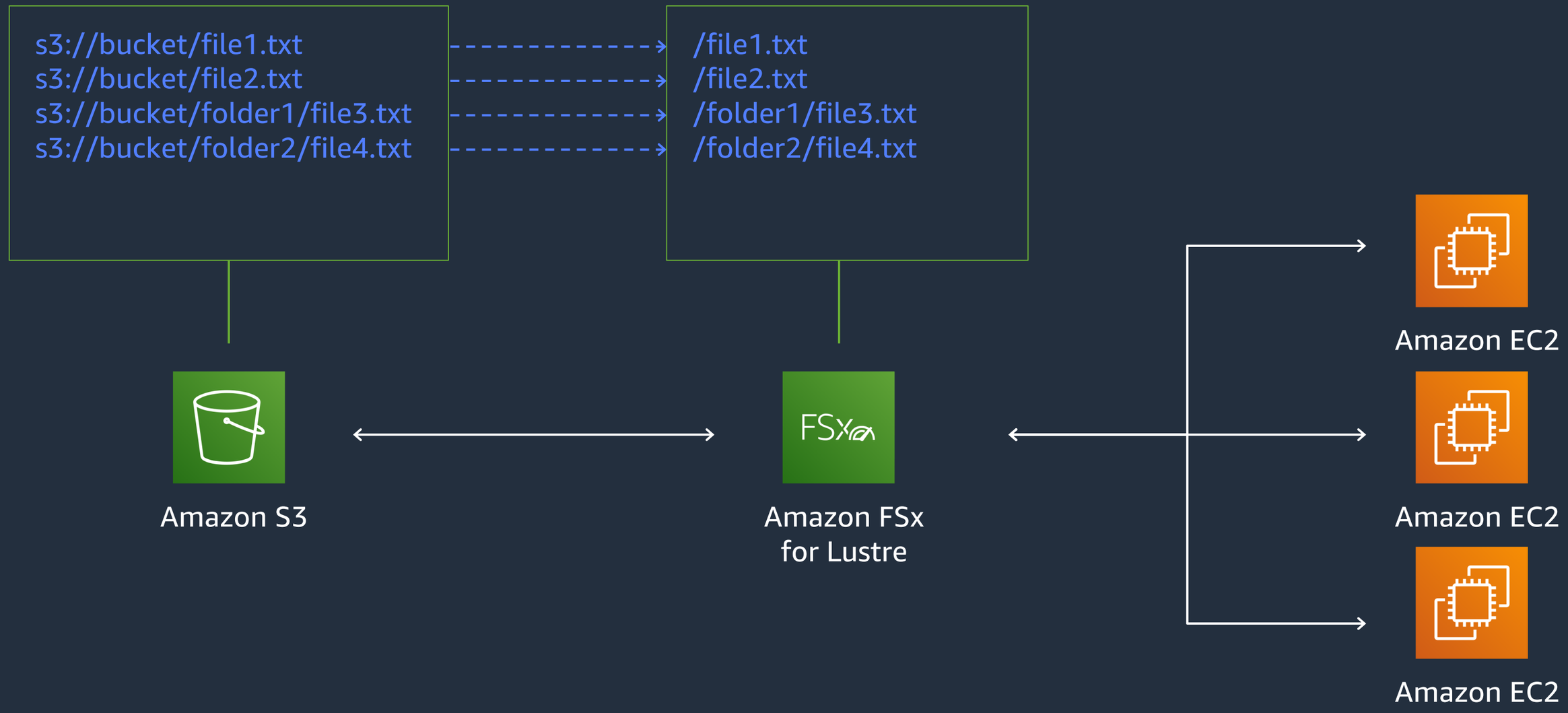


Supports concurrent access from hundreds of thousands of cores

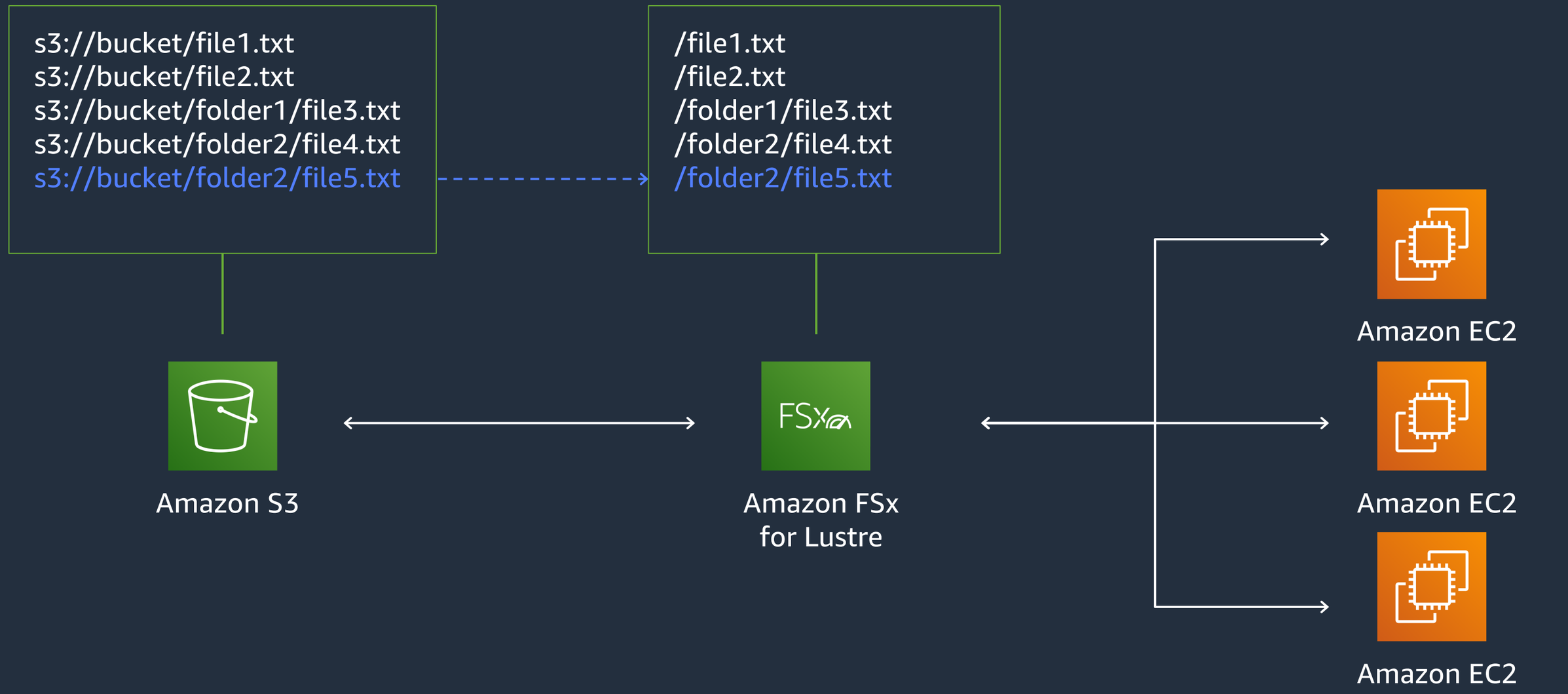
S3-linked file systems : Objects stored S3 appear on FSx file systems



S3-linked file systems : Objects stored S3 appear on FSx file systems



S3-linked file systems : Objects stored S3 appear on FSx file systems



S3-linked file systems : New files can be exported from FSx to S3

s3://bucket/file1.txt
s3://bucket/file2.txt
s3://bucket/folder1/file3.txt
s3://bucket/folder2/file4.txt
s3://bucket/folder2/file5.txt

/file1.txt
/file2.txt
/folder1/file3.txt (edit)
/folder2/file4.txt
/folder2/file5.txt
/folder2/file6.txt (new)

Amazon S3

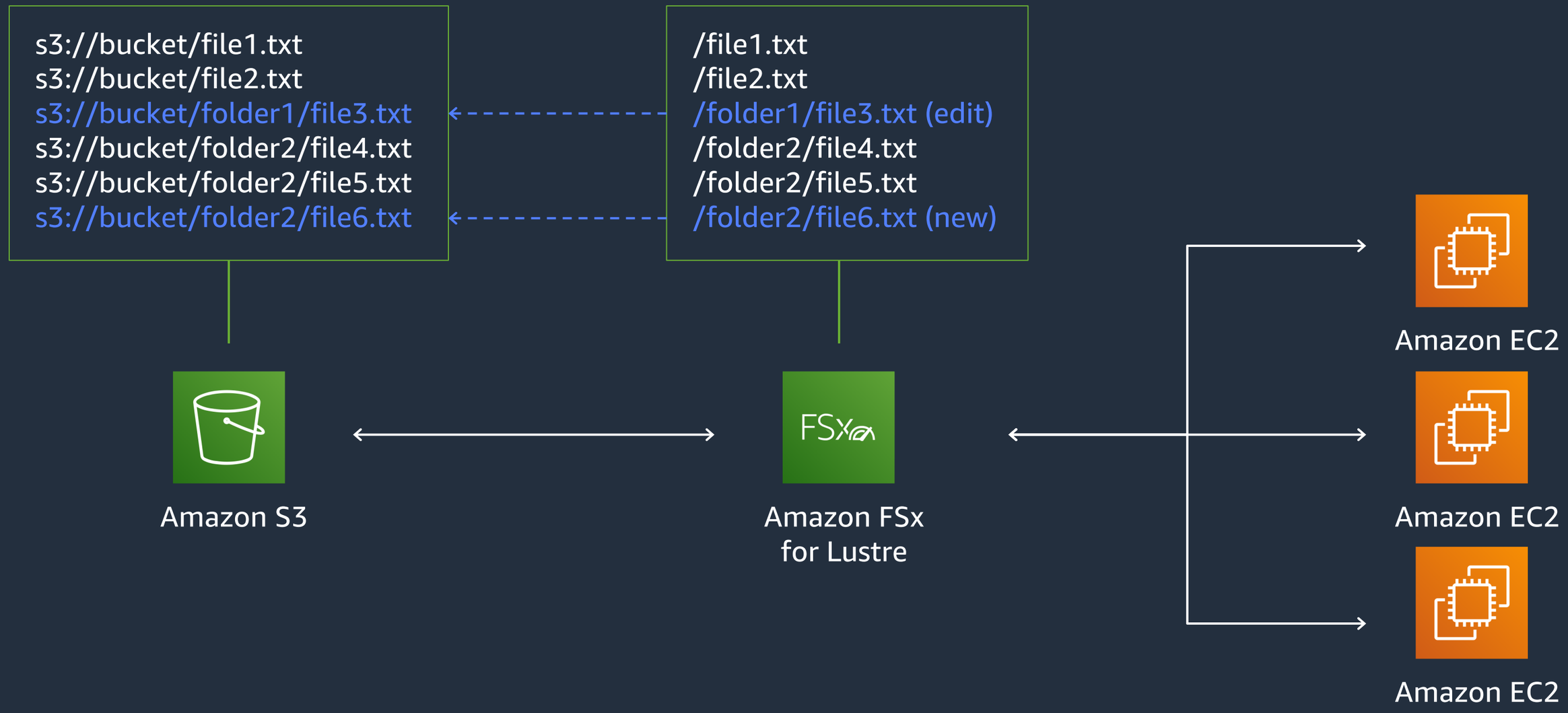
Amazon FSx
for Lustre

Amazon EC2

Amazon EC2

Amazon EC2

S3-linked file systems : New files can be exported from FSx to S3



S3-linked file systems : Spin up / spin down with compute resources

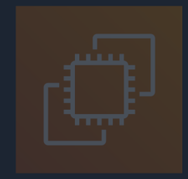
s3://bucket/file1.txt
s3://bucket/file2.txt
s3://bucket/folder1/file3.txt
s3://bucket/folder2/file4.txt
s3://bucket/folder2/file5.txt
s3://bucket/folder2/file6.txt

Amazon S3

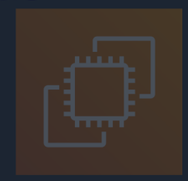
/file1.txt
/file2.txt
/folder1/file3.txt (edit)
/folder2/file4.txt
/folder2/file5.txt
/folder2/file6.txt (new)

Spin down resources between workloads

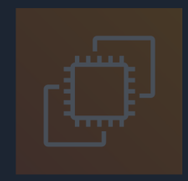
Amazon FSx
for Lustre



Amazon EC2



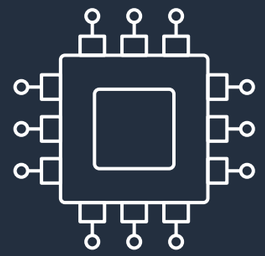
Amazon EC2



Amazon EC2

Key services and hardware that enable HPC on AWS

Compute



Amazon EC2

Bare metal performance
AWS Graviton 2
EC2 Ultra Clusters

Networking

Elastic Fabric Adapter (EFA)

MVAPICH2-X-AWS is designed for EFA

Storage

Amazon FSx for Lustre

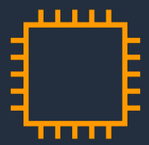
Workflow

AWS ParallelCluster

AWS Batch

AWS ParallelCluster

Set one configuration file with parameters and AWS ParallelCluster will provision an elastic HPC cluster on AWS for you



Login Node
(place to submit jobs, edit files, etc.)

+

Scheduler (e.g. Slurm)

Compute Nodes
Scalable set of compute resources (increases and decreases with workload in queue)

Shared Filesystem
Filesystem mounted on login node and all compute nodes

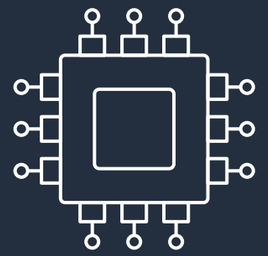
Make Viz Easy on login node

+ Add your own changes in either a base image, post-install script, or interactively (e.g., add MVAPICH2-X-AWS, specific tuning, application setup)

Let’s try it!

Key services and hardware that enable HPC on AWS

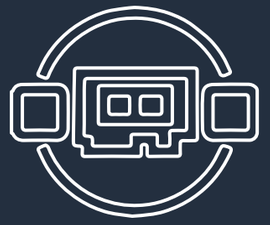
Compute



Amazon EC2

Bare metal performance
AWS Graviton 2
EC2 Ultra Clusters

Networking



Elastic Fabric Adapter (EFA)

MVAPICH2-X-AWS is designed for EFA

Storage

Amazon FSx for Lustre

Workflow

AWS ParallelCluster

AWS Batch

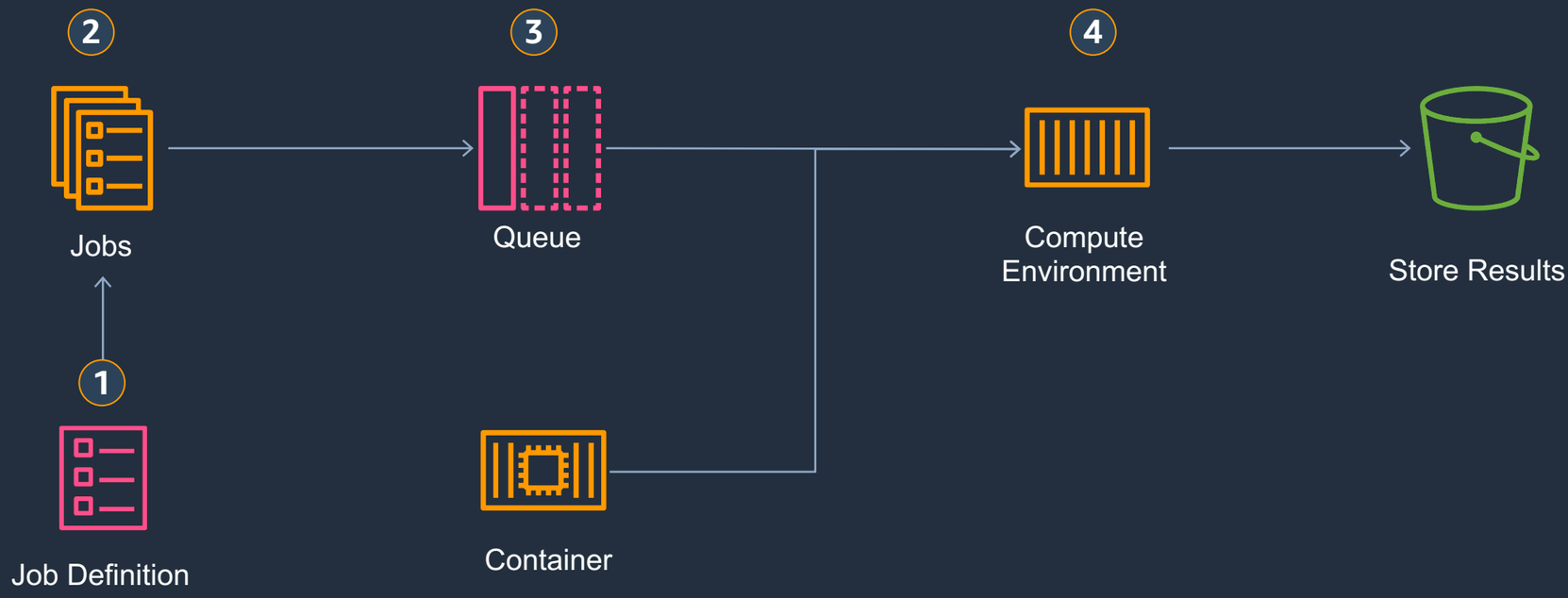
AWS Batch Overview

1. **Job Definition**
   Template that has common attributes (container image, IAM role, vCPU & memory requirements, ,…

2. **Job**
   Each job must reference a job definition, but many parameters may be overridden when submited

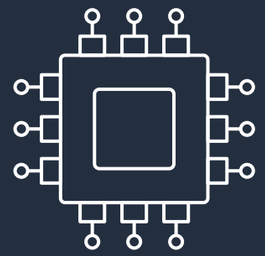
3. **Job Queue (JQ)**
   Queue determines priorities. Each JQ is connected to 1 or more CE

4. **Compute Environment (CE)**
   Resource Mix (defines On-demand vs. Spot and instance types. CE can be connected to more than one JQ



Key services and hardware that enable HPC on AWS

Compute



Amazon EC2

Networking

Elastic Fabric Adapter (EFA)

Storage

Amazon FSx for Lustre

Workflow

AWS ParallelCluster

AWS Batch

We're hiring LOTS of HPC Software Developers (and managers) in **Seattle and Asti (in Italy)** and will offer **relocation packages** to the right candidates as well as attractive comp and an incredibly fun work environment.

`http://hpc.news/jobs`

WE'RE HIRING … to change the world(s).

Thank you!

mkoop@amazon.com