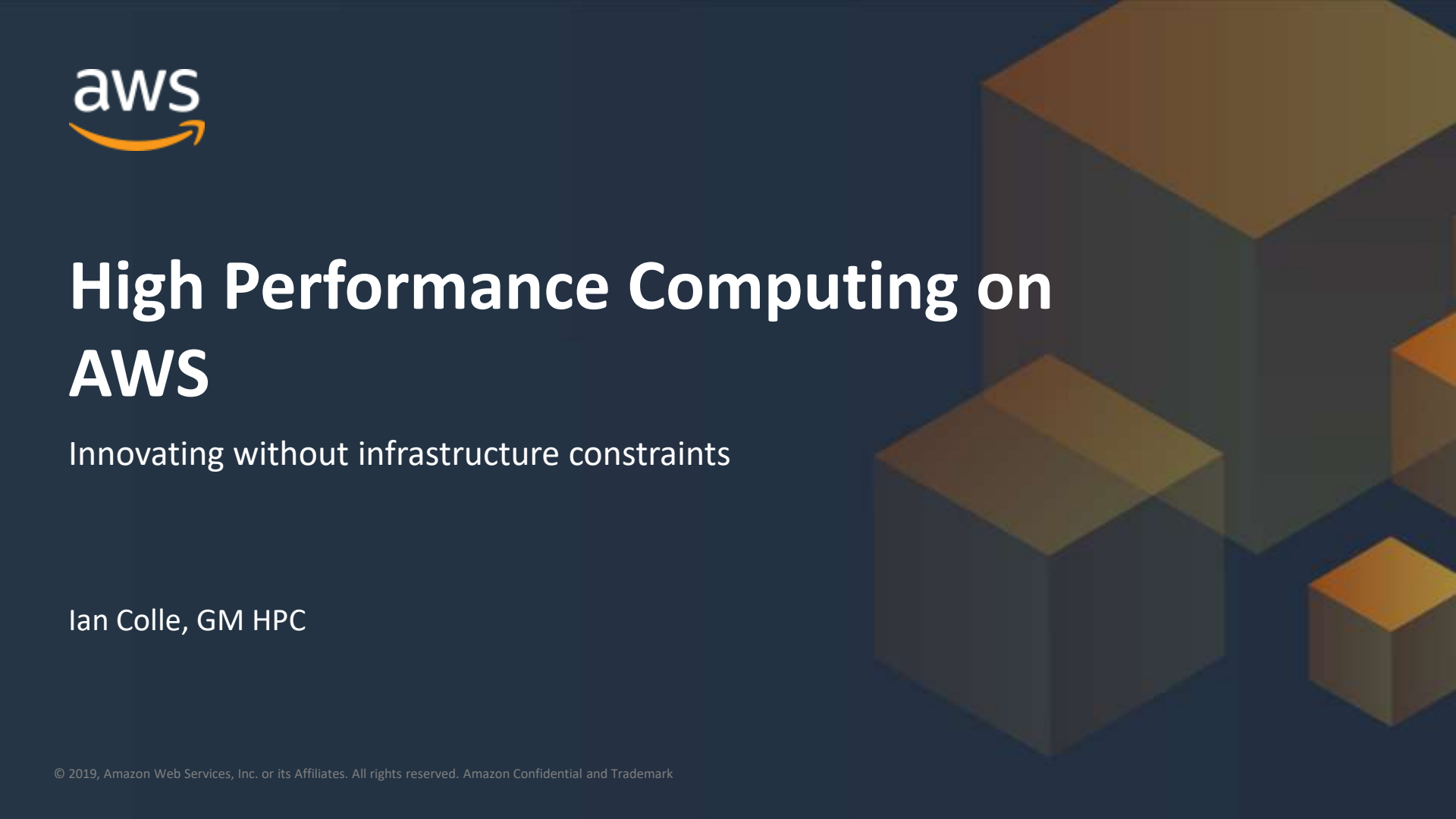
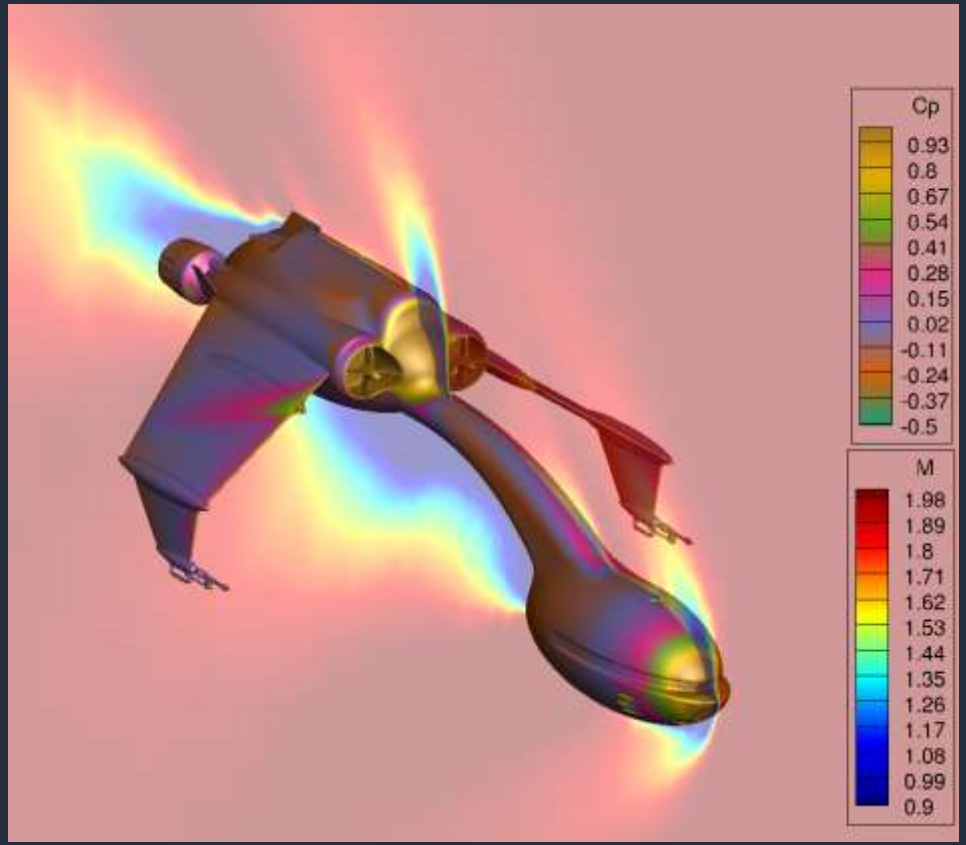


High Performance Computing on AWS

Innovating without infrastructure constraints

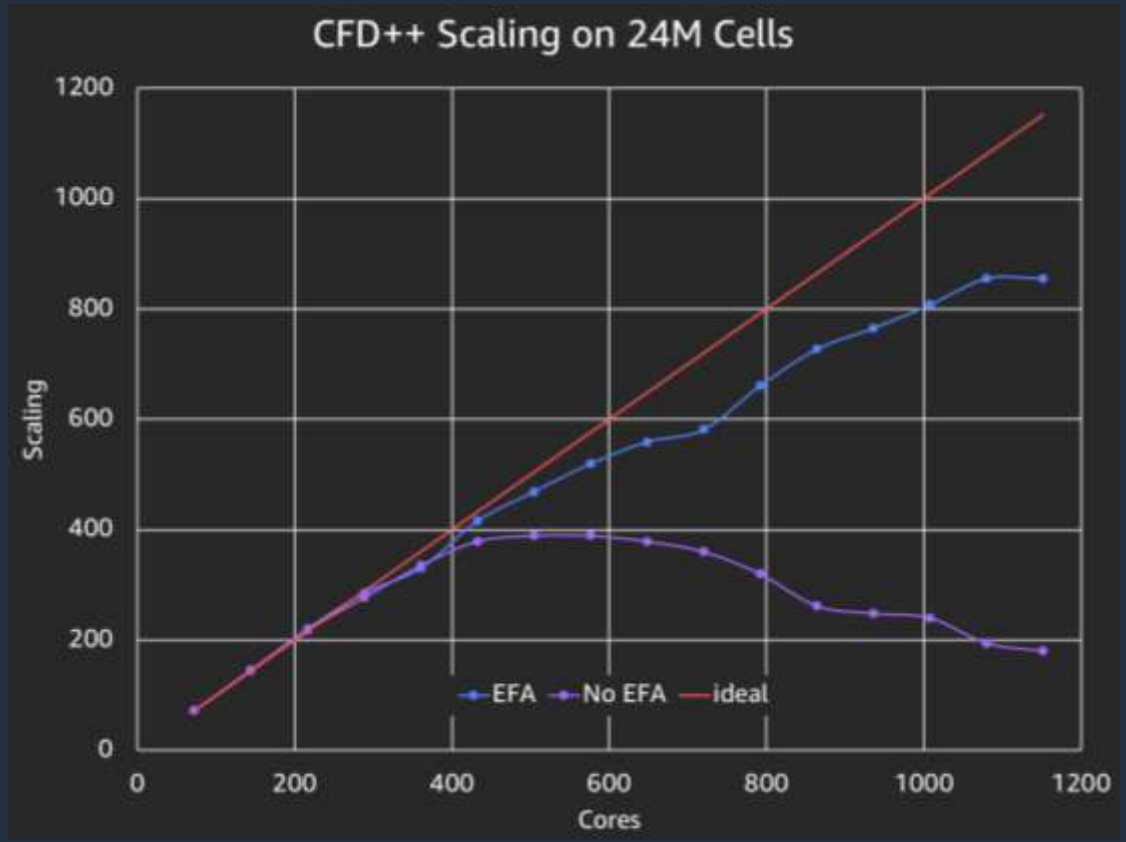
Ian Colle, GM HPC

What can HPC on AWS do?



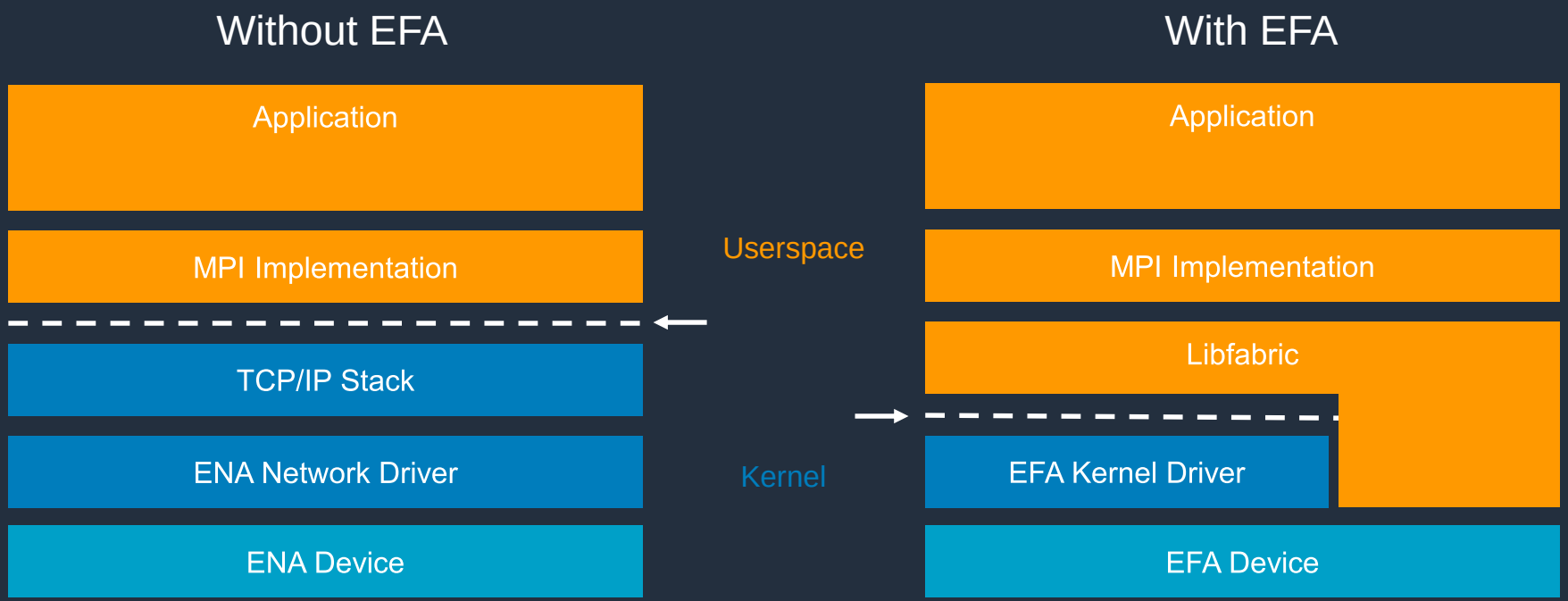
Thanks to Klingon Empire.

Can HPC Scale in the Cloud?



Thanks to Metacomp Technologies

HPC software stack in Amazon EC2



Enhanced Network for HPC and machine learning

Up to 100 Gbps network bandwidth

C5n

Most elastic and scalable HPC network

Custom Intel® Xeon® Scalable processor

P3dn

Fastest machine learning training in the cloud

NVIDIA V100 Tensor Core GPUs

Elastic Fabric Adapter for HPC

Best for scaling large HPC and ML workloads

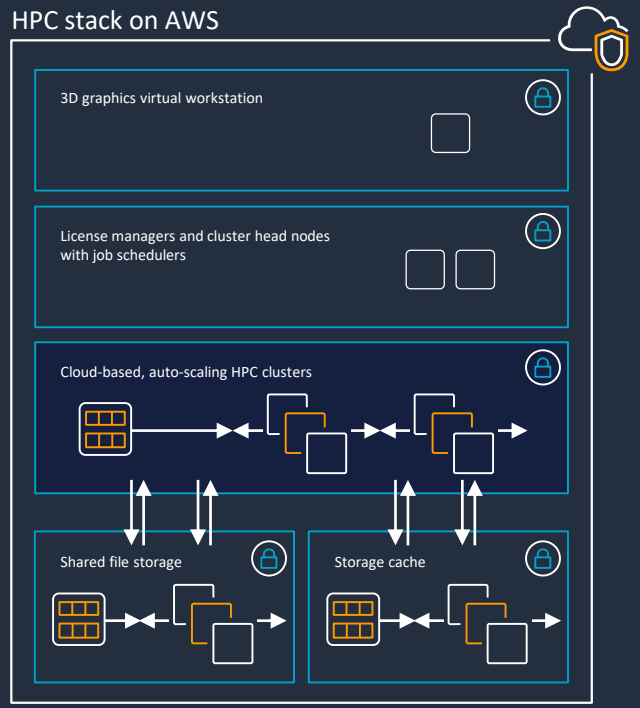
aws

High bandwidth compute instances: C5n

Massively scalable performance

- C5n Instances offer up to 100 Gbps of network bandwidth
- Up to 72 3.5GHz vCPUs / 192GiB Memory
- Significant improvements in maximum bandwidth, packet per seconds, and packets processing
- Custom designed Nitro network cards
- Purpose-built to run network bound workloads including distributed cluster and database workloads, HPC, real-time communications and video streaming

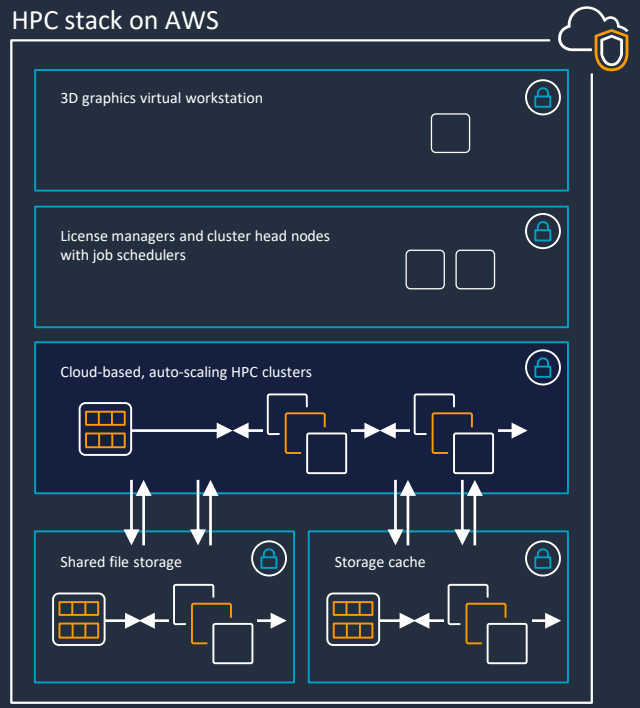
Featuring Intel Xeon Scalable (Skylake) processor



High bandwidth compute instances: P3dn

Optimized for distributed ML training

- One of the most powerful GPU instance available in the cloud
- Based on NVIDIA's Tesla V100 with 32GB of memory each
- Up to 8 GPUs / 256 GB GPU Memory, 96 vCPUs / 768 GB Mem
- Distributed machine learning training across multiple GPU instances
- 100 Gbps of networking throughput



Broadest and deepest platform choice

Categories

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

+

Capabilities

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 12 TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare Metal

Size
(Nano to 32xlarge)

+

Options

Elastic Block Store

Elastic Graphics

Elastic Inference

=

200+

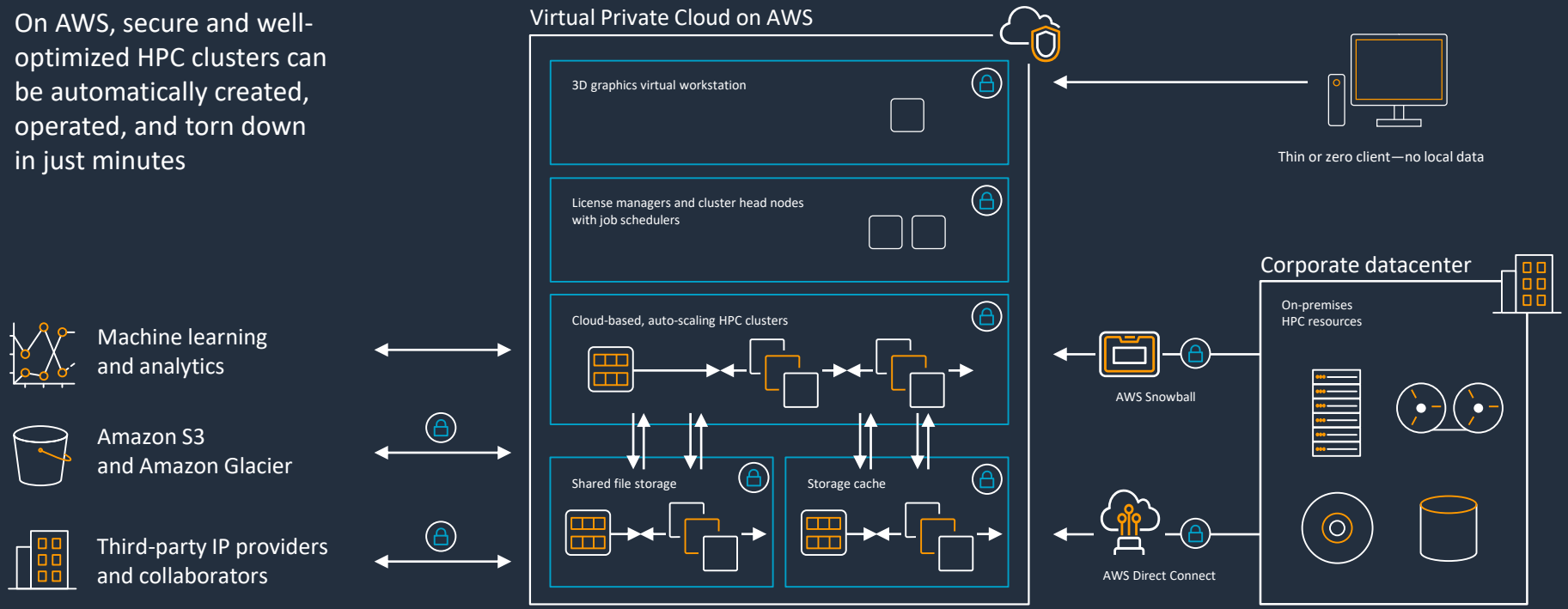
instance types

for virtually every workload and business need

aws

High Performance Computing (HPC) on AWS

On AWS, secure and well-optimized HPC clusters can be automatically created, operated, and torn down in just minutes



Comprehensive portfolio of high performance storage options

Block storage

Amazon EBS

Elastic, high performance block storage at any scale

File storage

Amazon EFS

Petabyte-scale, elastic file storage sharable across applications, instances and servers

Object storage

Amazon S3

Low cost, highly scalable cloud storage with 99.999999999% durability

Fully managed high performance shared file system: Amazon FSx for Lustre

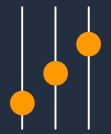
Massively scalable performance

100+ GiB/s throughput

Millions of IOPS

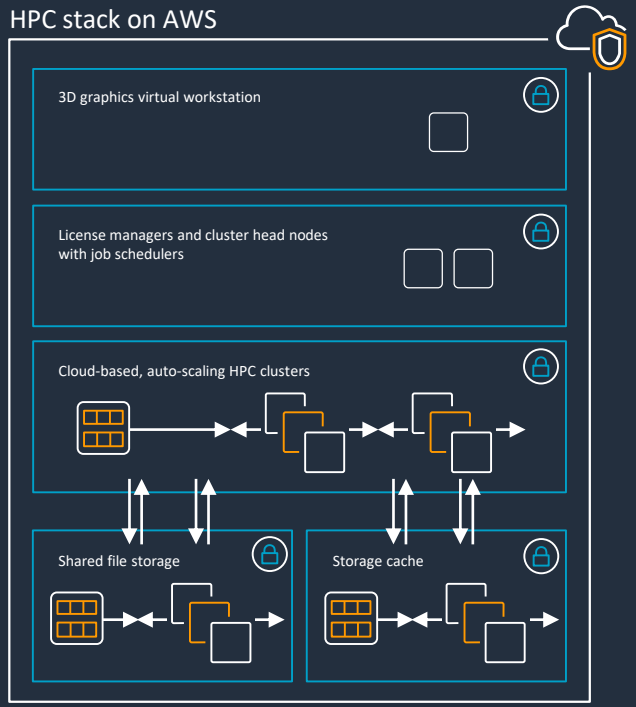
Consistent low latencies

High performing



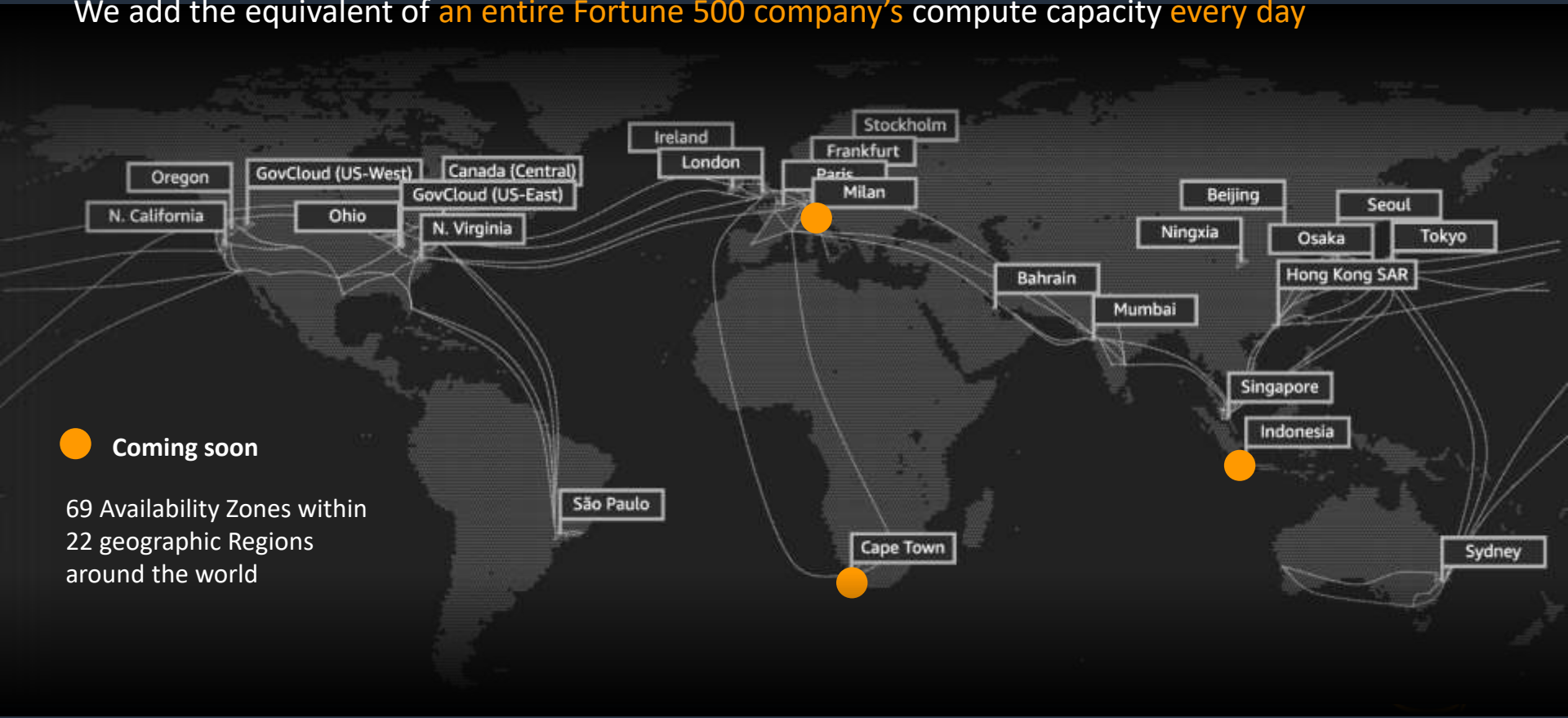
Parallel distributed file system

Tune complex performance parameters



Global Infrastructure

We add the equivalent of an entire Fortune 500 company's compute capacity every day



69 Availability Zones within
22 geographic Regions
around the world

Compliance programs

Global

SOC 1


SOC 2

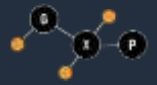

SOC 3


Europe


Asia Pacific

United States








aws

HPC workloads across industries

Life Sciences

Financial Services

Oil & Gas

Design & Engineering

Climate & Geosciences

Autonomous Vehicles

aws

HPC workloads with different compute and throughput characteristics

Tightly-coupled workloads

Loosely-coupled workloads

Accelerated computing

Visualization

AI/ML

High volume data analytics

Top500…In The Cloud?

Performance: #136 – Rmax 1,926.4 Tflop/s

Cores: 41,472 3GHz Skylake Processors

Memory: 157,824 GB

Cost: ~$5,000

"We didn't ask Amazon to give our engineers any special dispensation, discount, or custom planning or setup. We wanted to see if we could do this on our own, which if completed successfully, would also be a testament to the self-service model of AWS. Our team merely followed the standard steps and charged it to the company credit card. The potential for democratization of HPC was palpable since the cost to run custom hardware at that speed is probably closer to $20 to $30 million. Not to mention a 6–12 month wait time."

—Mike Warren, CTO, Descartes Labs

Descartes Labs

aws

Running HPC applications at extreme scale

Accelerating time to innovation

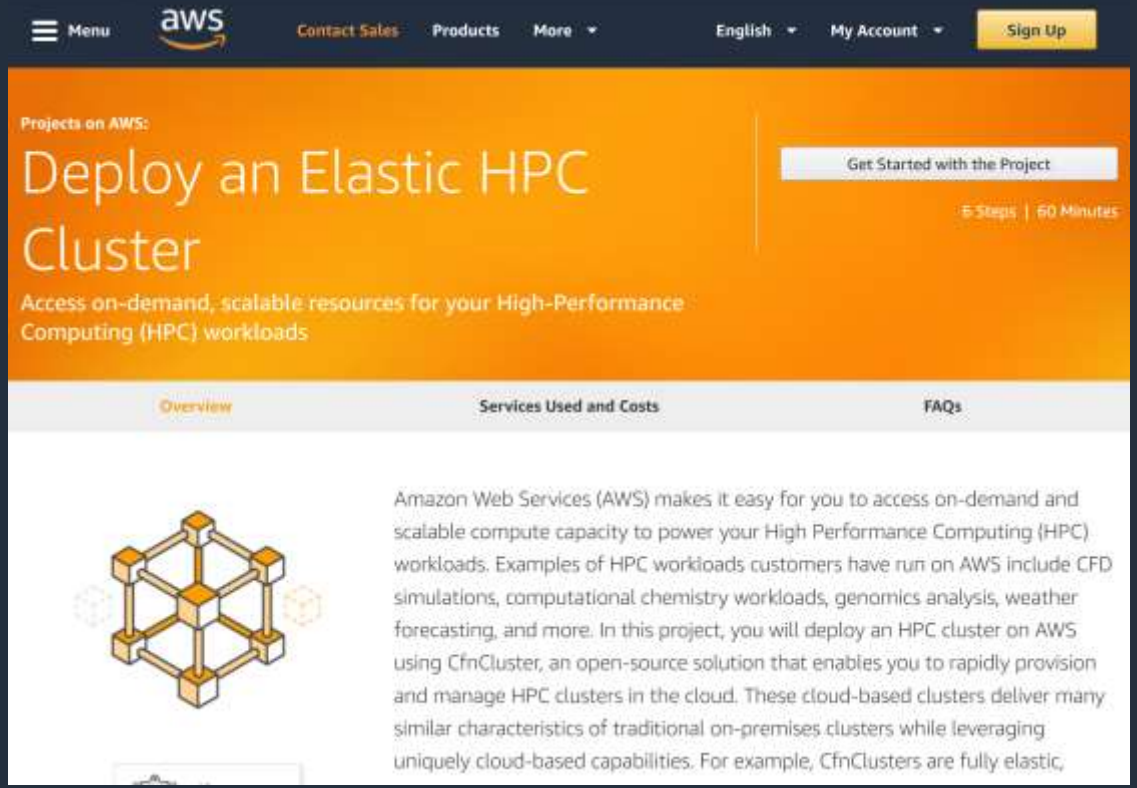
Over 2.3 million simulation jobs on a single HPC cluster of 1 million vCPUs
—built using Amazon EC2 Spot Instances

Time to results: 20 days → 8 hours

"Storage technology is amazingly complex and we're constantly pushing the limits of physics and engineering to deliver next-generation capacities and technical innovation. This successful collaboration with AWS shows the extreme scale, power and agility of cloud-based HPC to help us run complex simulations for future storage architecture analysis and materials science explorations. Using AWS to easily shrink simulation time from 20 days to 8 hours allows Western Digital R&D teams to explore new designs and innovations at a pace un-imaginable just a short time ago." —Steve Phillpott, CIO, Western Digital

aws

https://aws.amazon.com/getting-started/projects/deploy-elastic-hpc-cluster/



aws