

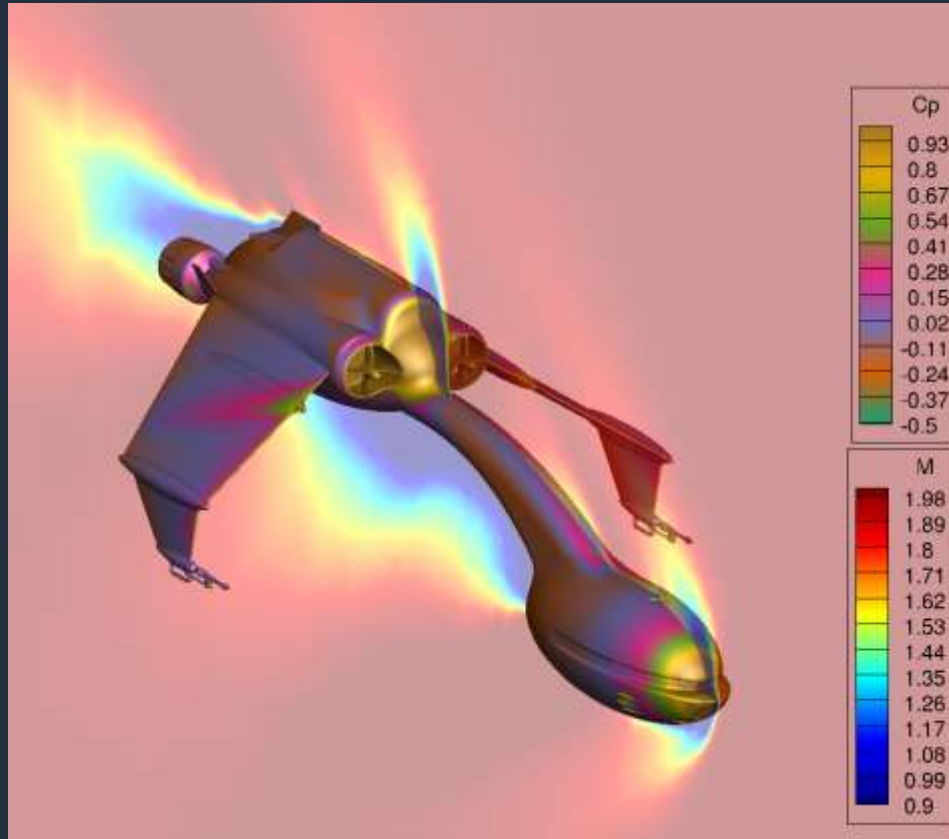


High Performance Computing on AWS

Innovating without infrastructure constraints

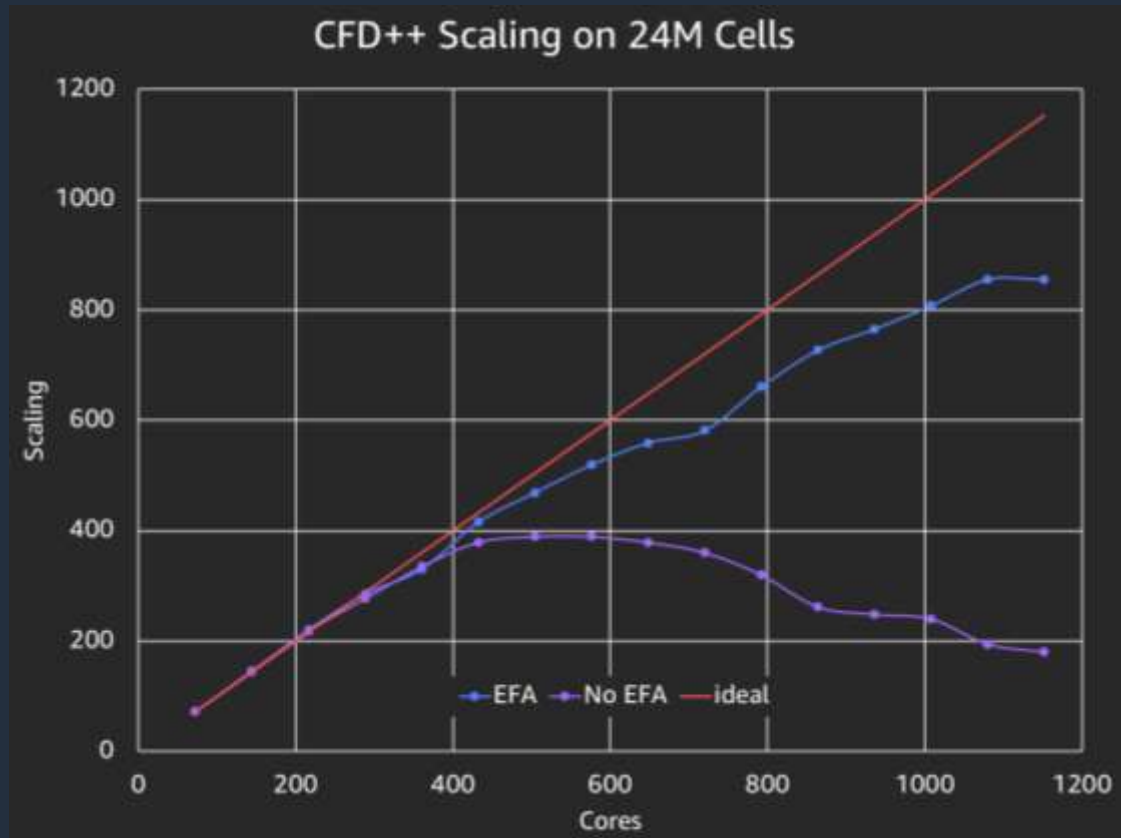
Ian Colle, GM HPC

What can HPC on AWS do?



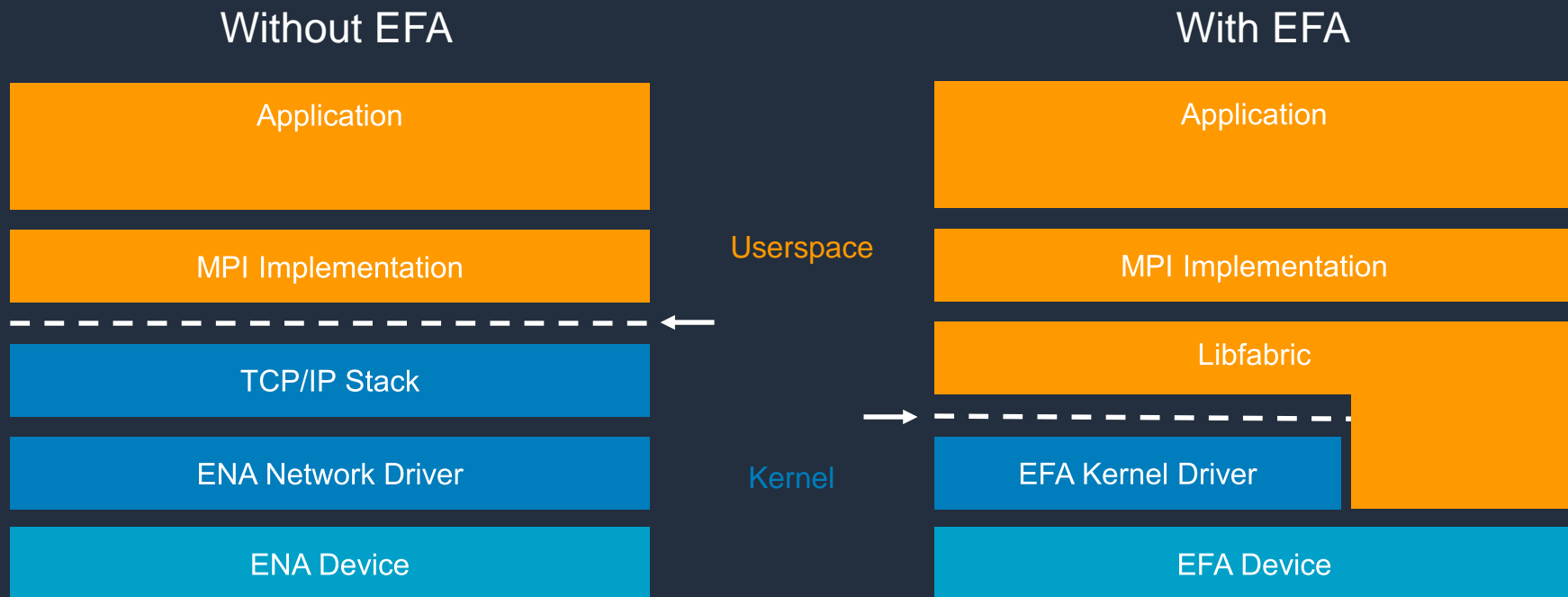
Thanks to Klingon Empire.

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

High bandwidth compute instances: C5n

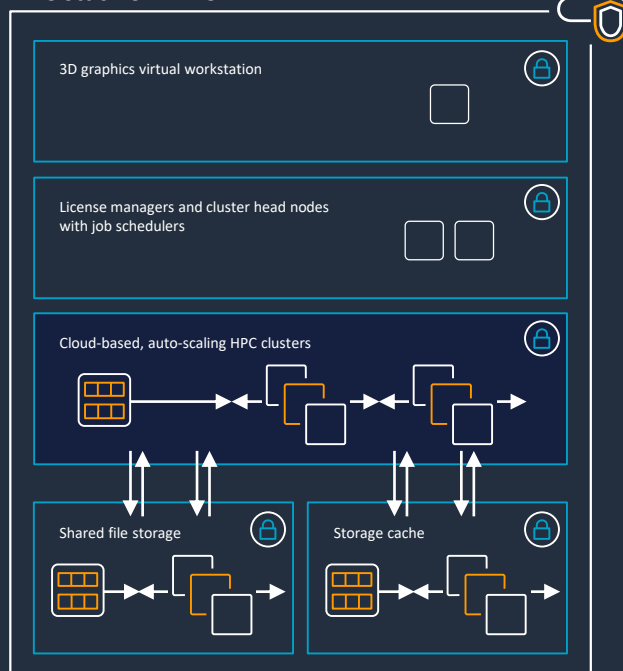
Massively scalable performance

- 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



HPC stack on AWS

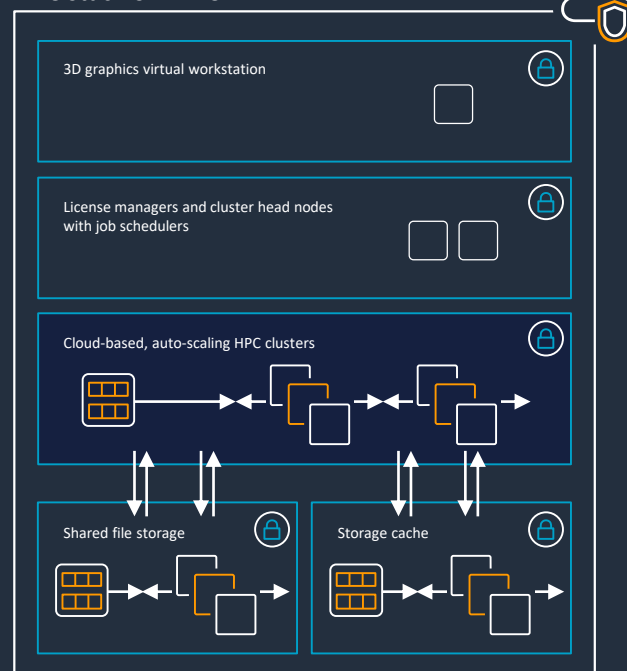


High bandwidth compute instances: P3dn

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

HPC stack on AWS



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



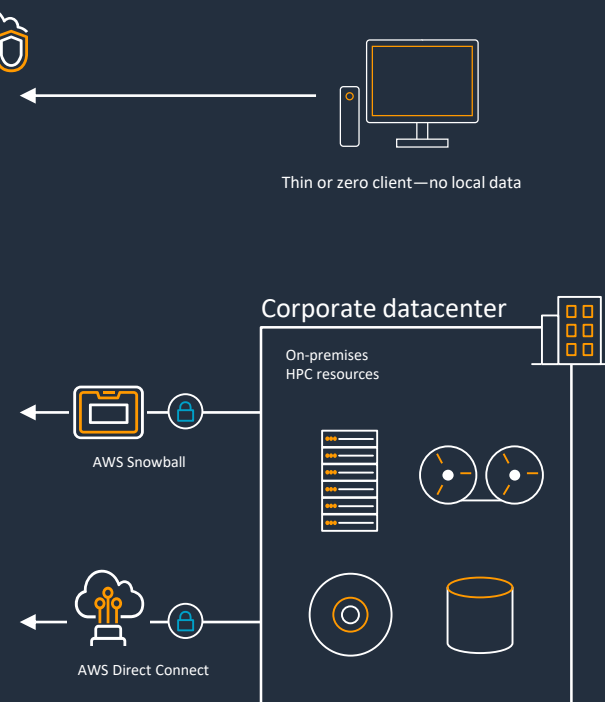
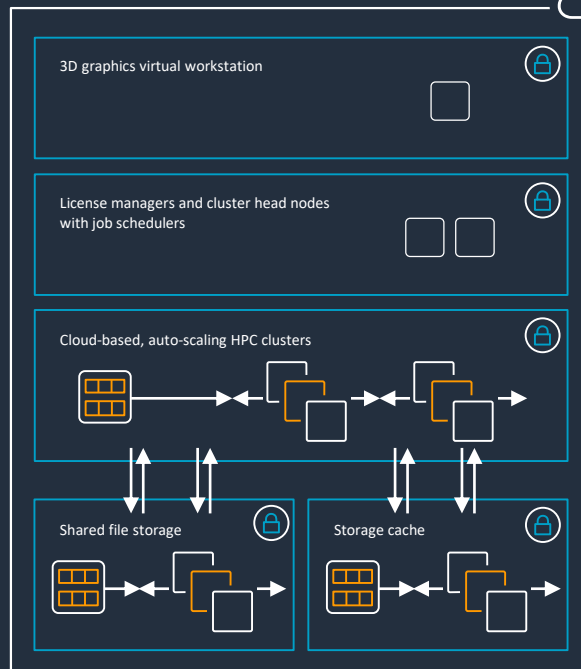
200+
instance types
for virtually
every workload
and business need

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

-  Machine learning and analytics
-  Amazon S3 and Amazon Glacier
-  Third-party IP providers and collaborators

Virtual Private Cloud on AWS



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
storage with
99.999999999% durability

cloud

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

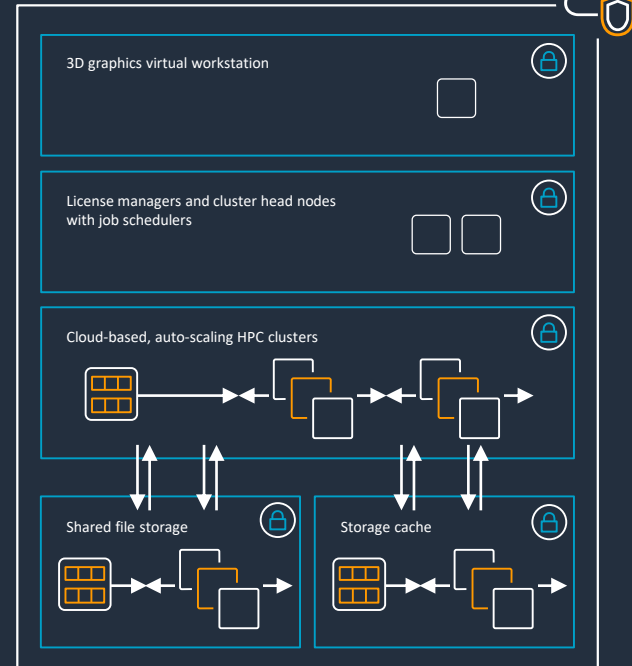


Parallel distributed file system



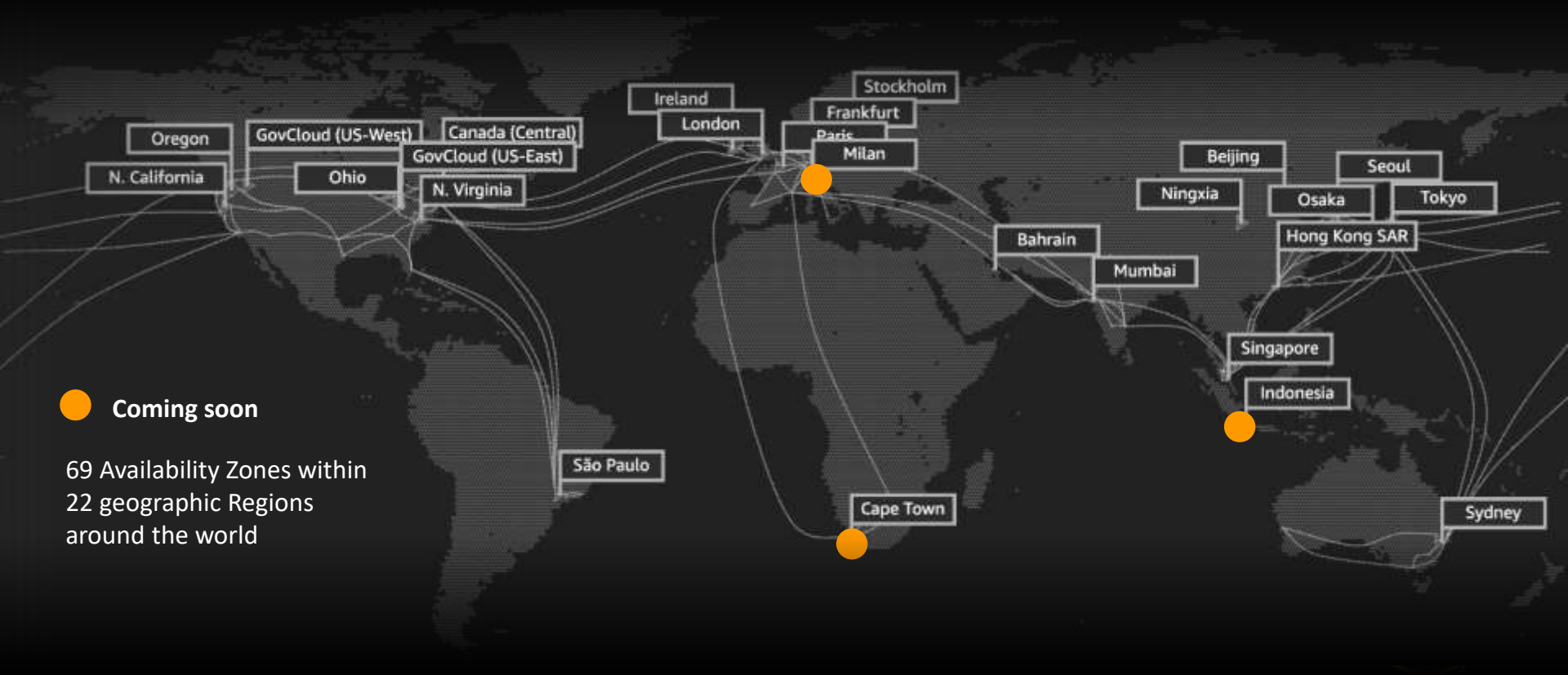
Tune complex performance parameters

HPC stack on AWS



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



PARTICIPATING ORGANIZATION

Europe



Asia Pacific



United States



HPC workloads across industries



Life Sciences



Financial Services



Oil & Gas



Design & Engineering



Climate & Geosciences



Autonomous Vehicles

HPC workloads with different compute and throughput characteristics

VOLKSWAGEN
GROUP

Tightly-coupled workloads

illumina®

Loosely-coupled workloads

SCHRÖDINGER.

Accelerated computing

mlk

Visualization



AI/ML

DigitalGlobe®

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



**Descartes
Labs**

“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

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

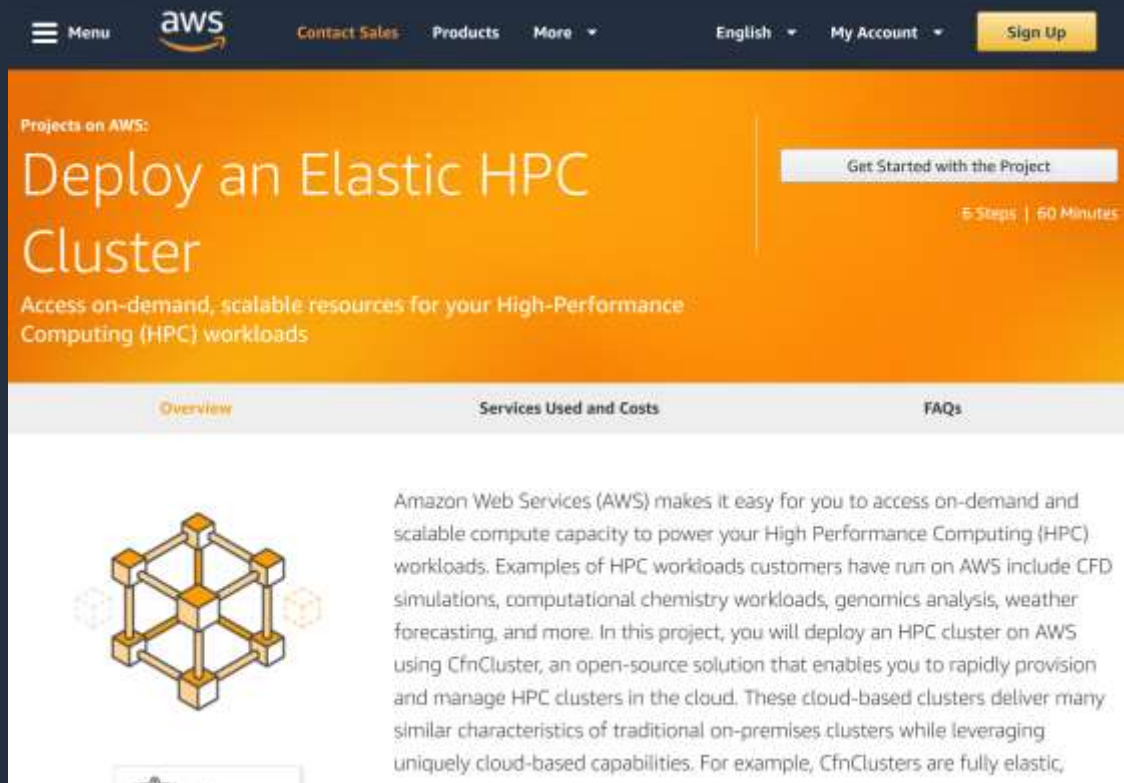
Western Digital

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**



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



The screenshot shows the AWS project page for 'Deploy an Elastic HPC Cluster'. At the top, there is a navigation bar with the AWS logo, 'Menu', 'Contact Sales', 'Products', 'More', 'English', 'My Account', and a 'Sign Up' button. The main content area has an orange header with the text 'Projects on AWS:' followed by the title 'Deploy an Elastic HPC Cluster' in large white font. Below the title is the subtitle 'Access on-demand, scalable resources for your High-Performance Computing (HPC) workloads'. To the right of the title is a 'Get Started with the Project' button and the text '5 Steps | 60 Minutes'. Below the header is a navigation bar with three tabs: 'Overview' (selected), 'Services Used and Costs', and 'FAQs'. The main content area features a 3D diagram of a cluster of server racks on the left and a text block on the right. The text block explains that AWS makes it easy to access on-demand and scalable compute capacity for HPC workloads, listing examples like CFD simulations, computational chemistry, genomics, and weather forecasting. It also mentions that the project involves using CfnCluster to provision and manage HPC clusters in the cloud.

Projects on AWS:

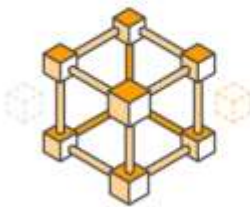
Deploy an Elastic HPC Cluster

Access on-demand, scalable resources for your High-Performance Computing (HPC) workloads

Get Started with the Project

5 Steps | 60 Minutes

Overview Services Used and Costs FAQs



Amazon Web Services (AWS) makes it easy for you to access on-demand and scalable compute capacity to power your High Performance Computing (HPC) workloads. Examples of HPC workloads customers have run on AWS include CFD simulations, computational chemistry workloads, genomics analysis, weather forecasting, and more. In this project, you will deploy an HPC cluster on AWS using CfnCluster, an open-source solution that enables you to rapidly provision and manage HPC clusters in the cloud. These cloud-based clusters deliver many similar characteristics of traditional on-premises clusters while leveraging uniquely cloud-based capabilities. For example, CfnClusters are fully elastic,