# AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

CMP202

# HPC on AWS: Innovating without infrastructure constraints

Ian Colle
General Manager, AWS Batch, HPC, and Visualization
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

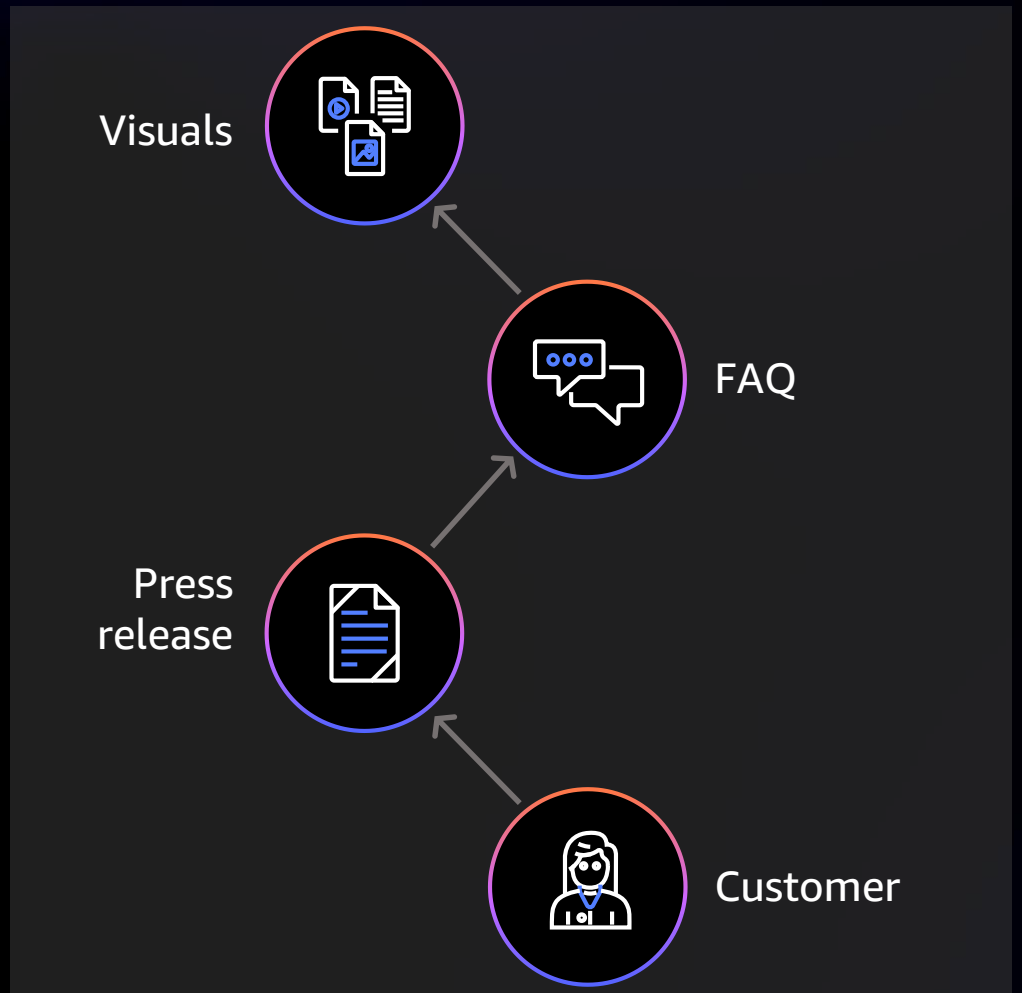
# 2020 was hard




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How AWS thinks about HPC

WORKING BACKWARD FROM THE CUSTOMER




## The questions we ask before development

- Who is your customer?
- What is the customer problem or opportunity?
- What is the most important customer benefit?
- How do you know what your customer needs or wants?
- What does the experience look like?

## Customers want choice and price-performance

- Over 200 AWS services
- Choice of Intel, AMD, and Arm-based processors
- Nearly 500 different instance types to optimize every workload
- Multiple purchase options (e.g., Spot, On-Demand, Savings Plans)

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# 20%* of HPC workloads are currently run in the cloud

On-premises HPC infrastructure limits engineers, scientists, and researchers from getting timely results and insights to answer the world's biggest questions

**Lost productivity and longer time to results**

**72.8%** of organizations that use HPC reported delayed or canceled HPC jobs*



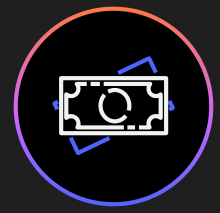
### Lost innovation

Questions are left **unasked**, experiments are left **undone**, and potential revenue is **left** on the table



### Outdated technology

Almost **20%** of the useful life of new technology/ hardware is **lost** in the procurement process



### Technical debt

Adapting **newer algorithms** to meet the requirements of an **existing infrastructure** = delays and **subpar** performance

*Source: Hyperion Research

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Cloud HPC creates value beyond cost savings



## Cost savings (TCO)

**What is it?**

Infrastructure cost savings/avoidance from moving to the cloud

**Examples**

50%+ reduction in TCO (GE)

Typical focus



## Staff productivity

Efficiency improvement by function on a task-by-task basis

Over 500 hours per year of server configuration time saved (Sage)



## Operational resilience

Benefit of improving SLAs and reducing unplanned outages

Critical workloads run in multiple Availability Zones and Regions for robust disaster recovery (Expedia)



## Business agility

Deploying new features/ applications faster and reducing errors

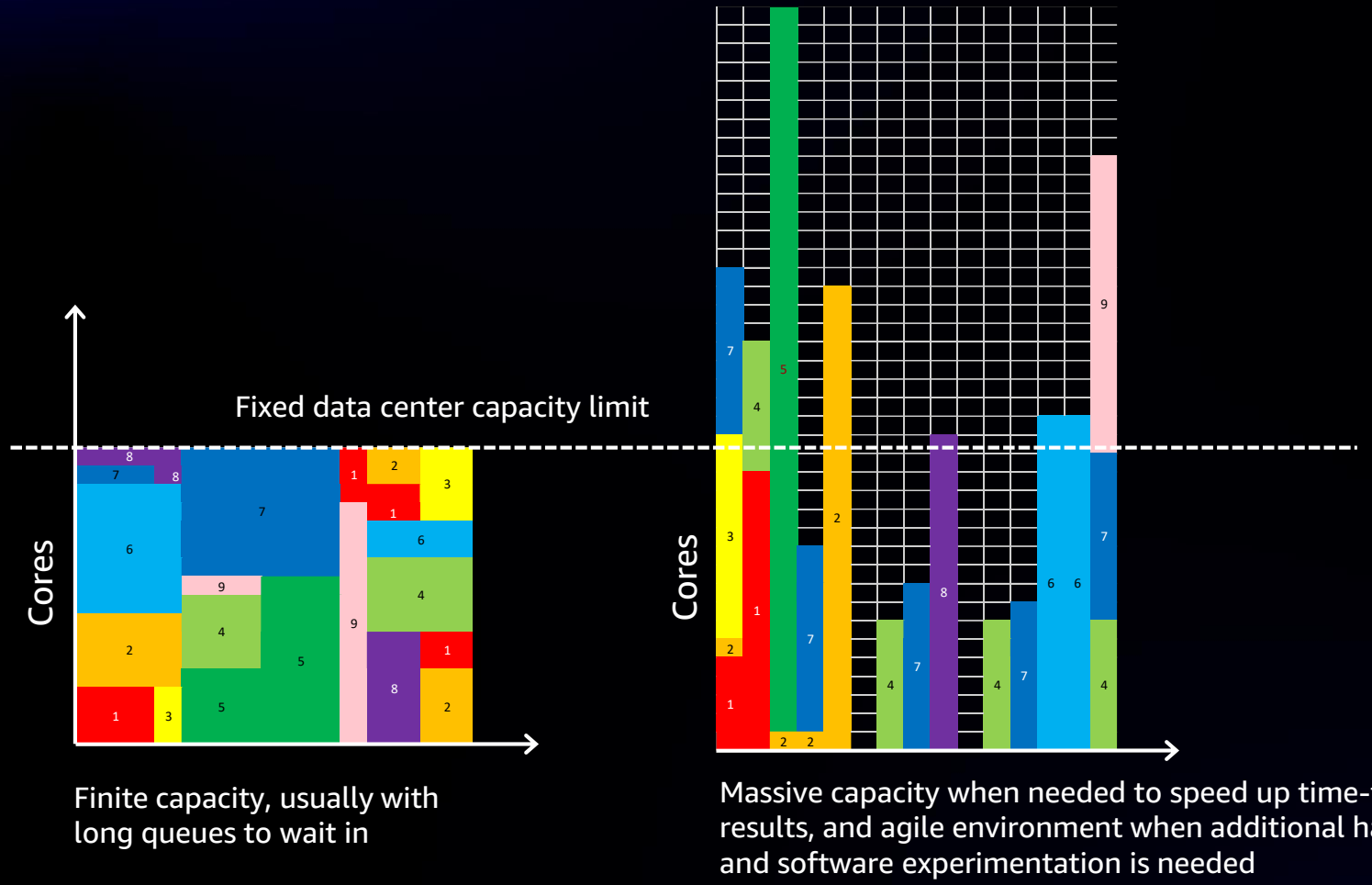
Buildouts and deployments happen 5x faster (Intuit)

Most compelling cloud benefits
Business/value

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Metric for success should be time-to-results




Finite capacity, usually with long queues to wait in

Massive capacity when needed to speed up time-to-results, and agile environment when additional hardware and software experimentation is needed

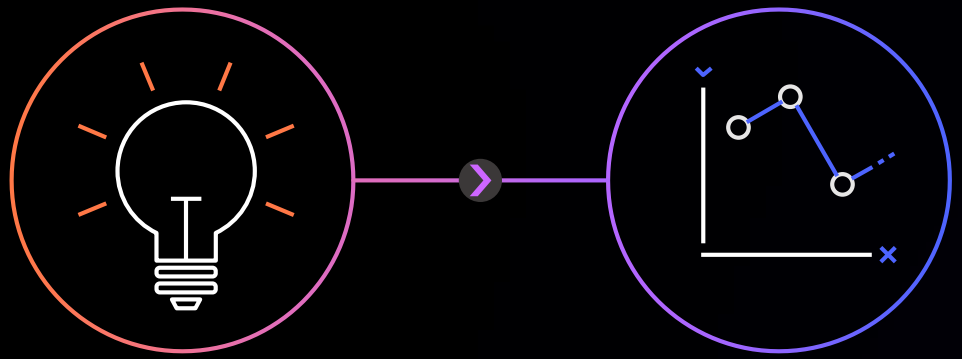
**“For every $1 spent on HPC, businesses see $507 in incremental revenues and $47 in incremental profit.”***

*Source: Hyperion ROI study, http://www.hyperionresearch.com/roi-with-hpc/

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# HPC is about making researchers, scientists, and engineers more productive

## Idea-to-market



"Using AWS helps us get results from our CFD work faster. In some cases, we get results in 24 hours that may have previously taken a week."

**Alex Stoll**
Aeromechanics Lead, Joby Aviation

FRED HUTCH
CURES START HERE®

"Our goal is to accelerate our research processes on AWS so we can get closer to developing therapeutics to fight cancer."

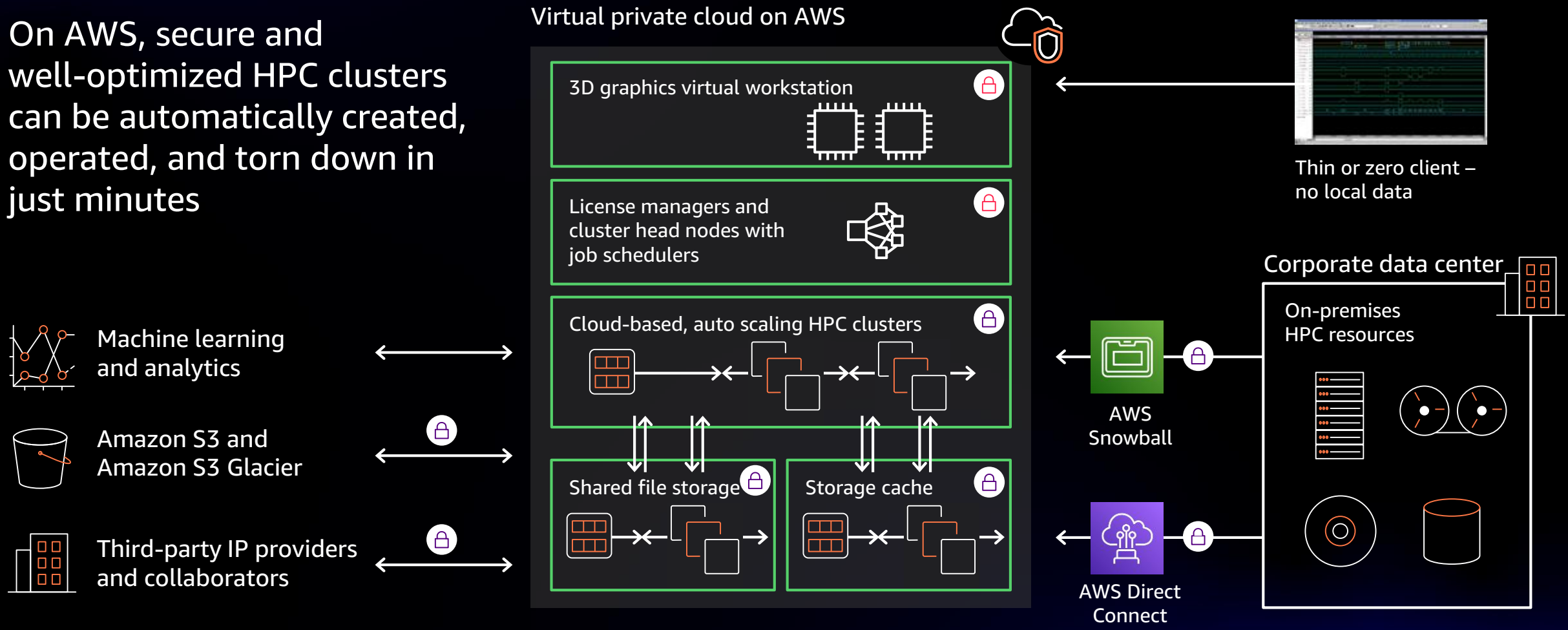
**Sam Minot**
PhD and Staff Scientist,
Fred Hutch Microbiome Research Initiative

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

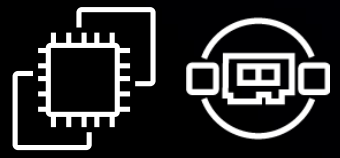
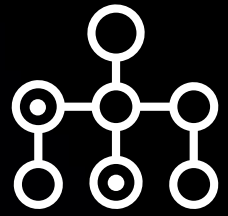
# HPC on AWS

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




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Key services that enable HPC on AWS

| Compute & networking | Management | Storage | Batch processing | Visualization |
|---|---|---|---|---|
|  |  |  |  |  |
| Amazon EC2<br>Elastic Fabric Adapter (EFA) | AWS ParallelCluster<br>NICE EnginFrame | Amazon FSx for Lustre | AWS Batch | NICE DCV |

**Run your HPC workloads with the price-performance you expect and the security you demand**

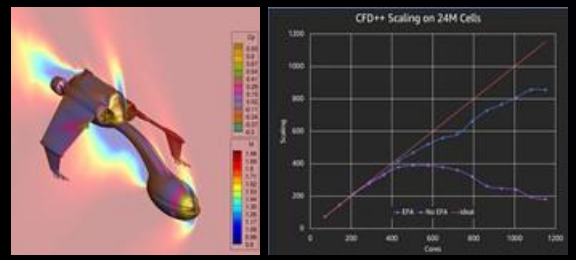
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Why HPC on AWS?

## Flexible compute to maximize performance

### Memory & compute optimized
2.5–3.5 Ghz, 2–16 GB/core, 100 Gbps, EFA



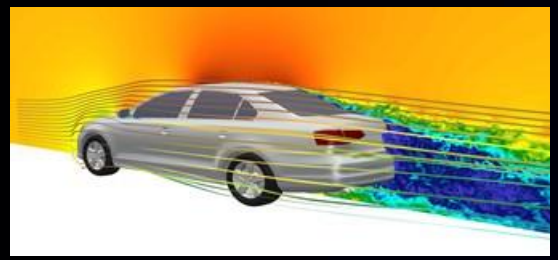

### Graphics and rendering
1–8 GPUs, up to 384 GB RAM, SSD



### High clock speed
4.5 Ghz, 192 GB RAM, 100 Gbps, EFA



### Accelerated computing
8 A100 GPUs, 1.1 TB RAM, SSD, 400 Gbps



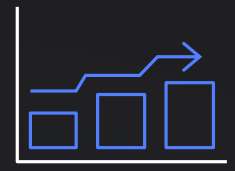
## Flexible pricing models to optimize cost

### On-Demand



Pay for compute capacity by the second with no long-term commitments

### Savings Plans & Reserved Instances



Make a commitment and to save up to 72% off compute

### Spot Instances



Spare EC2 capacity at savings of up to 90% off On-Demand prices

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Broadest choice of processors in the cloud







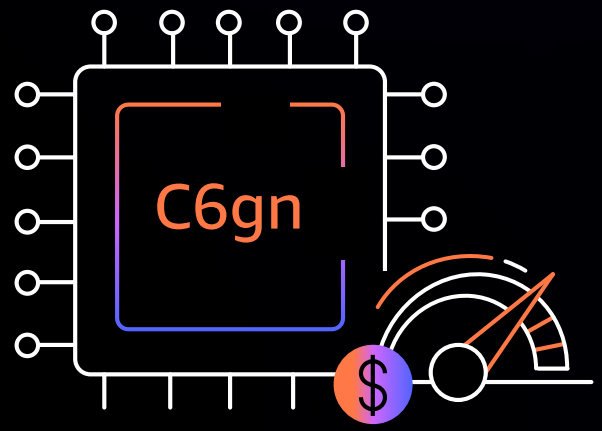



Modular building blocks for rapid design and delivery of EC2 instances

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 C6gn instances powered by AWS Graviton2 processors



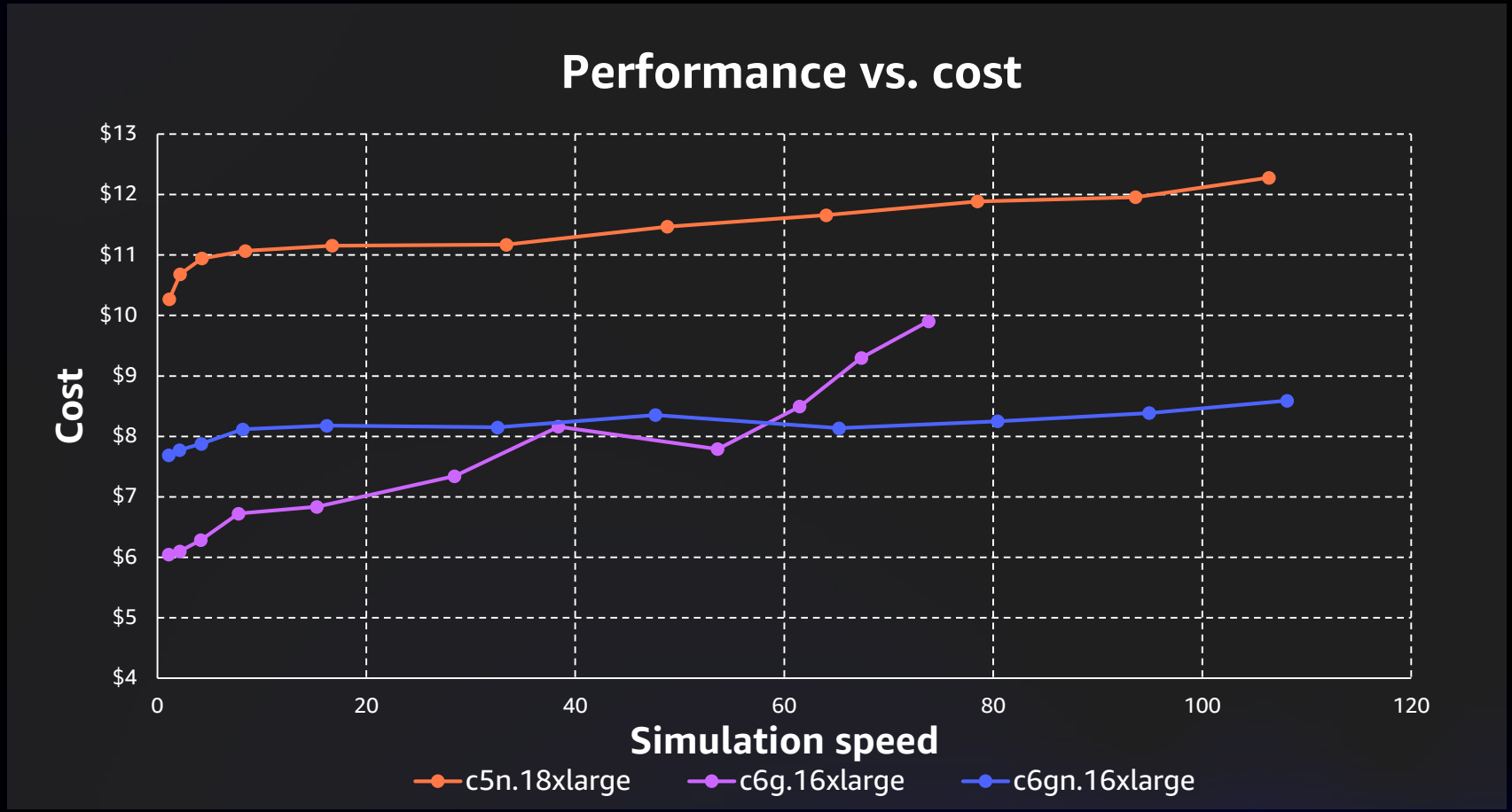

- Up to 40% better price-performance over C5n instances
- 100 Gbps networking
- Ideal for compute-optimized workloads that need high networking throughput and can use Arm

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

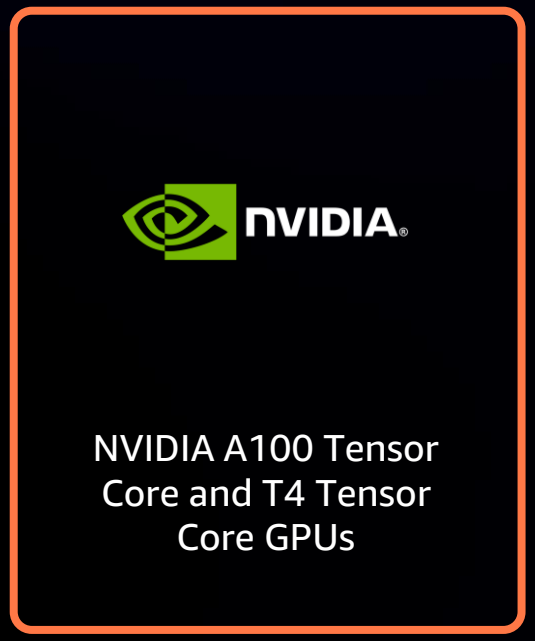
# C6gn price-performance for weather research forecasting simulations

- Scaling out the simulation to 112 nodes for each instance type
- C6g and C6gn are 40% and 30% less, respectively, compared to C5n
- C6gn is the most performant at the least cost




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Broadest choice of accelerators in the cloud









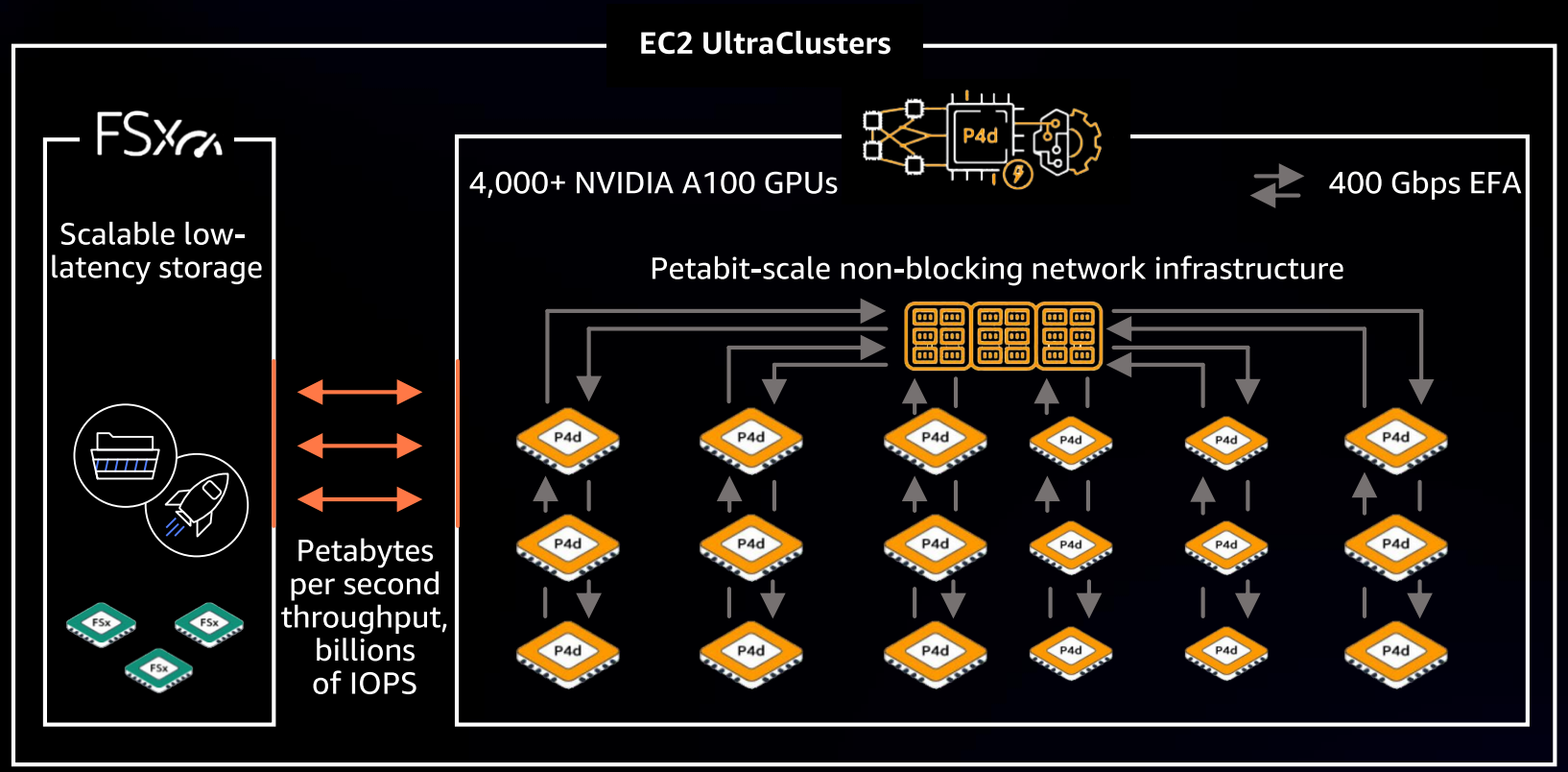

Right accelerator for any workload

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# EC2 UltraClusters of P4d

SUPERCOMPUTING-CLASS PERFORMANCE FOR DEEP LEARNING WORKFLOWS

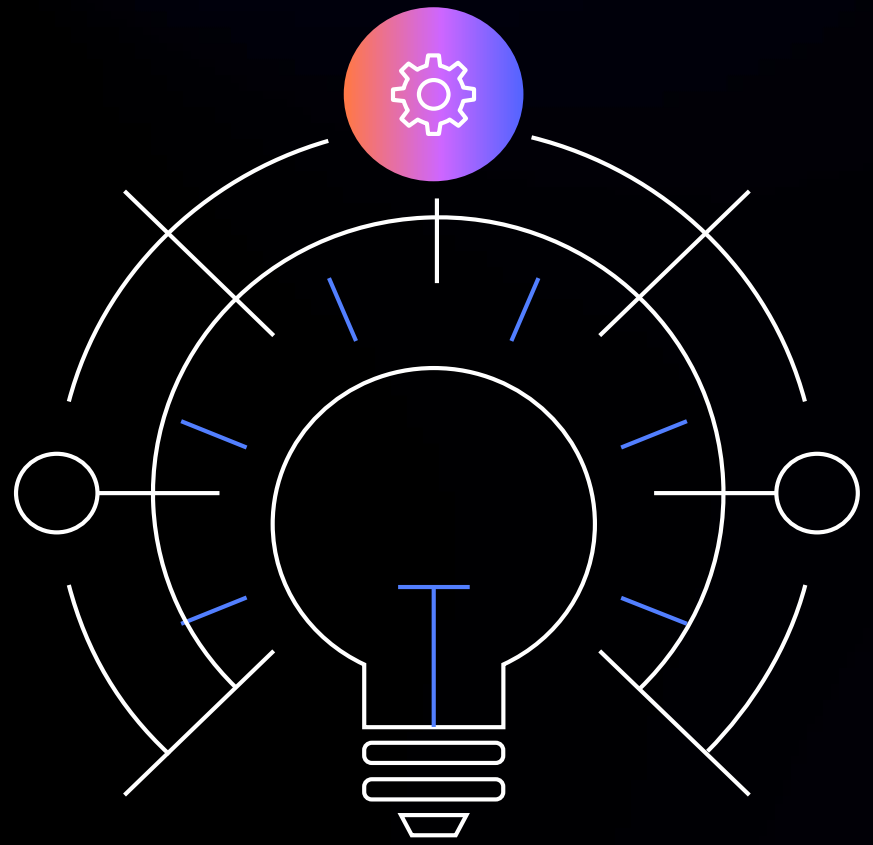



- Based on NVIDIA A100 GPUs
- Availability to scale out to large clusters for distributed training
- 2.5x better deep learning performance and 60% lower cost to train
- EC2 UltraClusters with EFA enables 400 Gbps and allows you to scale to over 4,000 GPUs

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Customers are looking for ways to economize their workflows



Focusing on innovation

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

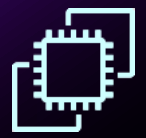
intel®

**15 years** of collaboration and innovation

## KEY HIGHLIGHTS

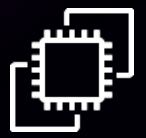
- Fastest processor in the cloud with up to 4.5 GHz all-core turbo
- First cloud to launch Ice Lake
- Deep engineering collaboration across Amazon portfolio
- Over 250 EC2 instances are powered by Intel Xeon processors

## RECENT AMAZON EC2 INSTANCES



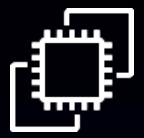
**I4**

STORAGE OPTMIZED



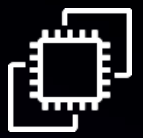
**C6I**

COMPUTE OPTIMIZED



**R6I**

MEMORY OPTIMIZED



**M6I**

GENERAL PURPOSE

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 C6i instances

NEW COMPUTE OPTIMIZED INSTANCES WITH UP TO 15% BETTER PRICE-PERFORMANCE THAN C5 INSTANCES

NEW!

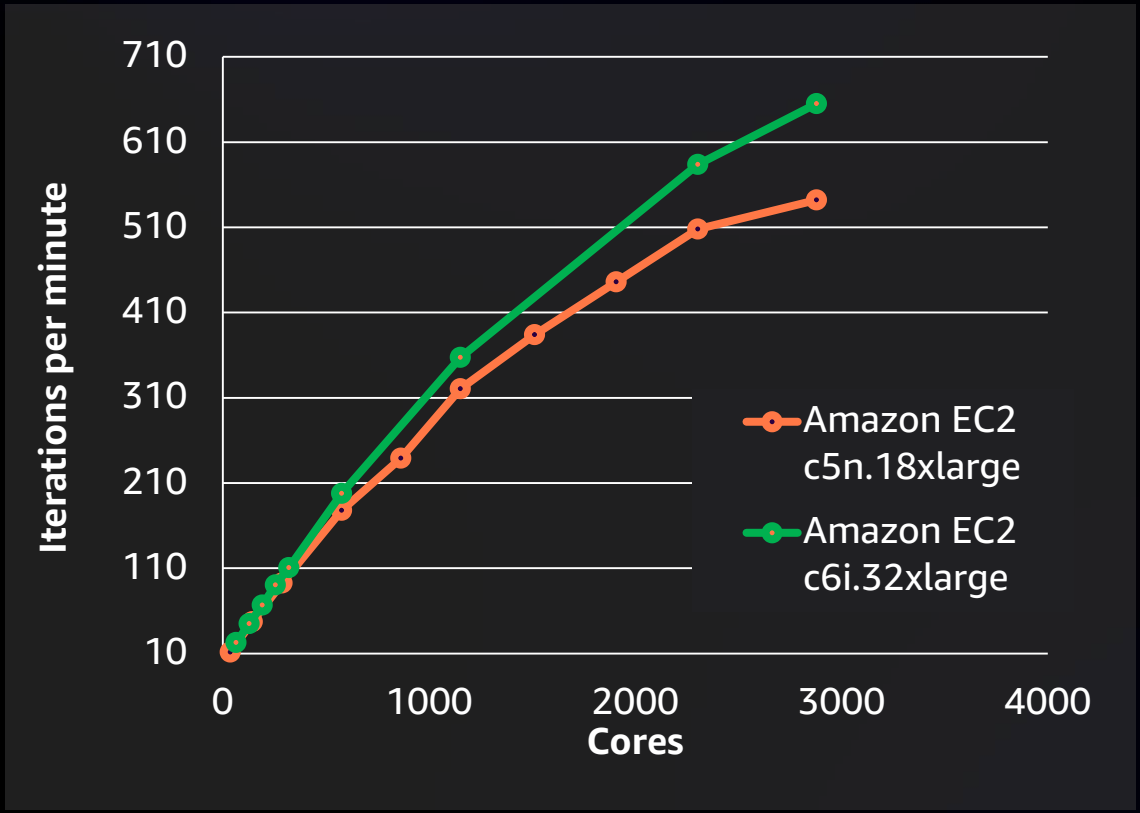


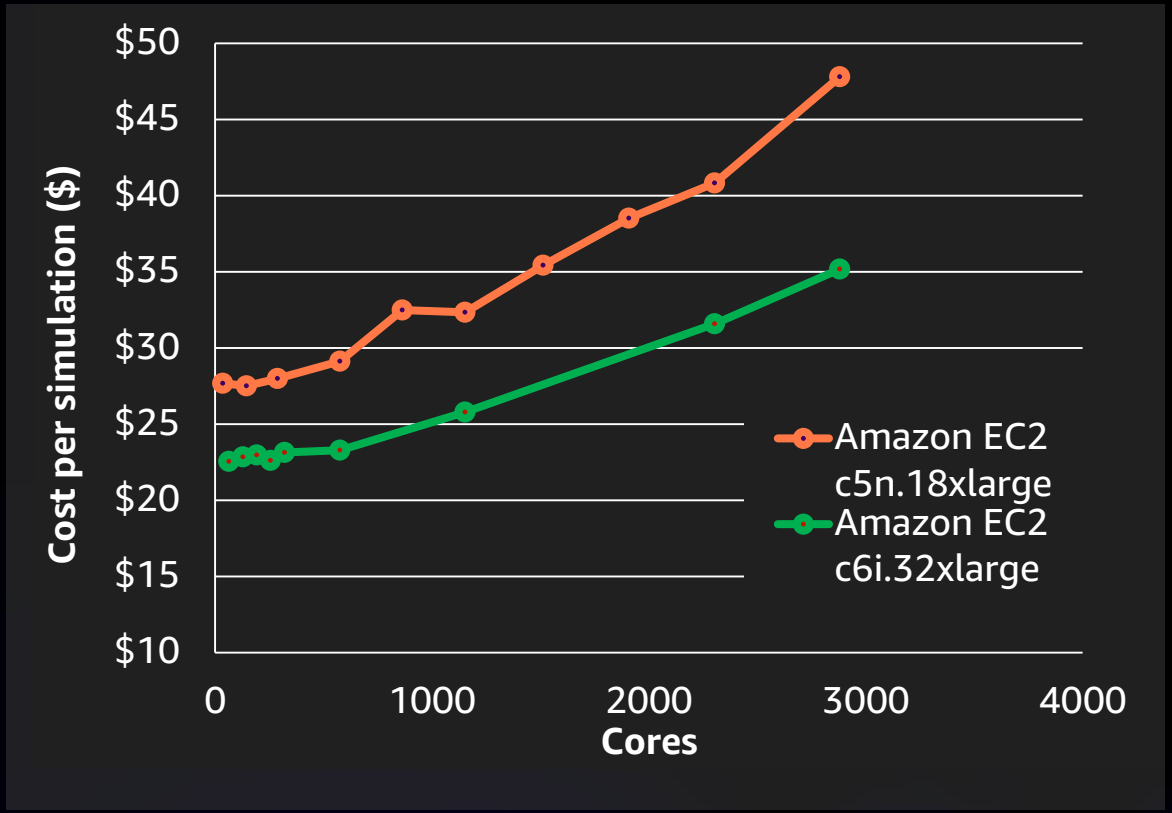

- Compute optimized instances powered by third-generation Intel Xeon Scalable processors, 64 cores, up to 3.5 GHz frequency, and 256 GB of RAM
- Up to 50 Gbps networking bandwidth (2x compared to C5)
- Up to 40 Gbps Amazon EBS throughput (2x compared to C5)
- ~10% more memory bandwidth/core over C5 instances
- Flexibility and choice – 9 instance sizes

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Siemens Simcenter STAR-CCM+ Le Mans benchmark




**20% performance improvement per core and 35% less cost**

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 G5 instances

- Featuring up to 8 NVIDIA A10G GPUs with 24GB of GPU memory
- Up to 3.3x better ML training performance and up to 3x better ML inference performance compared to EC2 G4dn instances
- Up to 15% lower cost-to-train than Amazon EC2 P3 instances
- Instance size flexibility to accommodate a broad range of customer requirements
- Get started using DLC, DL AMIs, Amazon EKS, or Amazon ECS

NEW!




| Instance sizes | GPU | vCPU | Memory | Storage | Network bandwidth | EBS bandwidth |
|---|---|---|---|---|---|---|
| **G5.xlarge - G5.48xlarge** | 1, 4, or 8 | up to 192 | up to 768 GB | up to 7600 GB | up to 100 Gbps | up to 19 Gbps |

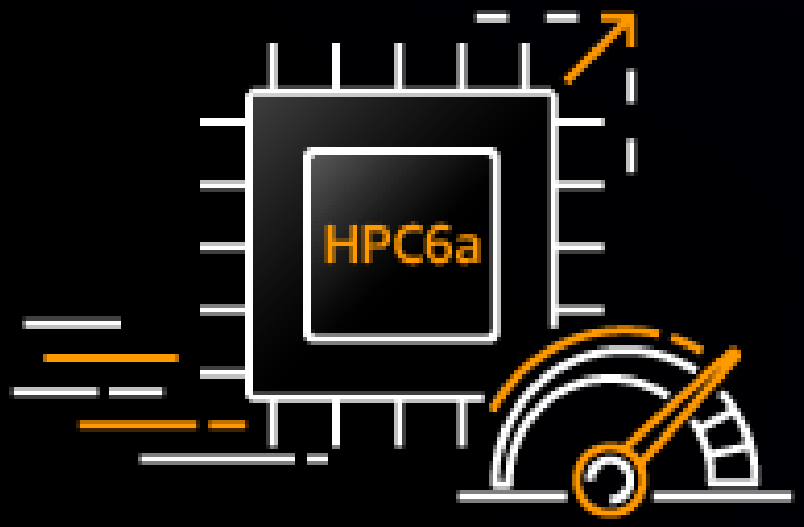
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 Hpc6a instances

Designed to deliver the best price performance for compute-intensive, high performance computing workloads in Amazon EC2

NEW!




AMD third-generation EPYC Milan processors, 96 cores, up to 3.6 GHz frequency, and 384 GB of RAM

Up to 65% better price performance over comparable Amazon EC2 x-86 based, compute-intensive instances

Elastic Fabric Adapter enabled by default for 100 Gbps networking for inter-instance communications

AWS Regions at launch: US East (Ohio) and GovCloud US West

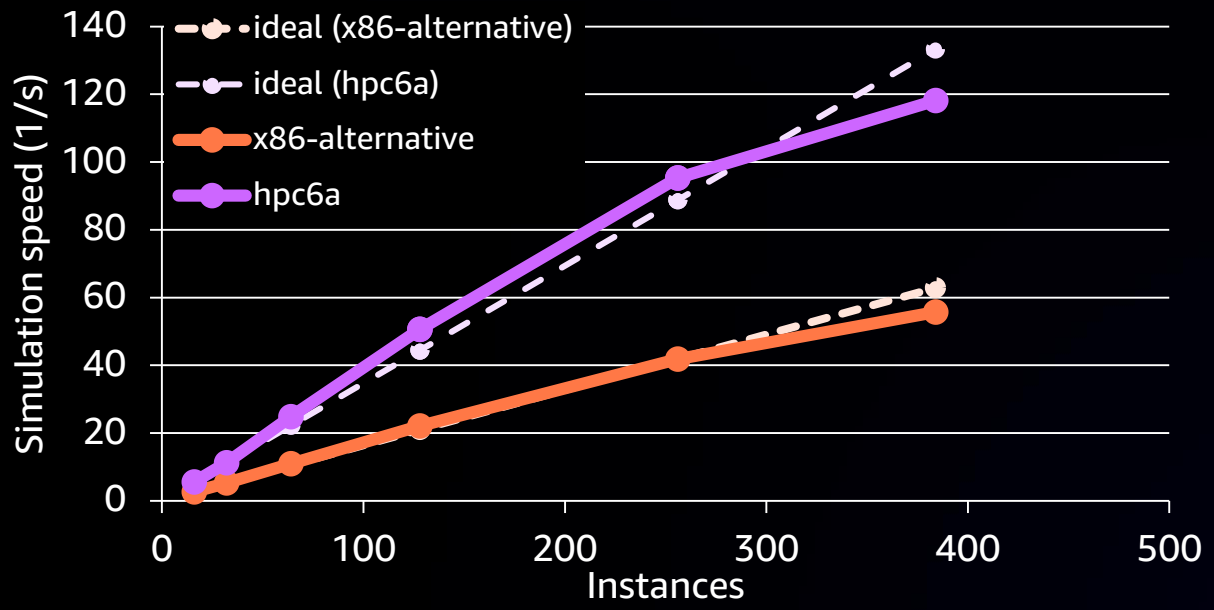
aws

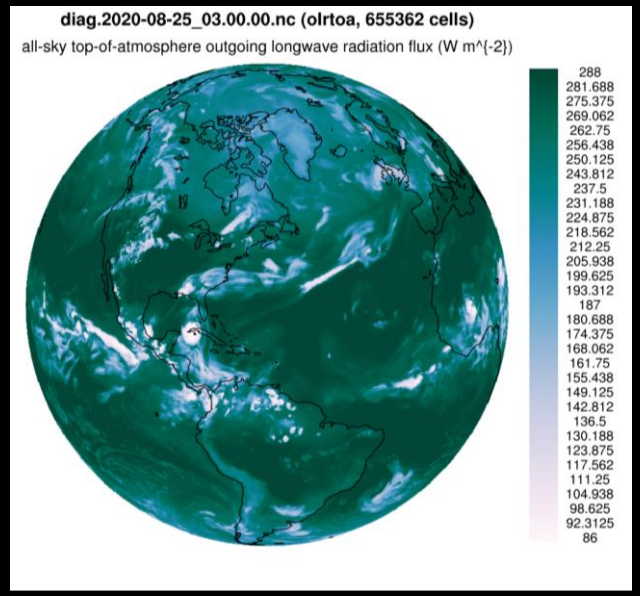
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

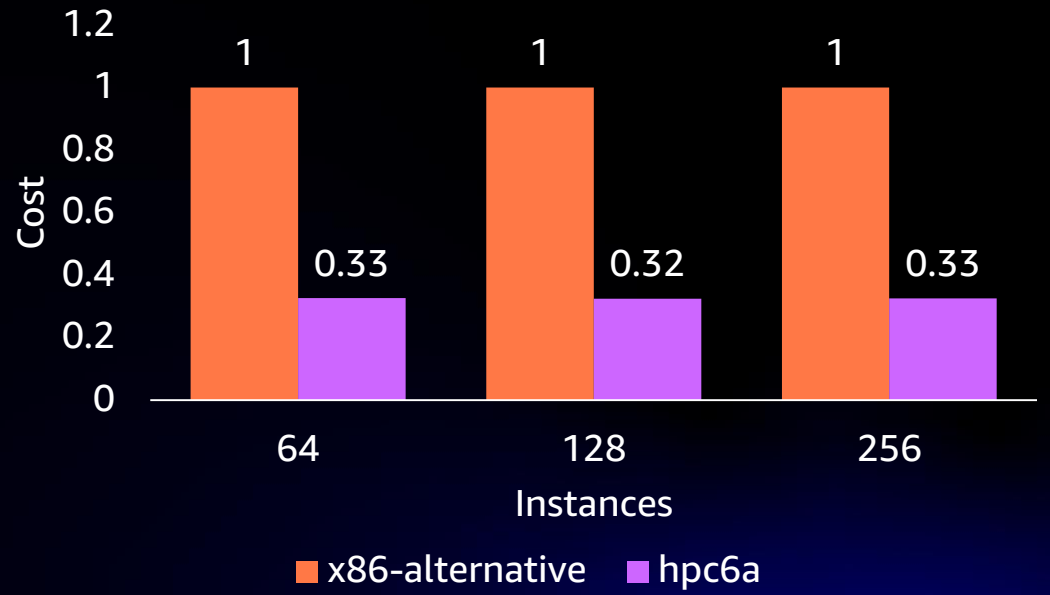
# Hpc6a enables high-resolution weather modeling

*"Our collaboration with AWS allows us to better serve our customers with high-resolution weather prediction systems that feed analytics engines. We're very excited to see the price/performance of Hpc6a and we expect this to be our go-to Amazon EC2 instance choice for HPC workloads going forward"*

*- Lars Ewe, Chief Technology Officer, DTN*




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Elastic Fabric Adapter



**SRD protocol**

Proving myths about latency constraints wrong



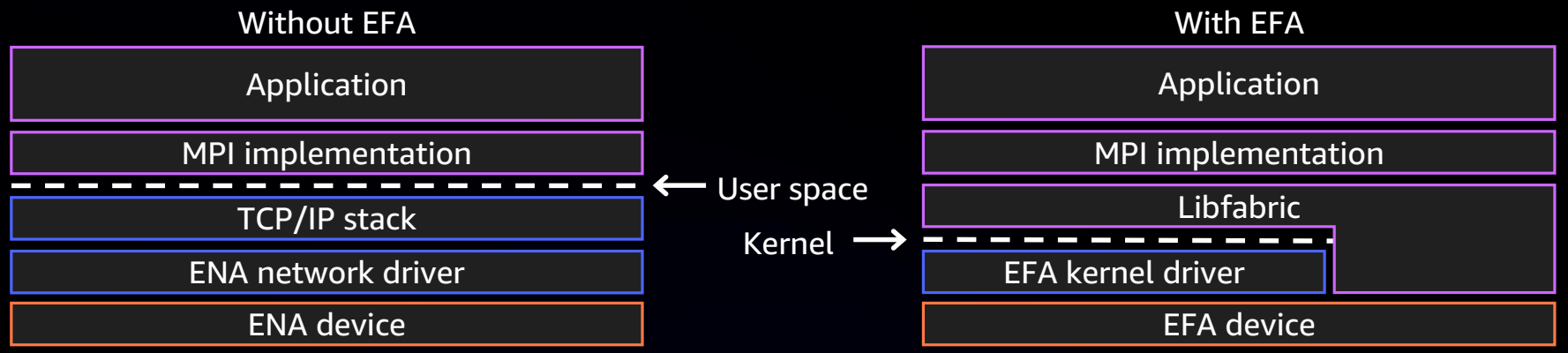
CFD



Seismic



Weather modeling



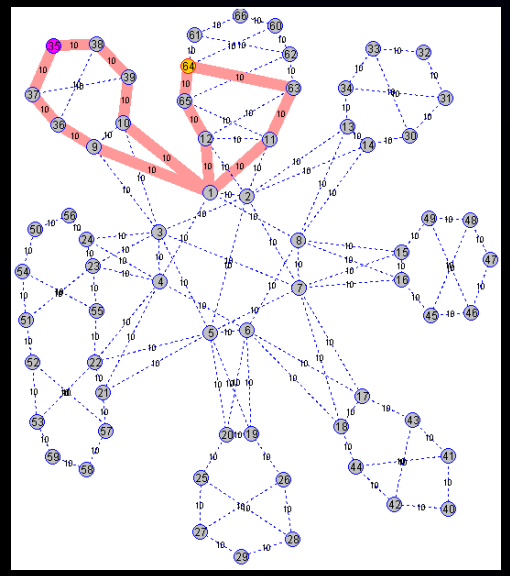

RIVIAN

Rivian uses EFA-enabled instances to improve software tool performance by 66%

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Purpose-built: EFA



## Elastic Fabric Adapter

OS bypass

GPUDirect and RDMA

Libfabric core supports wide array of MPIs and NCCL

## Scalable reliable datagram

OCMP-enabled packet spraying

Cloud-scale congestion control

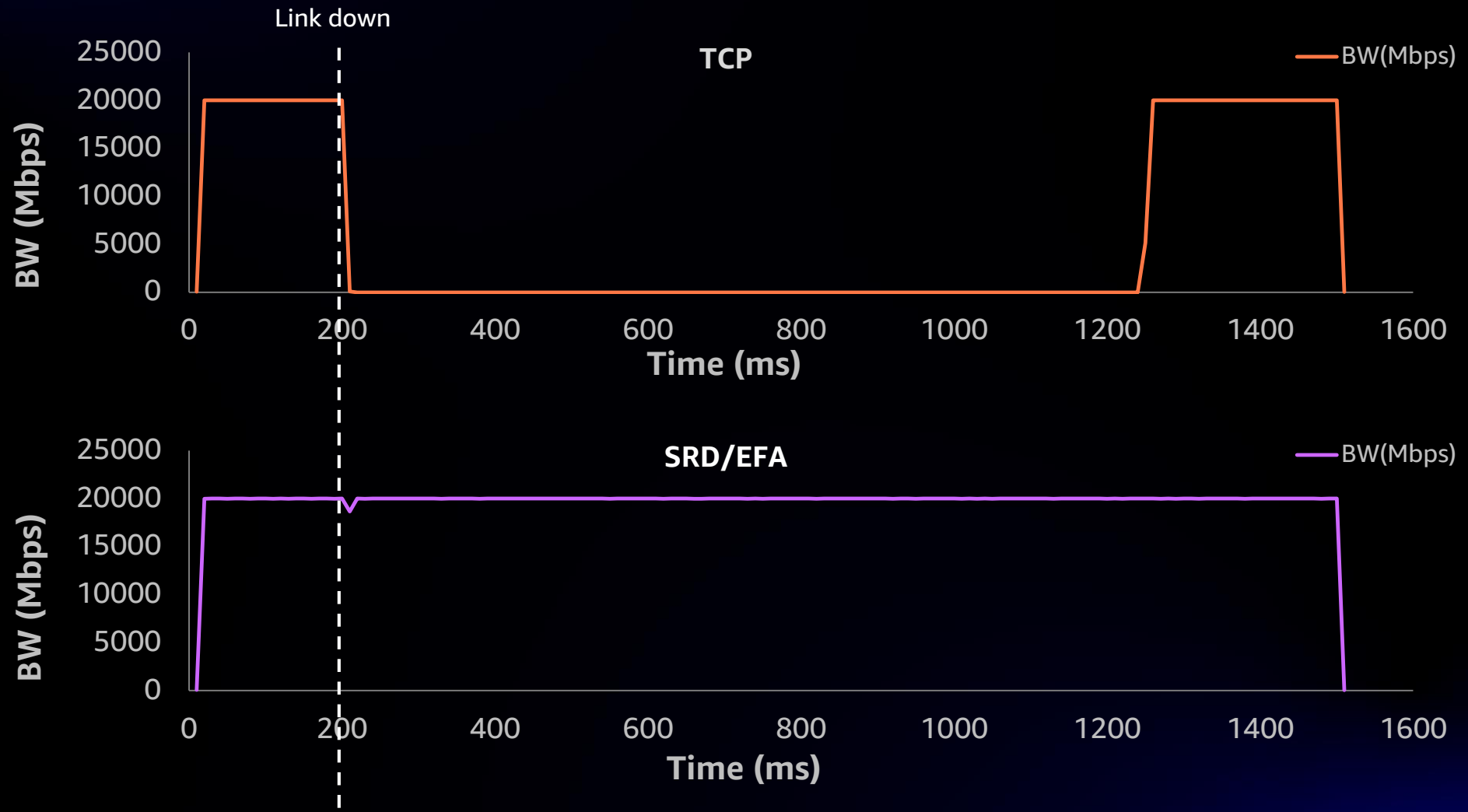
Fast recovery from packet loss or link failure

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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Graceful error recovery: Throughput




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# FLYING WHALES uses EFA and AWS ParallelCluster to accelerate CFD simulations

## Challenge

FLYING WHALES used an on-premises HPC cluster to perform Computational Fluid Dynamics (CFD) analyses

However, the company didn't have the scalability or flexibility it needed to support the workloads

## Solution

FLYING WHALES runs its HPC cluster on Amazon EC2 C5n.18xlarge instances with EFA, and the company uses AWS ParallelCluster to simplify the deployment and management of an HPC cluster

## Benefits

- Runs CFD workflow jobs 15 times faster
- Completes jobs in days instead of months
- Scales to support 600-core computational models




> "We can run CFD workflow jobs 15 times faster on AWS thanks to the computing power and inter-node network performance we get using the Amazon EC2 C5n.18xlarge instances and EFA."
>
> **Guillaume Martinat,** Lead Aerodynamics Engineer, FLYING WHALES

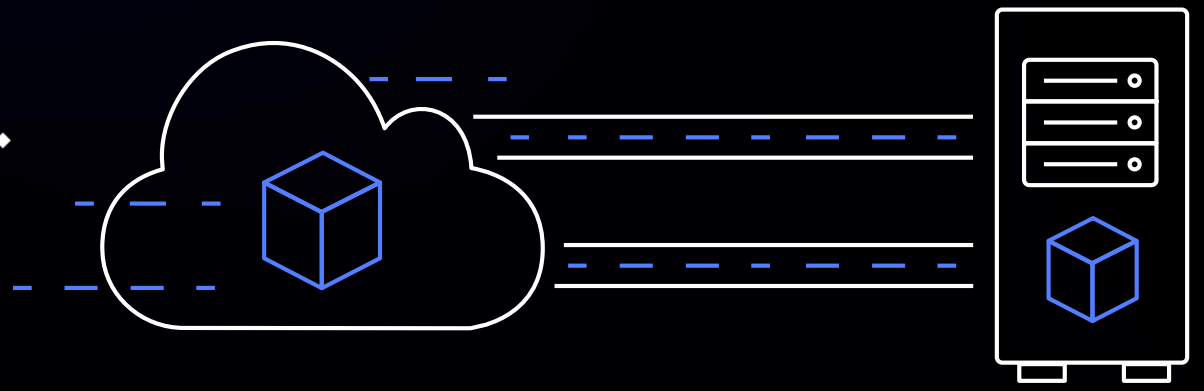
FLYING WHALES

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS ParallelCluster

intel select solution



One-stop shop to set up your HPC cluster

## Integrated with AWS services you need

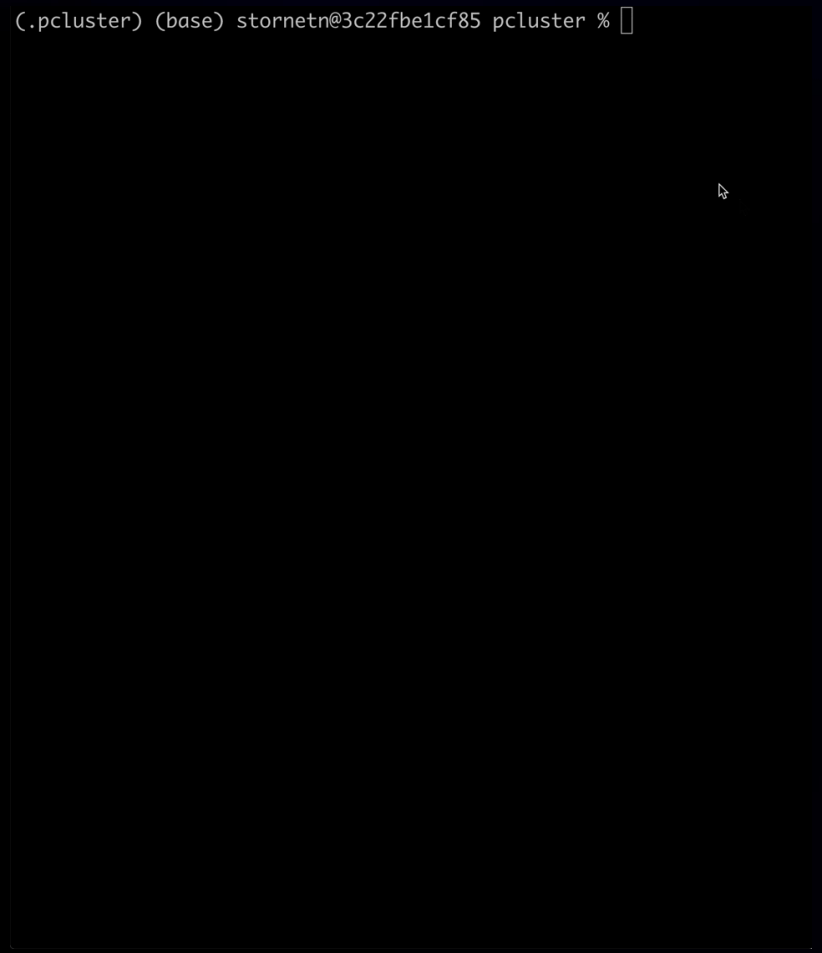
- Amazon FSx for Lustre
- Amazon EC2 instances
- EFA
- NICE DCV
- AWS Batch

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS ParallelCluster 3.0

## Infrastructure as code




## What's new in 3.0

NEW!

- A new flexible AWS ParallelCluster API
- Build custom AMIs with EC2 Image Builder
- A new configuration file format
- Simplified network configuration options
- Finer-grained IAM permissions
- Runtime customization scripts

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NICE EnginFrame with AWS HPC Connector

NEW!

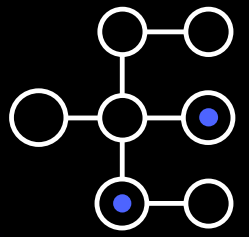
## Increase scale

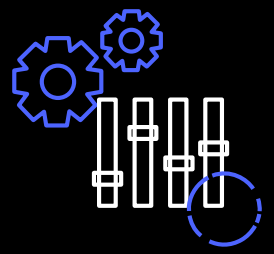


## Easy-to-use interface



## Simple job submission





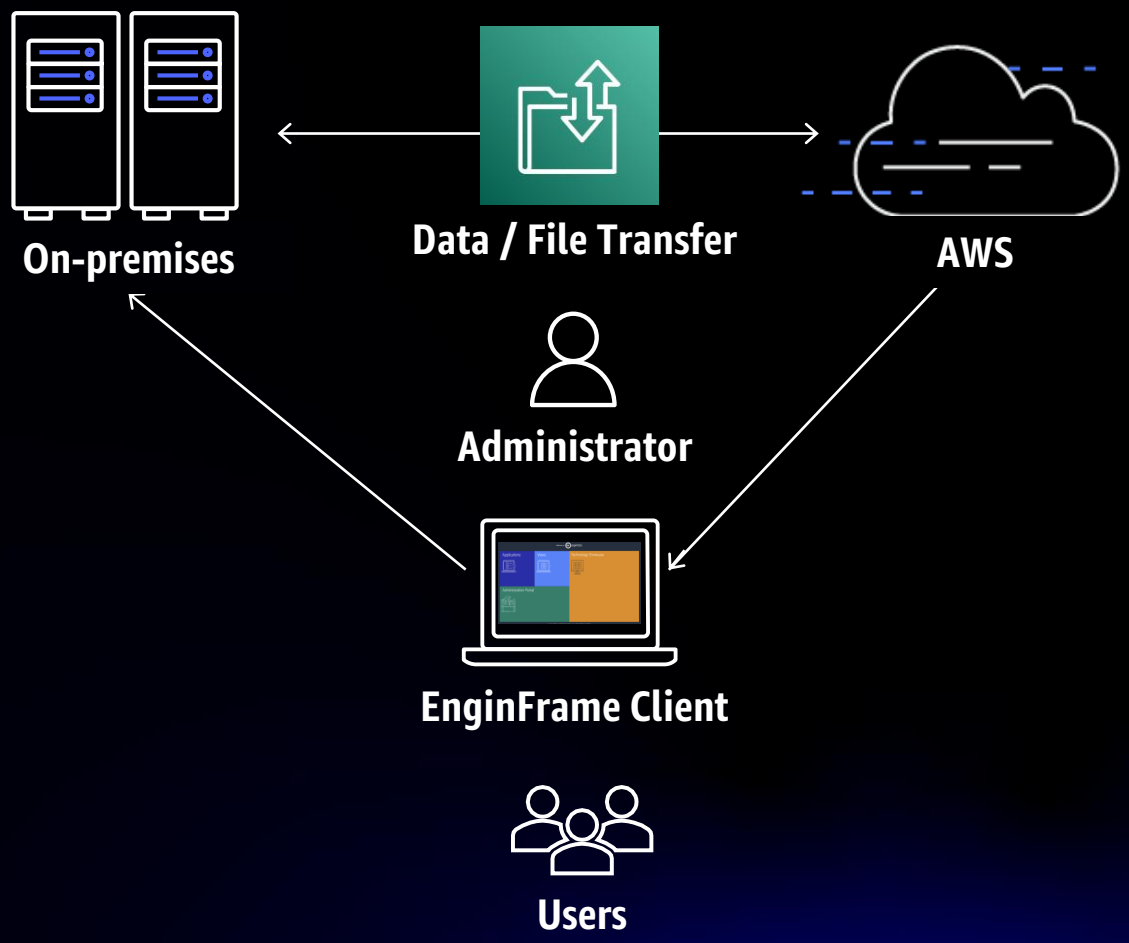
- Manage your existing on-premises infrastructure and access the virtually unlimited capacity on AWS
- Schedule jobs across on-premises and cloud environments through a single portal
- Access the broad range of cloud-based services like ML/AI and analytics

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

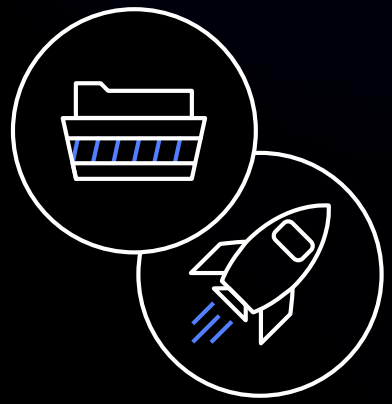
# A single interface to manage workloads across on-premises and AWS

- Consistent interfaces across on-premises and AWS
- Managed AWS clusters (using ParallelCluster) lower the barriers to entry for HPC in the cloud
- Built-in remote visualization to minimize need for data movement
- Tools for administrators to manage and mediate users' work with AWS



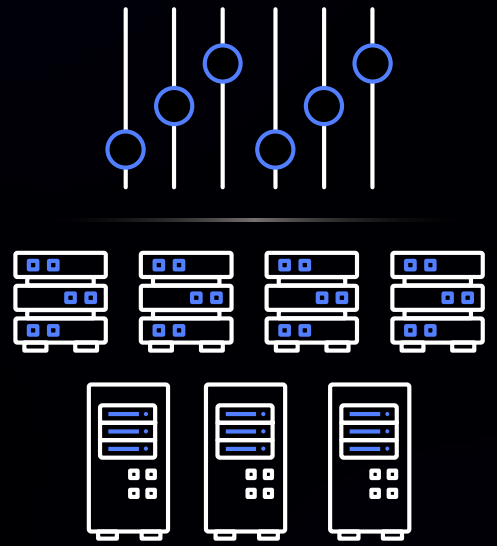

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon FSx for Lustre



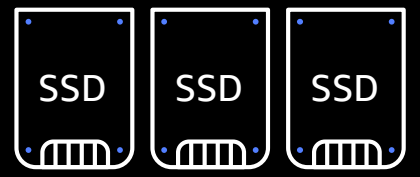
**High and scalable performance**

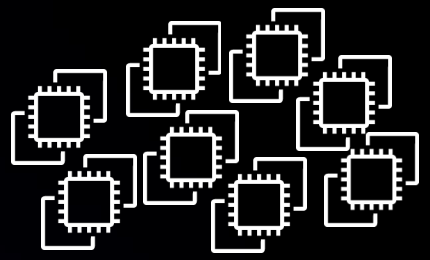
Parallel file system



+100 GiB/s throughput
Millions of IOPS
Consistent submillisecond latencies

SSD-based



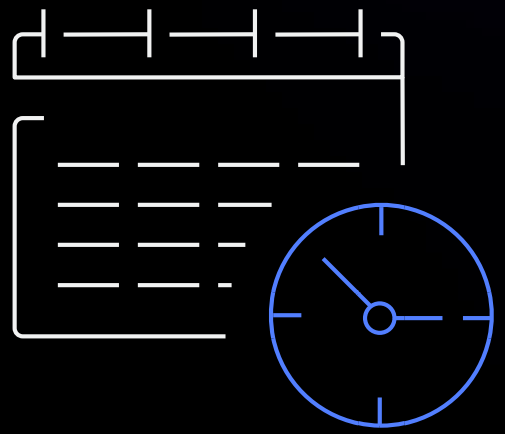



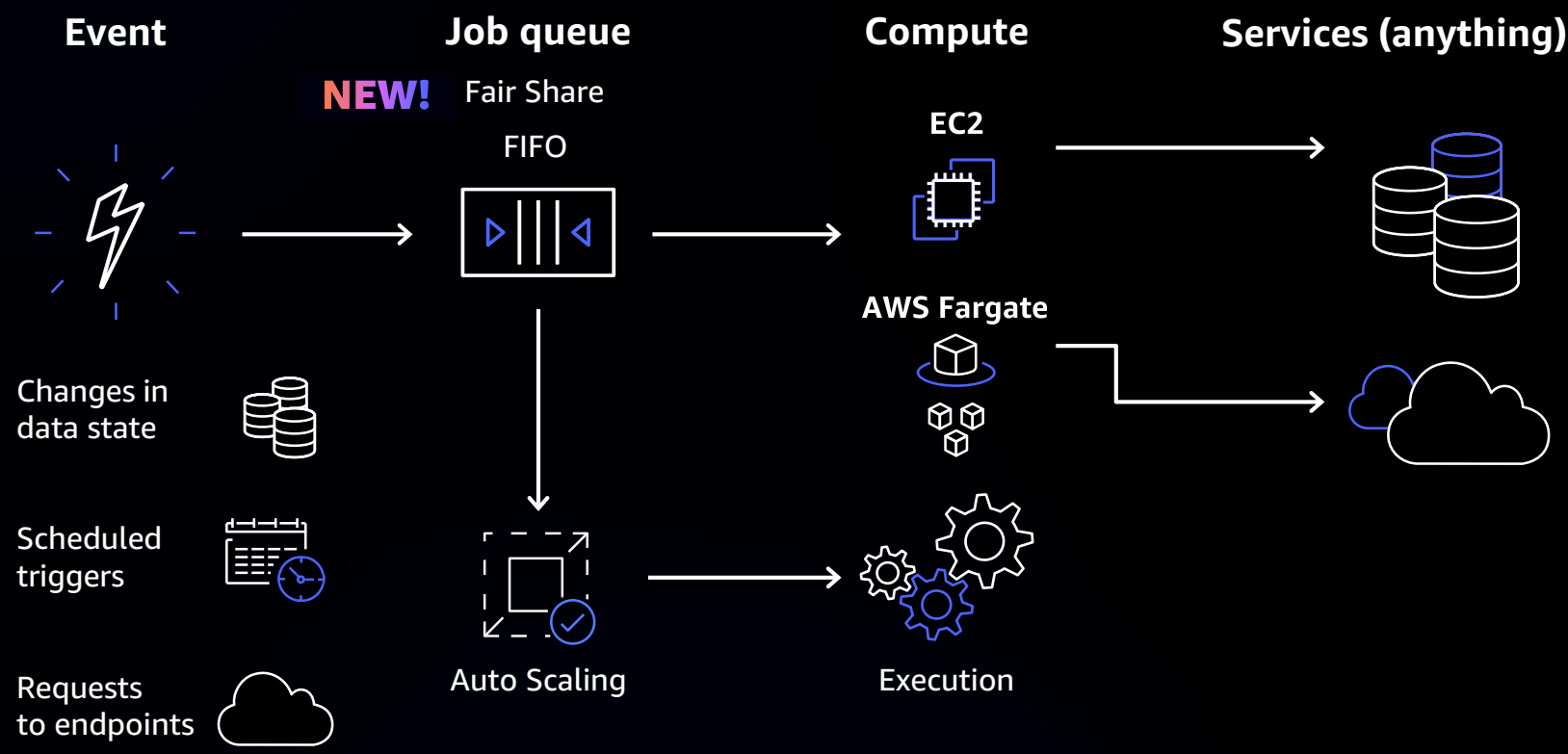
Supports concurrent access from hundreds of thousands of cores

MAXAR Maxar uses FSx for Lustre to generate weather forecasts 58% faster

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Batch






RELAY THERAPEUTICS

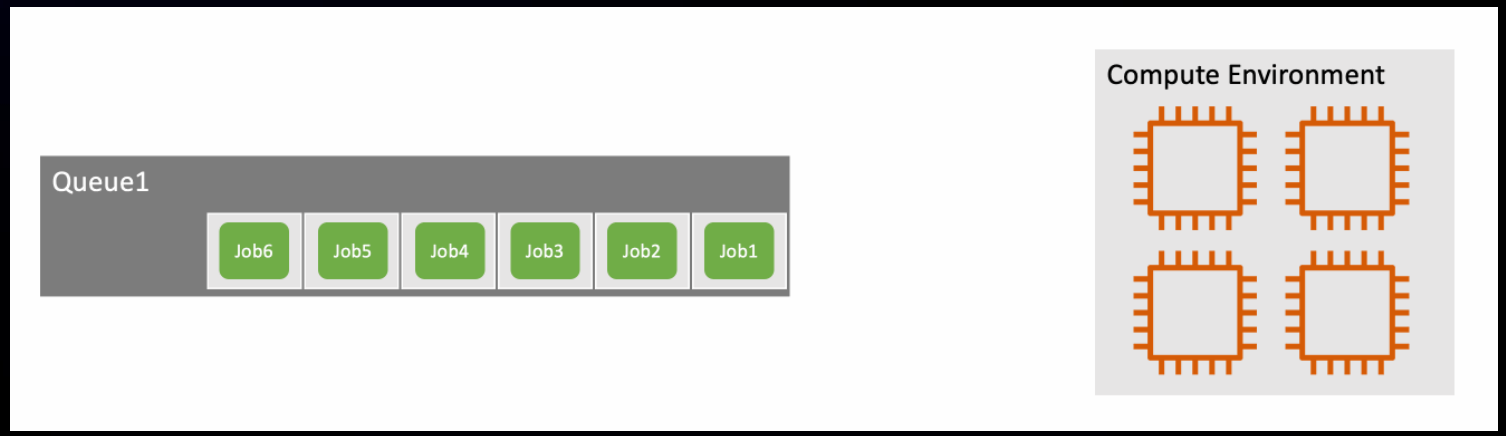
Relay Therapeutics uses AWS Batch to analyze 5 billion molecules in 1 day

aws

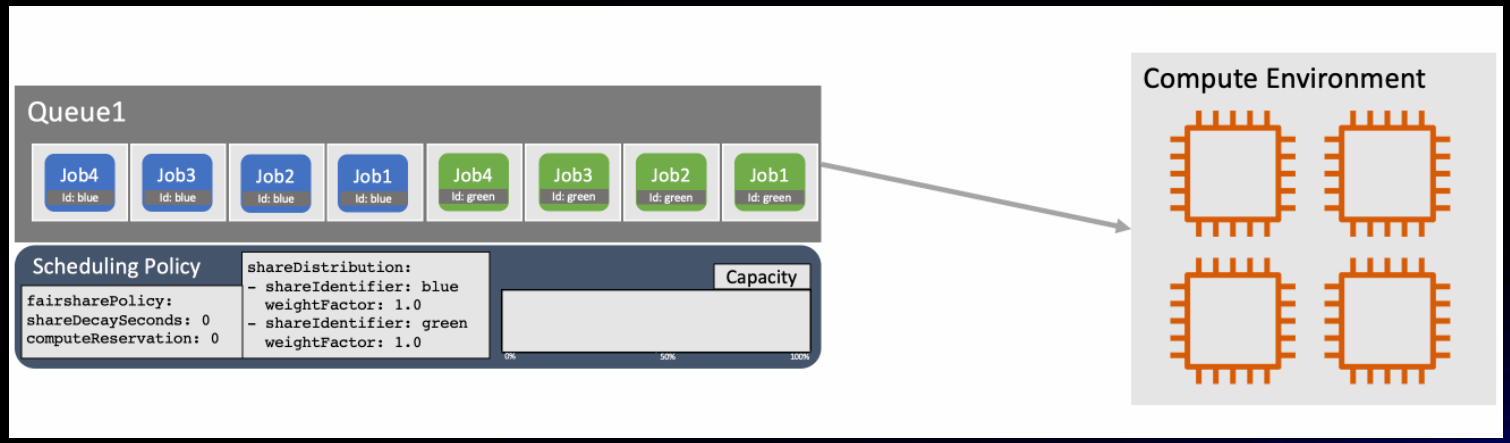
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Introducing fair-share scheduling

## FIFO




## Fair-share




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NICE DCV



Transfer pixels, not data



**Local clients**

**Remote servers**

Users can access, manipulate, and share business-critical information, regardless of their location, over LAN or WAN

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Volkswagen uses NICE DCV from AWS for 3D remote visualization

## Challenge

Volkswagen Passenger Cars needed to provide remote access to CAE applications for more than 1,000 automotive engineers

## Solution

The company's automotive engineers now use NICE DCV to remotely access CAE applications running in the company's on-premises, high performance computing (HPC) cluster

## Benefits

- Delivers reliable remote streaming of 3D applications to 1,000 engineers
- Enables flexibility for offsite employees
- Provides near-real-time responsiveness with high image quality



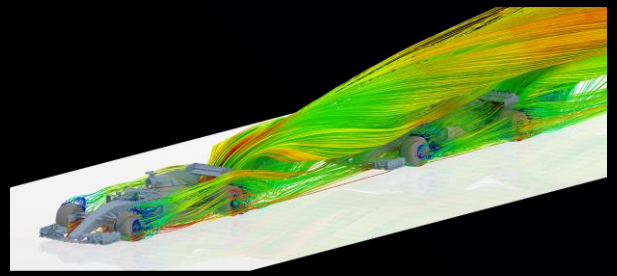
"Our automotive engineers can reliably access their high-end Linux workstations and **complete 3D simulations from home or other remote locations using NICE DCV**."

**Gunther Mayer,** IT Specialist, R&D, Volkswagen Passenger Cars

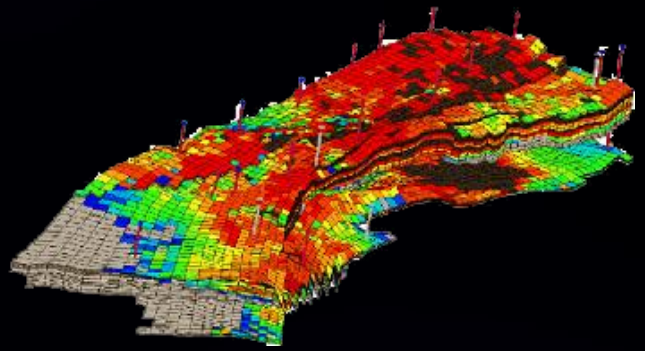
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

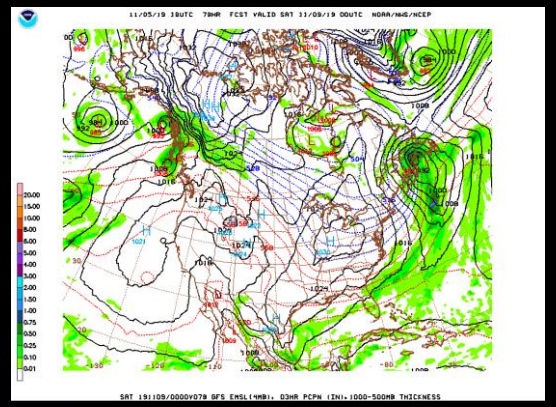
# Any type of HPC workload is possible on AWS



Computational fluid dynamics



Reservoir simulations



Weather forecasting

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# HPC workloads across industries




Life sciences




Financial services




Energy




Design & engineering




Climate & earth sciences




Autonomous vehicles

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# HPC needs scale

VirtualFlow

2.2M+ vCPUs to analyze billions of cancer proteins

Woodside

1.1M vCPUs to accelerate seismic processing workload by 150X

Descartes Labs

9.95 petaflop high-performance LINPACK (HPL) benchmark to rank #50 on the TOP500

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

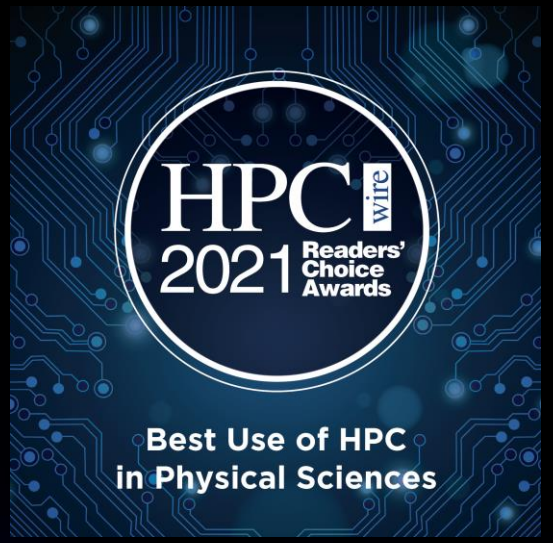
# AWS HPC continues to be recognized by the HPC community





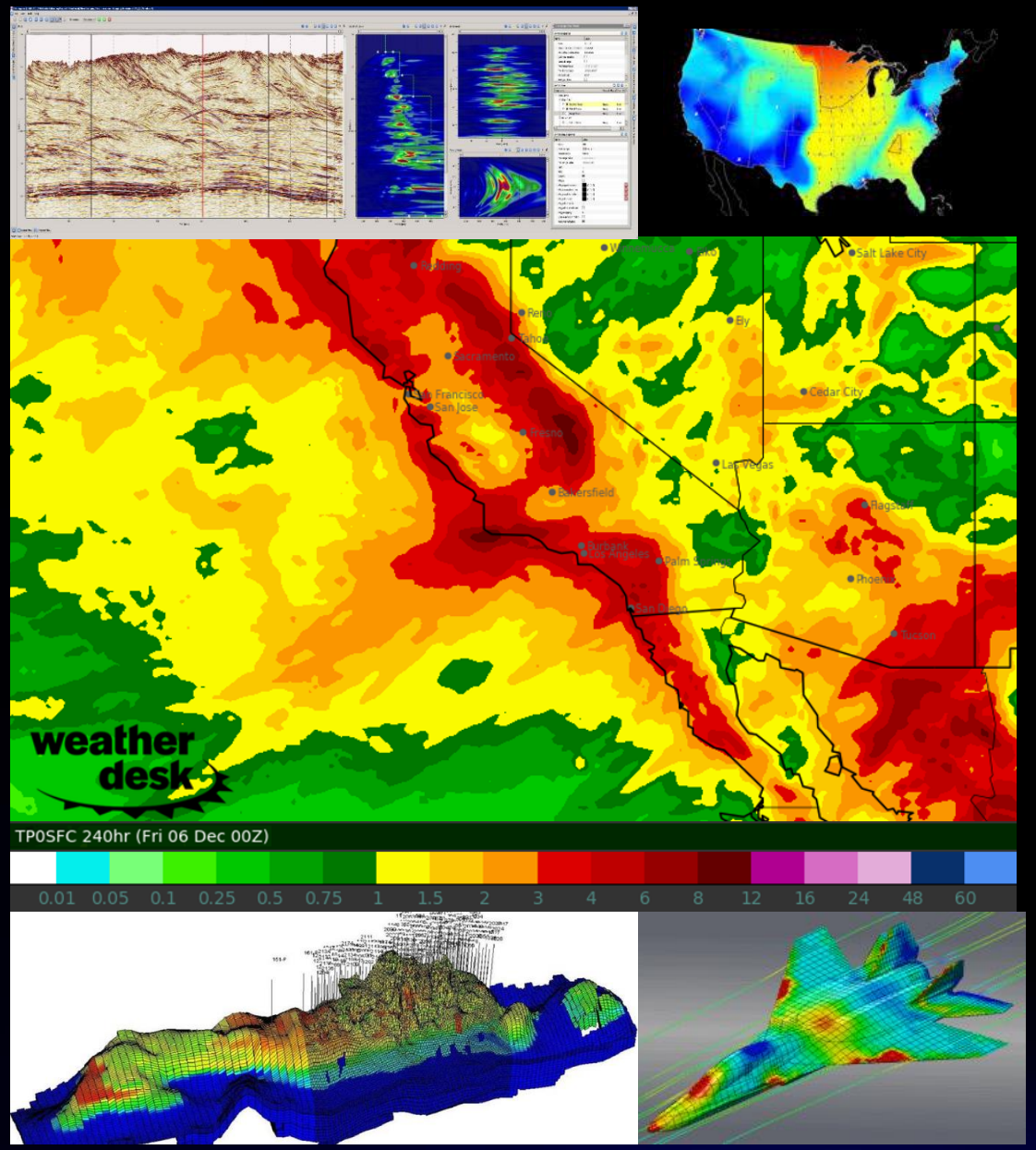





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# With AWS HPC solutions, how will you accelerate your innovation?




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.