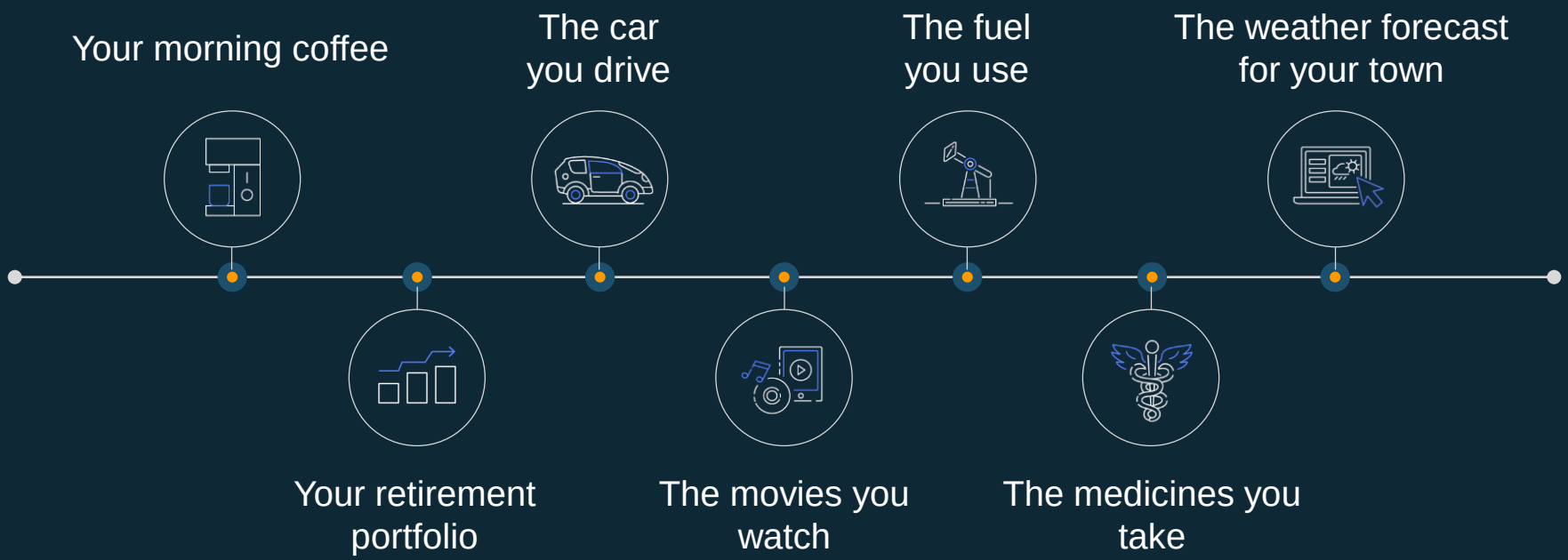
HPC at AWS

Innovating without infrastructure constraints

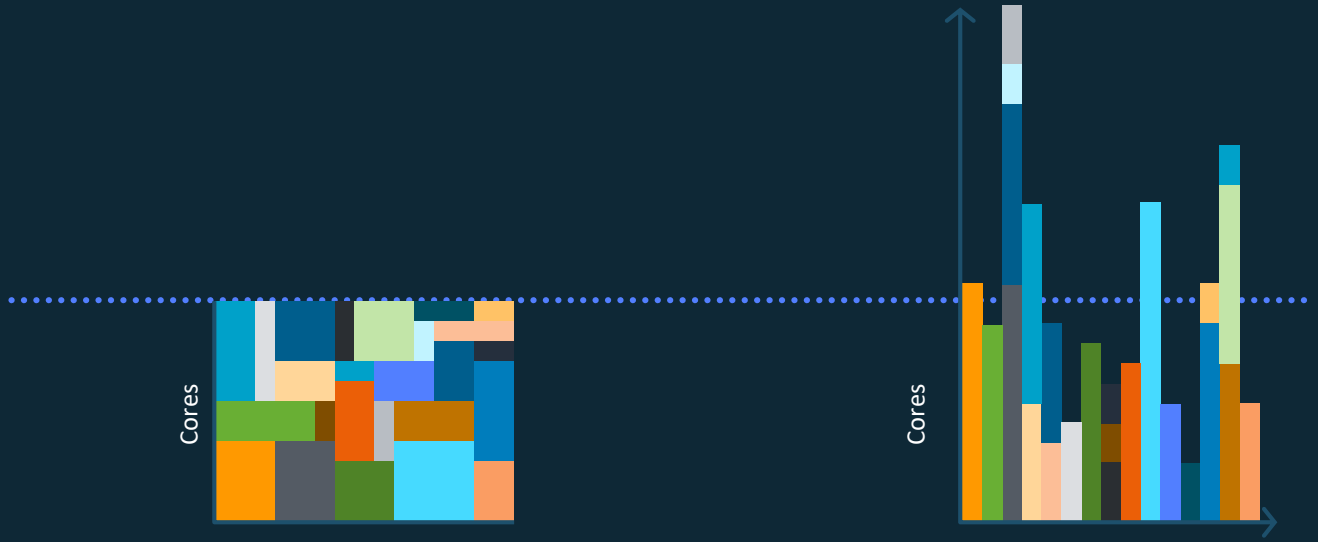
Gilles Tourpe, gtourpe@amazon.com, +33754844572

October 2020

HPC impacts your life every day



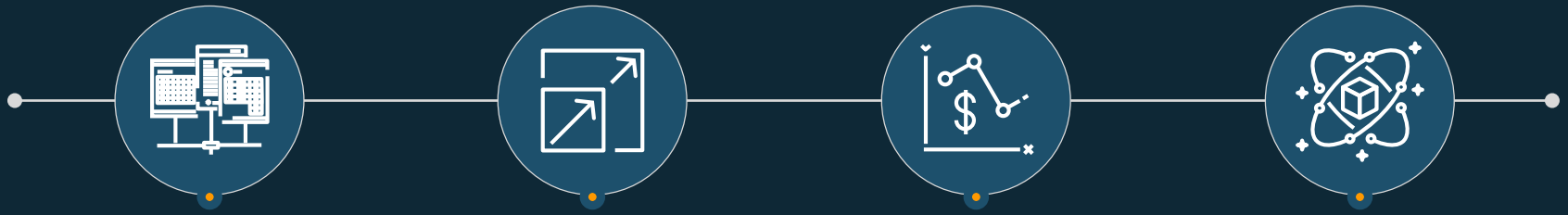
What if you could escape the bounds of on-premises?



Finite capacity
Queues
Engineering bottlenecks

Elasticity
Agility
Evolving ecosystem of services (e.g. AI/ML)

Taking advantage of the cloud for HPC workloads



Infrastructure choice

Scale

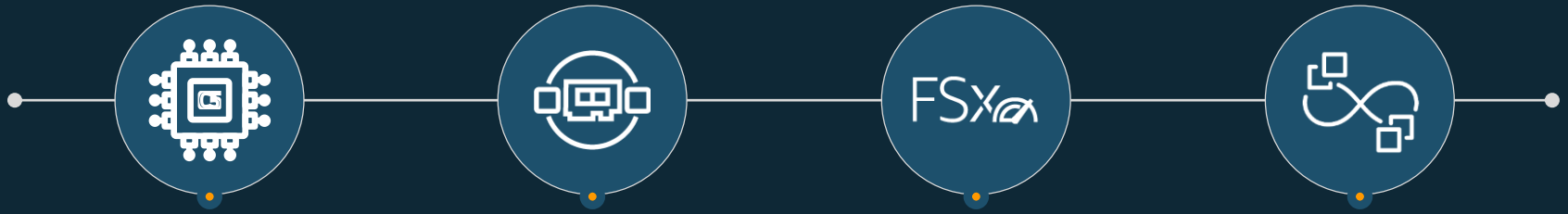
Pricing models

Agility

Rethink your workflow architecture

aws

Continuously improving services for HPC



Compute

EC2 instances for every need with Intel, NVIDIA, Graviton, AMD, Inferentia and much more

Networking & Data Transfer

Elastic Fabric Adapter
AWS DataSync
AWS Snowball
AWS Snowmobile

Data Management

Amazon FSx for Lustre
Amazon EBS
Amazon EFS
Amazon S3
Amazon Glacier

Orchestration, Optimization and Visualization

AWS Batch
AWS ParallelCluster
AWS EnginFrame
DCV & Appstream
AWS Spot

Elastic Fabric Adapter

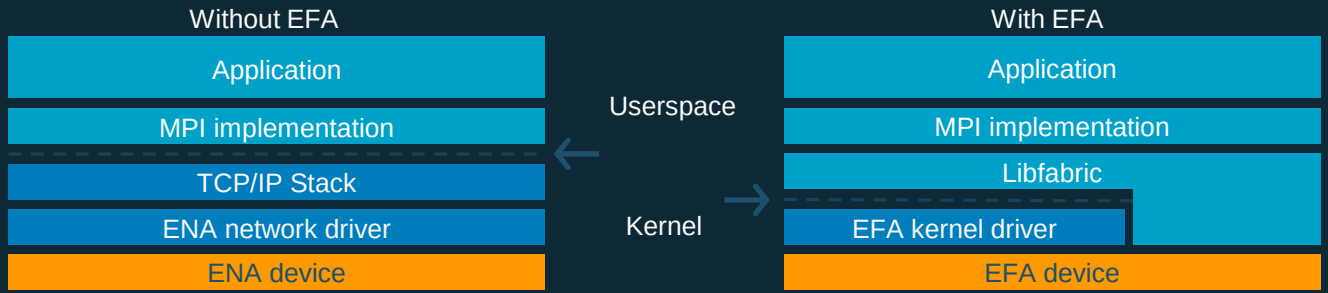
SRD protocol

Proving myths about latency constraints wrong

CFD

Seismic

Weather modeling



MAXAR

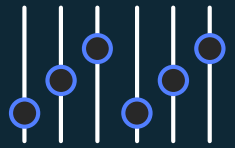
Scale tightly coupled HPC applications on AWS

aws

Amazon FSx for Lustre

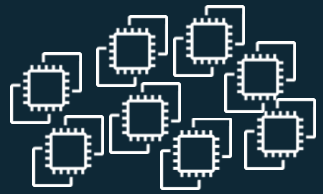
High and scalable performance

Parallel File System



SSD-based

100+ GiB/s throughput
Millions of IOPS
Consistent sub-millisecond latencies

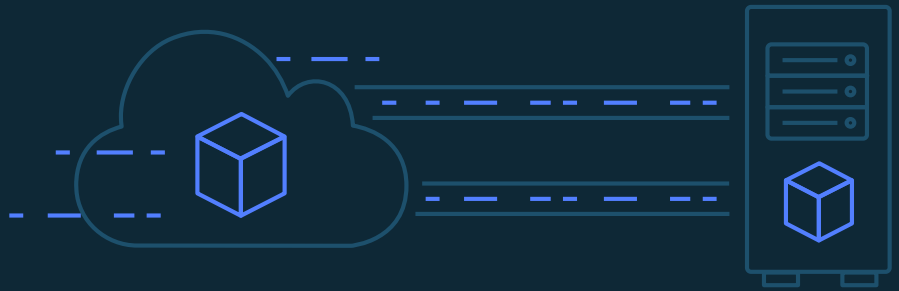


Supports concurrent access from hundreds of thousands of cores

CONDUCTOR

Conductor Technologies accelerates rendering workloads by up to 4X using Amazon FSx for Lustre

AWS ParallelCluster



One-stop shop to set up your HPC Cluster

Easy integration with AWS services

- Amazon FSx for Lustre
- Amazon EC2 Instances
- Elastic Fabric Adapter
- NICE DCV
- AWS Batch

MAXAR

Tightly coupled workloads

illumina®

Loosely coupled workloads

SCHRÖDINGER.

Accelerated computing

AUTODESK.

Visualization

F1™

AI/ML

FINRA

High-volume data analytics

aws

INEOS | TEAM UK

INEOS TEAM UK accelerates boat design for America's Cup using AWS

" Working with AWS has given us access to more and faster computational resources, which has proven crucial in developing the fastest race boat possible. It has helped the team push ahead as we continue to design and develop our race boat for the America's Cup. "

Sir Ben Ainslie *Team Principal and Skipper*

aws

Formula 1 / AWS Partnership – All about Data. Large-Scale CFD Simulations

- Simulation time reduced from **60hrs** to **18hrs** thanks to AWS.
- Two car turbulence simulations **increased from 1 to 5** on AWS
- 192 cores to 1440 cores (**C5n with EFA + AWS ParallelCluster**)
- For more on how Formula 1 is using AWS, **visit: aws.com/f1**

" This project with AWS was one of the most revolutionary in the history of Formula 1 aerodynamics "

***Pat Symonds**, Chief Technical Officer of Formula 1*

TOYOTA RESEARCH INSTITUTE

Toyota Research Institute chooses FSx for Lustre to reduce object recognition machine learning training times

“We needed a parallel file system for our ML training data sets and chose Amazon FSx for Lustre... the integration with AWS services, including S3, also made it the preferred option for our high performance file storage.”

David Fluck, Software Engineer
Toyota Research Institute

aws

Standard Chartered cuts Risk Grid Costs 60%

Re-Imagining the HPC Workflow

"Grid computing is an ideal workload for the cloud, so working with AWS technology was the easy part."

Richard Davis
Global Head of Technology Services, Standard Chartered

Innovative Cloud HPC architecture with Amazon Spot, Lambda, & RDS

aws

FactSet migrates its Real-Time Ticker Plant to AWS

"Moving a full ticker plant onto the cloud is the holy grail of market data engineers. Many people thought it would be impossible. … AWS provides a stable and secure environment for processing major exchange feeds. We are confident that this groundbreaking project will help us drive greater speed and efficiency to accelerate our clients' digital transformations."

***Gene Fernandez**, Chief Product and Technology Officer, **FactSet**.*

Challenge

FactSet's Ticker Plant, ingests and delivers live market data from exchanges globally

Solution

The company will rely on AWS services and AWS EC2 instances to build their **Real-Time Ticker Plant** service on AWS.

Benefits

- Increased scalability of the Service
- Increased overall performance
- First Global Ticker Plant of this kind on the Cloud

aws

AstraZeneca raising the bar running their Genome sequencing pipeline on AWS

- Goal: To analyze 2M genomes by 2026. Scale and Orchestration to build the fastest and most efficient sequence data pipeline in the industry.
- Decreased processing time by 2400% from 20 days to 20 hours for 20,000 samples
- Providing scientists with advanced access to the clinical effects of natural mutations in humans that mimic drug inhibition/suppression for the majority of human genes
- FPGA instances for compute and Step Functions, Lambda, S3, SQS and AWS Batch.

"Building our bespoke analysis pipeline on AWS with the support of AWS Professional Services has enabled us to achieve acceleration of key objectives towards the analysis of up to 2 million genomes (AstraZeneca's Genomics Initiative)."

Slavé Petrovski

PhD Vice President and Head of Genome Analytics and Bioinformatics, Discovery Sciences, R&D

Lyft increases simulation capacity, lowers costs using Amazon EC2 Spot Instances

Challenge

Rideshare company Lyft runs millions of compute-intensive simulations each year to improve the performance and safety of its self-driving system requiring computing power that could scale up and down at an affordable price.

Solution

The company significantly increased its AV simulation testing while reducing the corresponding computing costs by two-thirds with Amazon EC2 Spot Instances and Amazon EKS.

Benefits

- Reduced compute costs by two-thirds
- Scaled up computing capacity significantly
- Increased velocity of development for AVs

lyft Level 5

Company: Lyft

Industry: Transportation & Logistics

Country: United States

About

Lyft, one of the largest transportation networks in the United States and Canada, is on a mission: improve people's lives with the world's best transportation. It provides shared rides, electric scooters, bikeshare systems, and public transit partnerships.

"About 77% of our computing fleet is now on Amazon EC2 Spot Instances. We were able to scale up our computing capacity significantly **while reducing the overall cost of operation**."

Timothy Perrett, Level 5 senior staff engineer, Lyft

aws

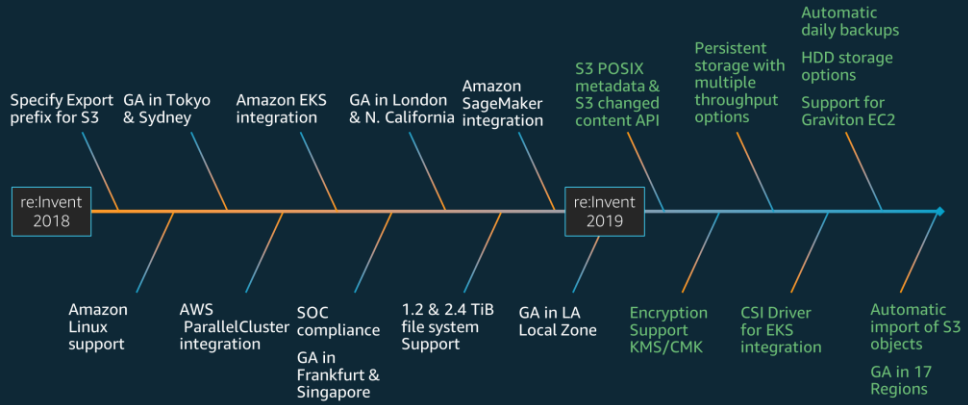
The pace for innovation never slows

Batch

- Graviton2, AMD, and Inferentia support
- Increased capability for high-throughput workloads
- Custom Log Drivers (Splunk, Fluentd),
- New console: visibility improvements, sorting, and filtering
- Custom Retry capabilities

ParallelCluster 2.9

- Multiple Job Queues
- Multiple instance types per queue
- Support for Graviton2
- DCV for visualization
- FSx for Luste Integration
- Cloudwatch Logs
- Amazon Linux 2, Ubuntu 18 support



FSx for Lustre Innovations

Interesting public resources

Come regularely!
It changes everyday!

https://aws.amazon.com/fr/hpc/

https://pages.awscloud.com/hpc-cae-credit-code-2020.html

https://d1.awsstatic.com/whitepapers/benchmarking-aws-and-hpc-services.pdf?did=wp_card&trk=wp_card

https://aws.amazon.com/blogs/opensource/category/compute/aws-parallel-cluster/

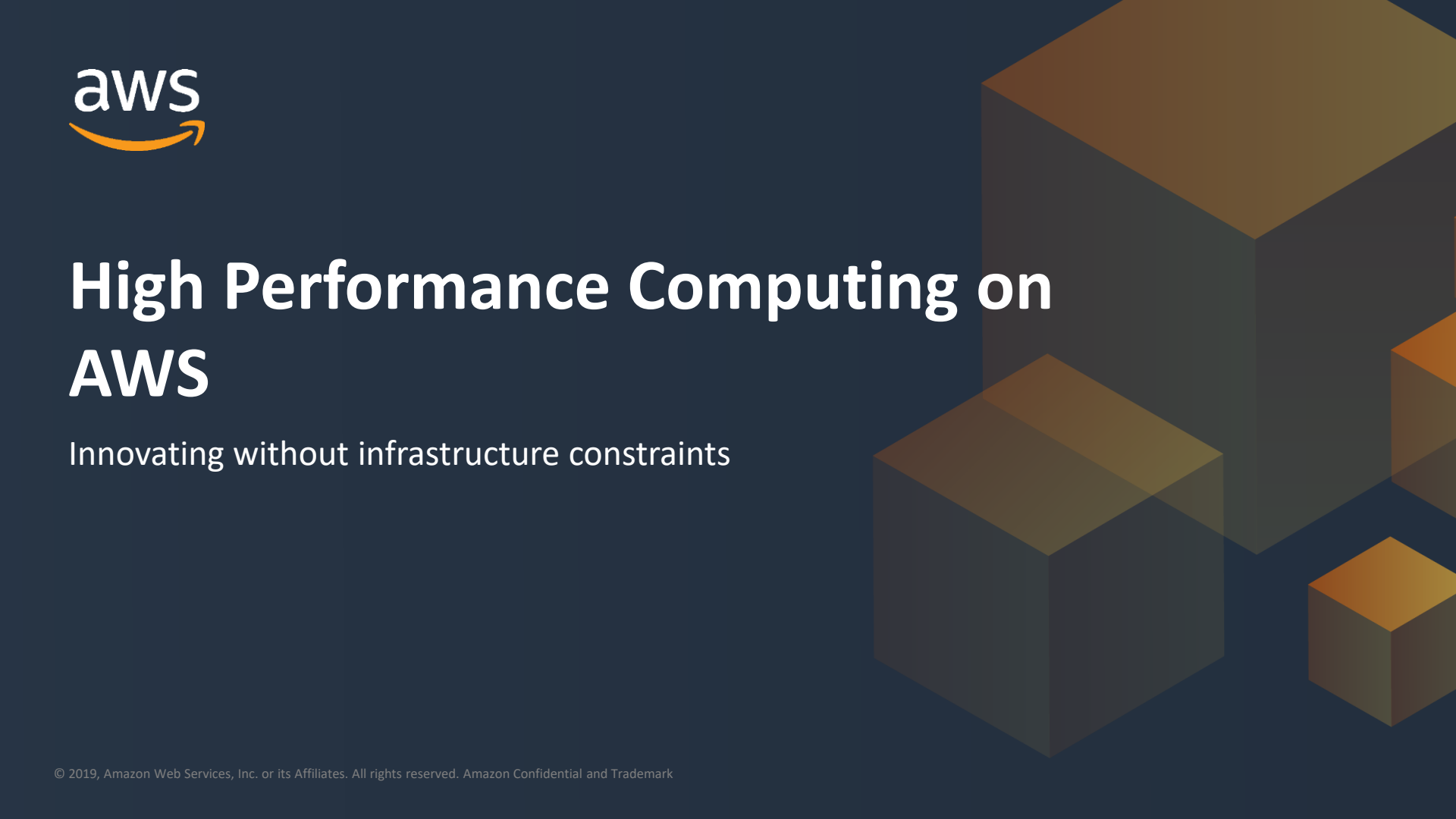
https://aws.amazon.com/blogs/storage/category/storage/amazon-fsx-lustre/

https://aws.amazon.com/blogs/compute/category/compute/aws-batch/

https://aws.amazon.com/blogs/compute/category/compute/high-performance-computing/

aws

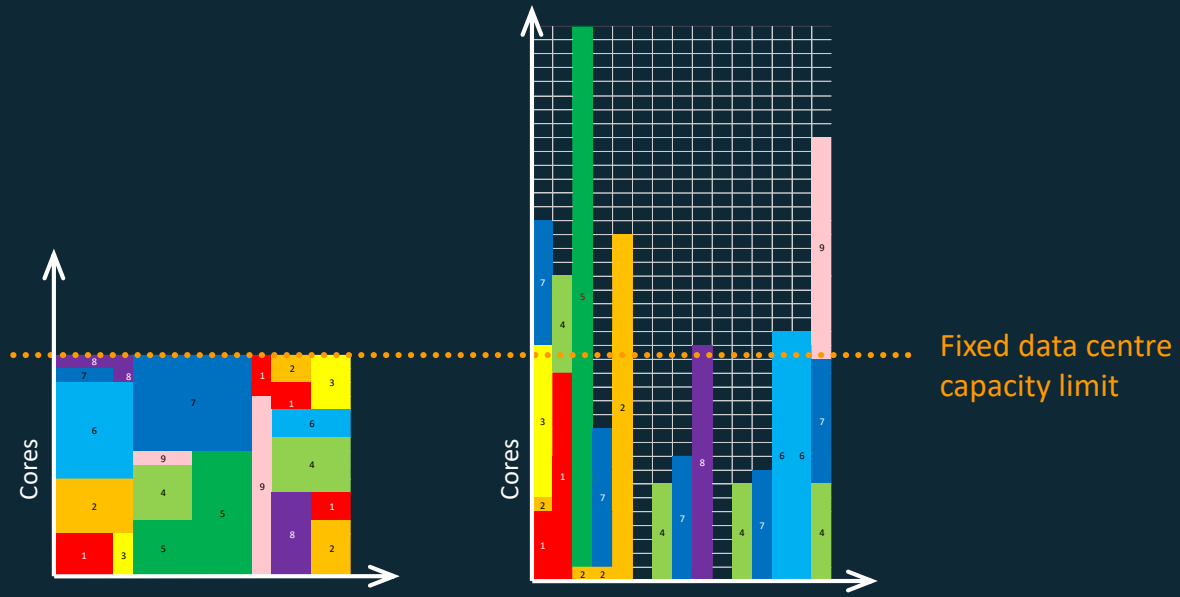
Thank You!



High Performance Computing on AWS

Innovating without infrastructure constraints

We think the metric for success for any business 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 $463 in incremental revenues and $44 in incremental profit."

—Hyperion Research, 2018

Because, a TCO analysis never tells the whole story

Lost productivity & longer time to results

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

Lost innovation

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

Outdated technology

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

Technical debt

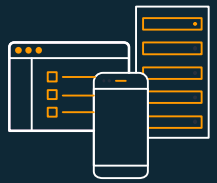
Adapting newer algorithms to meet the requirements of an existing infrastructure = delays, and below-par performance.

* Source: Hyperion Research, 2018

aws

HPC on AWS is a fundamental rethink of what is possible

From worrying about



Capex Capacity Technology

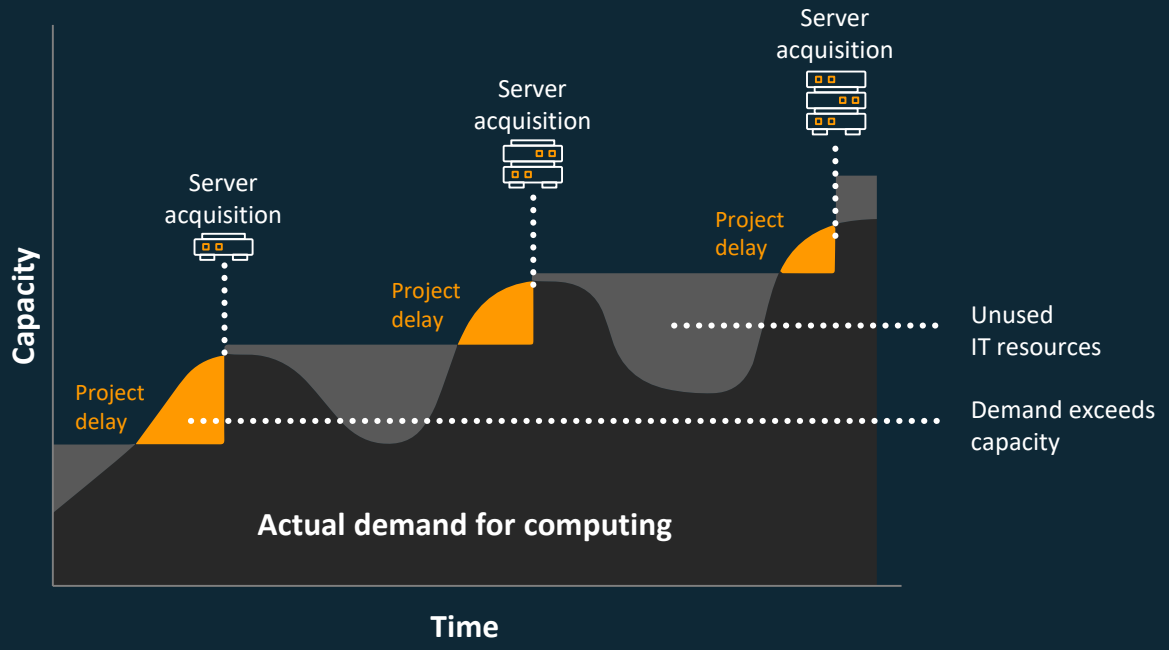
to

Focusing on innovation

aws

Matching demand to capacity is challenging

Attempting to match variable computing demand to static on-premise compute grids is extremely difficult, and adding capacity is time- and capital-intensive



And the choices poor

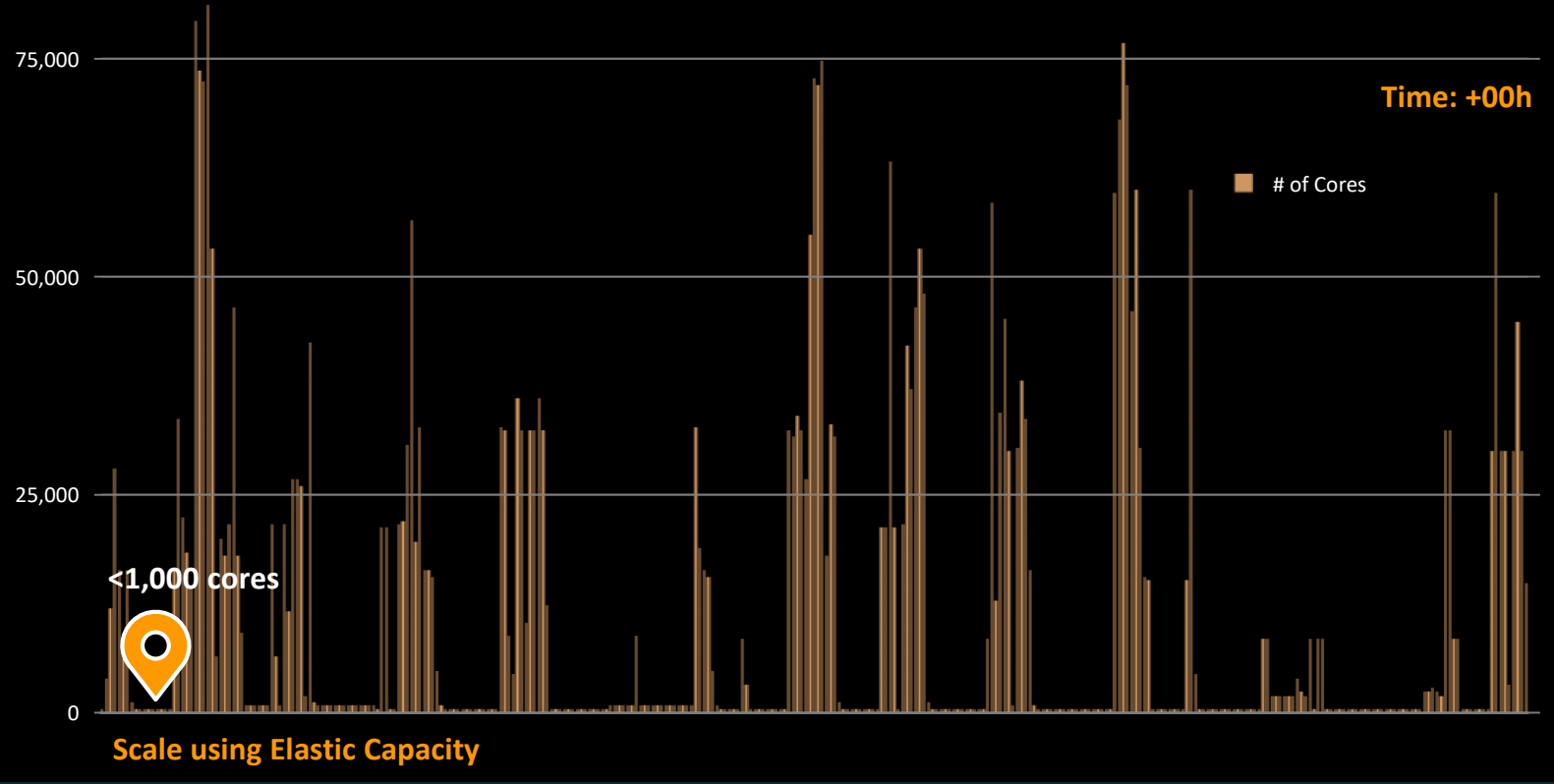
Optimize for availability

Usage remains low which drives up costs

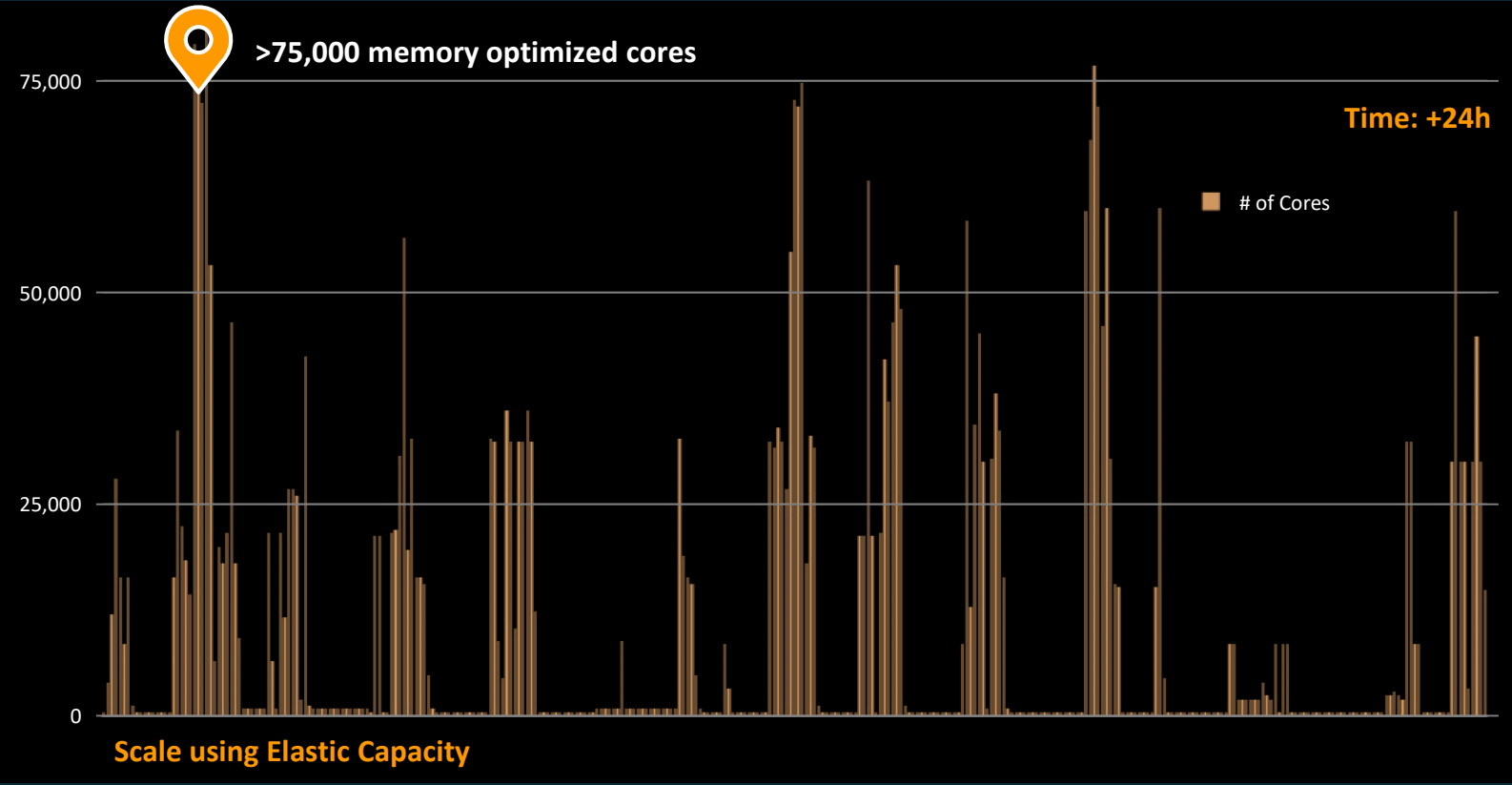
Optimize for cost

Utilization is high but users may wait a long time to access the system

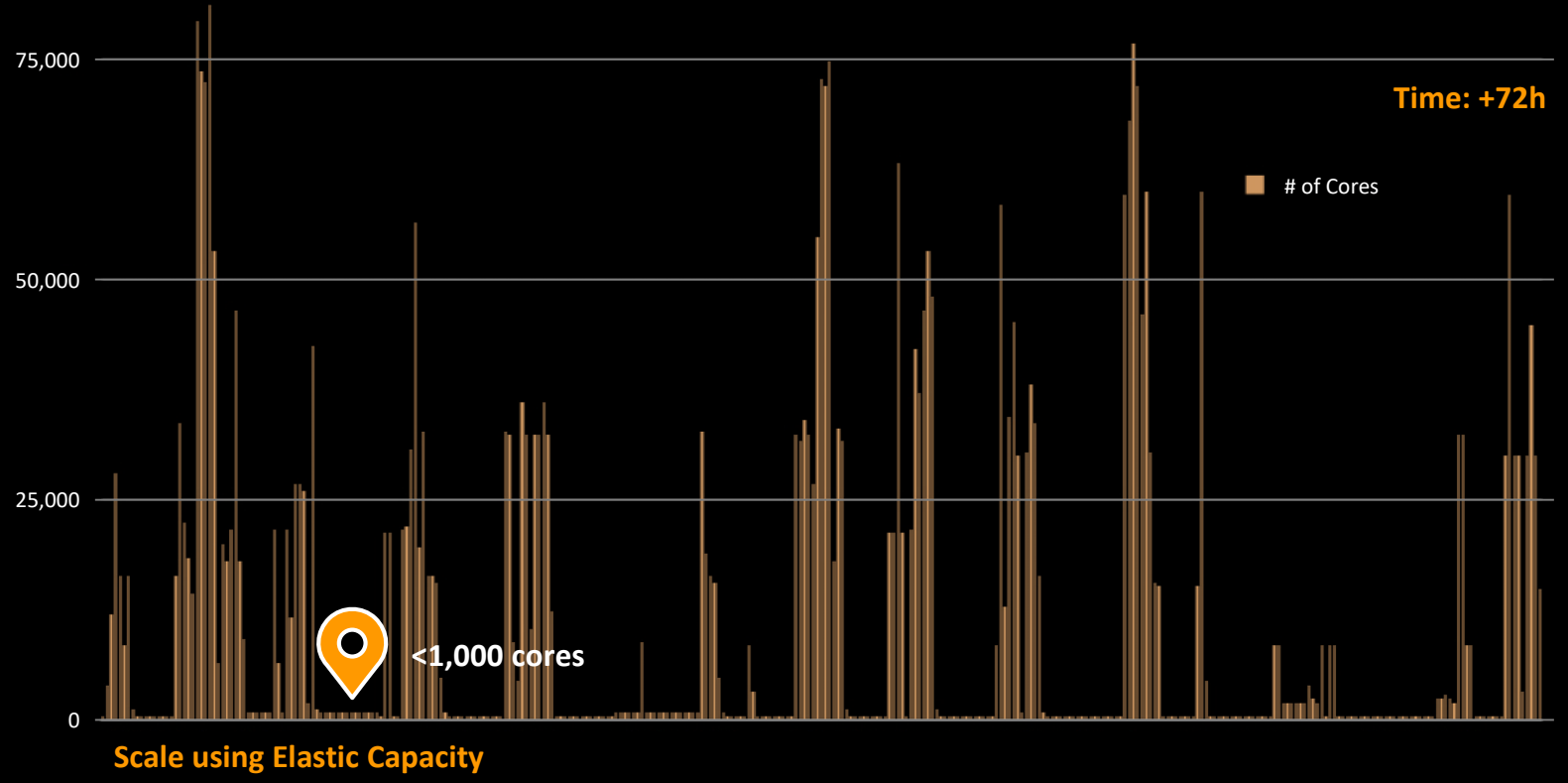
What does an ideal solution look like?



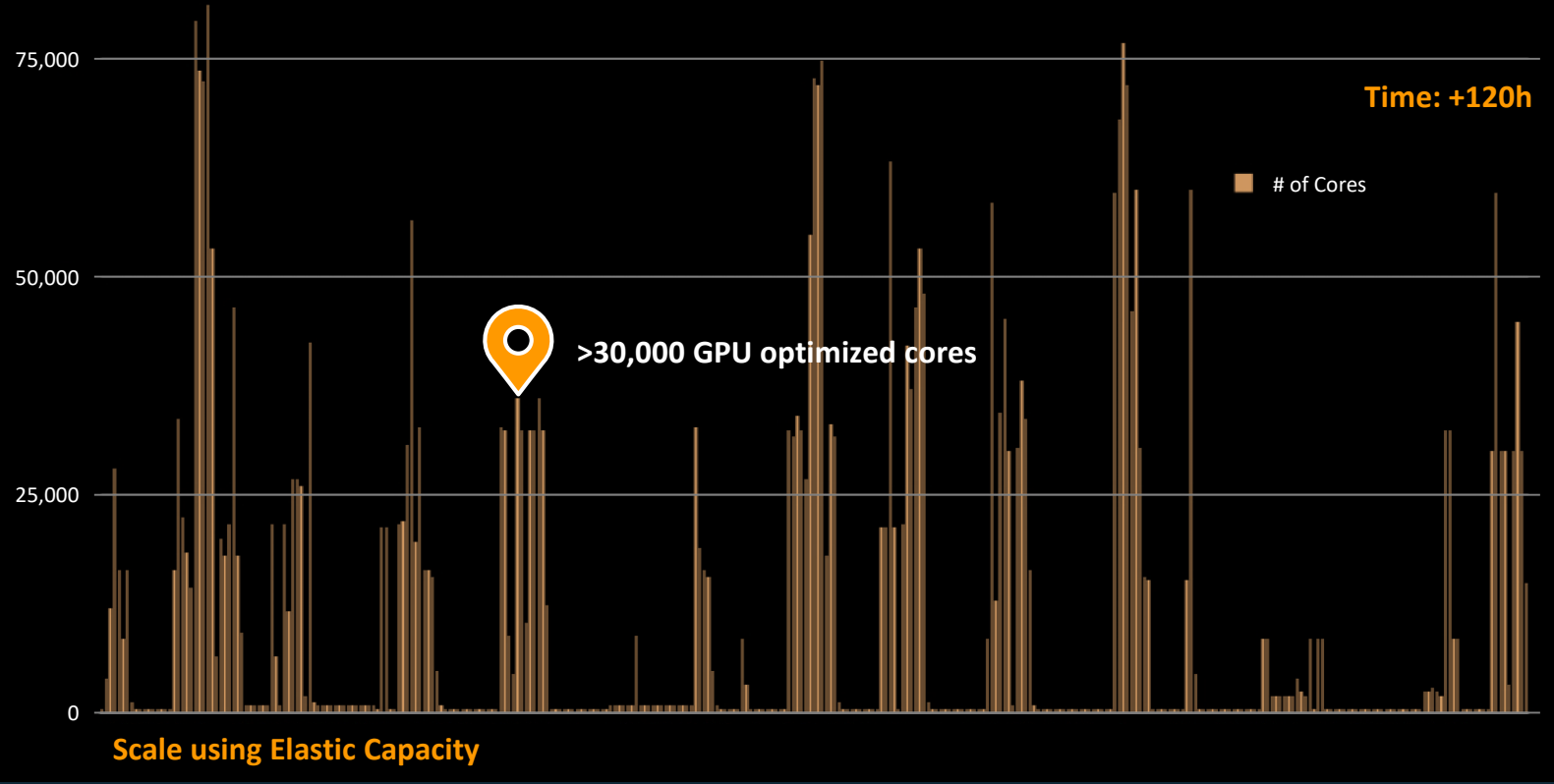
What does an ideal solution look like?



What does an ideal solution look like?



What does an ideal solution look like?



HPC on AWS

Accessible and on-demand elastic compute power

What if you could add thousands of cores of your choice to your HPC infrastructure as you needed them, remove them when you are done, and only pay what you used?

How would that kind of additional compute power improve your time-to-results ?

aws

Why HPC on AWS?

Virtually unlimited infrastructure enabling scaling and agility not attainable on-premises

Better ROI

Instant access to latest technologies with no lengthy procurement cycles or big capital investments

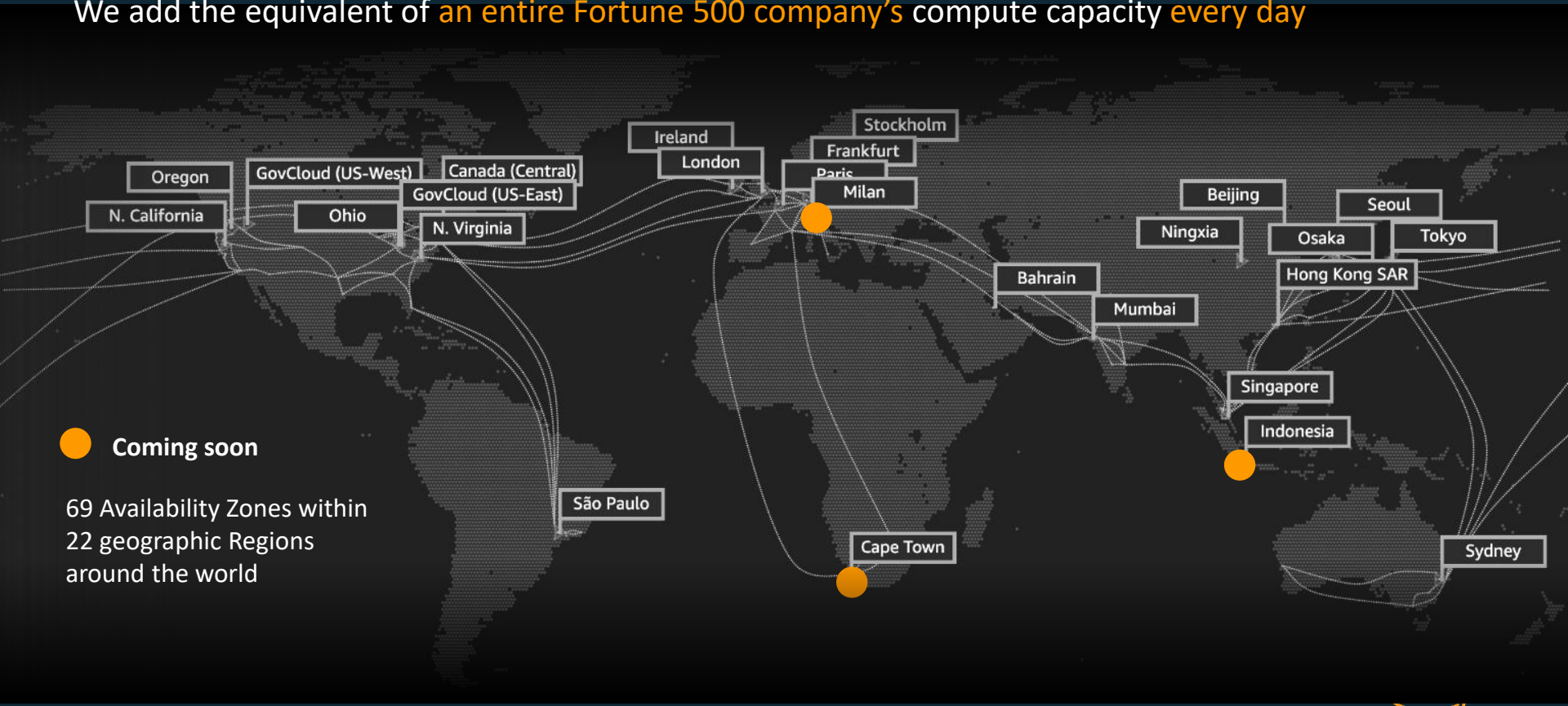
Faster time to results

Flexible configuration options quickly iterate resource selection and ensure cost optimization

aws

Global Infrastructure

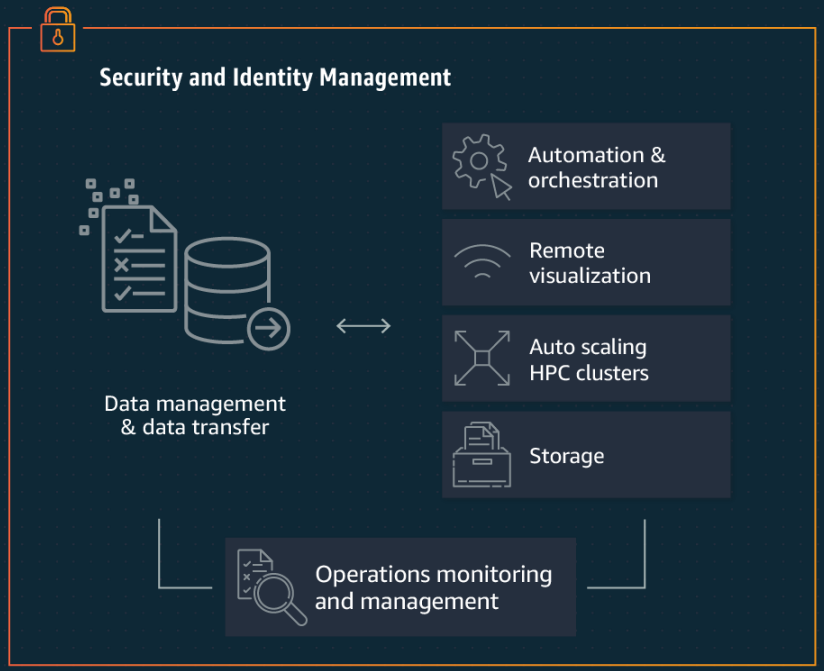
We add the equivalent of an entire Fortune 500 company's compute capacity every day



69 Availability Zones within
22 geographic Regions
around the world

High Performance Computing on AWS

AWS & Partner service options for every layer of the HPC stack



AWS Services to get started with HPC on AWS

Amazon CloudWatch

Data management & data transfer	Compute & networking	Storage	Automation & orchestration	Visualization
AWS DataSync	Amazon EC2 instances (CPU, GPU, FPGA)	Amazon EBS	AWS Batch	NICE DCV
AWS Snowball	Amazon EC2 Spot	with provisioned IOPS	AWS ParallelCluster	Amazon AppStream 2.0
AWS Snowmobile	AWS Auto Scaling	Amazon FSx for Lustre	NICE EnginFrame	
AWS DirectConnect	Placement groups	Amazon EFS		
	Enhanced networking	Amazon S3		
	Elastic Fabric Adapter			

Amazon IAM (Identity and Access Management)

AWS Budgets

aws

Broad HPC partner community

Application partners

ANSYS
MAXELER Technologies Maximum Performance Computing
aneo the other solution
Ridgeway Kite
DASSAULT SYSTEMES
RFD Rock Flow Dynamics
DNAnexus
SIEMENS Ingenuity for life
illumina
ZENOTECH SIMULATION UNLIMITED

Infrastructure partners

Altair
IBM
TIBCO
UNIVA

Technology partners

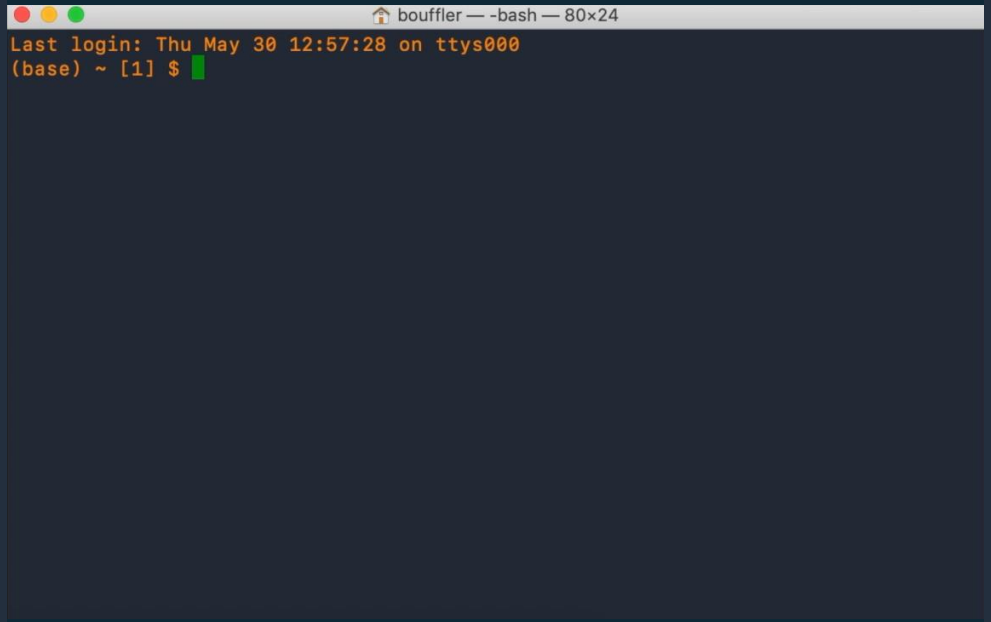
intel
NVIDIA

Consulting partners

CADFEM
DXC.technology
rescale
six nines
RONIN

aws

Infrastructure is code. Not a 5-yearly refresh



Iteratively decide on the best CPU, GPU, memory or I/O architecture for your workload

Test multiple options in **parallel** rather than sequentially

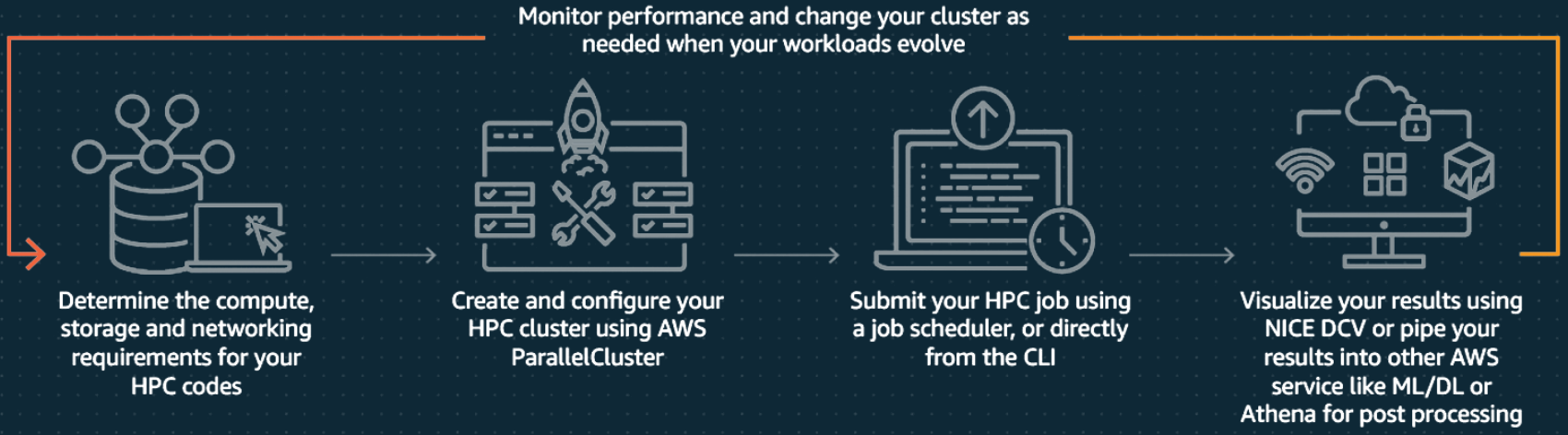
Dispose of what you don't need

Make **CI/CD** part of your HPC practice

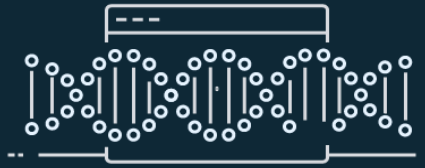
Link to tutorial <INSERT>

aws

Simple steps to get started



HPC workloads across industries



Life Sciences

Financial Services

Oil & Gas

Design & Engineering

Climate & Geosciences

Autonomous Vehicles

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

aws

HPC workloads with different compute and throughput characteristics

VOLKSWAGEN GROUP

Tightly-coupled workloads

illumina®

Loosely-coupled workloads

SCHRÖDINGER.

Accelerated computing

milk

Visualization

F1™

AI/ML

DigitalGlobe

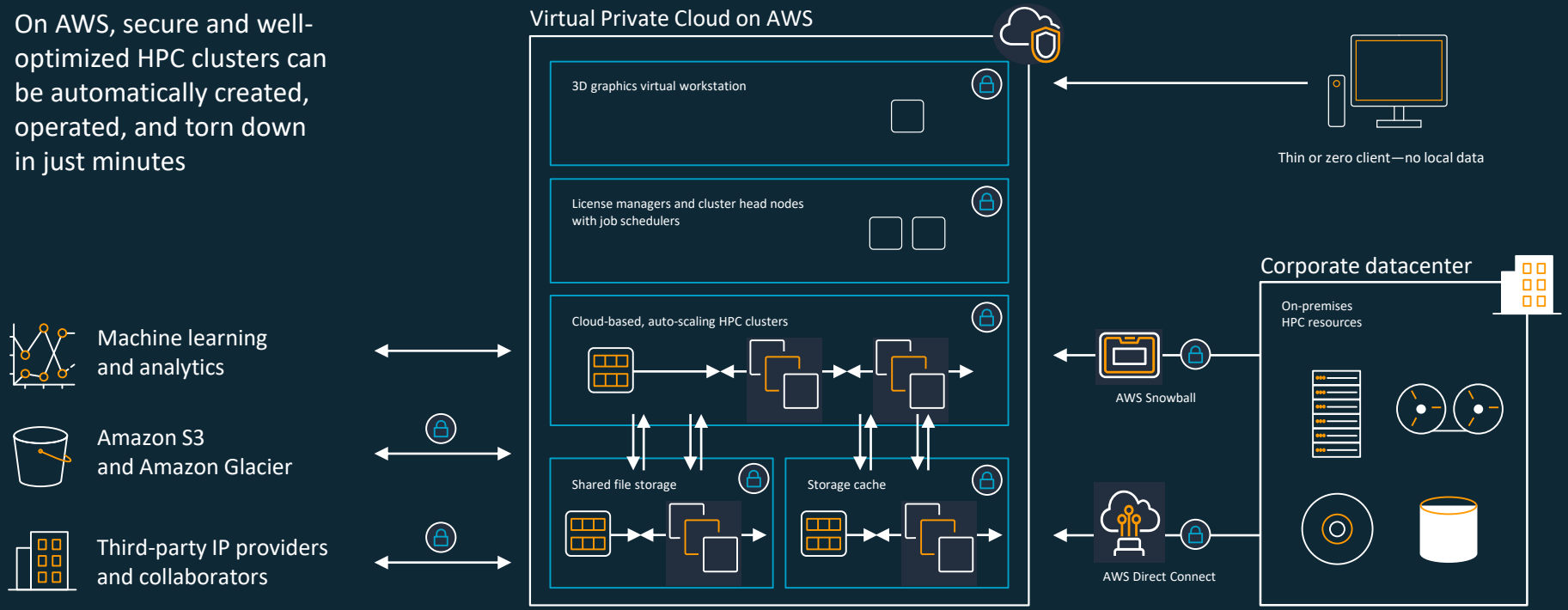
High volume data analytics

aws

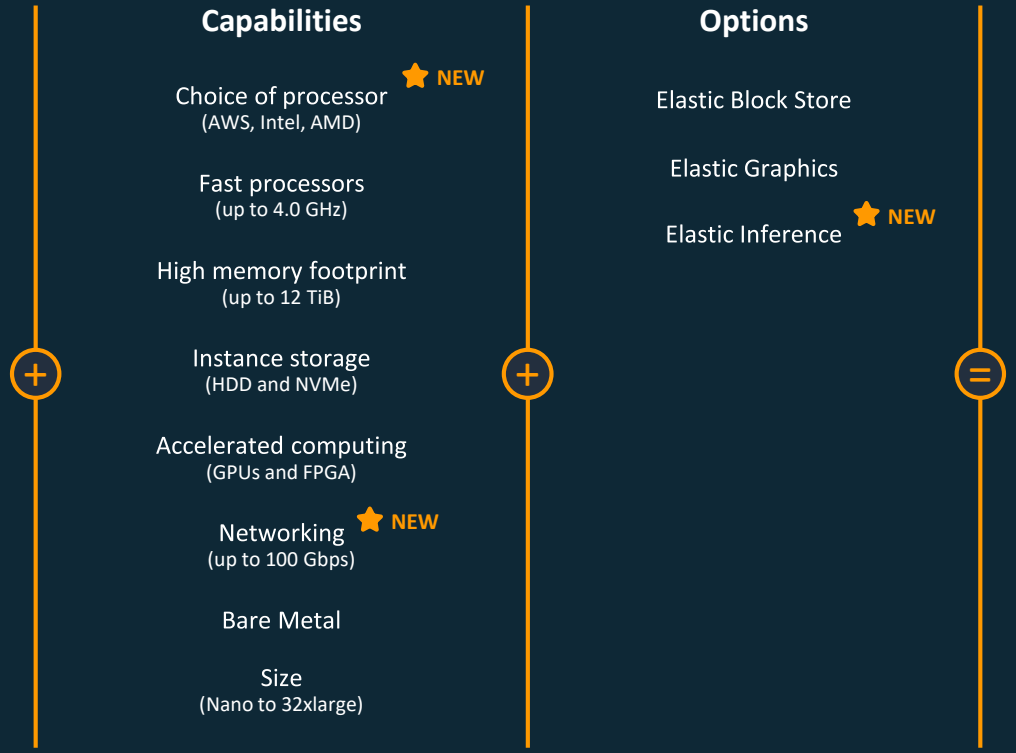
HPC on AWS: solution components

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



Broadest and deepest platform choice



200+

instance types

for virtually every workload and business need

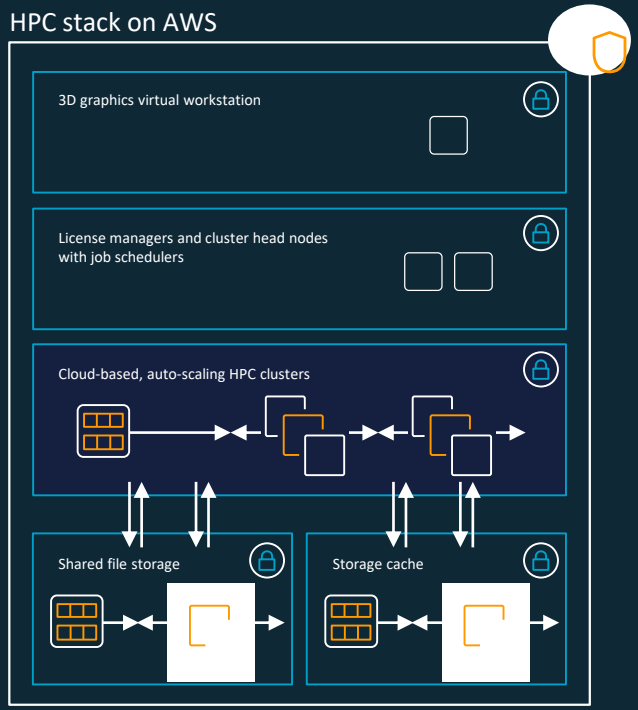
aws

High bandwidth compute instances: C5n

Massively scalable performance

- C5n Instances will offer up to 100 Gbps of network bandwidth
- 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

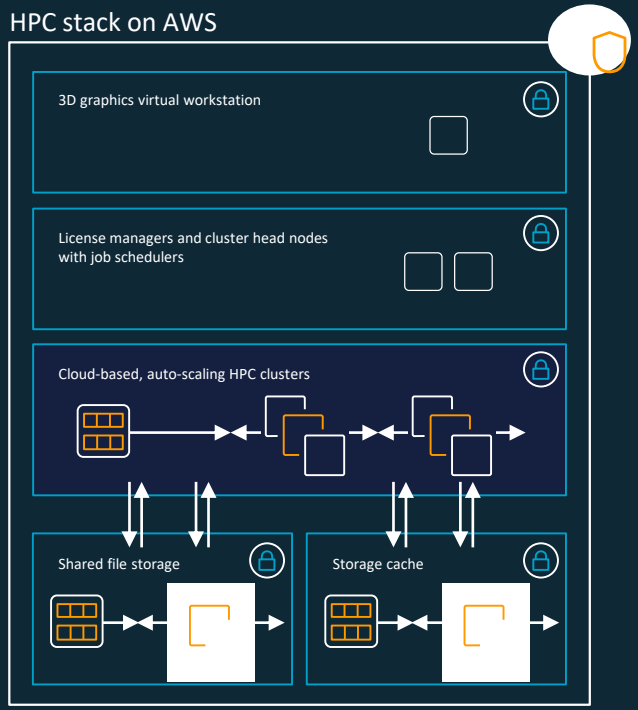
Featuring Intel Xeon Scalable (Skylake) processor



High bandwidth compute instances: P3dn

Optimized for distributed ML training

- One of the most powerful GPU instance available in the cloud
- Distributed machine learning training across multiple GPU instances
- 100 Gbps of networking throughput
- Based on NVIDIA's latest GPU Tesla V100 with 32GB of memory each
- The largest Amazon Elastic Compute Cloud (Amazon EC2) P3 instance size available

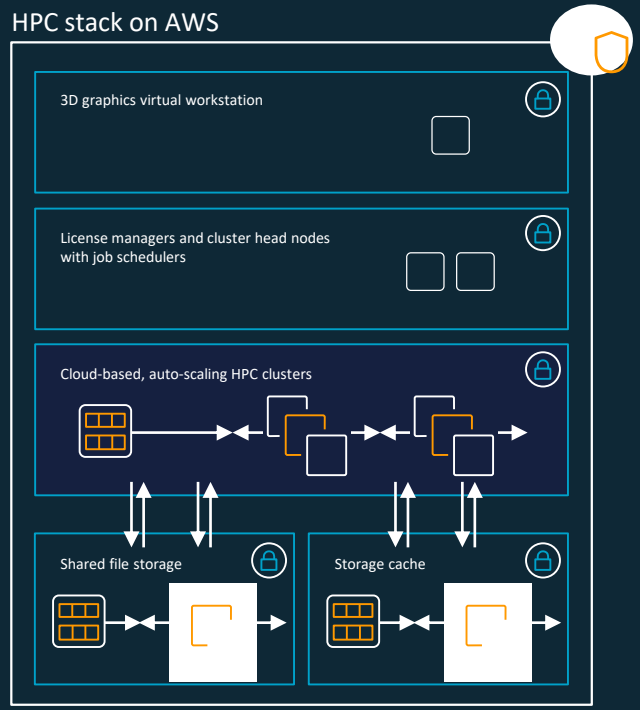


High clock speed compute instances: Z1d

Up to 4 GHz sustained, all-turbo performance

- Z1d instances are optimized for memory-intensive, compute-intensive applications
- Custom Intel Xeon Scalable processor
- Up to 4 GHz sustained, all-turbo performance
- Up to 385GiB DDR4 memory
- Enhanced networking, up to 25 GB throughput

Featuring Intel Xeon Scalable (Skylake) processor



Elastic Fabric Adapter (EFA)

Scale tightly-coupled
HPC applications on AWS

EFA

**Elastic Fabric Adapter,
best for large HPC workloads**

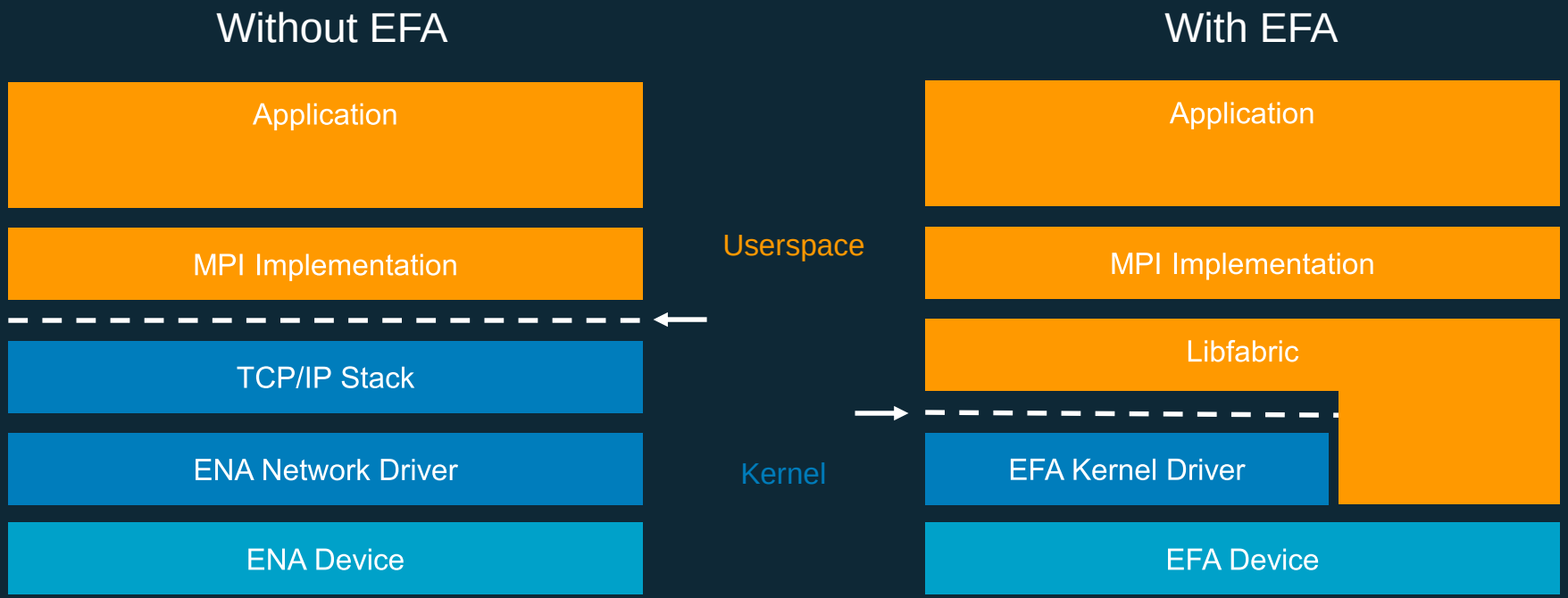
Scale tightly-coupled HPC
and ML workloads

100 Gbps network bandwidth

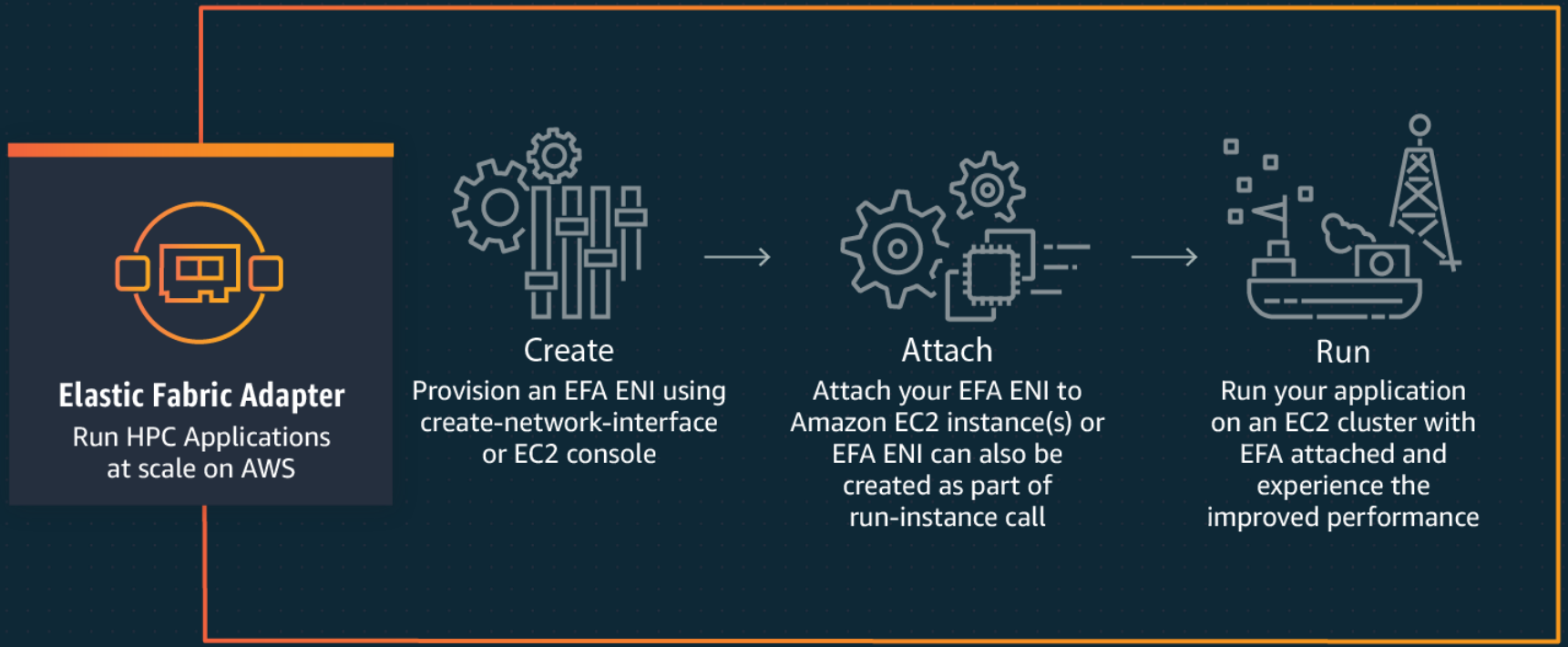
< 15 micro-seconds network latencies

aws

HPC software stack in Amazon EC2

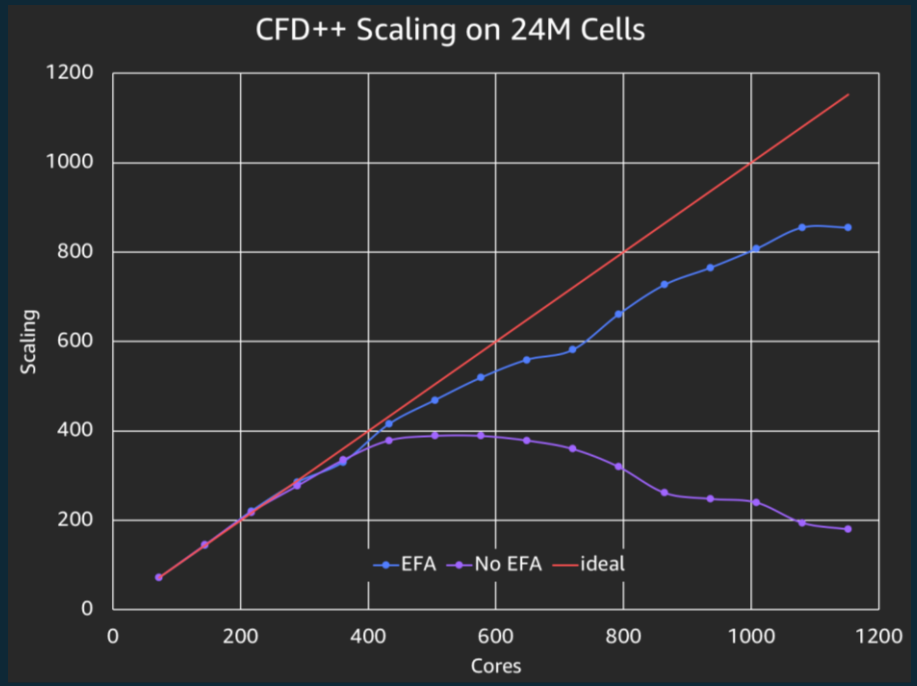


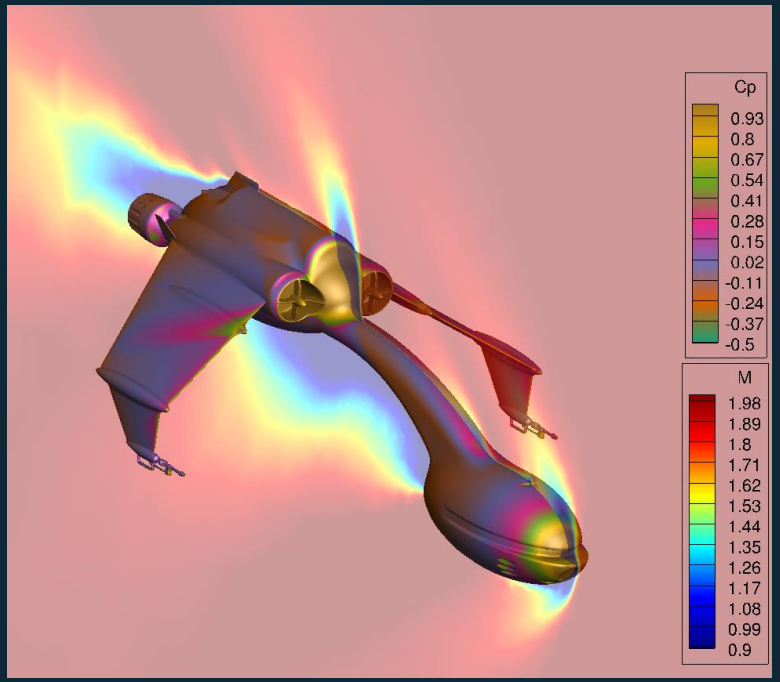
Getting started with EFA on AWS



aws

What can EFA do?





Thanks to Metacomp Technologies and the Klingon Empire.

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

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

aws

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

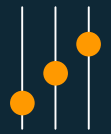
Massively scalable performance

100+ GiB/s throughput

Millions of IOPS

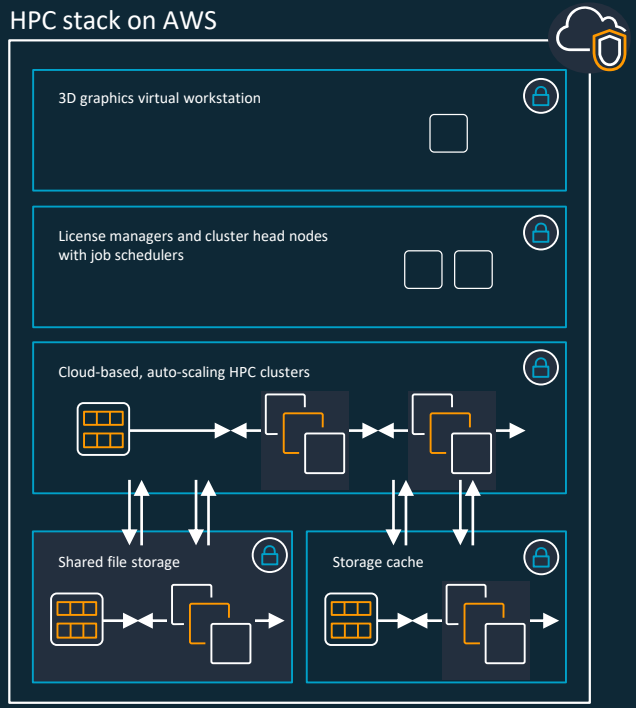
Consistent low latencies

High performing



Parallel distributed file system

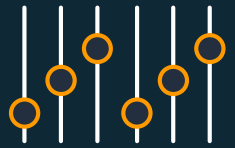
Tune complex performance parameters



High and scalable performance

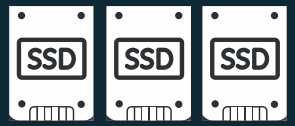
High and scalable performance

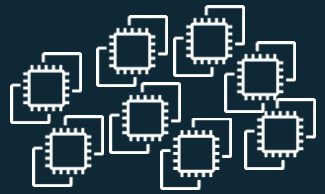
Parallel File System



100+ GiB/s throughput
Millions of IOPS
Consistent sub-millisecond latencies

SSD-based





Supports concurrent access from hundreds of thousands of cores

Each terabyte (TB) of storage provides 200 MB/second of file system throughput and ~5,000 IOPS

aws

File system throughput and IOPS scale linearly with storage capacity

Each TB of storage provides 200 MB/s of baseline throughput, and up to 12x burst throughput

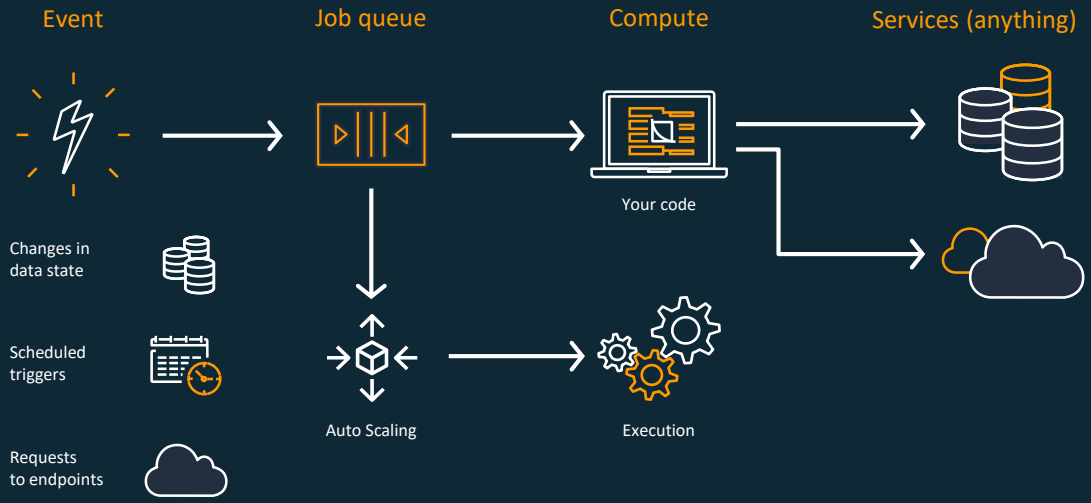
File systems can scale to hundreds of GB/s and millions of IOPS

Capacity	Baseline throughput	Burst throughput
1TB	200 MB/s	up to 2.4 GB/s
10TB	2 GB/s	up to 24 GB/s
50TB	10 GB/s	up to 120 GB/s
100TB	20 GB/s	up to 240 GB/s
1PB	200 GB/s	at least 240 GB/s

aws

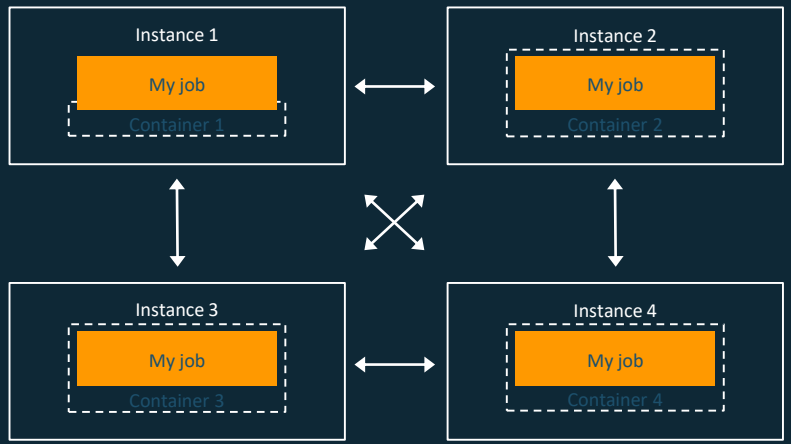
AWS Batch

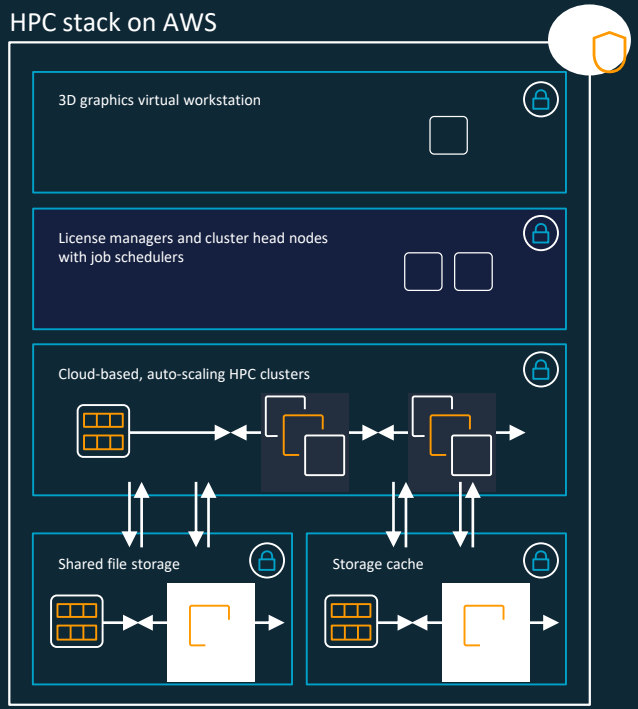
AWS Batch dynamically provisions resources, plans, schedules, and executes
No additional components to install



Efficient job scheduling: Multi-node parallel job support on AWS Batch

Simplify your compute clusters and scale jobs across multiple instances with AWS Batch support for Multi-node Parallel (MNP) jobs





Easy cluster management: AWS ParallelCluster

Simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers

Integrated with AWS Batch, Amazon FSx for Lustre and
Elastic Fabric Adapter

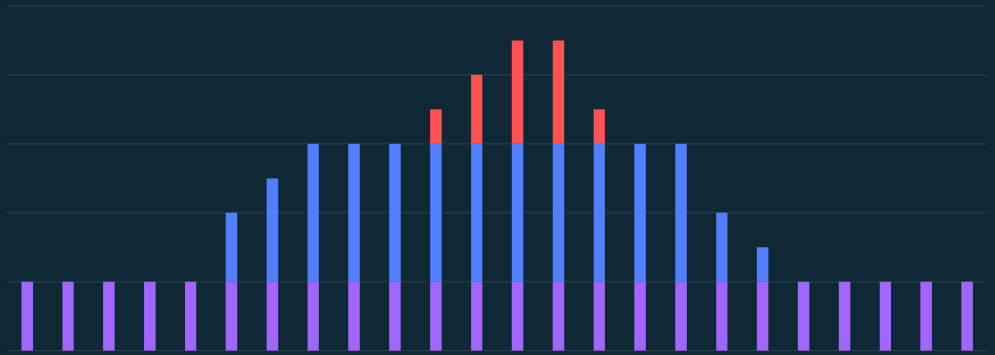
Link to tutorial <INSERT>

aws

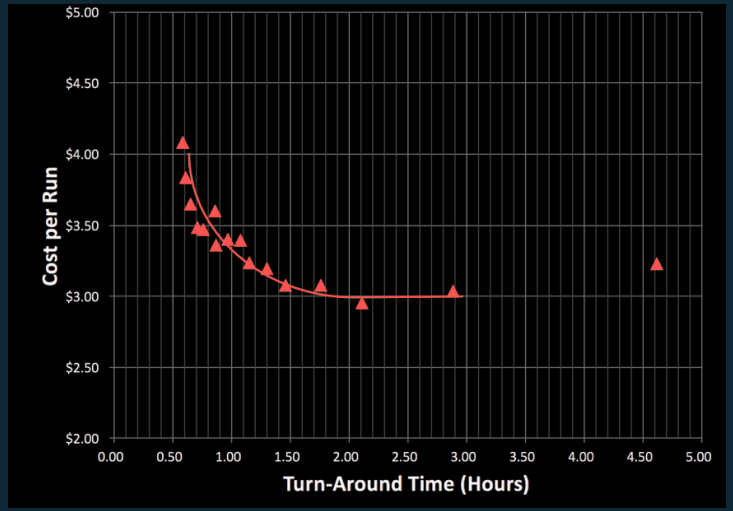
Simplifying capacity and cost optimization

Scale using Spot, On-Demand, or both

Use Reserved Instances for known/steady-state workloads

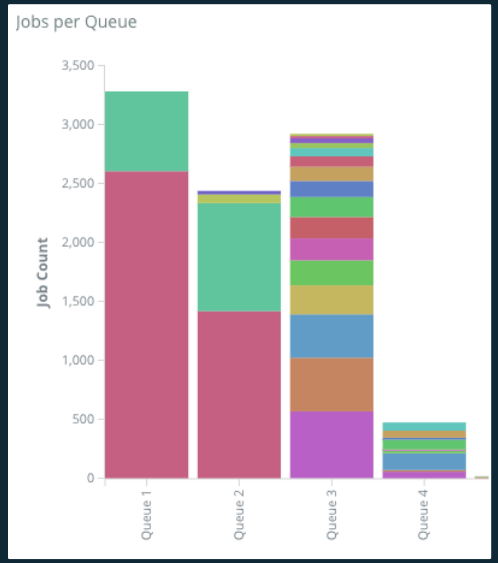


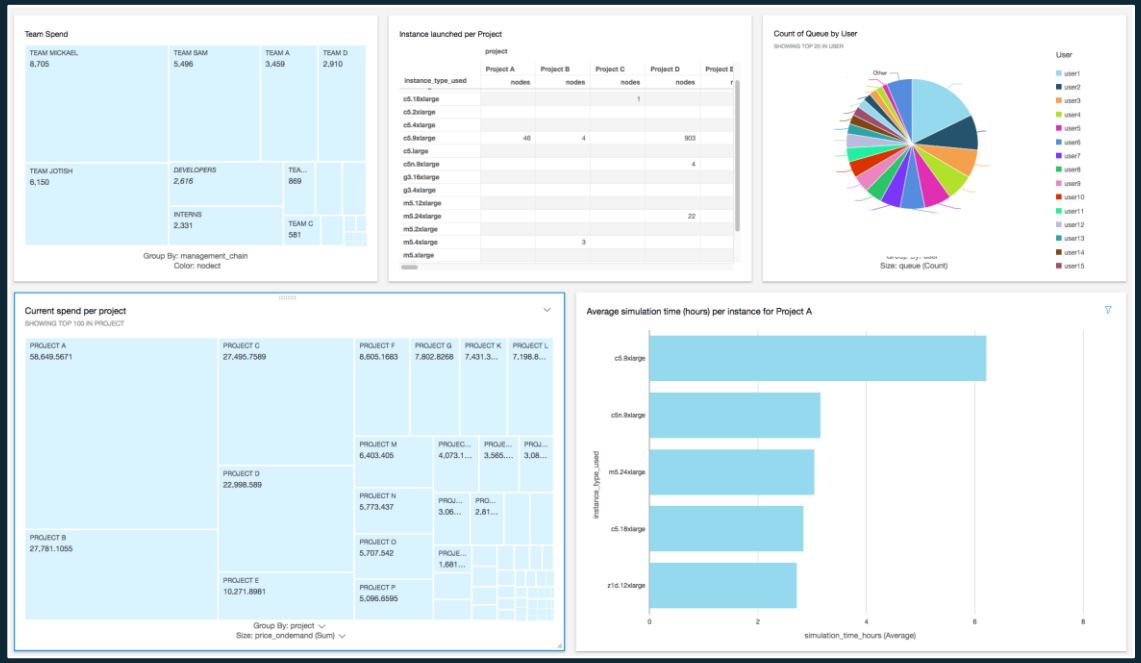
Evaluate the trade-off of time to solution vs. cost for scaling



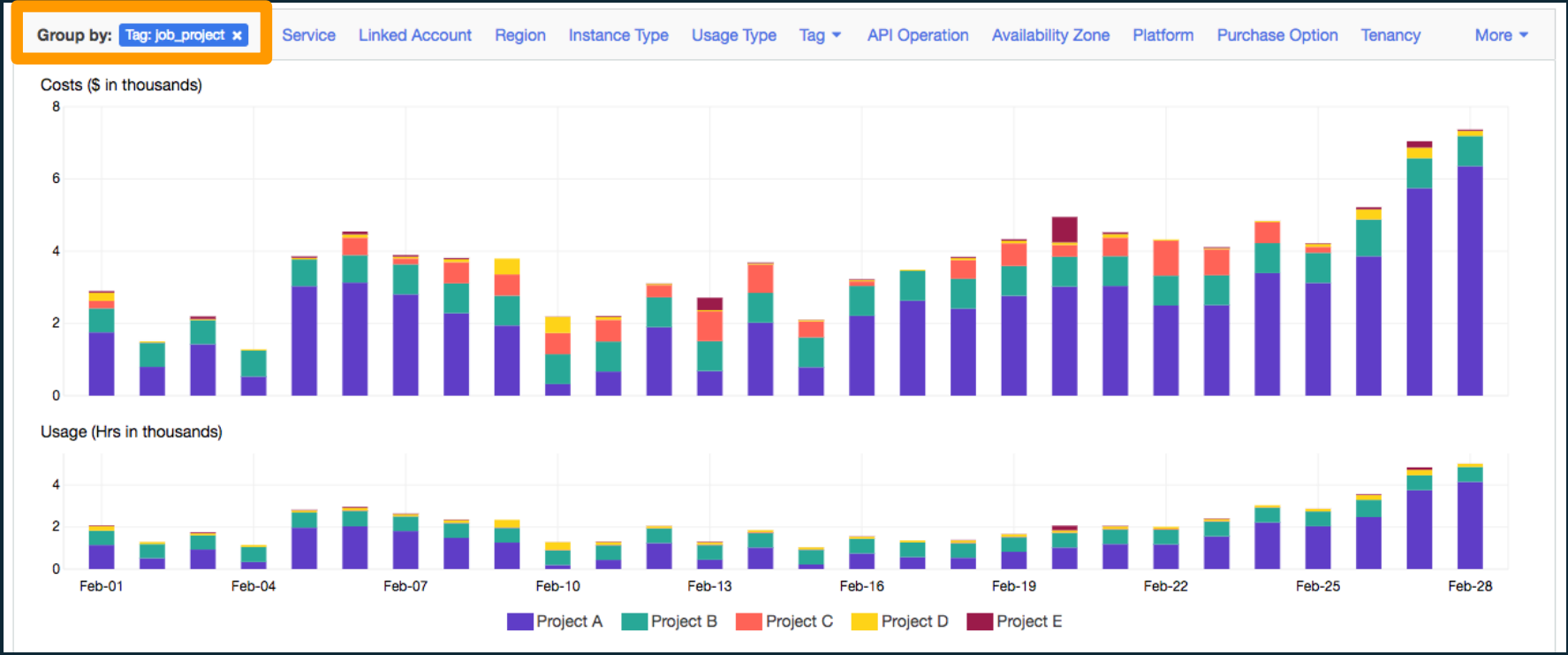
aws

Use AWS services to manage, analyze, and visualize operational metrics



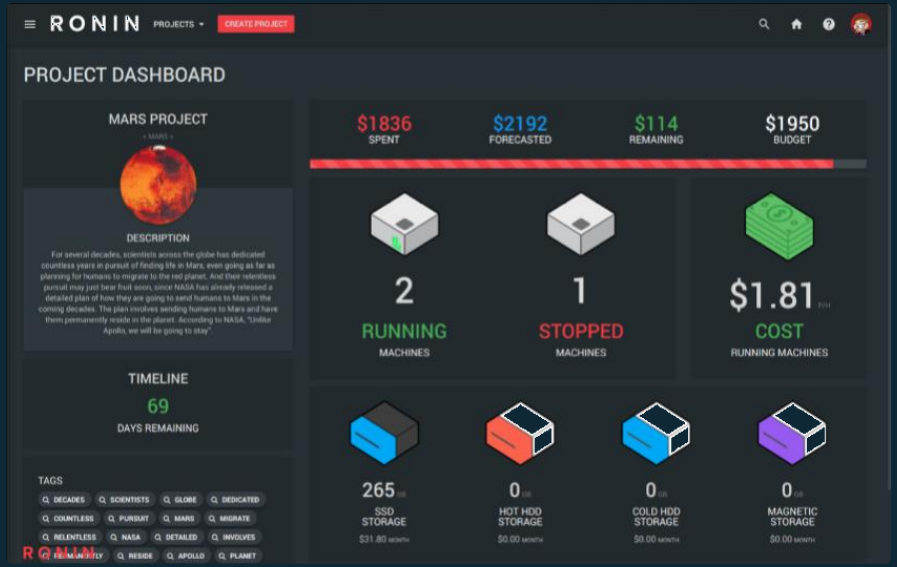


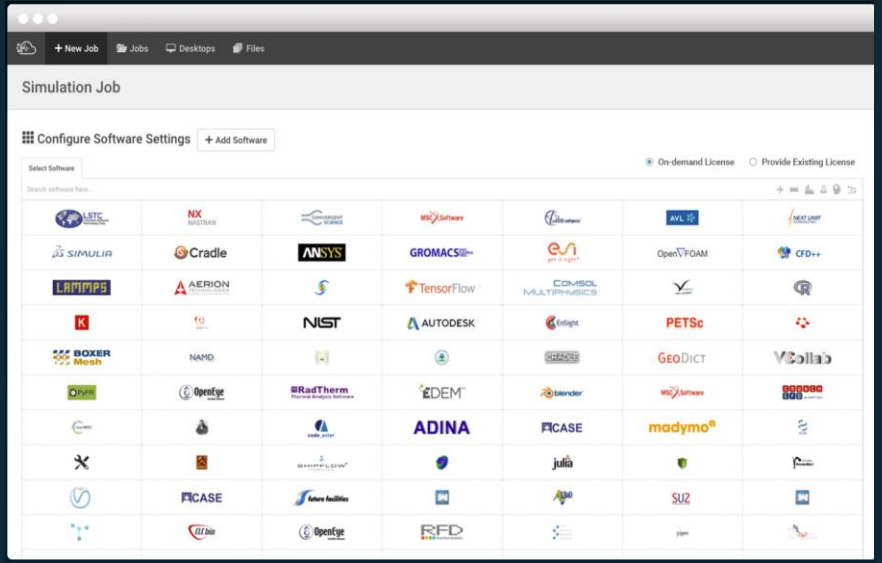
Track your spend with AWS Cost Explorer



Multiple partner options to get started

If DIY HPC clusters are not your thing, explore our rich partner community





Compliance programs

Global

Europe

Asia Pacific

United States

SOC 1

SOC 2

SOC 3

AWS is the first choice for highly regulated organizations

> We can be far more secure in the cloud and achieve a higher level of assurance at a much lower cost, in terms of effort and dollars invested. We determined that security in AWS is superior to our on-premises data center across several dimensions, including patching, encryption, auditing and logging, entitlements, and compliance.
>
> – John Brady, CISO, FINRA

Over 50 global compliance certifications and accreditations

AWS industry-leading security teams: 24/7, 365 days a year

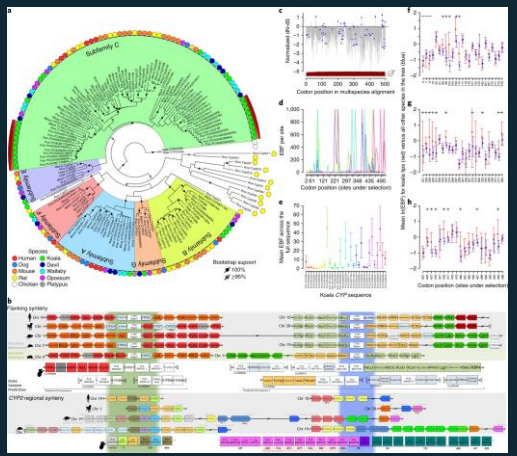
Security infrastructure built to satisfy military, global banks, and other high-sensitivity organizations

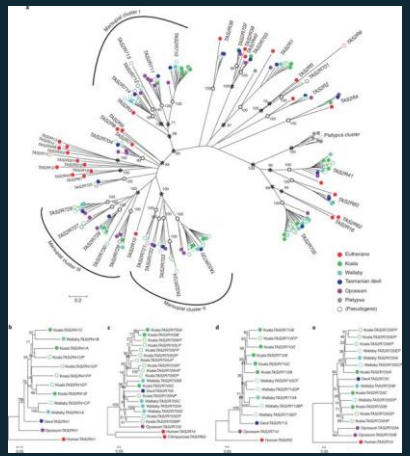
Security enhancements from 1M+ customer experiences

Driving business outcomes and making a difference

Complete sequencing of 3.24 billion base pairs

3 million core-hours of Amazon EC2 Spot capacity





https://www.nature.com/articles/s41588-018-0153-5

aws

Helping financial institutions model investment risks

Run risk models

4,000 times faster

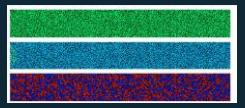
In hours, instead of months

Manage 50X the number of securities

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

HPC on AWS

Flexible configuration and virtually unlimited scalability
to grow and shrink your infrastructure as your HPC workloads
dictate, not the other way around