Accelerating Product Design with High Performance Computing on AWS

Don’t let your innovation get stuck in a queue !

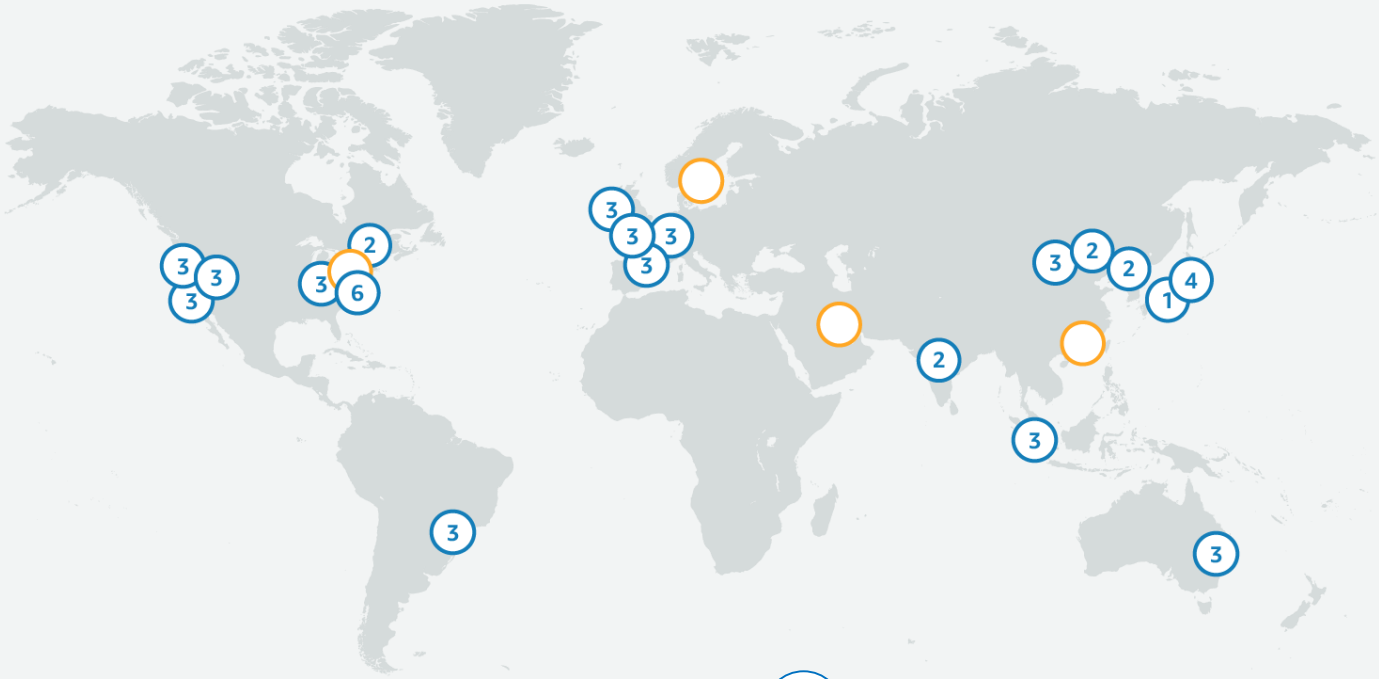
Dave Pellerin
Global Head of Business Development, Infotech/Semiconductors
dpelleri@amazon.com

Agenda

- **Overview of AWS Infrastructure**
- HPC Solution Overview
- HPC Use Cases in Design and Engineering
- Customer Success Stories
- Best Practices for Performance, Scalability, and Cost
- Additional Resources

AWS Global Infrastructure

18 Regions – **55** Availability Zones

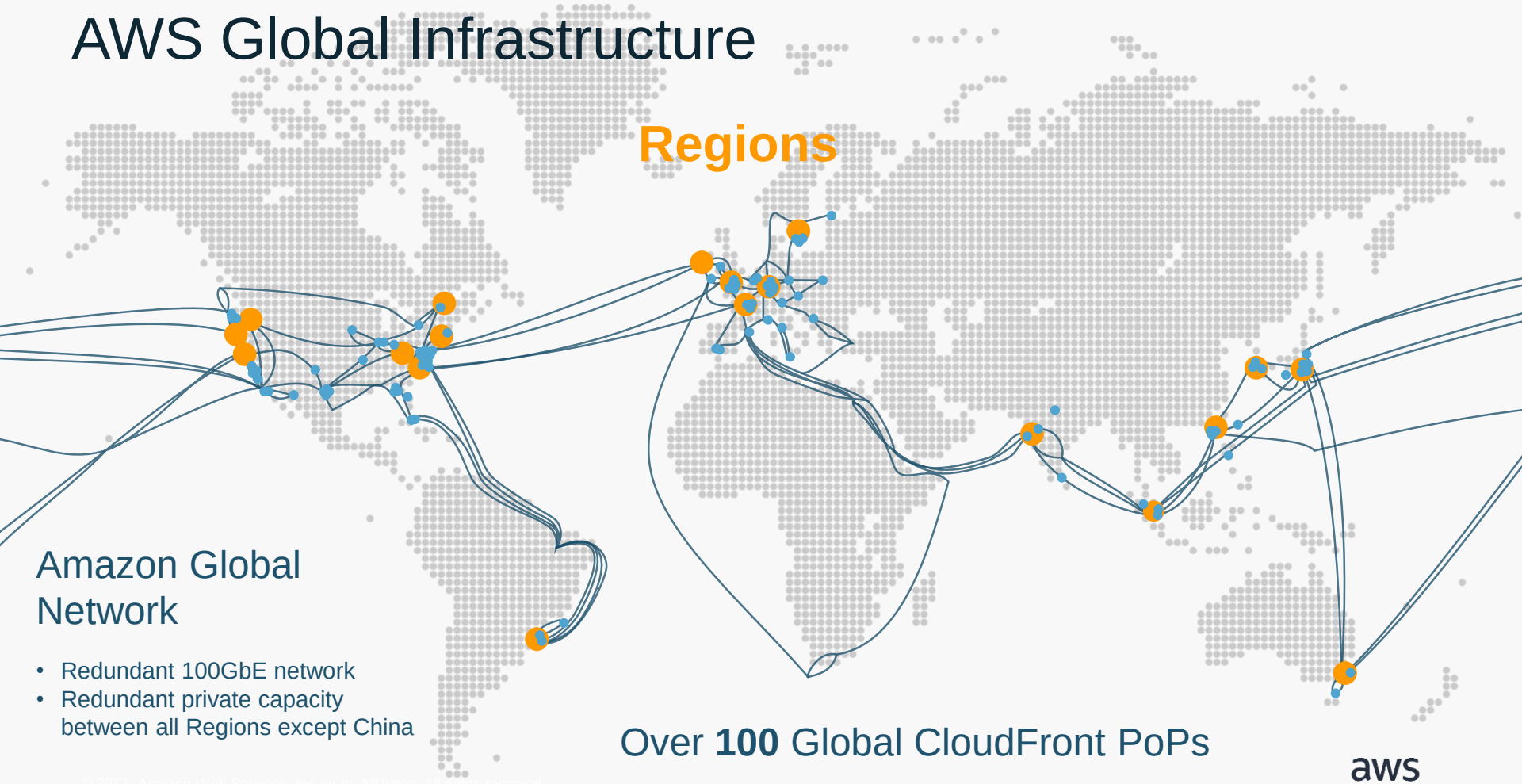


The AWS Cloud spans **55 Availability Zones** within **18 geographic Regions** and 1 Local Region around the world,

Announced plans: 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.

# Indicates # of AZs in the Region

AWS Global Infrastructure



Amazon Global Network

- Redundant 100GbE network
- Redundant private capacity between all Regions except China

Over **100** Global CloudFront PoPs

Agenda

Overview of AWS Infrastructure

HPC Solution Overview

HPC Use Cases in Design and Engineering

Customer Success Stories

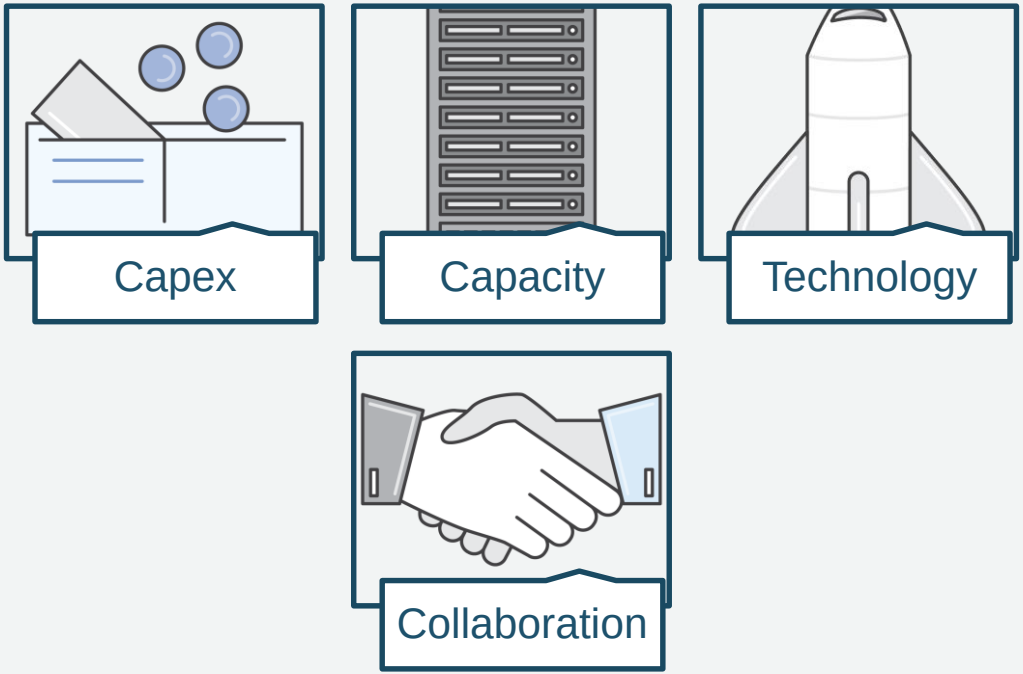
Best Practices for Performance, Scalability, and Cost

Additional Resources

aws

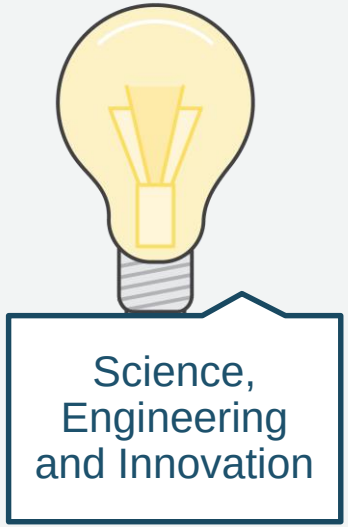
HPC on AWS: Fundamental Rethink of What is Possible

From worrying about

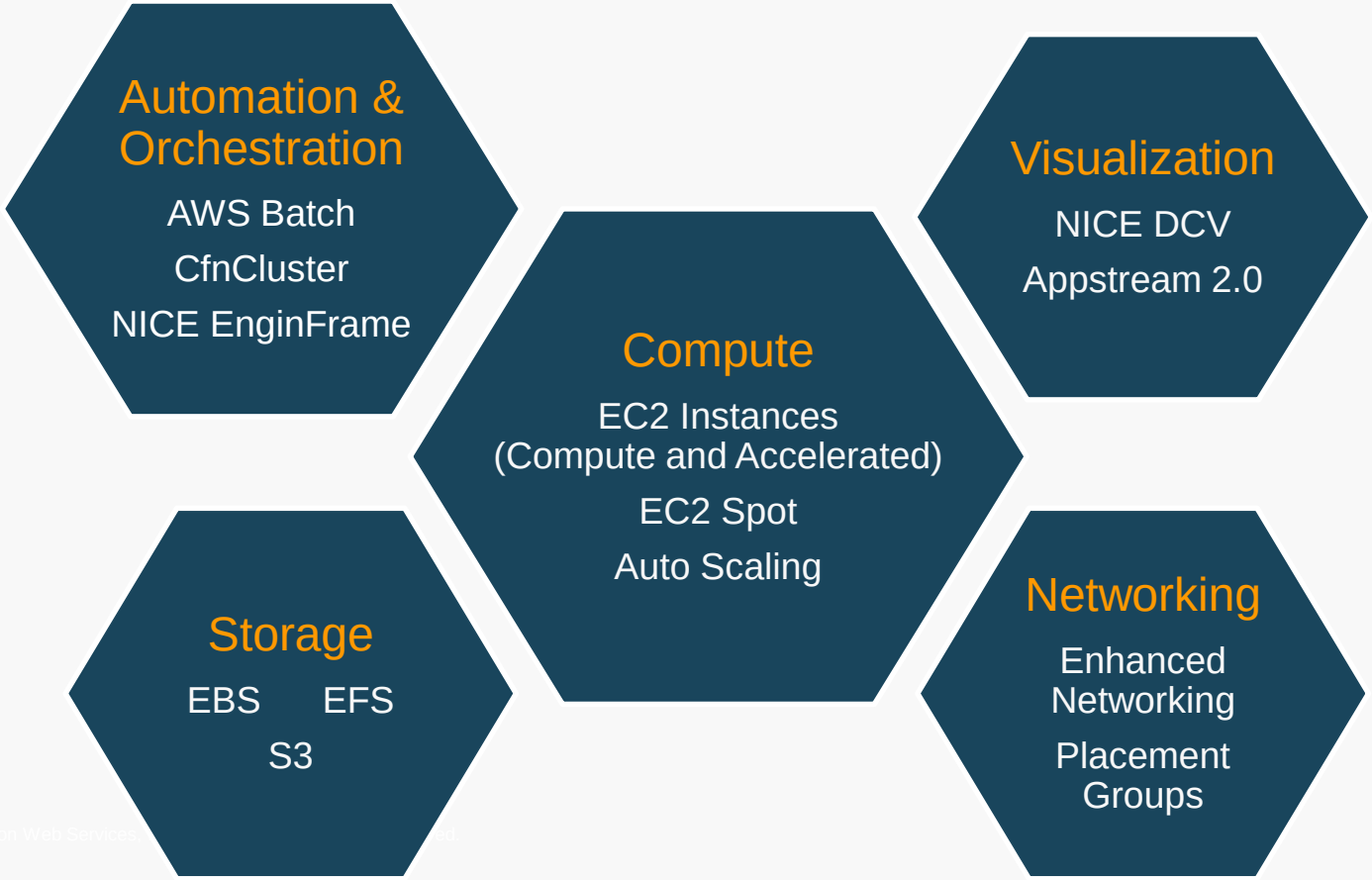


to

Focusing on

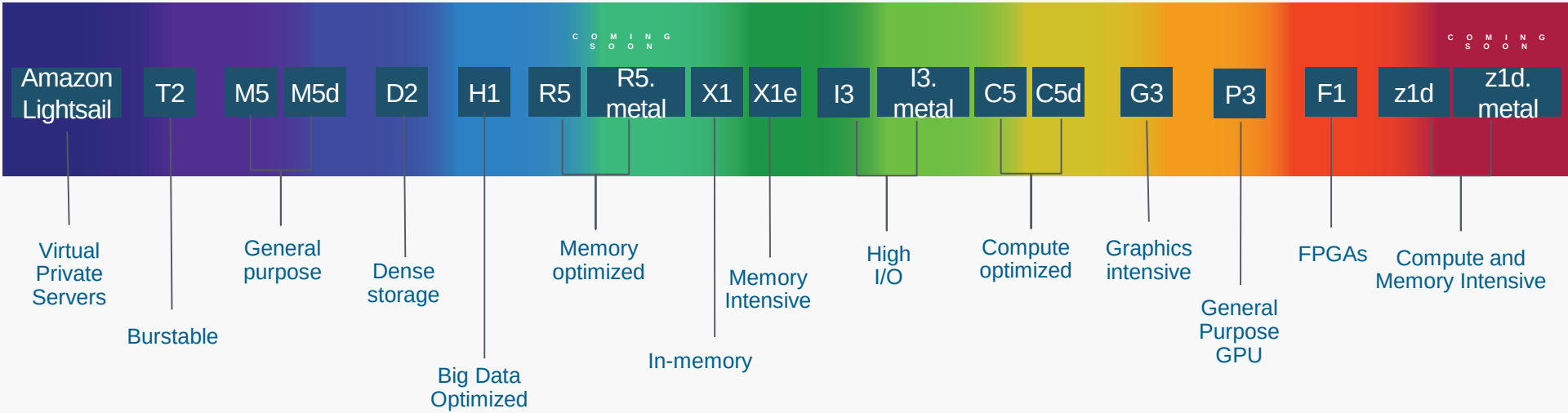


AWS HPC Solution Components

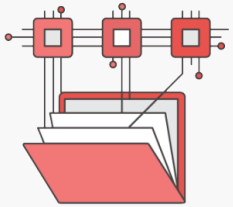


Amazon EC2 Instances

- Select compute that best fits the workload profile; Match the architecture to the job, not vice versa
- Optimize price/performance of your HPC Workloads with widest range of compute instances
- Benefit from the AWS pace of innovation

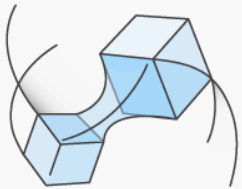


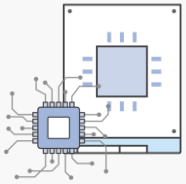
AWS Storage is a Platform



Amazon EFS

File

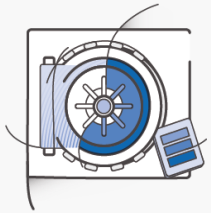




Amazon EBS

Amazon EC2 Instance Store

Block

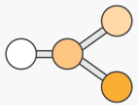


Amazon S3 / S3-IA

Amazon Glacier

Object

Data Transfer



Internet/ VPN

AWS Direct Connect

Amazon CloudFront

S3 Transfer Acceleration

ISV Connectors

Storage Gateway

AWS Snowball

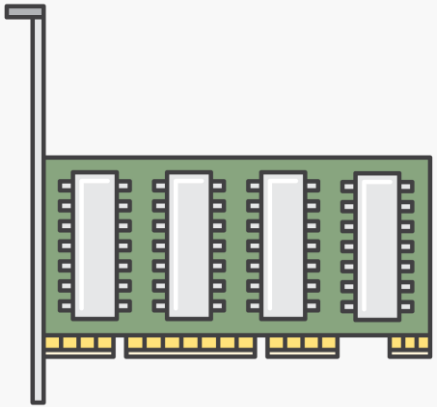
Amazon Kinesis Firehose

aws

High Speed Networking

Elastic Network Adaptor (ENA)

- Latest generation of Enhanced Networking
  - Hardware Checksums
  - Multi-Queue Support
  - Receive Side Steering
- Up to 25Gbps in a Cluster Placement Group
- Open Source Amazon Network Driver
- Compatible with MPI libraries including OpenMPI 3.0
- A cluster placement group is a logical grouping of instances within a single AZ
- Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both



aws

AWS Batch

- AWS Batch is a set of fully managed batch primitives
- Focus on your applications (shell scripts, Linux executables, Docker images) and their resource requirements
- We take care of the rest!

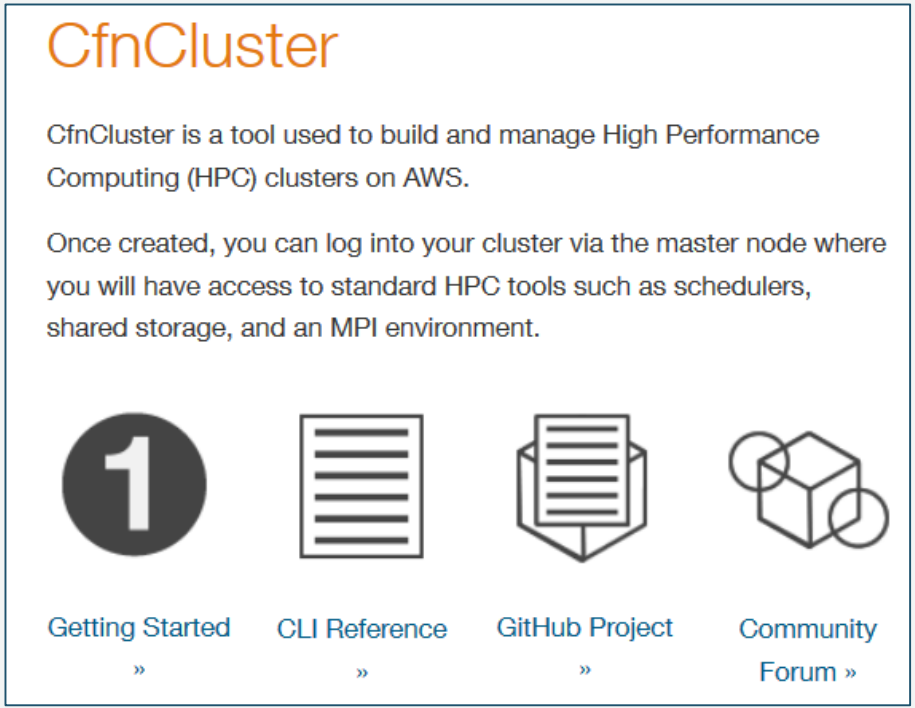
aws

AWS Batch Concepts

- The Scheduler evaluates when, where, and how to run jobs that have been submitted to a job queue.
- Jobs run in approximately the order in which they are submitted as long as all dependencies on other jobs have been met.
- There is no charge for AWS Batch; you only pay for the underlying resources that you consume!

aws

HPC Automation with CfnCluster

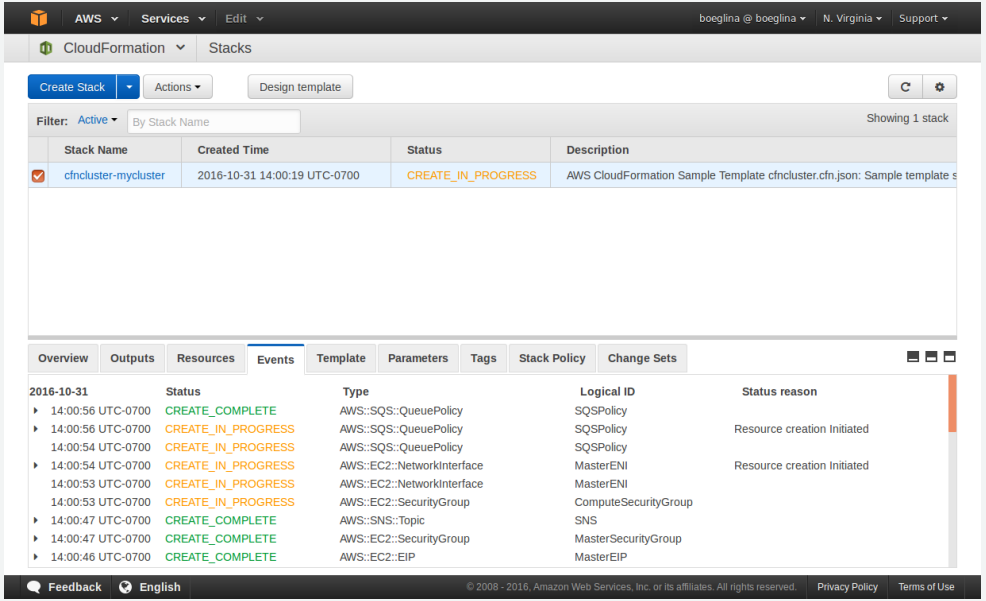


- CfnCluster simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers (SGE, Torque, Slurm)
- Built on AWS CloudFormation, easy to modify to meet specific application or project requirements
- CfnCluster will handle the automatic addition of compute nodes when there are pending jobs in the queue

aws

Launch a Cluster in Minutes

- Cluster creation usually takes ~15 minutes
- Completely managed by CloudFormation

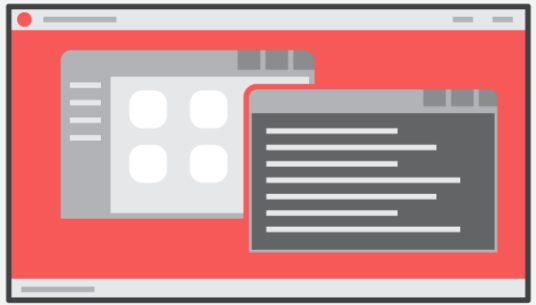



```
$ cfncluster create mycluster
```

CfnCluster Configuration Options

- Operating System
  - Amazon Linux
  - Centos 6
  - Centos 7
  - Ubuntu 14.04
- Scheduler
  - Sun Grid Engine (SGE)
  - PBS/Torque
  - SLURM

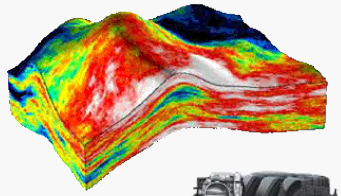
- Storage Size & IOPS
- EBS & Instance Store Encryption
- Scaling Speed & Limits
- Provisioning Scripts

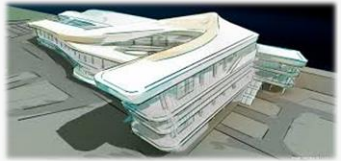


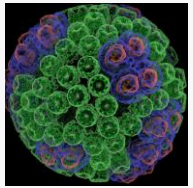
aws

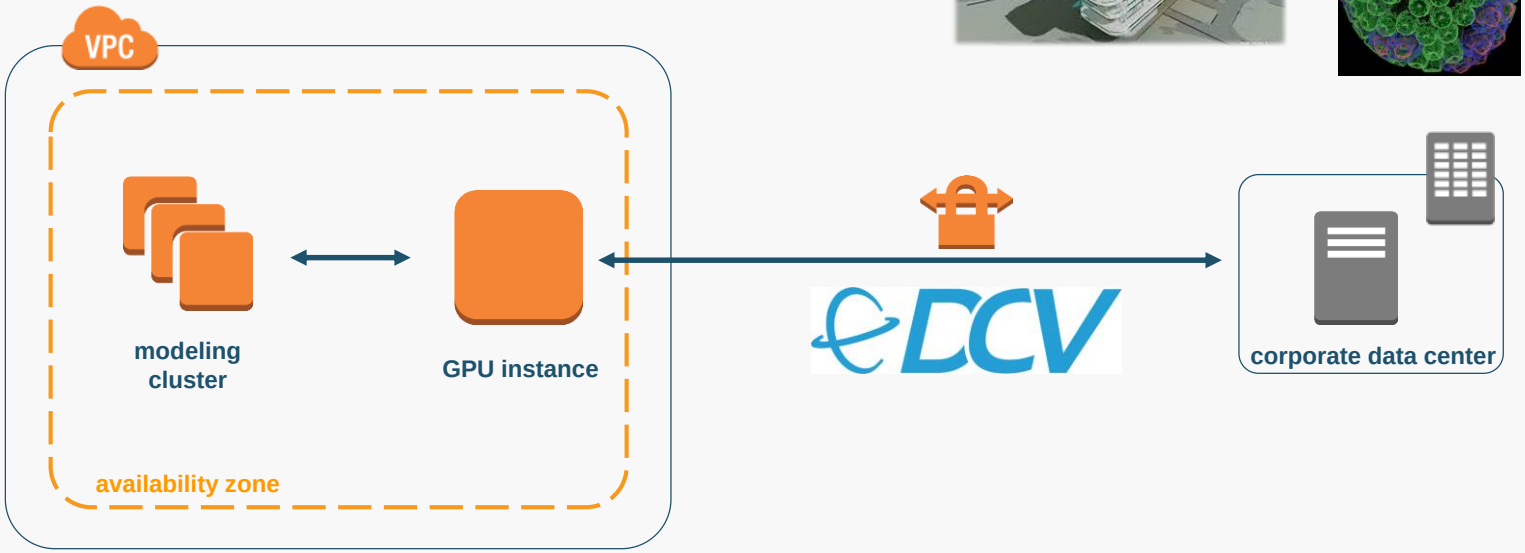
Remote Visualization

- Using a GPU optimized instance (G3) and NICE DCV to visualize results







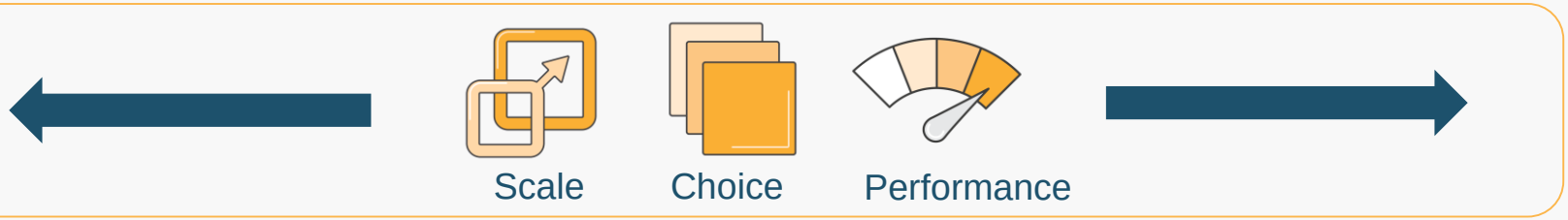


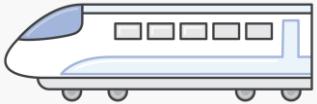
aws

Partnerships that enable a seamless migration

aws

AWS Advantages for HPC Workload Types



Tightly Coupled Parallel Computing	Loosely Coupled Parallel Computing	Accelerated Computing	Visualization and Interpretation	High Performance Data Storage and Analytics
				
Skip the Queue	EC2 Spot Pricing	Early Access to Technology	View results instantly	Derive unique insights with AI/ML

aws

High Performance Computing on AWS

- **Innovate faster** with virtually unlimited infrastructure enabling scaling and agility not attainable on-premises
- **Optimize cost** with flexible resource selection and pay per use
- **Increase collaboration** with secure access to clusters around the world

aws

AWS Directly Benefits Design and Engineering

Accelerated Time to Results

- Scale higher and run more simulations to quickly converge on more efficient, safer, cost-effective products, and get to market faster
- No large upfront investments in time, infrastructure, money

Scalability and Dynamic Resourcing

- Efficiently scale resources to meet shifting demands throughout the product lifecycle
- Encourages more experimentation, more frequent product launches, higher quality manufacturing

Secure Environment for Collaboration

- Streamlined, repeatable, and secure simulation and development environments
- Global regions with traceability helps satisfy regulations and audits, while enabling supply chain collaboration

Global, Fault-Tolerant Infrastructure

- 55 AWS Availability Zones in 18 Regions worldwide means high availability
- Allows for seamless, controlled information-sharing between global stakeholders, and enables reduced latency for interactive engineering use-cases

Agenda

Overview of AWS Infrastructure

HPC Solution Overview

HPC Use Cases in Design and Engineering

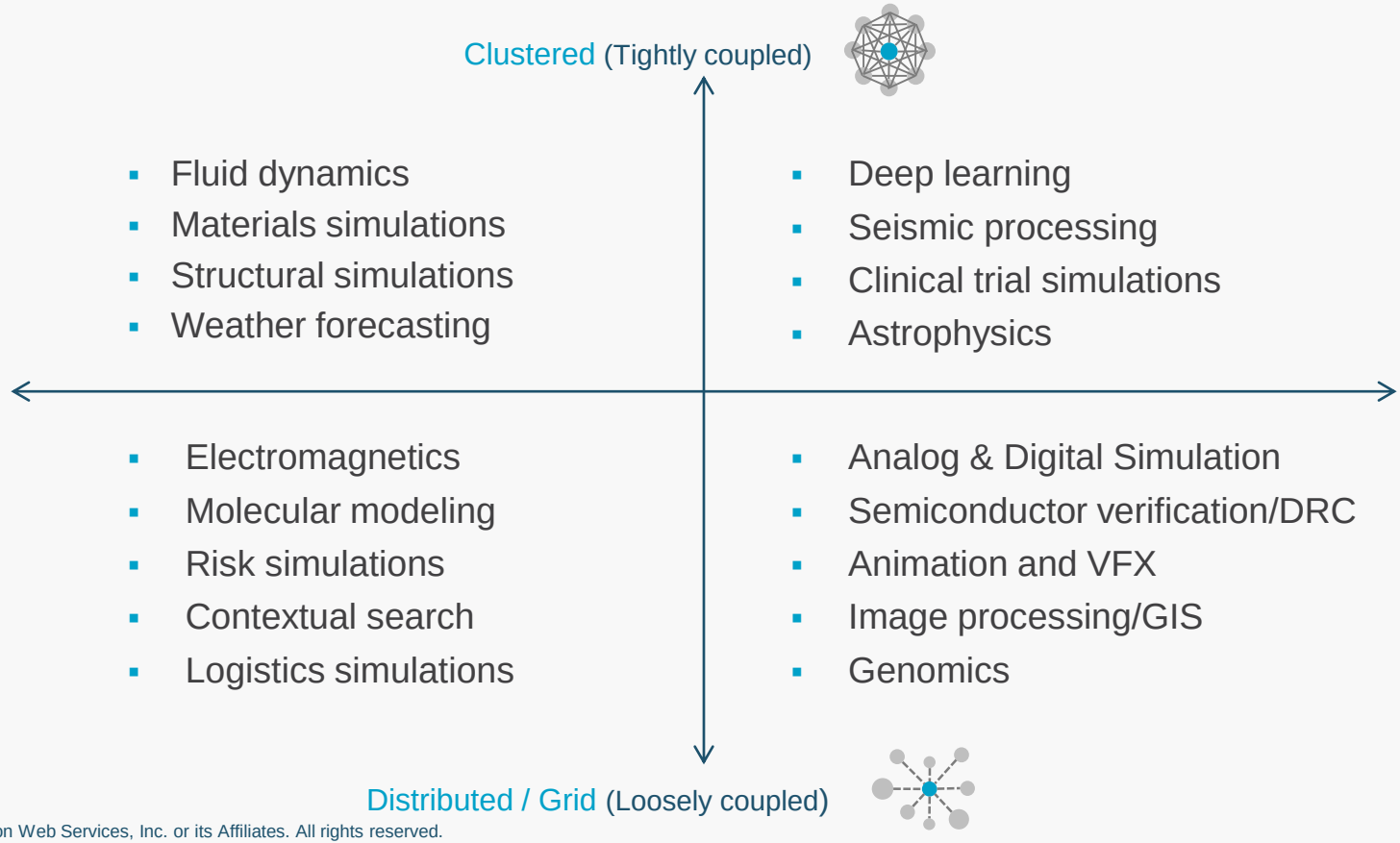
Customer Success Stories

Best Practices for Performance, Scalability, and Cost

Additional Resources

aws

Sample HPC Use Cases



HPC Use Cases in Product Engineering

Clustered (Tightly coupled)

- **Fluid dynamics**
- **Materials simulations**
- **Structural simulations**
- Weather forecasting

- **Deep learning**
- Seismic processing
- Clinical trial simulations
- Astrophysics

Data Light

Minimal requirements for high performance storage

Data Heavy

Benefits from access to high performance storage

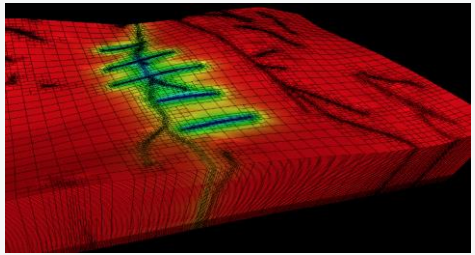
- **Electromagnetics**
- **Molecular modeling**
- **Risk simulations**
- Contextual search
- Logistics simulations

- **Analog & Digital Simulation**
- **Semiconductor verification/DRC**
- Animation and VFX
- Image processing/GIS
- Genomics

Distributed / Grid (Loosely coupled)

aws

Graphics for Design and Engineering



Energy

HALLIBURTON
Landmark

“The exploration and production models are increasingly complex with very large datasets, 3D and dynamic algorithms, security, and global reach… Amazon EC2 G3 instances enable Landmark to deliver value to our clients in ways that were not possible before.”

Chandra Yeleshwarapu
Global Head of Services and Cloud
Landmark, Halliburton

Automotive

ZeroLight

“Amazon EC2 G3 instances will enable us to continue to deliver our unique, real-time, high-quality 3D experiences for automotive customers through every channel.”

Darren Joblin
Chief Executive Officer
ZeroLight

Media

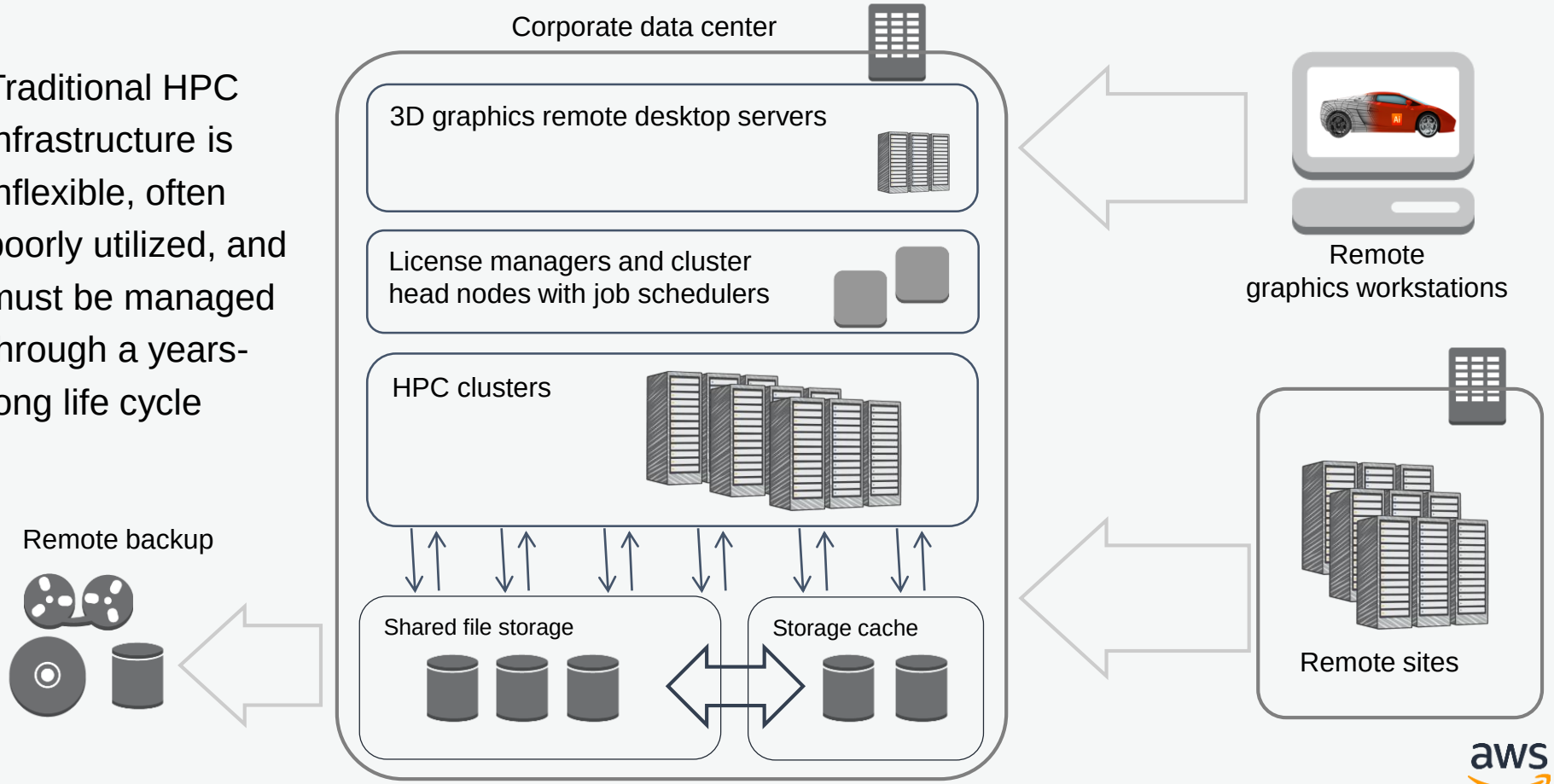
BEBOP TECHNOLOGY

“The advanced graphics features of Amazon EC2 G3 instances help our customers edit their work with the same quality and fidelity as a local workstation, deliver results faster, and much more cost effectively than ever before.”

David Benson
Chief Technology Officer
BeBop Technology

Traditional HPC stack for engineering & science applications

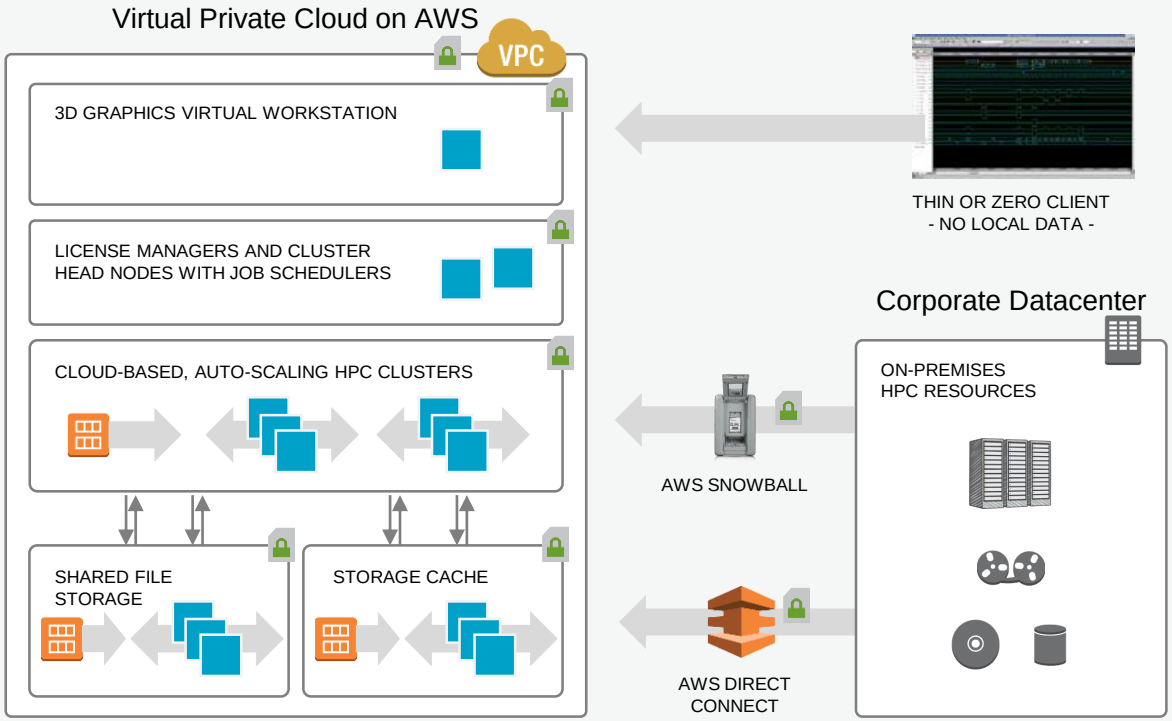
Traditional HPC infrastructure is inflexible, often poorly utilized, and must be managed through a years-long life cycle



Migrating HPC to 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



Encryption everywhere – with your own keys!

Agenda

Overview of AWS Infrastructure

HPC Solution Overview

HPC Use Cases in Design and Engineering

Customer Success Stories

Best Practices for Performance, Scalability, and Cost

Additional Resources

HPC in aerospace

Boom leverages Rescale and AWS to enable supersonic travel

- Simulated vortex lift with 200M cell models on 512+ cores
- Increased simulation throughput: 100 jobs in parallel with 6x speedup per job → 600x speedup
- Eliminated IT overhead, including server capital costs & in-house IT and software teams
- Elastic HPC capacity and pay-as-you-go AWS clusters allow business agility & ability to scale

boom

"Rescale's ScaleX cloud platform is a game-changer for engineering. It gives Boom computing resources comparable to building a large on-premise HPC center. Rescale lets us move fast with minimal capital spending and resources overhead."

Josh Krall
CTO & Co-Founder

aws

Big Data Meets HPC

Big Data Platform (BDP)

- Data from across global manufacturing sites are collected into a cloud-based Big Data Platform
- The BDP enables operational/logistics tracking of millions of hard drives produced each year
- Allows analysts to visualize data across JMP, Tableau, IBM SPSS, and SAS to support all phases of engineering, manufacturing, testing and support

High Performance Computing for Design and Engineering

- Cloud-based HPC accelerates product optimization, using clusters of CPUs and GPUs to perform millions of drive-head and disk interface simulations, and to improve storage magnetics product capacities
- Cloud-based HPC is the foundation for future storage architecture analysis, materials science explorations and machine learning based investigations for multiple storage product families

"The IT organization has been driving a massive digital transformation and optimization of business capabilities across the organization. IT has been leading these changes by creating rich environments for the data to thrive, ensuring improvements in productivity and collaboration across the massively global organization."

- Steve Phillpot, CIO
  Western Digital

Western Digital.

HGST
a Western Digital brand

SanDisk®

aws

HPC in design and manufacturing

Applications for engineering:

- Molecular dynamics, CAD, CAE, EDA
- Collaboration tools for engineering
- Big data for manufacturing yield analysis

HGST
a Western Digital brand

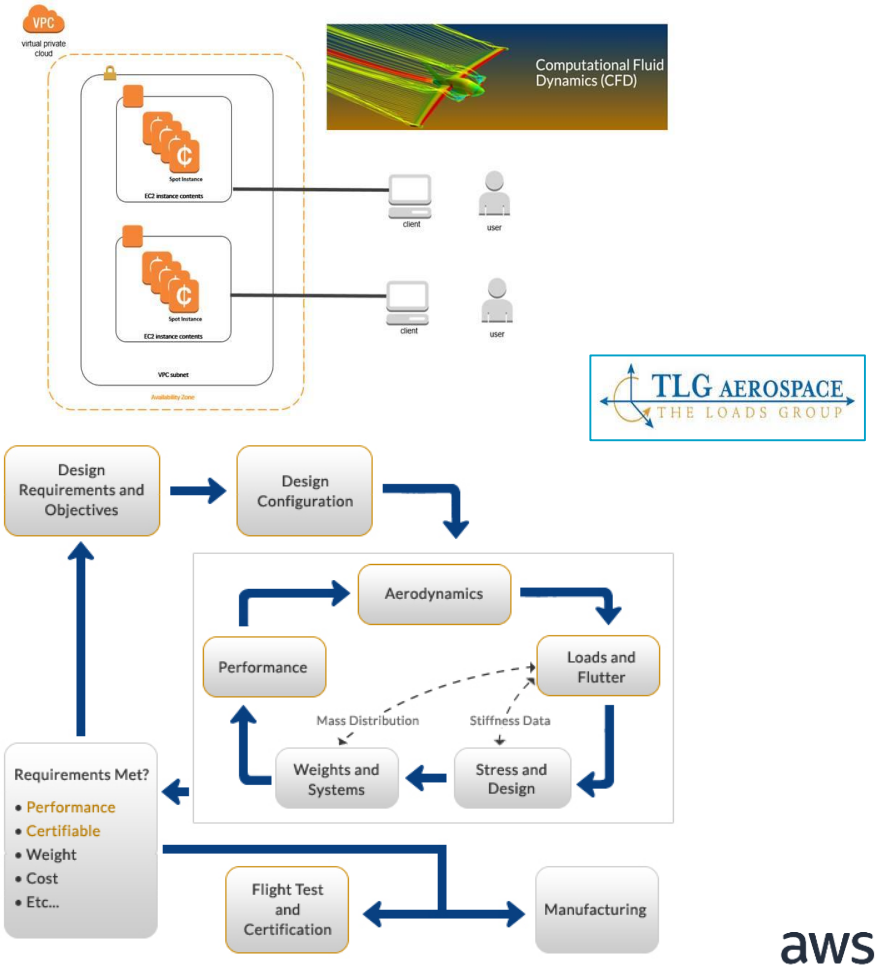
Running drive-head simulations at scale:

Millions of parallel parameter sweeps, running months of simulations in just hours

Over 85,000 Intel cores running at peak, using Spot Instances

HPC in aerospace

- Running parallel CFD studies using Siemens STAR-CCM+
  - Goal: shorten the time between Design Requirements and Configuration, and Flight Testing
- 1000+ cores per CFD study, multiple studies required for each workflow iteration
- Job-level optimizations:
  - Enhanced Networking, Placement Groups
  - Amazon Linux, Hyper-threading disabled
- Workflow optimizations:
  - Spot instances, multiple clusters
  - Multiple parallel studies for faster throughput

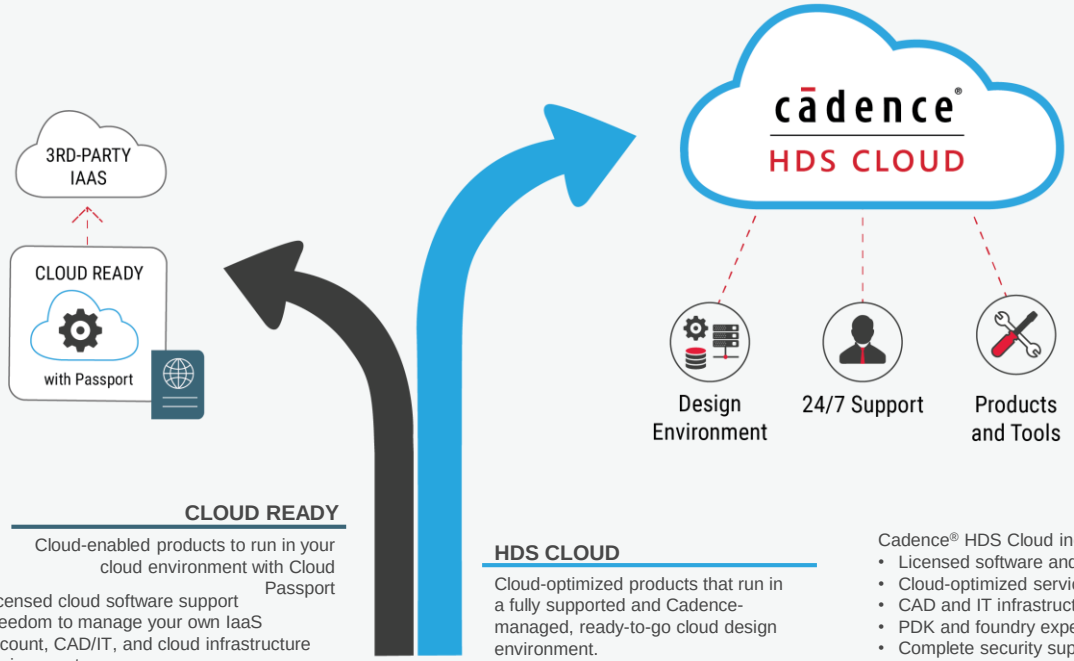


Semiconductor Design: Cadence Cloud Products

For electronic design automation (EDA)

Customer Managed

Cadence Managed



CLOUD READY

Cloud-enabled products to run in your cloud environment with Cloud Passport

- Licensed cloud software support
- Freedom to manage your own IaaS account, CAD/IT, and cloud infrastructure environment

HDS CLOUD

Cloud-optimized products that run in a fully supported and Cadence-managed, ready-to-go cloud design environment.

Cadence® HDS Cloud includes:

- Licensed software and support
- Cloud-optimized services
- CAD and IT infrastructure support
- PDK and foundry expertise
- Complete security support

Agenda

Overview of AWS Infrastructure

HPC Solution Overview

HPC Use Cases in Design and Engineering

Customer Success Stories

Best Practices for Performance, Scalability, and Cost

Additional Resources

Important Factors for Performance and Throughput

- Compute performance – CPUs, GPUs, FPGAs
- Memory performance – high RAM requirements in many applications
- Network performance – throughput, latency, and consistency
- Storage performance – including shared filesystems
- Automation and cluster/job management
- Remote graphics for interactive applications
- ISV support – including license management

…and SCALE

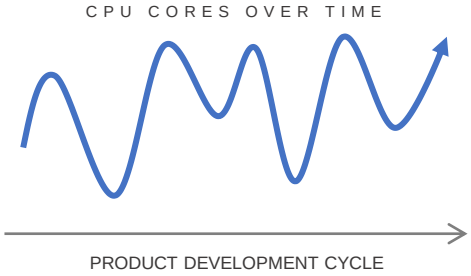
Accelerate Engineering with Rapid, Massive Scaling

Scale up when needed, then scale down

- In a traditional HPC datacenter, the only certainty is that you always have the wrong number of servers – too few, or too many
- Every additional HPC server launched in the cloud can improve speed of innovation – if there are no other constraints to scaling
- Overnight or over-weekend workloads reduced to an hour or less

Think BIG

What if you could launch 1 million concurrent jobs?

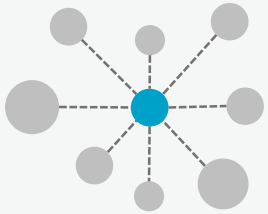


Use a “Grids of Clusters” Strategy for Scale

Cluster

Tightly coupled, latency-sensitive applications

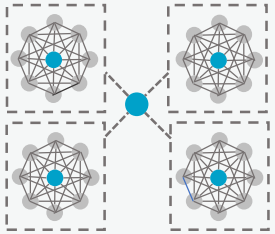
Use larger EC2 compute instances, placement groups, enhanced networking, HPC job schedulers to **“Scale-up”**



Grid

Loosely coupled, pleasingly parallel

Use a variety of EC2 instances, multiple AZs, Spot, Auto Scaling, to **“Scale-Out”**

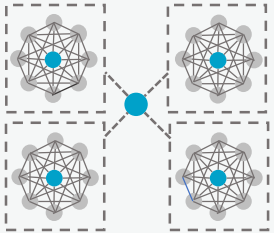


Grids of Clusters

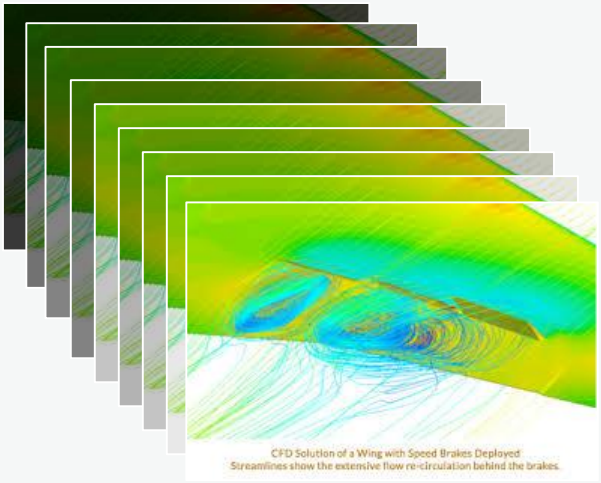
Running parallel cluster jobs, parameter studies

Use a grid strategy on the cloud to run a group of parallel, individually-clustered HPC jobs

Expand the solution space

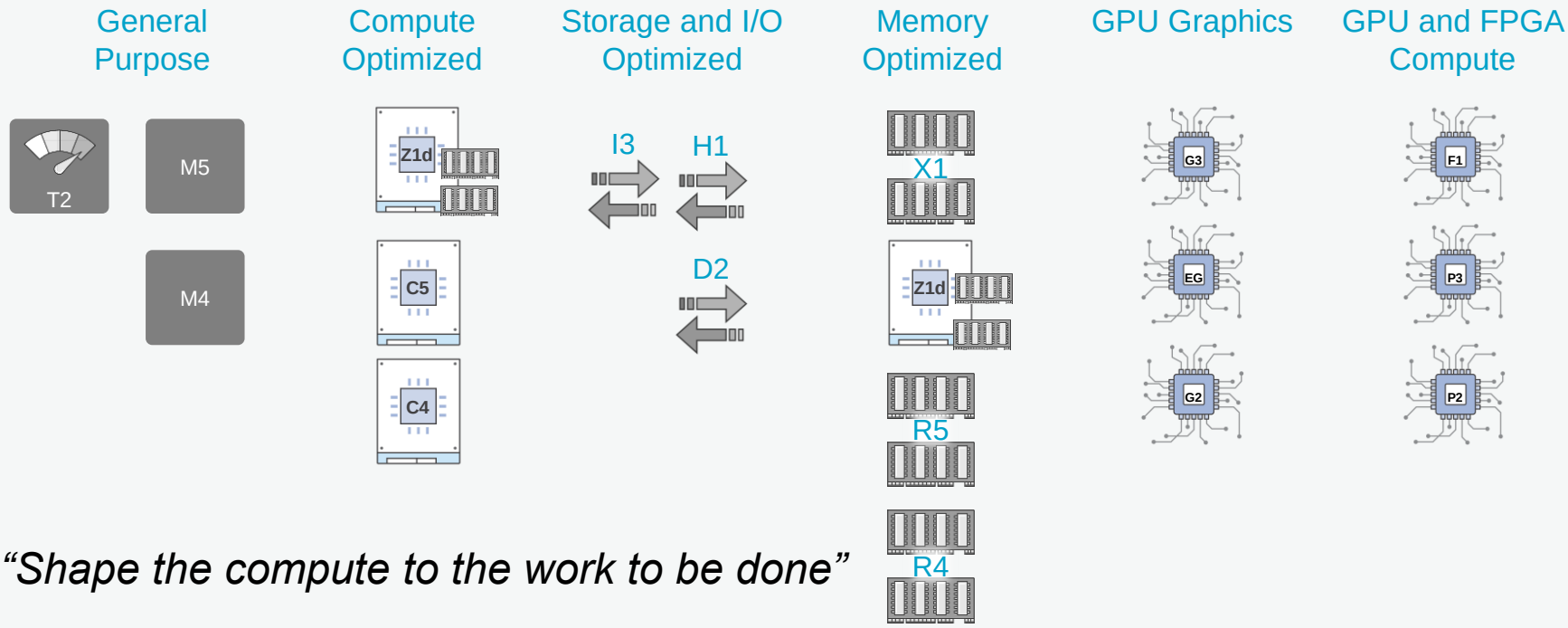


Using a “grid of cluster” strategy results in faster, higher quality of results for design exploration using parameter sweeps and other parallel methods



Courtesy of TLG Aerospace

Optimize Using AWS Compute Instance Types



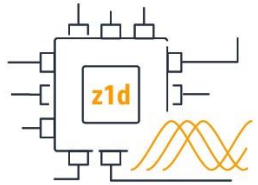
"Shape the compute to the work to be done"

Instance types: Z1d example

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

Choose an instance type and size to meet the unique needs of each application



Model	vCPU	Memory (GiB)	Instance Storage (GiB)	Networking Performance
z1d.large	2	16	1 x 75 NVMe SSD	Up to 10,000 Mbps
z1d.xlarge	4	32	1 x 150 NVMe SSD	Up to 10,000 Mbps
z1d.2xlarge	8	64	1 x 300 NVMe SSD	Up to 10,000 Mbps
z1d.3xlarge	12	96	1 x 450 NVMe SSD	Up to 10,000 Mbps
z1d.6xlarge	24	192	1 x 900 NVMe SSD	10,000 Mbps
z1d.12xlarge	48	384	2 x 900 NVMe SSD	25,000 Mbps

M5: Next Generation General Purpose Instance

Also Available: M5d with local NVMe-based SSD storage

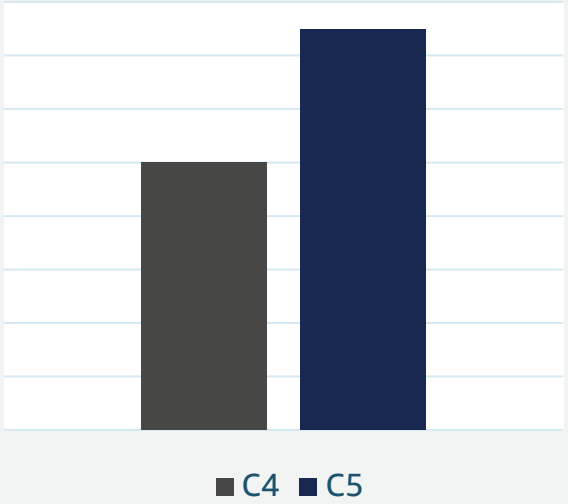
- Powered by 2.5 GHz Intel Xeon Scalable Processors (Skylake)
- New larger instance size—m5.24xlarge with 96 vCPUs and 384 GiB of memory (4:1 Memory:vCPU ratio)
- AWS Support for Intel AVX-512 offering up to twice the performance for vector and floating point workloads
- **Use case: FEA Implicit**

aws

C5: Compute-Optimized Instances Based on Intel Skylake

Also available: C5d with local NVMe-based SSD storage

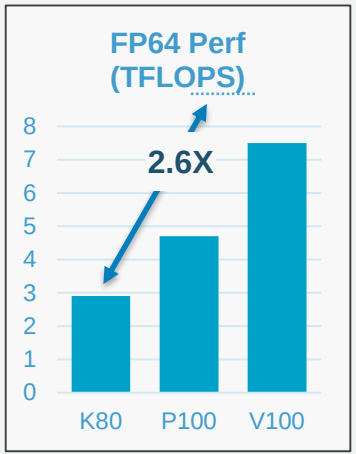
25% price/performance improvement over C4

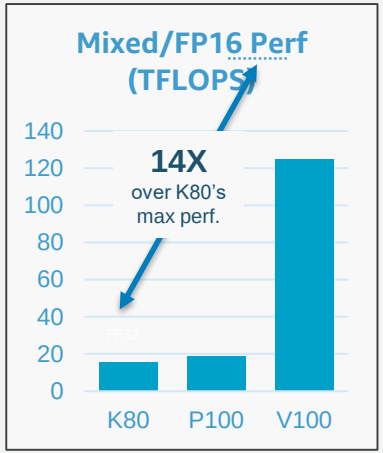


- Based on 3.0 GHz Intel Xeon Scalable Processors (Skylake)
- Run up to 3.5 GHz using Intel Turbo Boost Technology
- Up to 72 vCPUs and 144 GiB of memory (2:1 Memory:vCPU ratio)
- 25-Gbps NW bandwidth
- Support for Intel AVX-512
- **Use case: CFD, FEA Explicit**

aws

P3: Next Generation of GPU Compute Instances

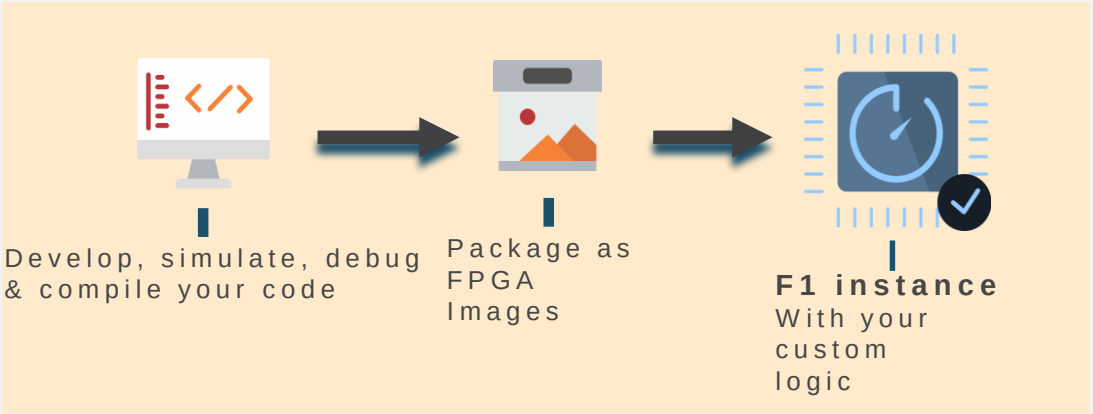




- Up to eight NVIDIA Tesla V100 GPUs
- 1 PetaFLOPs of computational performance – Up to 14x better than P2
- 300 GB/s GPU-to-GPU communication (NVLink) – 9X better than P2
- 16GB GPU memory with 900 GB/sec peak GPU memory bandwidth
- **Use cases: ML/AI, Accelerated Engineering**

aws

F1 Instances: First Cloud Instance with FPGA



- **Use cases:**
- Engineering simulations
- Image and video processing
- Big data and ML

Speed up applications over 30x

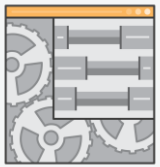
FPGA acceleration using F1

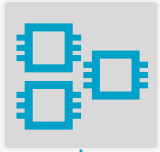
Amazon FPGA Image (AFI)

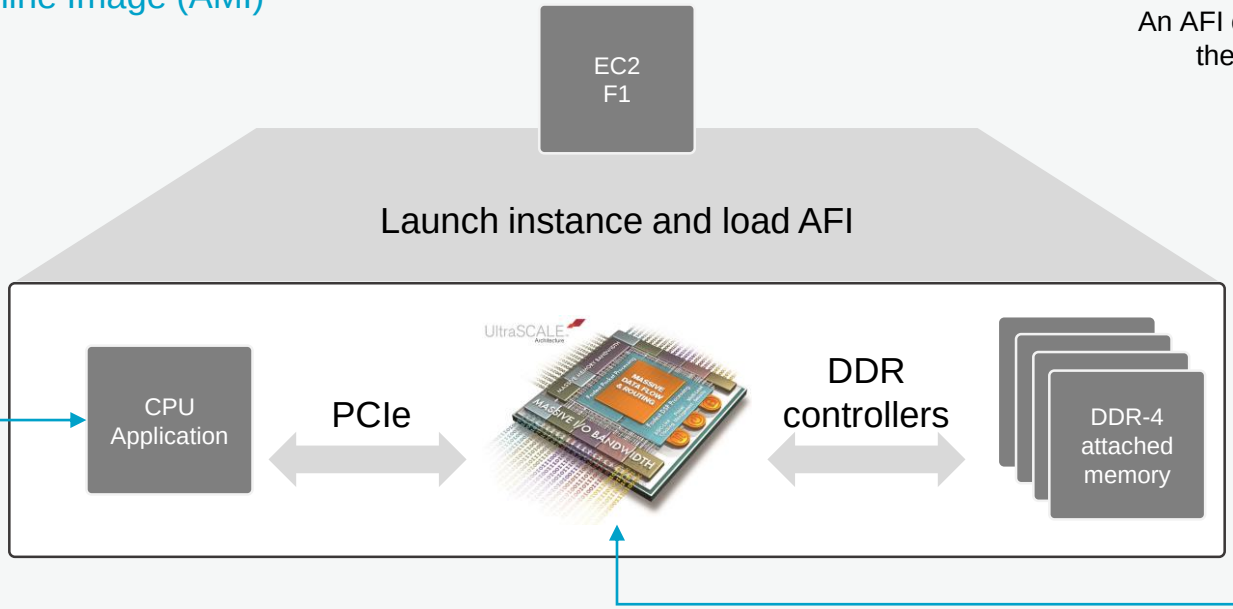
An F1 instance can have any number of AFIs

An AFI can be loaded into the FPGA in seconds

Amazon Machine Image (AMI)

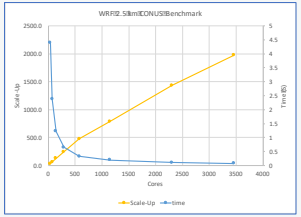






Performance considerations

For tightly-coupled cluster workloads



Test using real-world examples

- Use large cases for testing: do not benchmark scalability using only small examples

Domain decomposition

- Choose number of cells per core for either pre-core efficiency or for faster results

MPI libraries

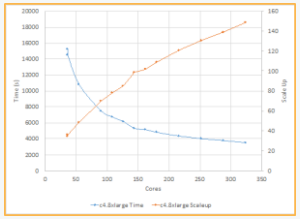
- Test with Intel MPI and OpenMPI 3.0, and make use of available tunings

Network

- Use a placement group
- Enable enhanced networking

Performance considerations

For all HPC workloads



OS version

- Use Amazon Linux, RHEL 7.5, or an updated 3.10+ kernel – 4.0+ if using NVME on Z1d, R5d, F1, I3, etc

Processor states

- Use P-states to reduce processor variability

Instance types

- Always test with the latest EC2 instances – and with earlier generation instance types as well

Hyper-threading and affinity

- Test with Hyper-threading (HT) on and off – usually off is best, but not always
- Use CPU affinity to pin threads to CPU cores when HT is off

Optimize Using EC2 Purchasing Options

On-Demand

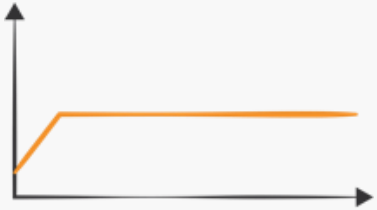
Pay for compute capacity **by the second** with no long-term commitments

Spiky workloads, to define needs

Reserved

Make a **1 or 3 Year commitment** and receive a significant discount off On-Demand prices

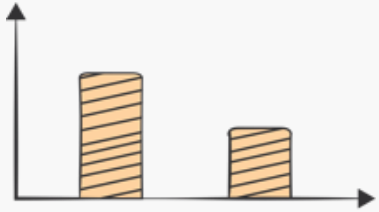
Committed, steady-state usage



Spot

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

Fault-tolerant, dev/test, time-flexible, stateless workloads



Per Second Billing for EC2 Linux instances & EBS volumes

aws

Cost optimization

On-Demand

Pay for compute capacity by the hour, with per-second billing and no long-term commitments

For spiky workloads,
or to define needs

Reserved

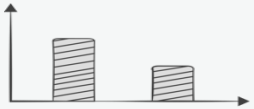
Make a low, one-time payment and receive a significant discount on the hourly charge

For committed utilization

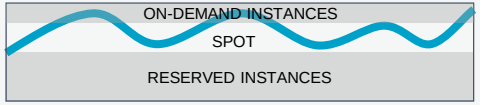
Spot

Use unused capacity, charged at a Spot price that fluctuates based on supply and demand

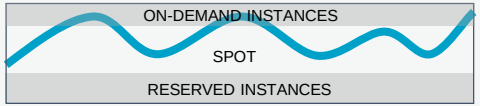
For high-scale, time-flexible workloads



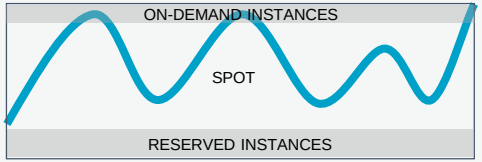
Conservative:



Optimized:



Optimized with scale-out (magnify the peak):



Agenda

Overview of AWS Infrastructure

HPC Solution Overview

HPC Use Cases in Design and Engineering

Customer Success Stories

Best Practices for Performance, Scalability, and Cost

Additional Resources

aws

Further reading

White Papers

- Introduction to HPC on AWS
- Optimizing Electronic Design Automation (EDA) Workflows on AWS

Reference Architecture

- HPC Lens - Well Architected Framework

Blogs

- Ansys - Getting Faster, Cost-effective Simulation on the Cloud
- Real World Scalability with AWS
-

Docs

- Processor State Control for Your EC2 Instance
- Optimizing CPU Options
- Big data
- EFS performance
- Monitor performance

- AWS Online Tech Talk

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

Q&A

Thank you

Learn More:

http://aws.amazon.com/hpc

Numbers are like pictures

They are worth a thousand words ..

Numbers that made a difference

aws