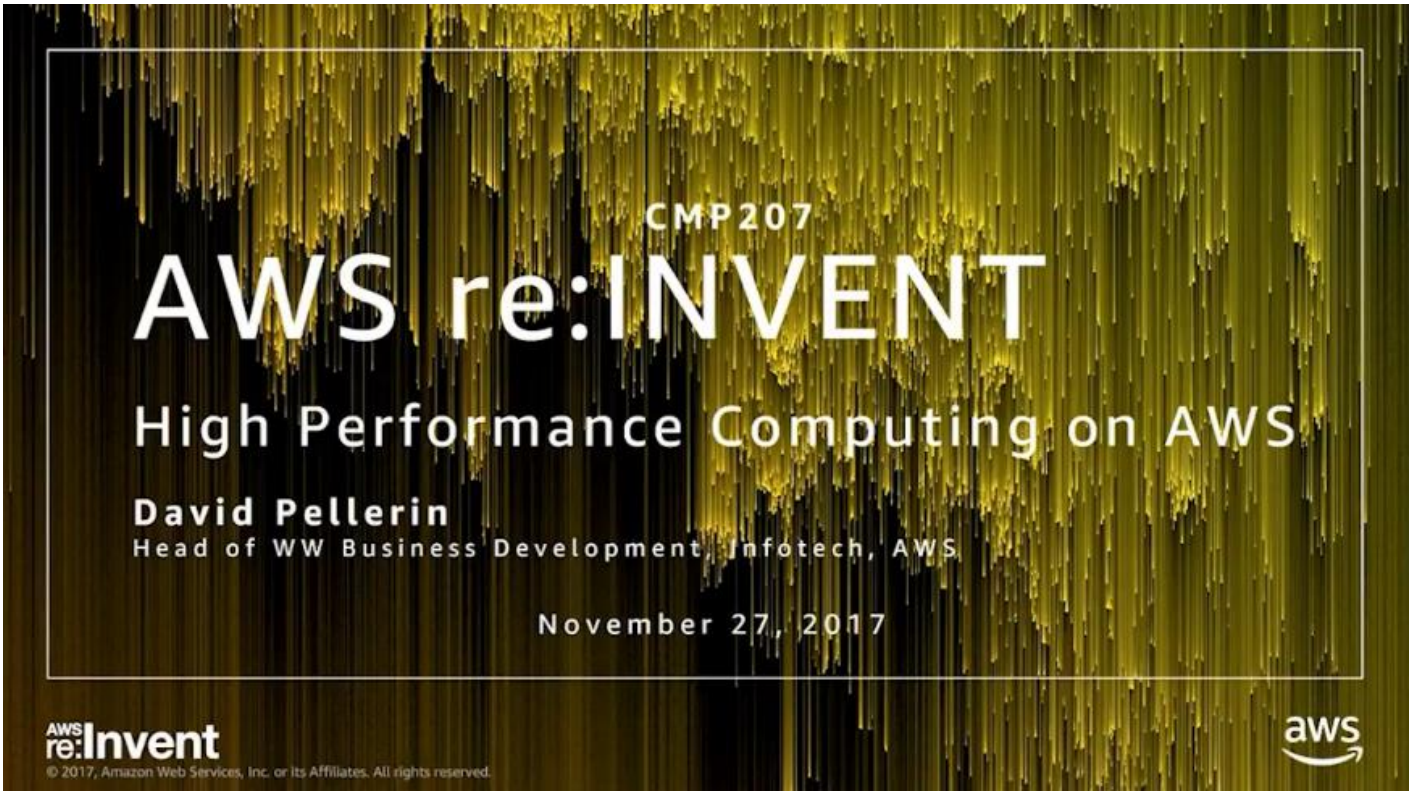
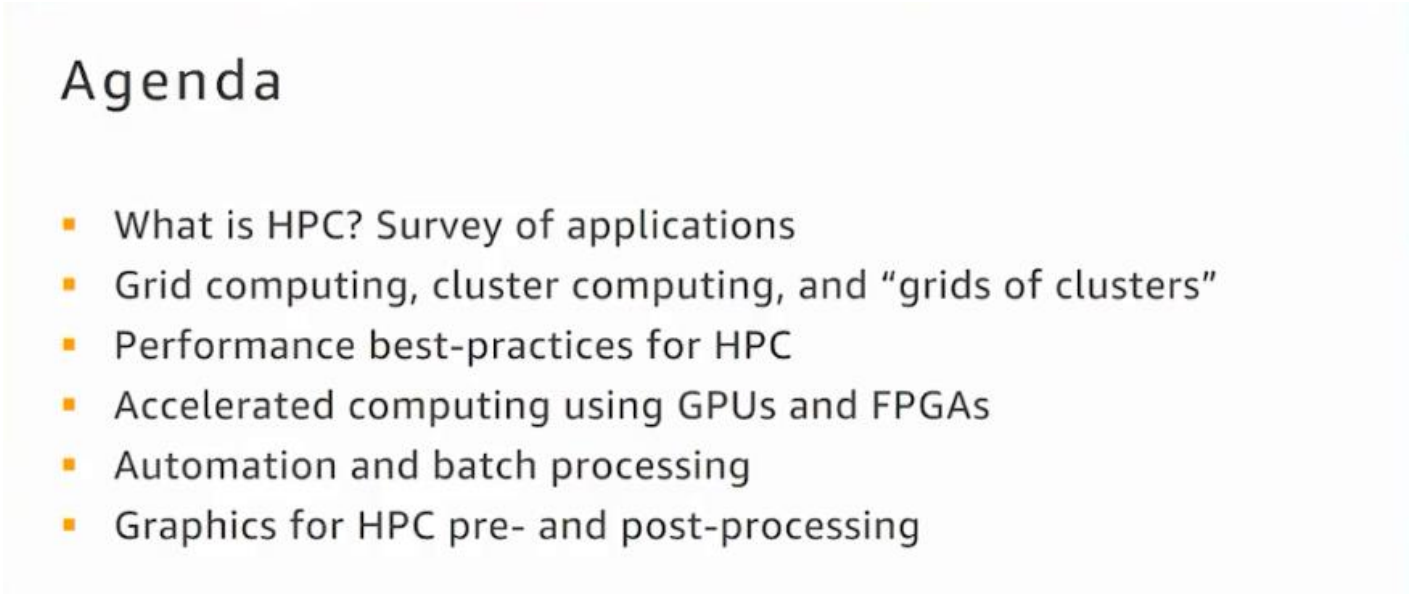
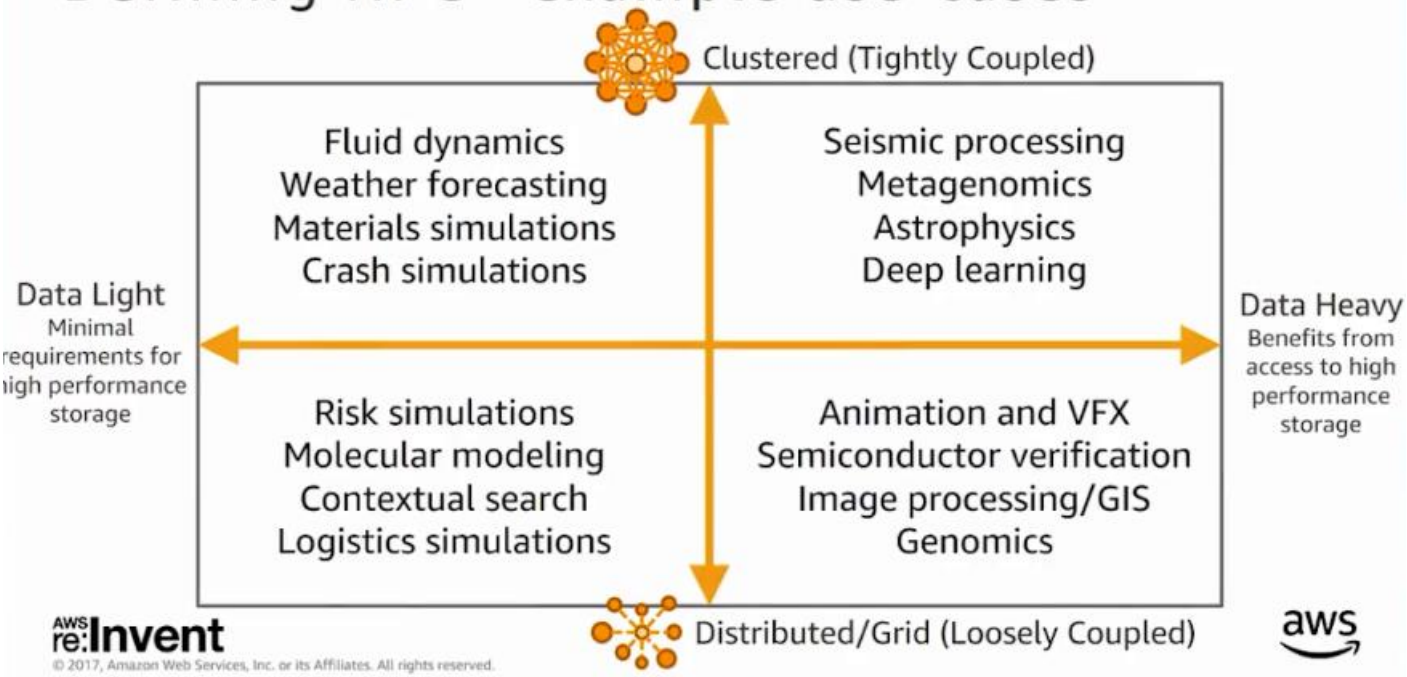


High-performance computing (HPC) in the cloud enables high scale compute- and graphics-intensive workloads across a range of industries—from aerospace, automotive, and manufacturing to life sciences, financial services, and energy. AWS provides application developers and end users with unprecedented computational power for massively parallel applications in areas such as large-scale fluid and materials simulations, 3-D content rendering, financial computing, and deep learning. In this session, we provide an overview of HPC capabilities on AWS. We describe the newest generation of accelerated computing instances, and we highlight customer and partner use cases across industries. Attendees learn best practices for running HPC workflows in the cloud, including graphical pre- and post-processing, workflow automation, and optimization. Attendees also learn about new and emerging HPC use cases, in particular, deep learning training and inference, large-scale simulations, and high-performance data analytics.



Defining HPC—example use-cases



Important enablers in HPC

- 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
- Graphics for pre- and post-processing

...and SCALE

Cluster and grid HPC in the cloud

Cluster HPC

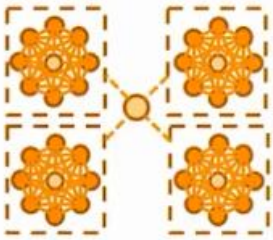
Tightly coupled, latency-sensitive applications

Use larger Amazon EC2 compute instances, placement groups, enhanced networking, HPC job schedulers

Grid HPC

Loosely coupled, pleasingly parallel

Use a variety of EC2 instances, multiple AZs, Spot, Auto Scaling, Amazon SQS, AWS Batch



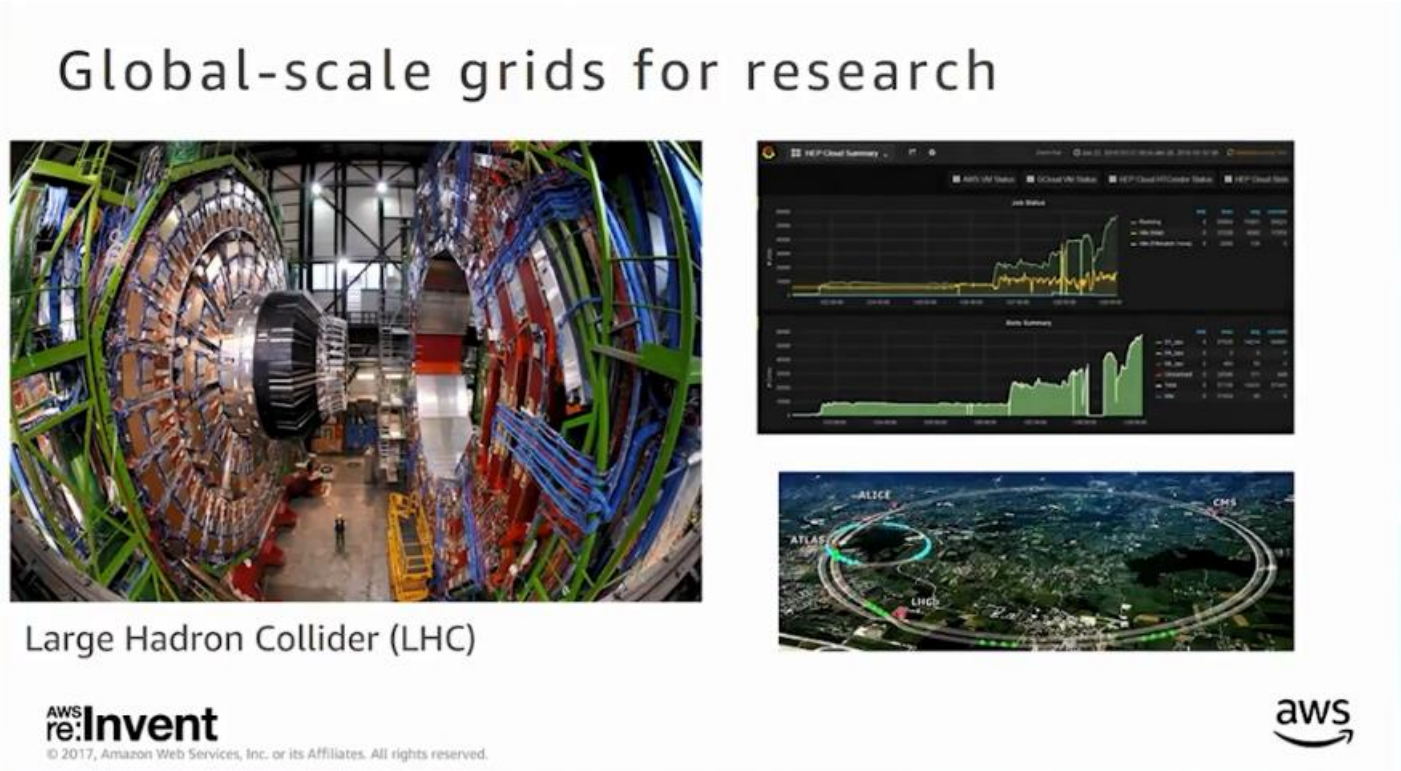
Grids of clusters

Running parallel cluster jobs, parameter studies

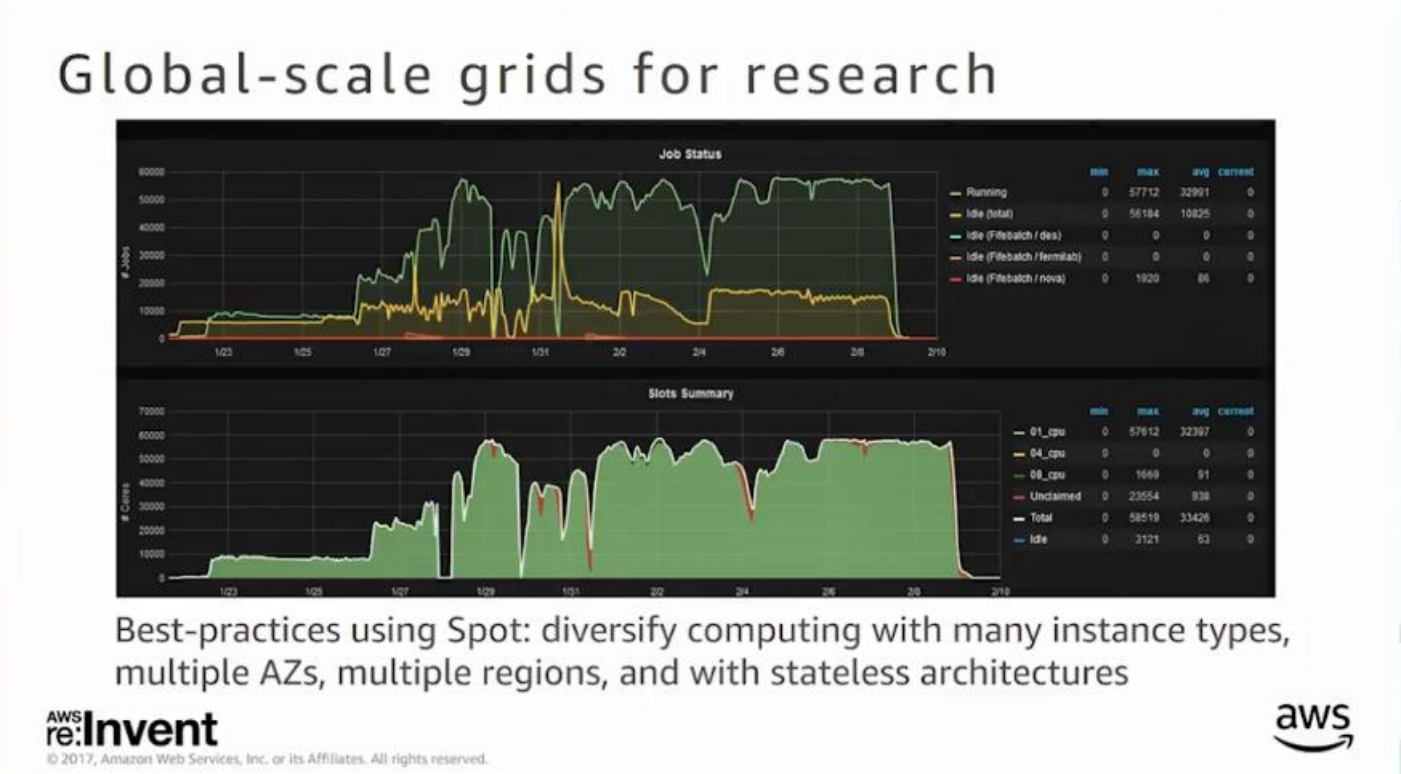
Use a grid strategy on the cloud to run a group of parallel, individually clustered HPC jobs

Grid computing examples

(Scale-out)



Generally, they run experiments that generate a huge amount of data and they analyze the data for patterns, when they find some patterns they then have to run several simulations to try and figure out the causality of the pattern they are seeing in the data.



This shows how many CPU cores they can launch using spot instances over the course of running simulations. This is for stochastic simulation for energy physics that are very parallel in form.

1.1M vCPUs for machine learning

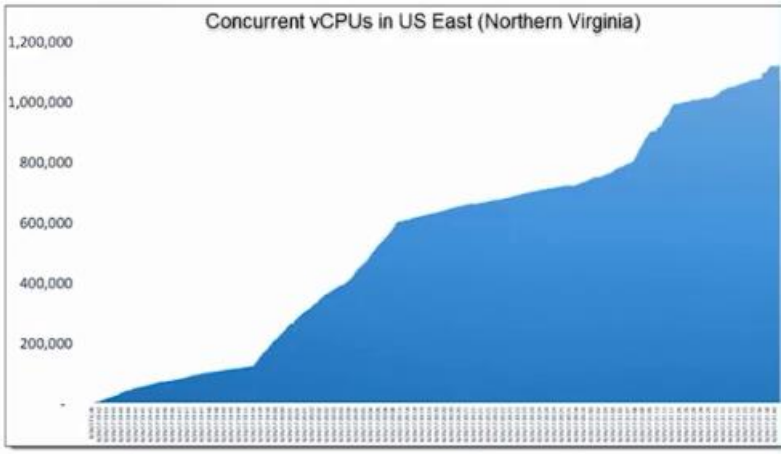
Natural Language Processing at Clemson University – 1.1 Million vCPUs & EC2 Spot Instances

by Jeff Barr | on 28 SEP 2017 | in Customer Success, EC2 Spot Instances | Permalink | Comments | Share

My colleague Sanjay Padhi shared the guest post below in order to recognize an important milestone in the use of EC2 Spot Instances.

-- Jeff;

A group of researchers from Clemson University achieved a remarkable milestone while studying topic modeling, an important component of machine learning associated with natural language processing, breaking the record for creating the largest high-performance cluster in the cloud by using more than 1,100,000 vCPUs on Amazon EC2 Spot Instances running in a single AWS region. The researchers conducted nearly half a million topic modeling experiments to study how human language is processed by computers. Topic modeling helps in discovering the underlying themes that are present across a collection of documents. Topic models are important because they are used to forecast business trends and help in making policy or funding decisions. These topic models can be run with many different parameters and the goal of the experiments is to explore how these parameters affect the model outputs.



Provisioning and Workflow Automation Framework

aws

HPC grids in financial services

Using GPU Acceleration

"Using AWS helps us **reduce a 10-day process to 10 minutes**. That's transformative: it broadens our ability to discover."

–Peter Phillips
Managing Director, Aon Benfield Securities

AON

The challenge
Spinning up large numbers of GPUs quickly and inexpensively to meet ABSI's customers financial modeling and reporting needs

ABSI uses proprietary algorithms (Monte Carlo simulations) running millions of times

The solution
ABSI moved its infrastructure to AWS and deprecated its co-located data center

ABSI built a front end on AWS for its processing solution, automatically running GPU instances on Amazon EC2 using Amazon EBS in an Amazon VPC for security.

The result
Can be as much as 500 times more efficient in terms of performance per dollar for some clients

AWS re:Invent

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

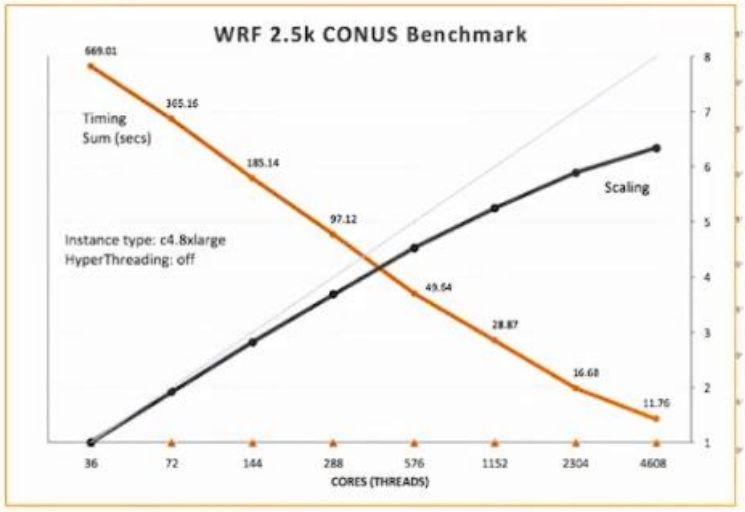
aws

aws re:Invent

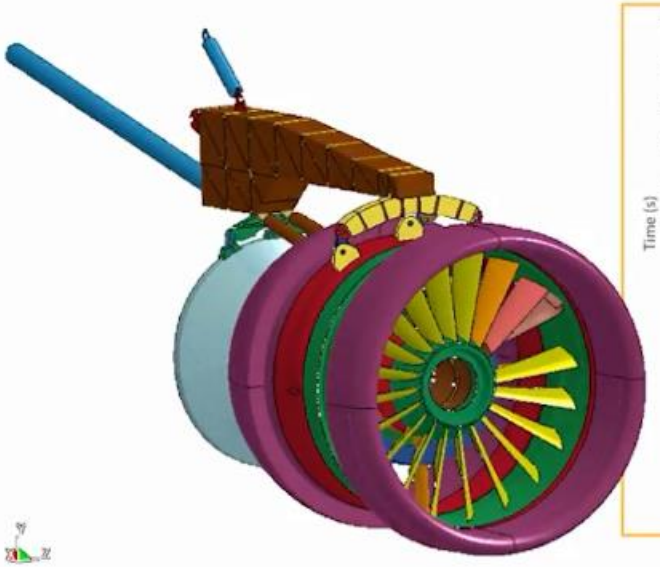
Cluster computing examples

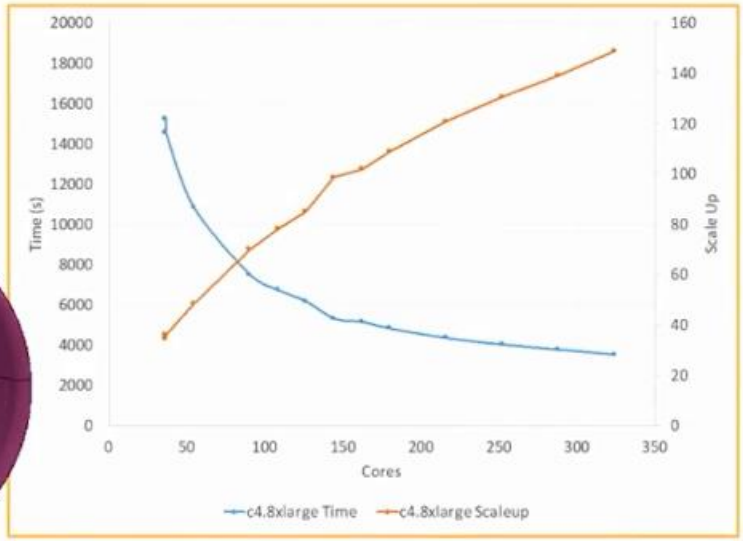
(Scale-up)

Tightly coupled HPC—weather



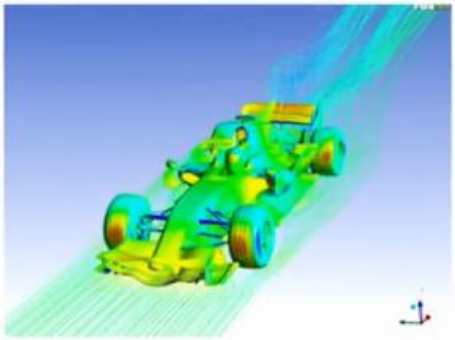
Structural simulation

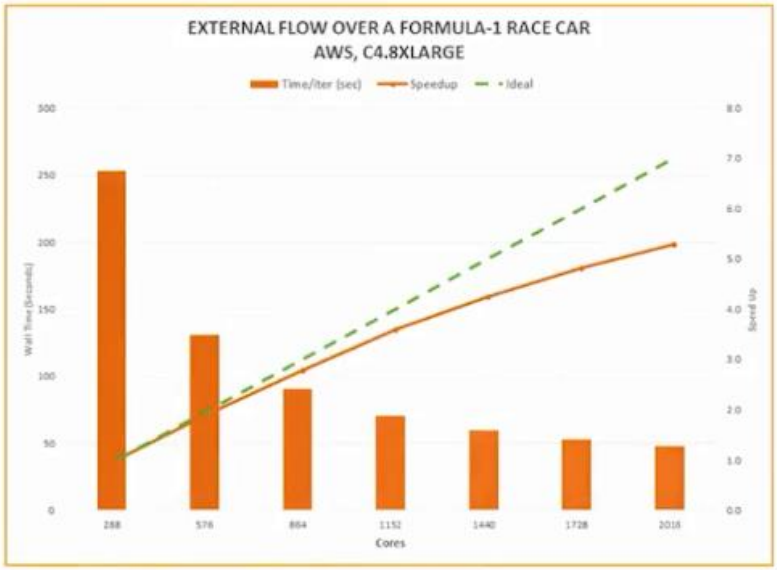




Fluid dynamics—Ansys Fluent

- C4.8xlarge instance type
- 140M cell model
- F1 car CFD benchmark





aws re:Invent

aws

HPC in aerospace

Boom leverages rescale and AWS to enable supersonic travel

boom

- 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, and in-house IT and software teams
- Elastic HPC capacity and pay-as-you-go AWS clusters allow business agility and ability to scale

"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 & cofounder

aws

Performance considerations

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 per-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 or an updated 3.10+ kernel–4.0+ if using NVME on F1 or I3

Processor states

- Use P-states to reduce processor variability

Instance types

- C5, C4, M4, R4 are the best choices today—but always test with the latest EC2 instances

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

Choice of AWS instances for HPC

General purpose	Compute optimized	Storage and IO optimized	Memory optimized	GPU, FPGA accelerated
M4	C5, C4	I3, D2	X1, R4	F1, P2

Instance sizes: R4 example

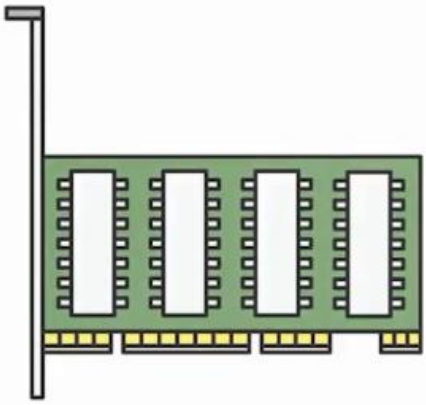
R4 instances are optimized for memory-intensive applications

- Xeon E5-2686 v4 processors
- DDR4 Memory
- Enhanced Networking, up to 25 Gb throughput

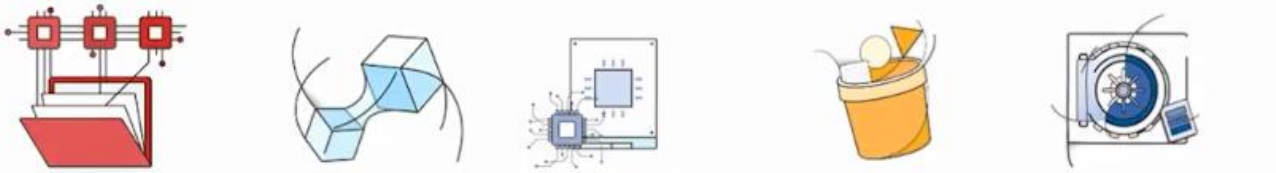
Model	vCPU	Mem (GiB)	Networking Performance	SSD Storage (GB)
r4.large	2	15.25	Up to 10 Gigabit	EBS-Only
r4.xlarge	4	30.5	Up to 10 Gigabit	EBS-Only
r4.2xlarge	8	61	Up to 10 Gigabit	EBS-Only
r4.4xlarge	16	122	Up to 10 Gigabit	EBS-Only
r4.8xlarge	32	244	10 Gigabit	EBS-Only
r4.16xlarge	64	488	25 Gigabit	EBS-Only

Elastic Network Adaptor (ENI)

- Latest generation of enhanced networking
  - Hardware checksums
  - Multi-queue support
  - Receive side steering
- 25 Gbps in a placement group
- Open source amazon network driver



AWS Storage is a platform



Amazon EFS	Amazon EBS	Amazon EC2 Instance Store	Amazon S3/S3-IA	Amazon Glacier
File	Block		Object	

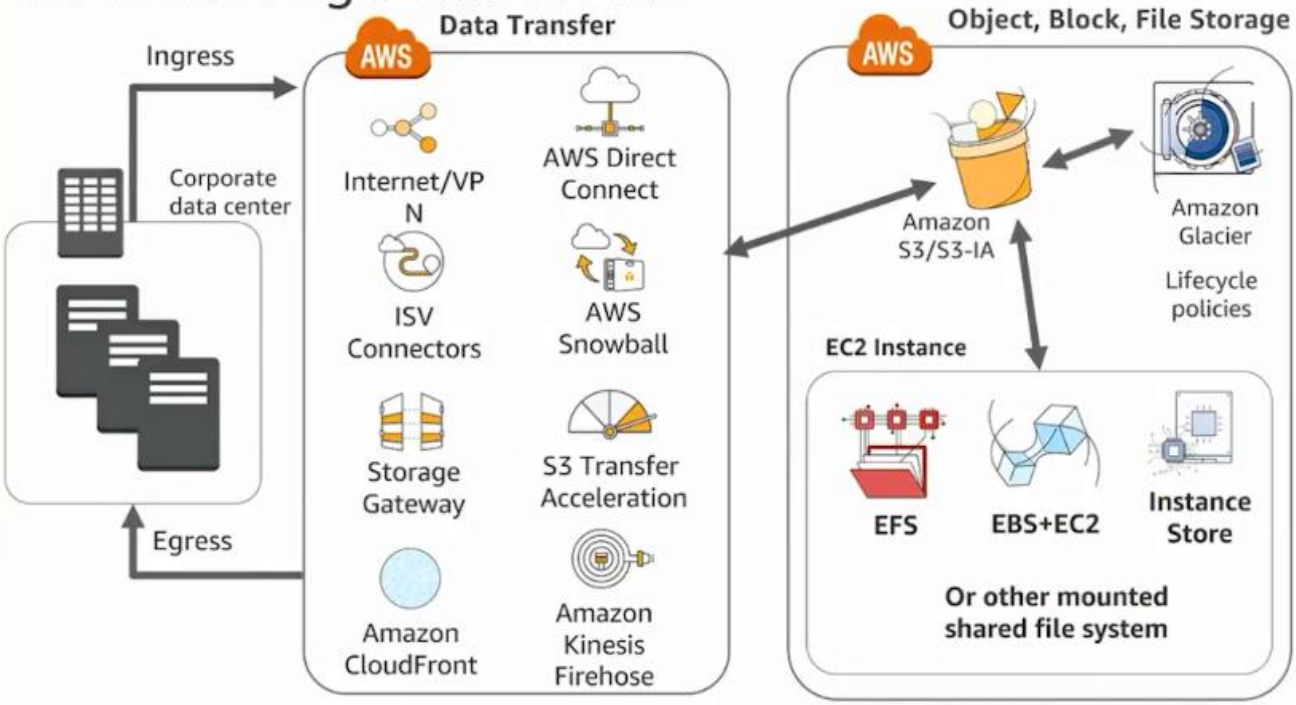
Data Transfer

Internet/ VPN | AWS Direct Connect | Amazon CloudFront | S3 Transfer Acceleration | ISV Connectors | Storage Gateway | AWS Snowball | Amazon Kinesis Firehose

Optimize HPC storage

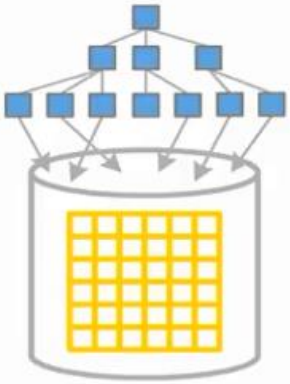
Amazon EFS	EBS+EC2	Amazon S3	Amazon Glacier
Highly available, multi-AZ, fully managed network-attached elastic file system.	Create a single-AZ, shared file system using EC2 and EBS, with third-party or open source software (ZFS, Weka.io, Avere, Intel Lustre, etc.).	Secure, durable, highly scalable object storage. Fast access, low cost.	Secure, durable, long term, highly cost-effective object storage.
For near-line, highly-available storage of files in a traditional NFS format (NFSv4).	For near-line storage of files optimized for high IOPS.	For long-term durable storage of data, in a readily accessible get/put access format.	For long-term storage and archival of data that is infrequently accessed.
Use for read-often, temporary working storage	Use for high-IOPS, temporary working storage	Primary durable and scalable storage for critical data	Use for long-term, lower-cost archival of critical data

HPC storage dataflow

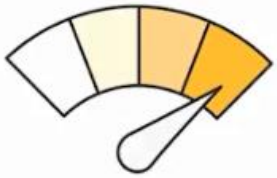


High-performance NFS on AWS

- EC2+EBS is the most performant method of creating scale-up file servers on AWS
- Build your own NFS or CIFS implementation or use a partner solution
- EC2 instances as fileservers, using EBS for block storage—tuned for application needs
- Single fileserver performance up to 25 Gb/s over the network



Accelerated computing



Accelerated computing on AWS

Parallelism increases throughout

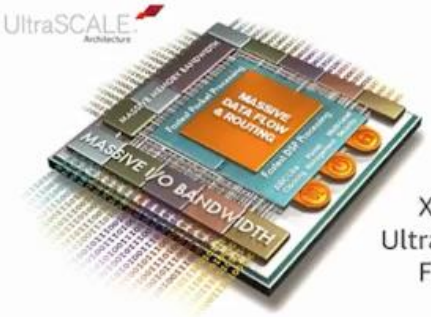
CPU: High speed, highly flexible

GPU/FPGA: High throughput, high efficiency

GPUs and FPGAs can provide massive parallelism and higher efficiency than CPUs for many categories of applications

Accelerated computing on AWS

NVIDIA GPU



Xilinx
UltraScale+
FPGA

P2: GPU-accelerated computing

- Enabling a high degree of parallelism—each GPU has thousands of cores
- Consistent, well documented set of APIs (CUDA, OpenACC, OpenCL)
- Supported by a wide variety of ISVs and open source frameworks

F1: FPGA-accelerated computing

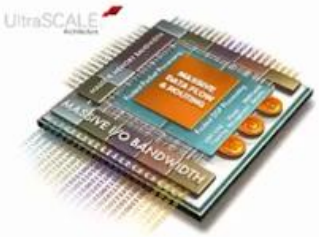
- Massively parallel—each FPGA includes millions of parallel system logic cells
- Flexible—no fixed instruction set, can implement wide or narrow datapaths
- Programmable using available, cloud-based FPGA development tools

AWS re:Invent

aws

Parallel processing: GPU and FPGA

A GPU is effective at processing the same instruction in parallel, for example, calculating pixel values in parallel for graphics shading, or running many parallel financial computations. A GPU has a well-defined instruction set, and fixed word sizes.



An FPGA is effective at processing the same or different instructions in parallel, for example, creating a complex pipeline of parallel, multistage operations on a video stream, or performing a sequence of dependent calculations and data manipulations for genomics processing. An FPGA does not have a predefined instruction set, or a fixed data width.

AWS re:Invent

aws

4D medical imaging on GPU

ARTERYS

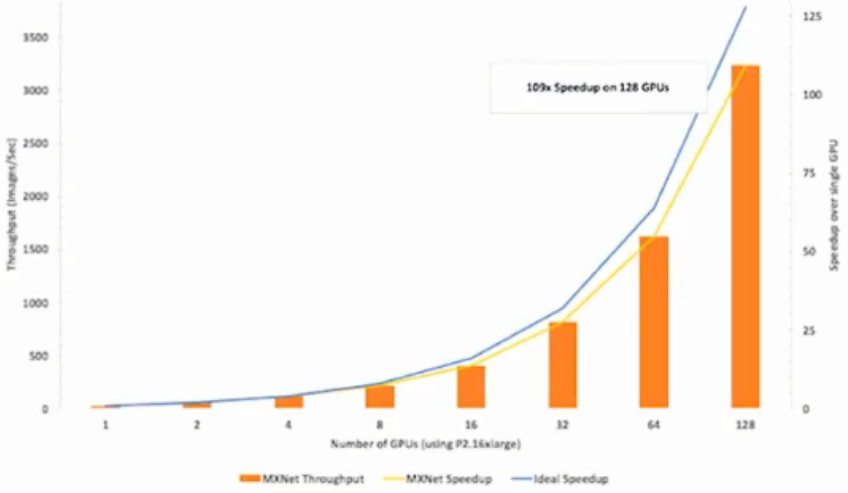
AWS re:Invent

aws

Deep learning on GPU

MXNet training on EC2 P2 instances:

- Training of a popular image analysis algorithm, Inception v3, using MXNet and running on P2 instances
- Scaling efficiency of 85%

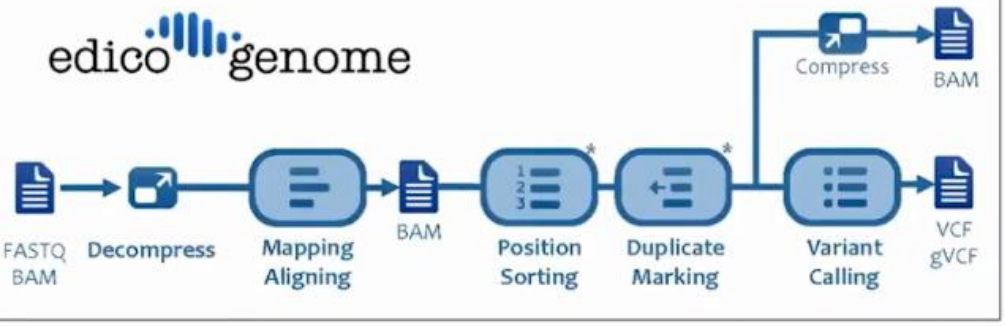


Genomics processing on FPGA

awsmarketplace

DRAGEN Genome Pipeline

DRAGEN's Genome Pipeline is now available in AWS Marketplace. The DRAGEN Genome Pipeline enables ultra-rapid analysis of Next Generation Sequencing (NGS) data, accelerating analysis time from many hours, with pipelines including BWA+GATK or Isaac, to minutes, while also improving accuracy for both SNPs and INDELs. This pipeline harnesses the tremendous power of the DRAGEN Bio-IT Platform and includes highly optimized algorithms for mapping, aligning, sorting, duplicate marking, haplotype variant calling, compression and decompression. Running on a 16xlarge instance, the Genome Pipeline takes ~25 minutes, and ~60 minutes on a 2xlarge instance.

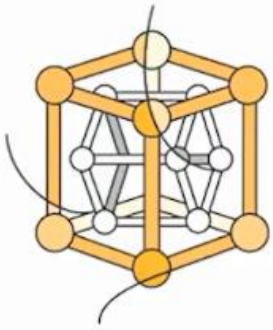


FPGA use-cases and F1 partners

- Financial computing
- Genomics sequencing
- Test and measurement
- Image and video processing
- Big data and machine learning
- Security, compression
- …and more

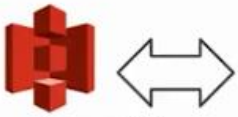
edico genome
Reconfigure.io
Mipsology
MAXELER Technologies
RYFT
Atomic Rules
NATIONAL INSTRUMENTS
NGCODEC
Falcon COMPUTING
TERADEEP
Titan IC

HPC deployment and automation

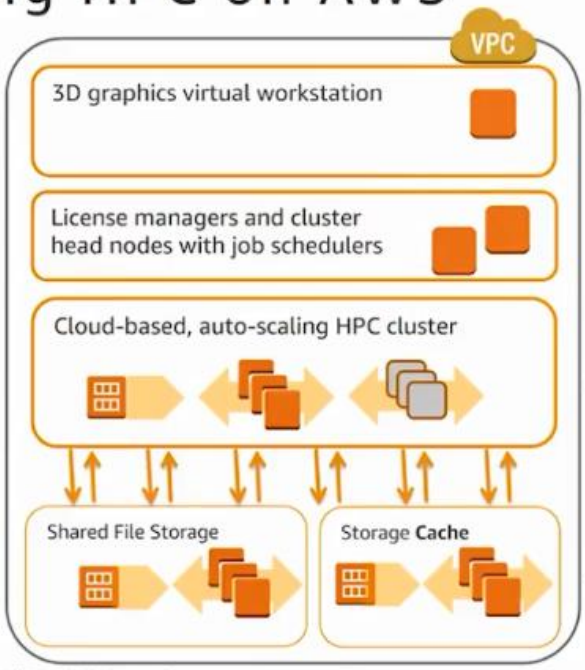


Deploying HPC on AWS

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



Amazon S3 and Amazon Glacier



Thin or Zero Client
- No local data -

Corporate datacenter

AWS Snowball

On-Premises HPC Resources

AWS Direct Connect

aws

AWS re:Invent

HPC automation with CfnCluster

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

Built on AWS CloudFormation, easy to modify to meet specific application or project requirements

CfnCluster

CfnCluster is a tool used to build and manage High Performance Computing (HPC) clusters on AWS.

Once created, you can log into your cluster via the master node where you will have access to standard HPC tools such as schedulers, shared storage, and an MPI environment.

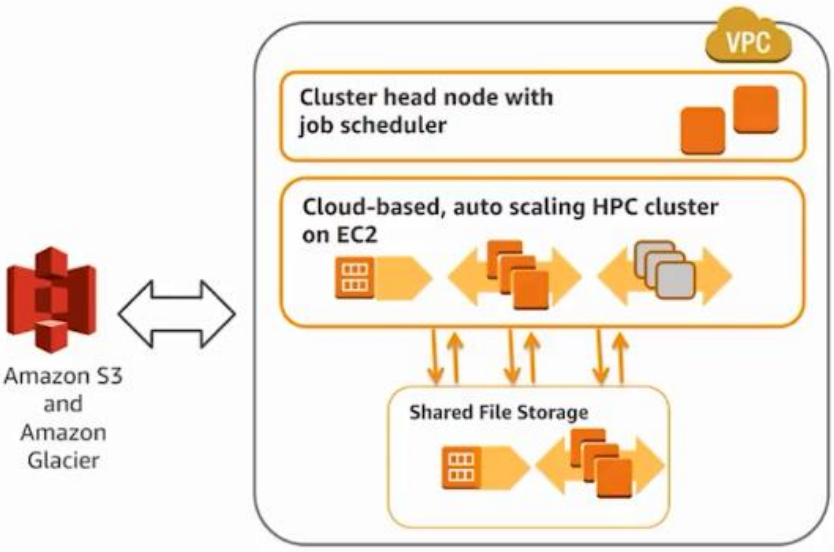
Getting Started »

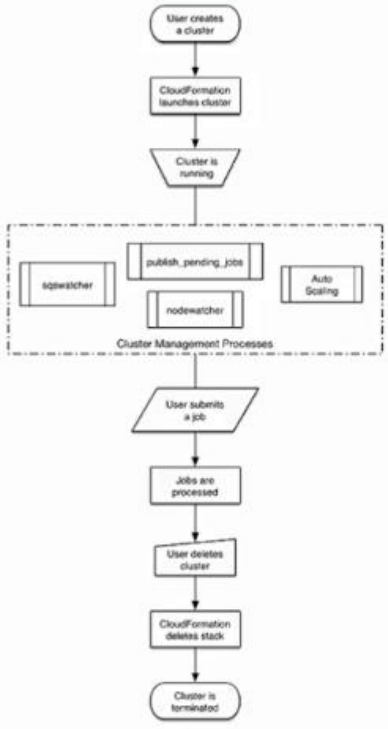
CLI Reference »

GitHub Project »

Community Forum »

CfnCluster HPC stack





AWS Batch for HPC workloads

Fully managed

No software to install or servers to manage. AWS Batch provisions, manages, and scales your infrastructure.

Integrated with AWS

Natively integrated with the AWS Platform, AWS Batch jobs can easily and securely interact with services such as Amazon S3, DynamoDB, and Amazon Rekognition.

Cost-optimized resource provisioning

AWS Batch automatically provisions compute resources tailored to the needs of your jobs using Amazon EC2 and EC2 Spot.

Hybrid HPC
Using Amazon EC2 Systems Manager

- Manage thousands of Windows and Linux nodes running on Amazon EC2 or on premises
- Control user actions and scope with secure, granular access control
- Safely execute changes with rate control to reduce blast radius
- Audit every user action with change tracking

Capabilities

Run Command

State Manager

Inventory

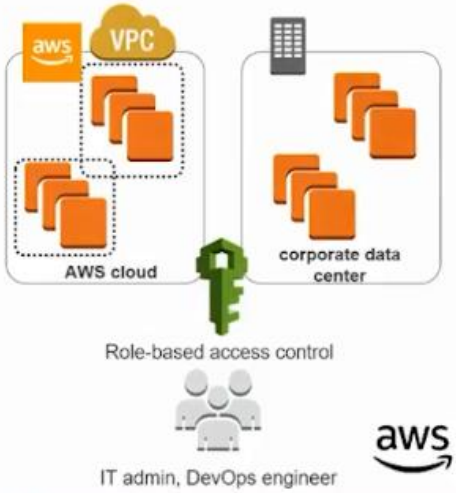
Maintenance Window

Patch Manager

Automation

Parameter Store

Documents



Graphics for HPC applications

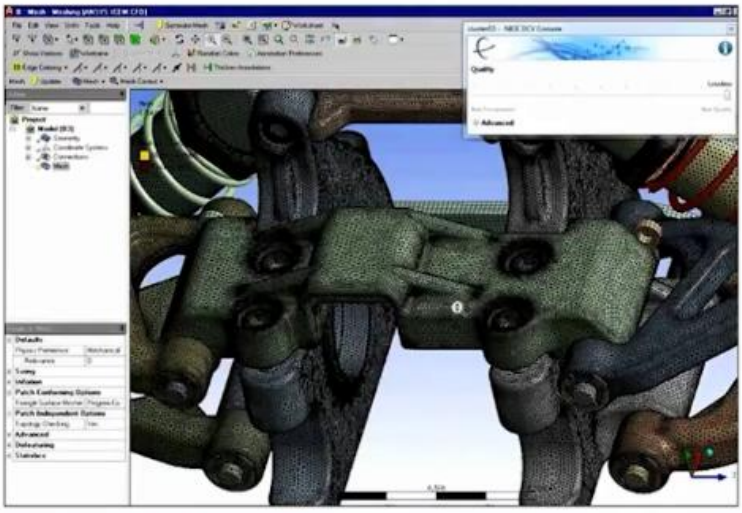
Secure graphics and collaboration

Cloud can be used for pre-and post processing as well as HPC

- Use GPUs in the cloud for remote rendering and remote desktops

Cloud is more secure for collaboration

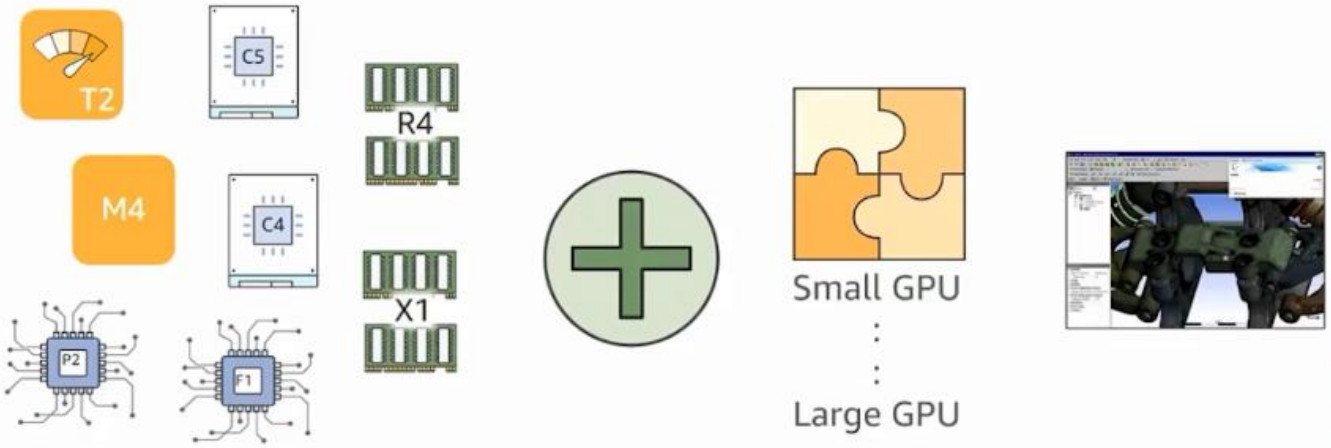
- Encrypt the data in flight and at rest
- Manage your own keys and credentials
- Deliver pixels to your collaborators, not the actual data



AWS re:Invent

HPC workstations in the cloud

EC2 + Elastic GPU



Attach Elastic GPU to an instance at launch, similar to attaching an EBS volume

Elastic GPU for HPC applications

- Wide range of available instance types depending on application needs, for example, high-memory
- Graphics attachment in any of a range of sizes depending on size of the project or model
- More cost-effective, higher performance than a one-size-fits-all technical workstation

Desktop application streaming
with Amazon AppStream 2.0

Run desktop apps in a web browser

Stream desktop applications securely to any web browser

Secure apps and data

Pay as you go

Scale globally

aws

AppStream 2.0 graphics support

- Multiple Instance types—including General Purpose, Compute Optimized, Memory Optimized, Graphics Design, Graphics Pro, and Graphics Desktop
- Always-On or On-Demand pricing models
- Support for OpenGL, DirectX, OpenCL and CUDA

aws

Summary

- HPC encompasses a wide range of applications—if you need more computing to solve a problem than one desktop or server, you need HPC
- HPC on AWS provides flexibility, speed of deployment, automation, and scale
- Graphics is key for delivering and viewing HPC results—HPC on AWS is not just about batch processing