AWS State, Local, and Education Learning Days

Washington, DC

aws Learning Days
State, Local, and Education

AWS ParallelCluster and Amazon SageMaker for Research

Run High Performance Computing (HPC) and Artificial Intelligence/Machine Learning for research

Jianjun Xu, Ph.D.

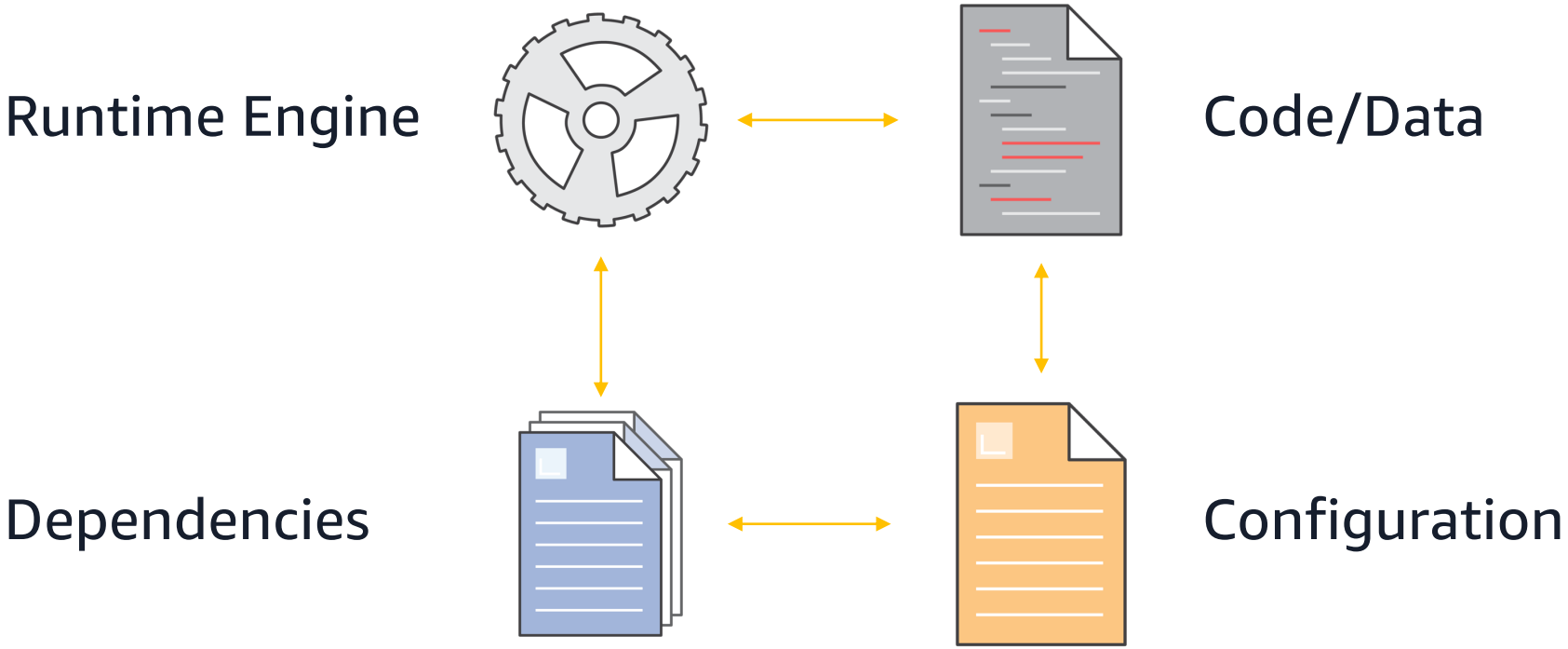
Principal Solutions Architect
Amazon Web Services
jianjx@amazon.com

Niris Okram

Sr. Solutions Architect
Amazon Web Services
niris@amazon.com

aws Learning Days
State, Local, and Education

Run application/tool/program/simulation

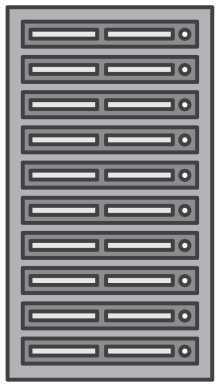


Moving to another environment is hard

v7.0.0

Local Laptop

v6.0.0

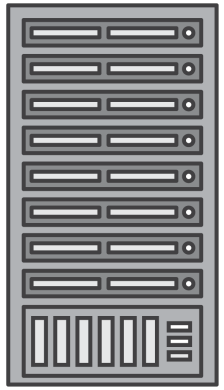


Lab

v4.0.0

Collaborator

v5.0.0



On-prem

Reproducing is hard

Current

Local Laptop

Time of publishing

Local Laptop

Time after publishing

Collaborator

Scaling is hard

Need more CPUs/GPUs/Memory/Storage

- Vertical scaling
- Newer generation of chips
- Software and data

Need more instances

- Horizontal scaling
- On-prem cluster limited size
- Job queues

Scheduling is hard

Cluster capacity

- Wait time is long
- Hardware availability limitation
- Hardware capability limitation
- Fixed size of nodes and storage

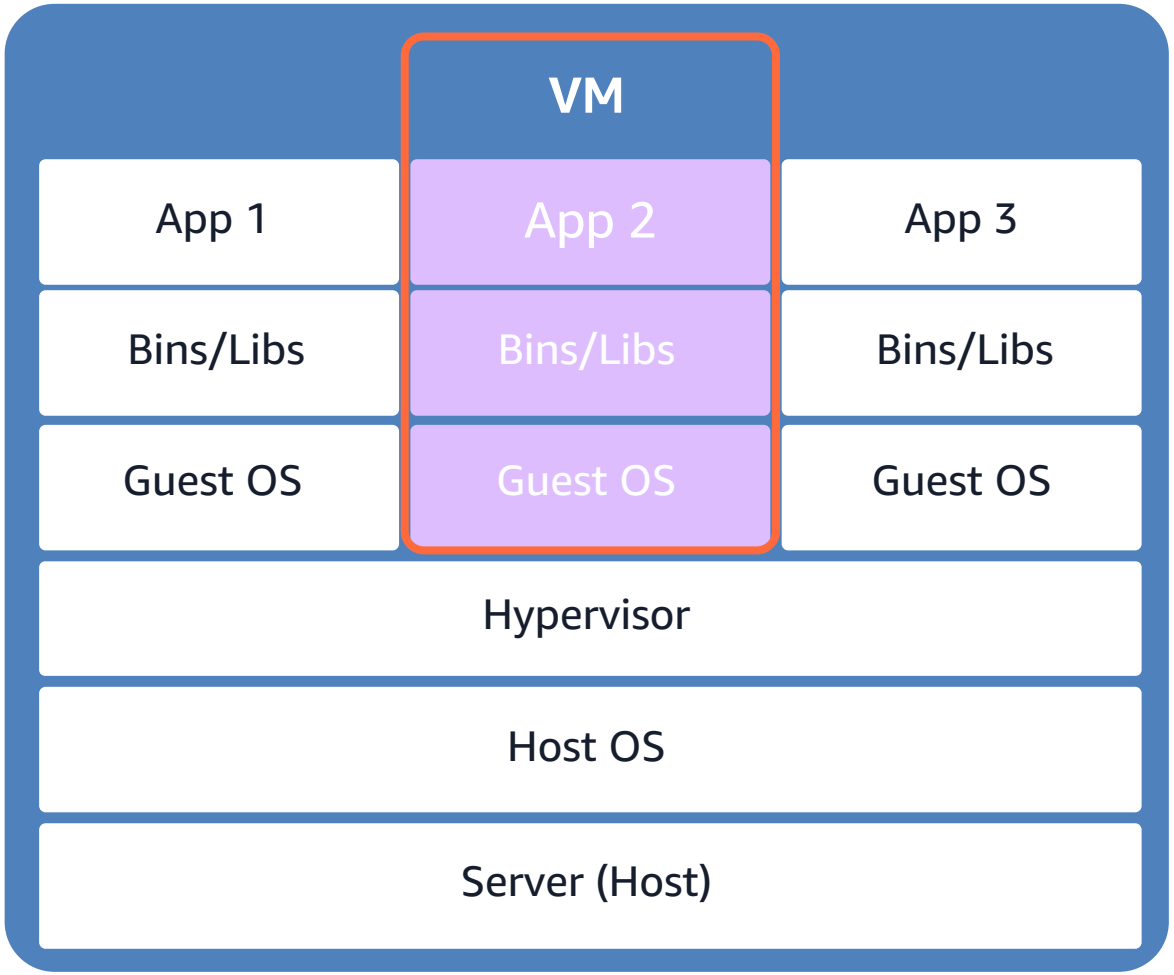
Right tool for the right job

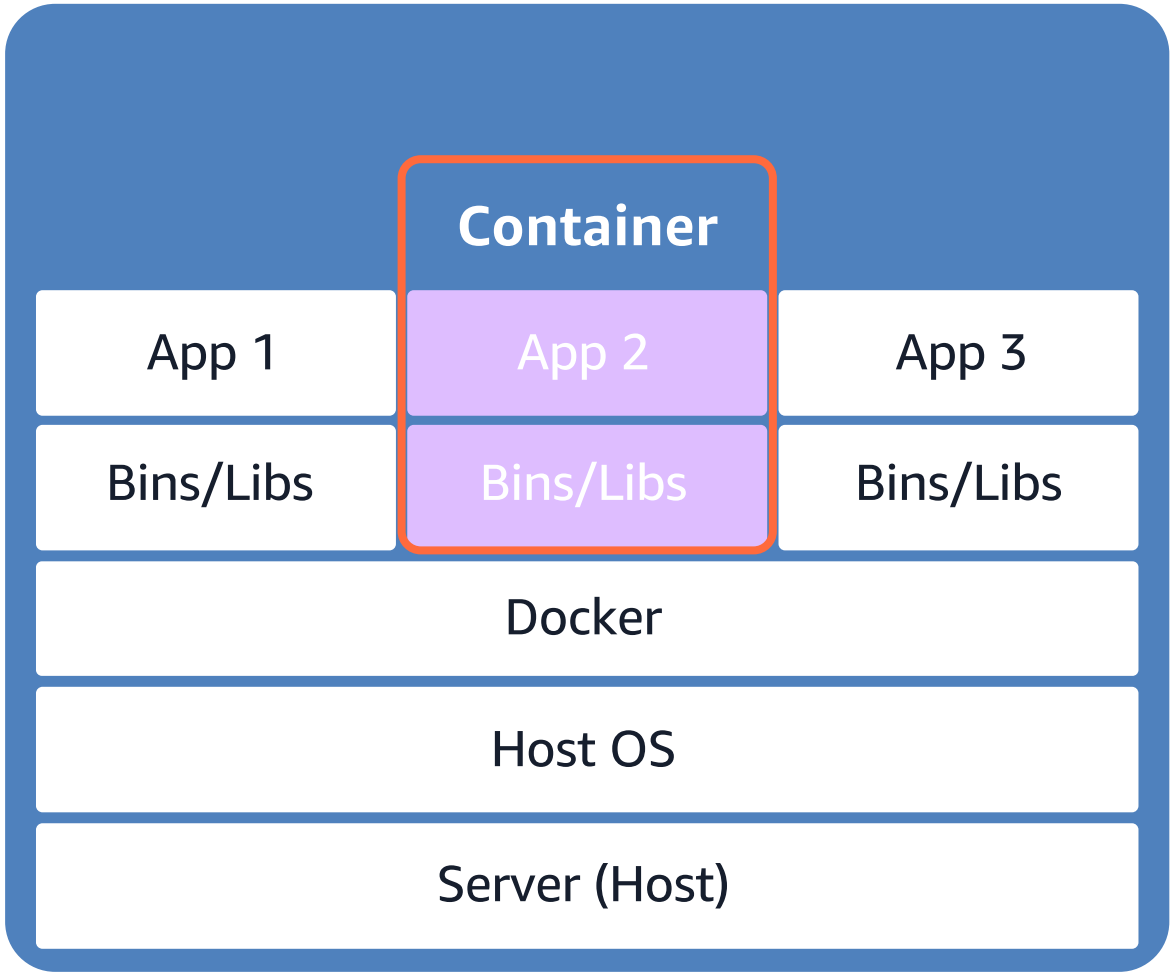
- Homogeneous compute types
- Homogeneous storage types

Bursting capability

- One-off tasks
- Tasks require specialty hardware
- Tasks need to complete quicker

Virtualization to the rescue



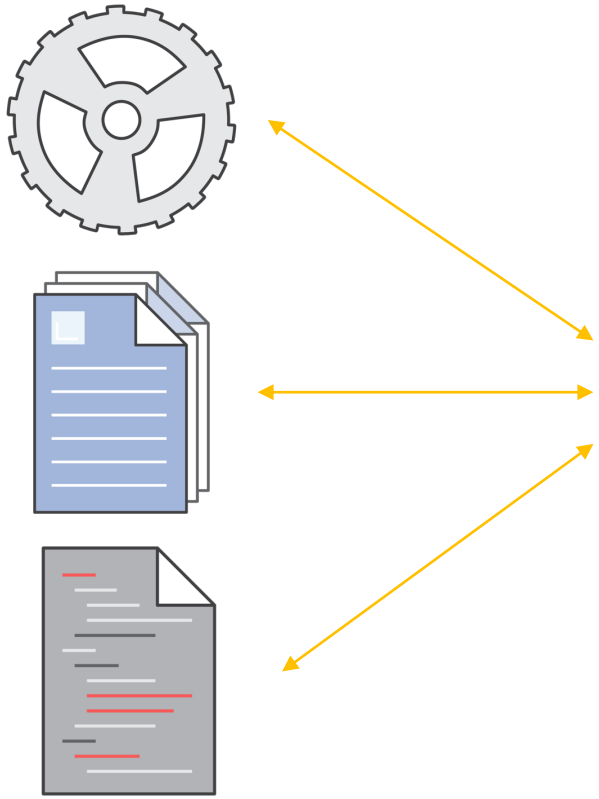


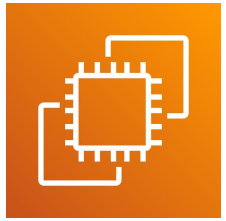
Virtual Machine – Amazon Elastic Compute Cloud (EC2)

Runtime Engine

Dependencies

Code





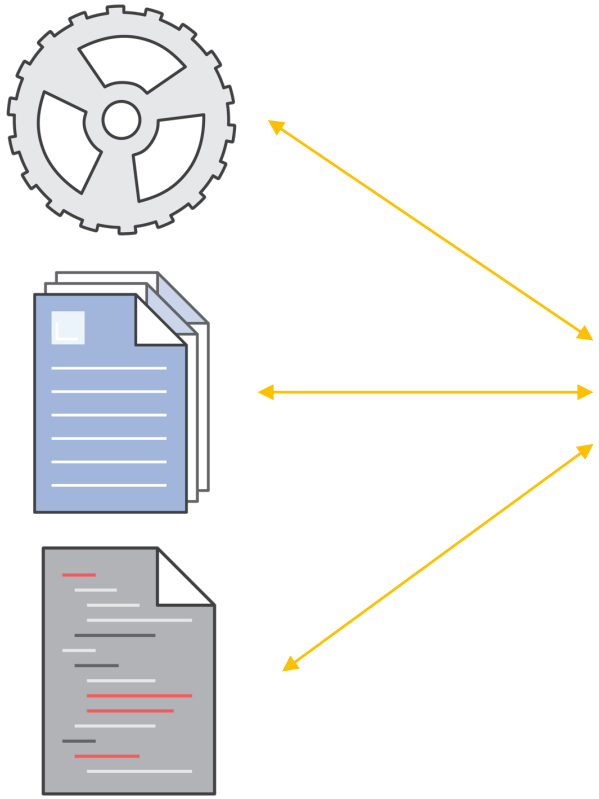
Amazon Elastic Compute Cloud (Amazon EC2)

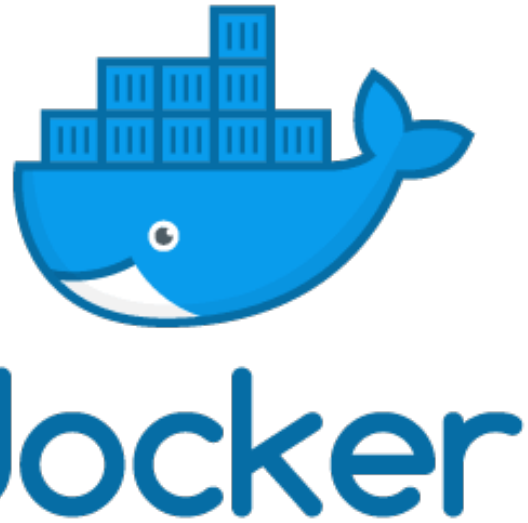
Container

Runtime Engine

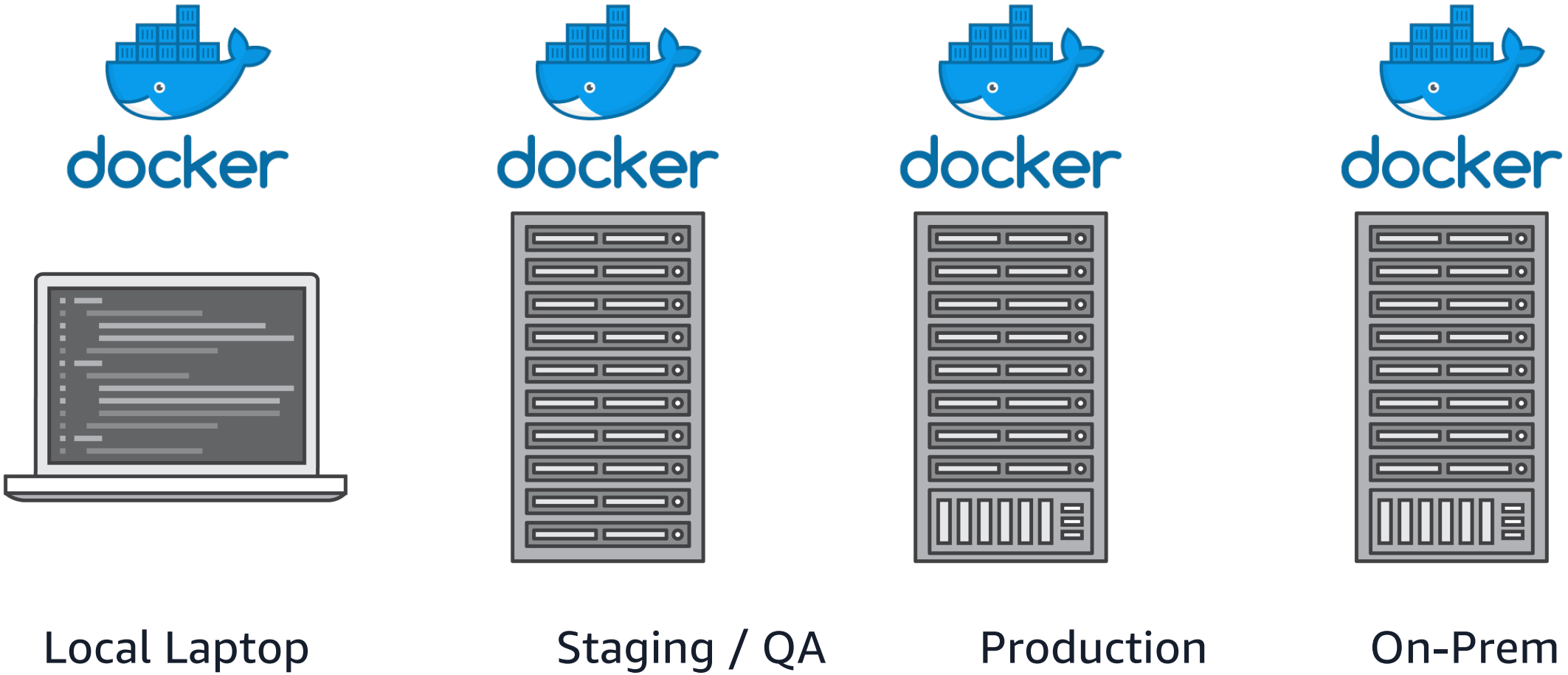
Dependencies

Code





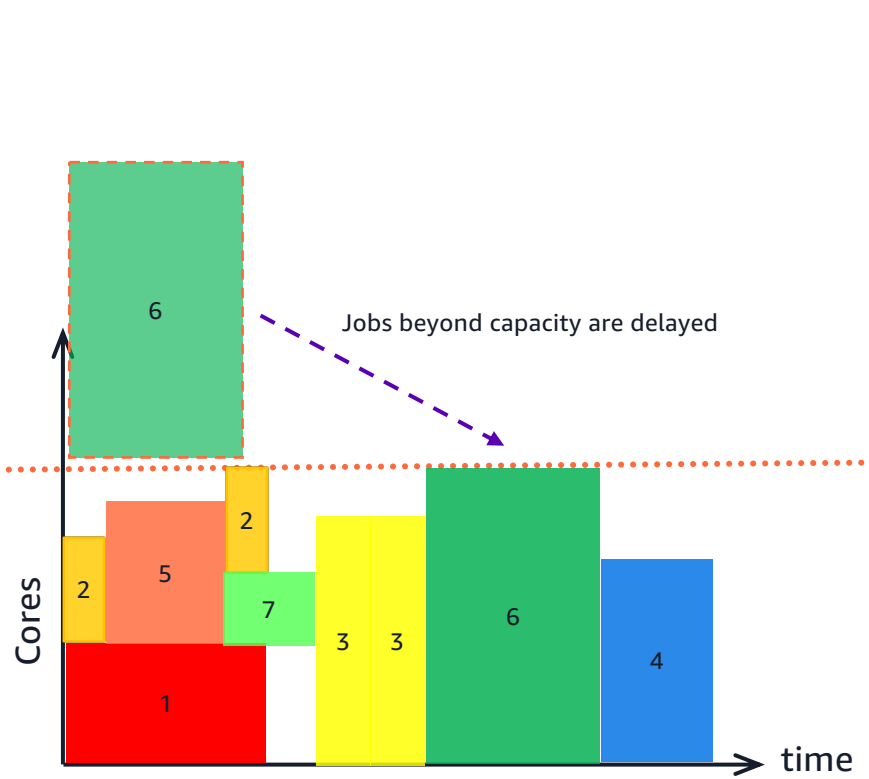
Four environments, same container



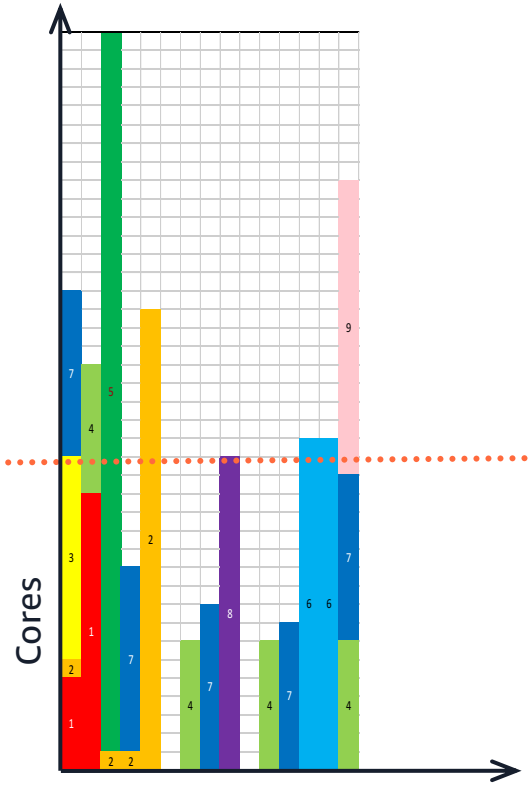
Owning hardware can be expensive

- Hardware not used 100%
- Upgrade and patching
- Compute and storage must be over-provisioned
- Archived data can take up valuable storage
- Require frequent refresh

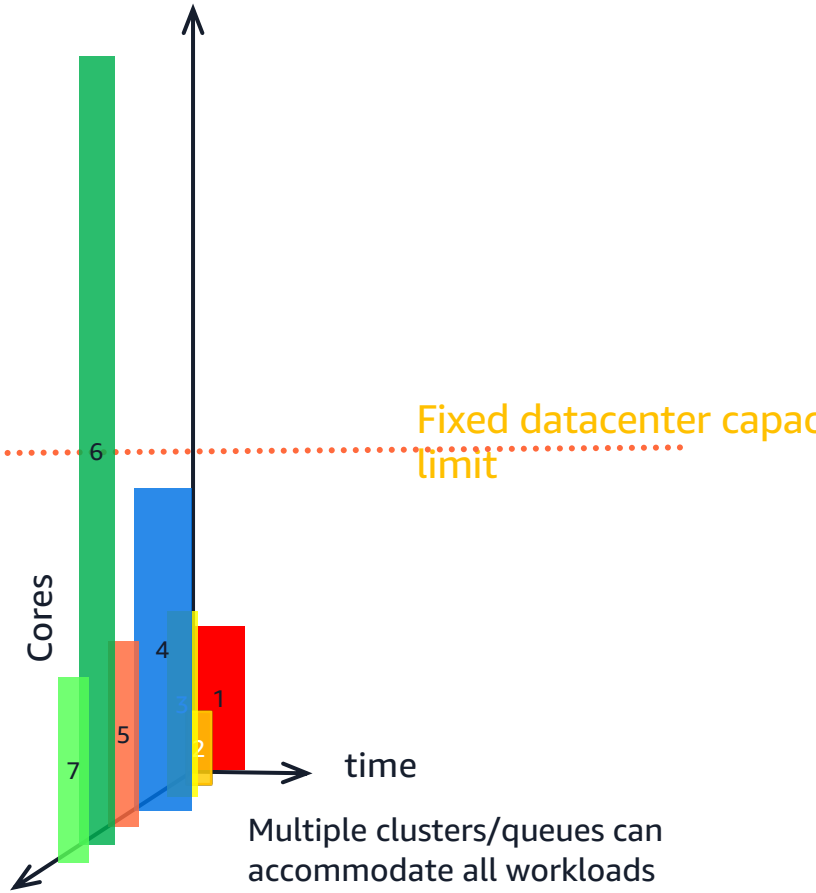
We think the metric for success for any researcher should be time-to-results



Finite capacity, usually with long queues to wait in
- on-prem cluster



Single cluster with massive capacity can shorten job run time with more cores



Multiple clusters/queues can accommodate all workloads at the same time – even faster to results

High Performance Computing (HPC) in the Cloud

Native and containerized application in a managed environment

Why HPC on AWS?
Capacity and capability

- Virtually unlimited infrastructure enabling scaling and agility not attainable on-premises
- Instant access to latest technologies with no lengthy procurement cycles or big capital investments
- Flexible configuration options quickly iterate resource selection and ensure cost optimization
- Virtual storage and compute enabling snapshot/restore for reproducibility

Better ROI

Faster time to results

Why HPC on AWS?
Simplicity

- Built-in security (IAM , encryption etc.)
- Easy to use – with managed infrastructure
- Many ways to run containers in fully managed environment
- Quick to provision cluster solutions for Slurm, Hadoop, containers

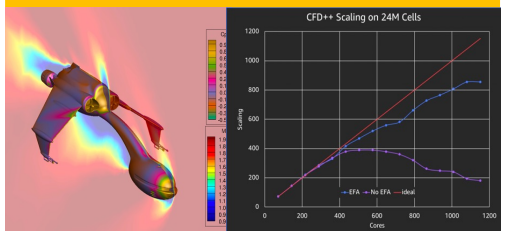
Better ROI

Faster time to results

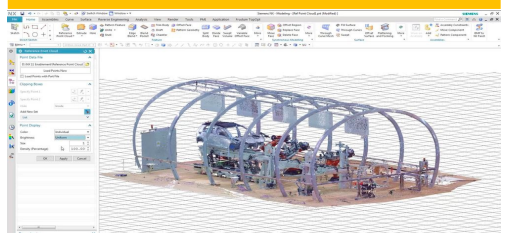
Flexible compute options/purchase models

Flexible compute to maximize performance

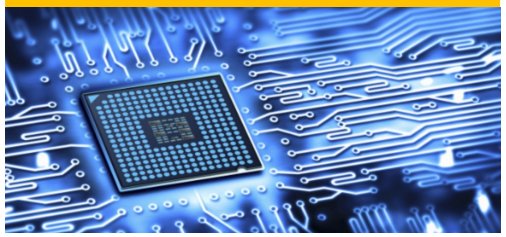
Memory & compute optimized
2.5-3.5Ghz, 2-16GB/core, 100Gbps, EFA



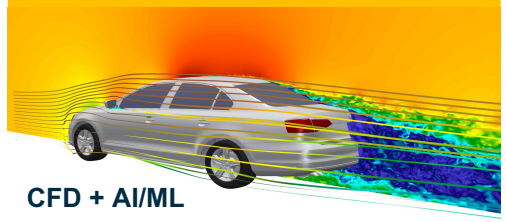
Graphics and rendering
1-8 GPUs, up to 384GB RAM, SSD



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



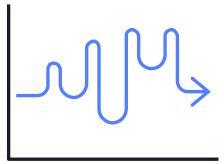
Accelerated computing
8 A100 GPUs, 1.1TB RAM, SSD, 400Gbps



CFD + AI/ML

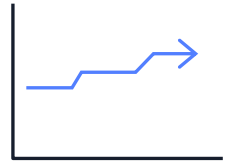
Flexible pricing models to optimize cost

On-Demand



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

Savings Plan & Reserved Instances

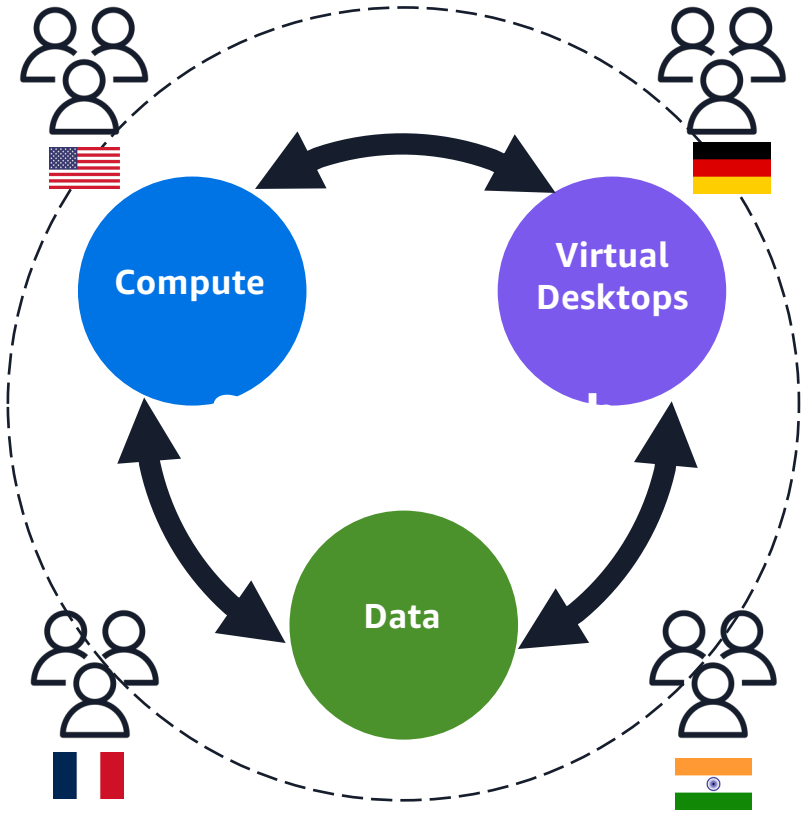


Make a commitment and to save up to 72% off compute.

Spot Instances

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

A Secure Research portal in the cloud



200+ Fully Managed Services, 1000's of Features

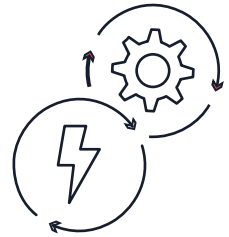
- Servers (600+ Types)
- Graphics Servers
- Gigabit Interconnect
- Machine Images
- Serverless Apps
- Infrastructure as Code
- Auto Scaling
- Auto Formation
- Activity Tracking
- Resource Monitoring
- Control Tower
- Dashboards
- License Manager
- Disaster Recovery
- Access Management
- Console
- Cluster Orchestration
- Job Scheduling
- HPC Web Interface
- Visualization Service
- Application Streaming
- Work Spaces
- Block Storage
- Object Storage
- File System
- High Perf. File System
- Data Syncing
- Backups
- Intelligent Tiering
- Cost Monitioring
- Budget Management
- Notification Service
- WAN Connections
- VPC
- Subnets
- Route Tables
- NAT Gateways
- Internet Gateways
- Transit Gateways
- Access Control Lists
- Security Groups
- VPN
- Firewalls
- Security Inspection
- Threat Detection
- Identity Management
- Active Dir. Connection
- Credential Mgmt.

. . . and many more

HPC workloads with different compute and throughput characteristics

Tightly-coupled workloads

Loosely-coupled workloads



Accelerated computing

Visualization

AI/ML

High volume data analytics

HPC on AWS

Amazon CloudWatch

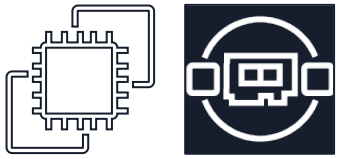
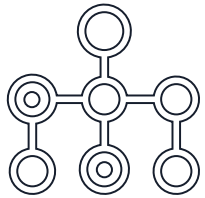
Applications and Services

Automation and orchestration	Storage	Compute	Visualization	Networking
AWS Batch	Amazon EBS	Amazon EC2 instances	NICE DCV	Enhanced networking
AWS ParallelCluster	Amazon FSx for Lustre	Amazon EC2 Spot	Amazon AppStream 2.0	Placement groups
AWS Paralllel Computing Services	Amazon EFS	AWS Auto Scaling		Elastic Fabric Adapter
Research and Engineering Studio	Amazon S3			

Amazon IAM (Identity and Access Management)

AWS Budgets

Key services that enable HPC on AWS

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

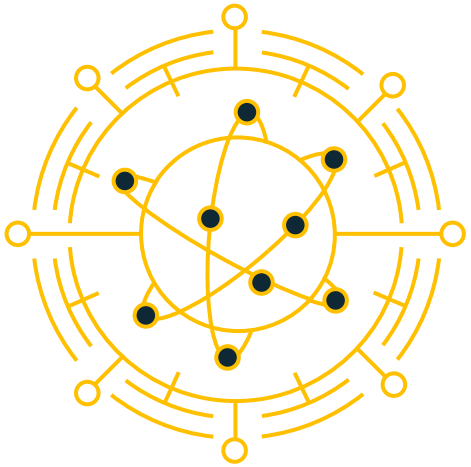
Run your HPC workloads with the price performance you expect and the security you demand

Cloud Computing Basics
Compute and Storage

Amazon EC2

The compute platform for every workload

Nearly 700+ instance types



Machine Learning

High-Performance Computing

Media Rendering

Containers

Web Based Apps

Batch Processing

Big Data

Scale tightly and loosely-coupled HPC applications on AWS

- Choice of processor (e.g. Intel, AMD, Arm)
- Scale tightly-coupled HPC and ML workloads

  Up to 400 Gbps network bandwidth

  < 15 micro-seconds network latencies

Compute, Memory, and Networking

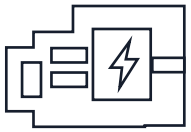
C7gn | C7a | C7i | M7i | R7i | M7a | M7i | M5zn | C7g | M7g | R8g

Accelerators

P5 | G5 | DL2q | Trn1 | F1 | Inf2 | VT1 | P4dn | G5g

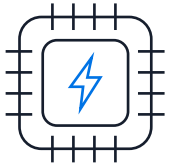
Nitro System is the foundation of AWS

Nitro Card



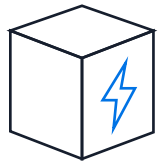
Local NVMe storage
Elastic Block Storage
Networking, monitoring, and security

Nitro Security Chip



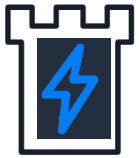
Integrated into motherboard
Protects hardware resources

Nitro Hypervisor



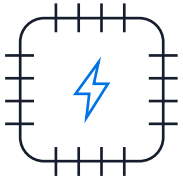
Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance

Nitro Enclaves



Isolated environments for highly sensitive data processing
Utilizes EC2's Isolation Technology

Nitro SSD

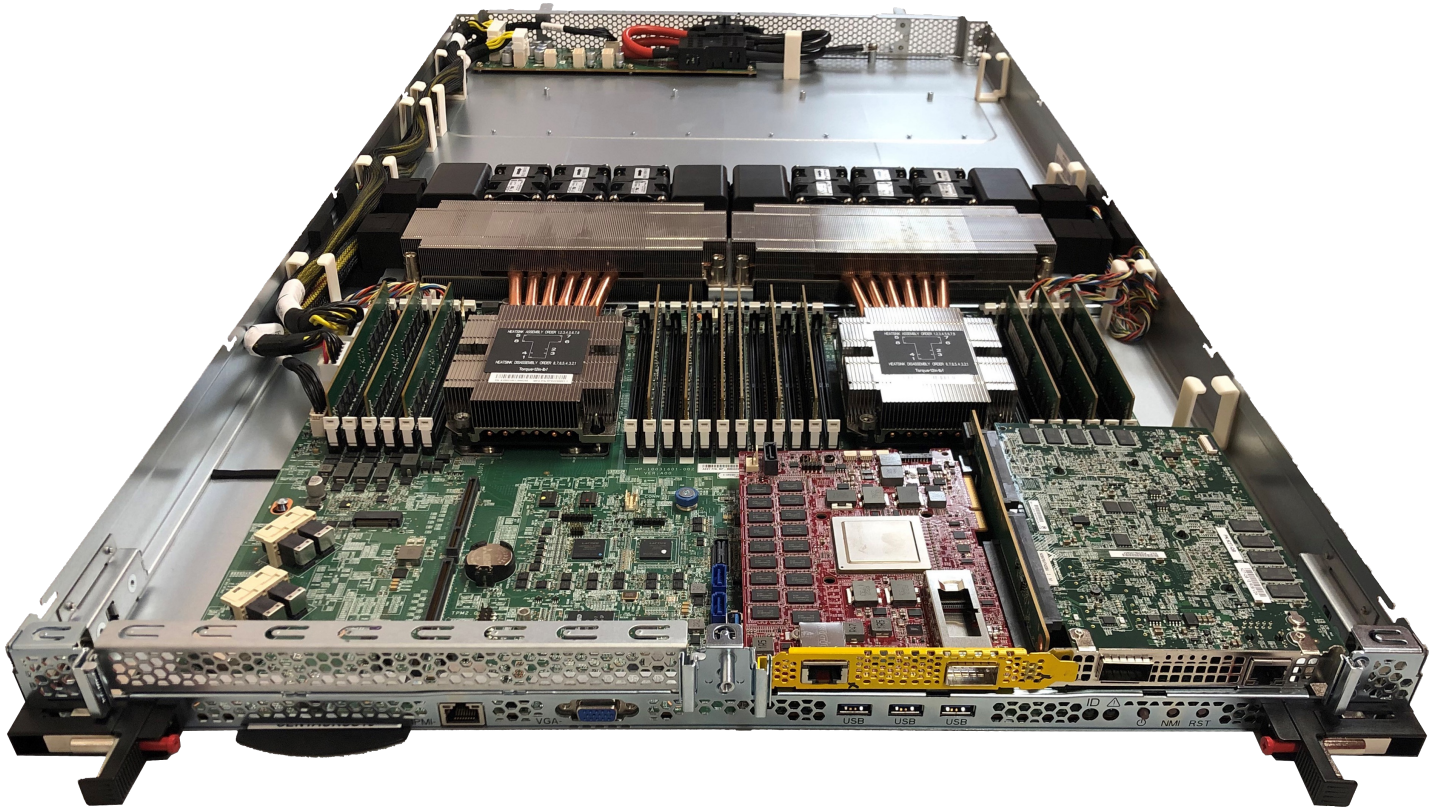


60% lower I/O latency
Firmware Upgrades w/o Interruption
Encryption at rest

Nitro TPM

TPM 2.0 specification
Cryptographic attestation of instances integrity

Nitro-Powered EC2 Server



C5n

Nitro Cards for VPC networking

VPC data-plane offload

Encapsulation, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

Transparent end-to-end 256-bit encryption

ENA Express

Improved Network Latency and Per-Flow Performance on EC2 w/ SRD

Elastic Network Adapter (ENA)

Extendible host interface

Elastic Fabric Adapter (EFA)

Kernel-bypass

Low latency at scale

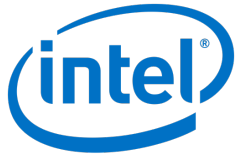
Multipathing (SRD)

High Memory compute instances: U7i

Massively compute and memory

- U7inh Instances will offer up to 1,920 vCPUs and 32 TB of DDR5 instance memory
- Powered by AWS Nitro System
- Ideal for running large enterprise databases, including SAP HANA in-memory database in AWS and other enterprise scale applications.

Featuring Intel Xeon Scalable (Sapphire Rapids) processor

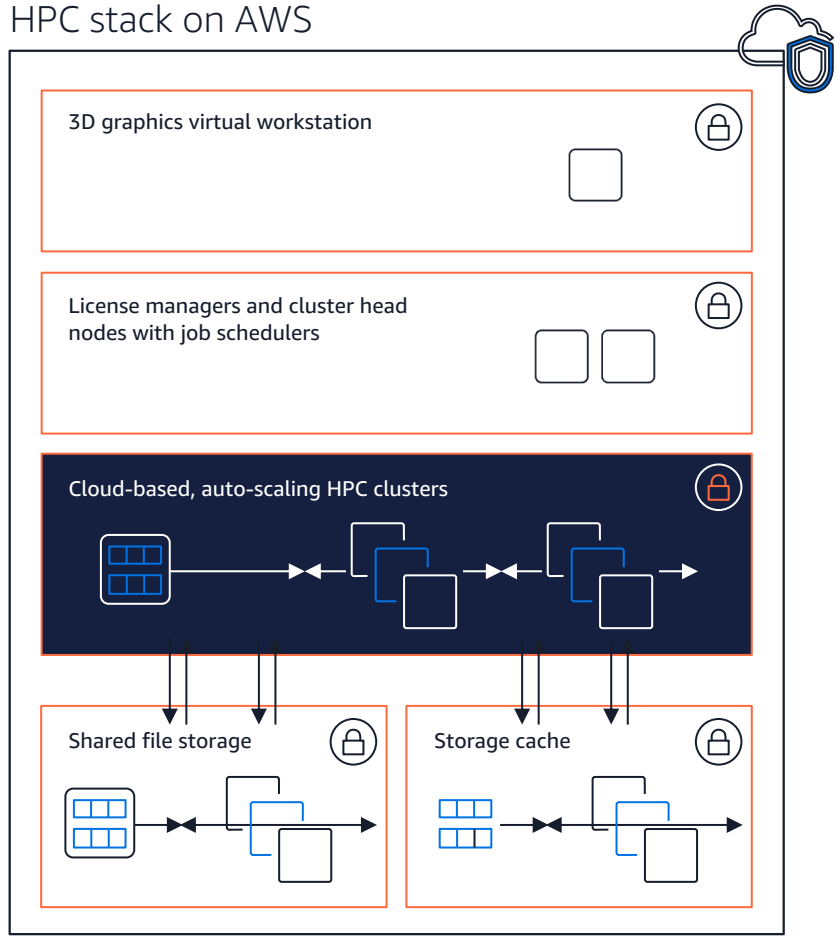


High bandwidth compute instances: hpc7a

Massively scalable performance

- Up to 192 physical cores, 768 GiB memory
- hpc7a Instances will offer up to 300 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 4th gen AMD EPYC 9R 14 processors



Amazon EC2 C8g/C7gn instances powered by AWS Graviton4/3E processors

NEW!

C8g – Up to 192 vCPUs and 384 GiB memory

C7gn - Up 200 Gbps network bandwidth

First in the cloud to feature DDR5 memory

50% higher memory bandwidth (vs DDR4)

Ideal for compute-optimized workloads such as HPC, video encoding, gaming, CPU-based machine learning inference

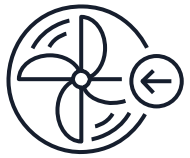
Elastic Fabric Adapter

SRD protocol

NEW!

Up to 400 Gbps networking bandwidth

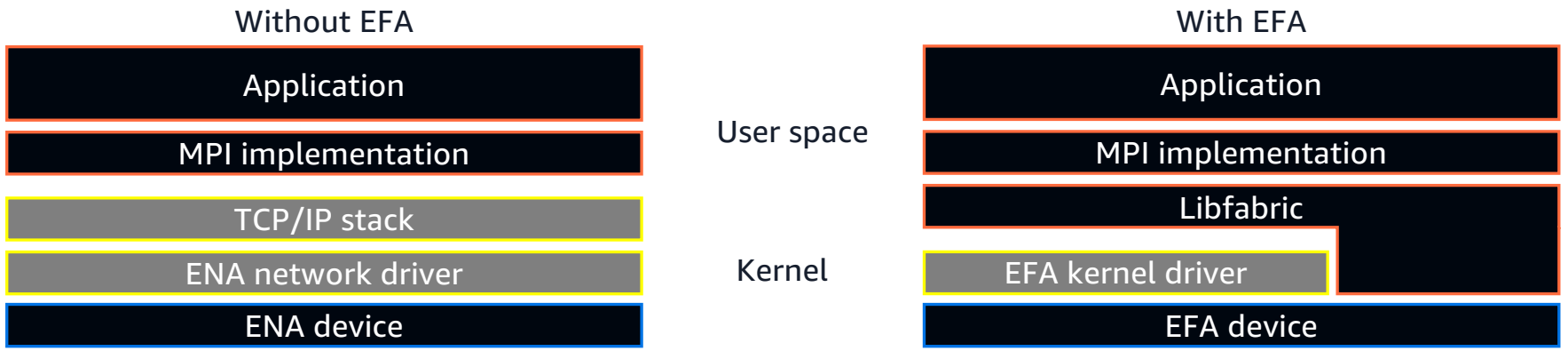
Proving myths about latency constraints wrong



CFD

Seismic

Weather modeling



INEOS | TEAM UK

INEOS uses EFA to run CAE / CFD workloads for boat designs

S3 Object Storage – a new kind of storage

Amazon S3

S3 Standard

- Active, frequently accessed data
- Milliseconds access
- ≥ 3 AZ
- $0.0210/GB

S3 Intelligent-Tiering

- Data with changing access patterns
- Milliseconds access
- ≥ 3 AZ
- $0.0210 to $0.0125/GB ($0.004 to $0.00099/GB Archive)
- No retrieval fees
- Monitoring fee per Obj.
- Min storage duration
- Min object size

S3 Standard-IA

- Infrequently accessed data
- Milliseconds access
- ≥ 3 AZ
- $0.0125/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 One Zone-IA

- Re-creatable, less accessed data
- Milliseconds access
- 1 AZ
- $0.0100/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Instant Retrieval

- Archive data instant retrieval
- Milliseconds access
- ≥ 3 AZ
- $0.0040/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Flexible Retrieval

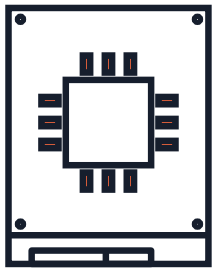
- Archive data
- Select minutes or hours
- ≥ 3 AZ
- $0.0036/GB – ($4.10/TB)
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Deep Archive

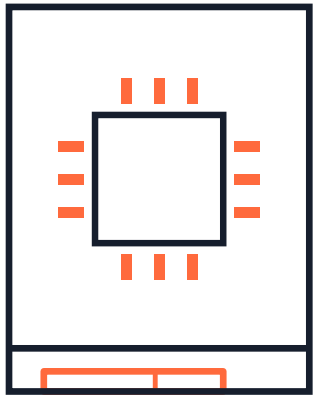
- Archive data
- Select 12 or 48 hours
- ≥ 3 AZ
- $0.00099/GB - ($1.01/TB)
- Retrieval fee per GB
- Min storage duration
- Min object size

Frequent ← *Access Frequency* → *Infrequent*

EBS Block Storage – think hard disks

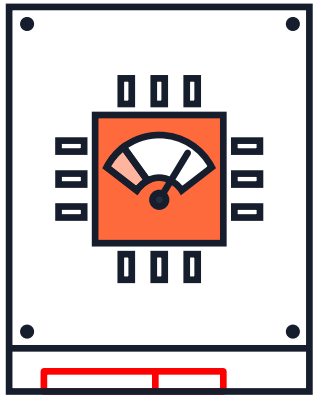


SSD



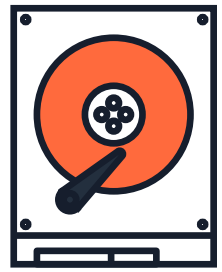
gp2 – gp3

General Purpose SSD

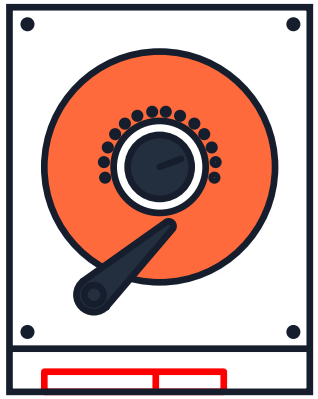


io1 – io2

Provisioned IOPS SSD

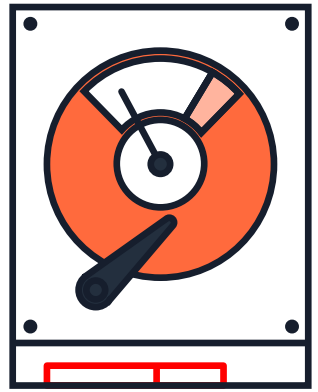


HDD



st1

Throughput Optimized HDD



sc1

Cold HDD

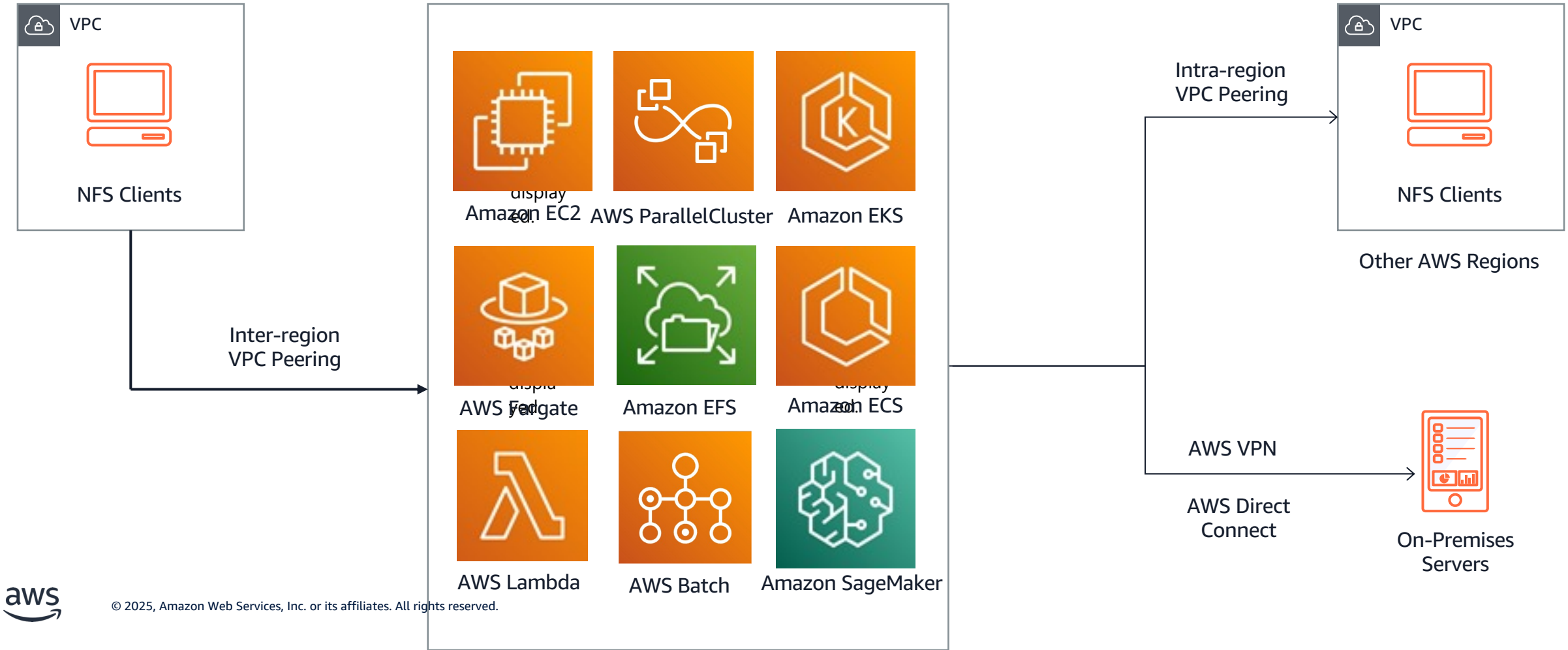
EBS and S3 – great combo

Low cost
Incremental backups do not duplicate data and reduce storage costs

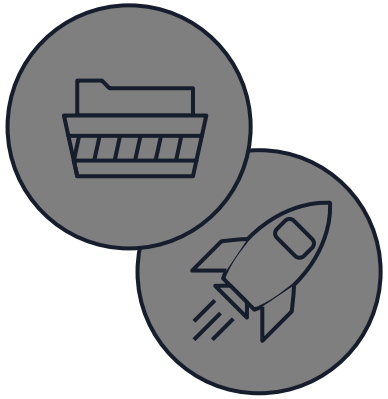
Protection
Snapshots are stored in Amazon S3

Agility
Quickly restore volumes across Availability Zones within a region

Elastic File System (EFS) – Unlimited network storage

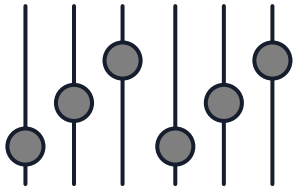


Amazon FSx for Lustre



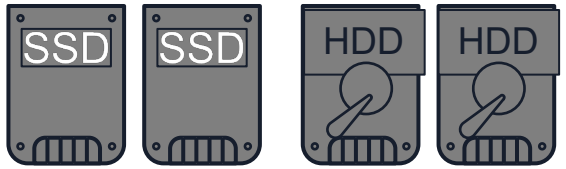
High and scalable performance

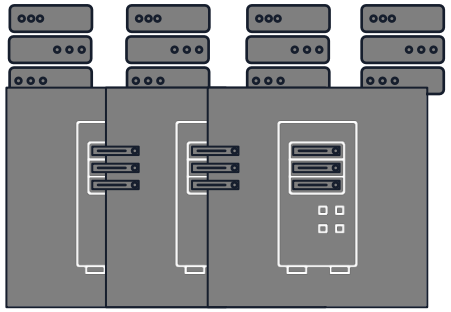
Parallel file system



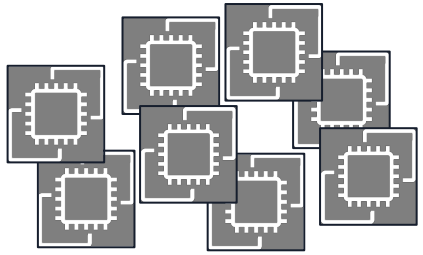
Native S3 Integration

SSD-based | HDD-based





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



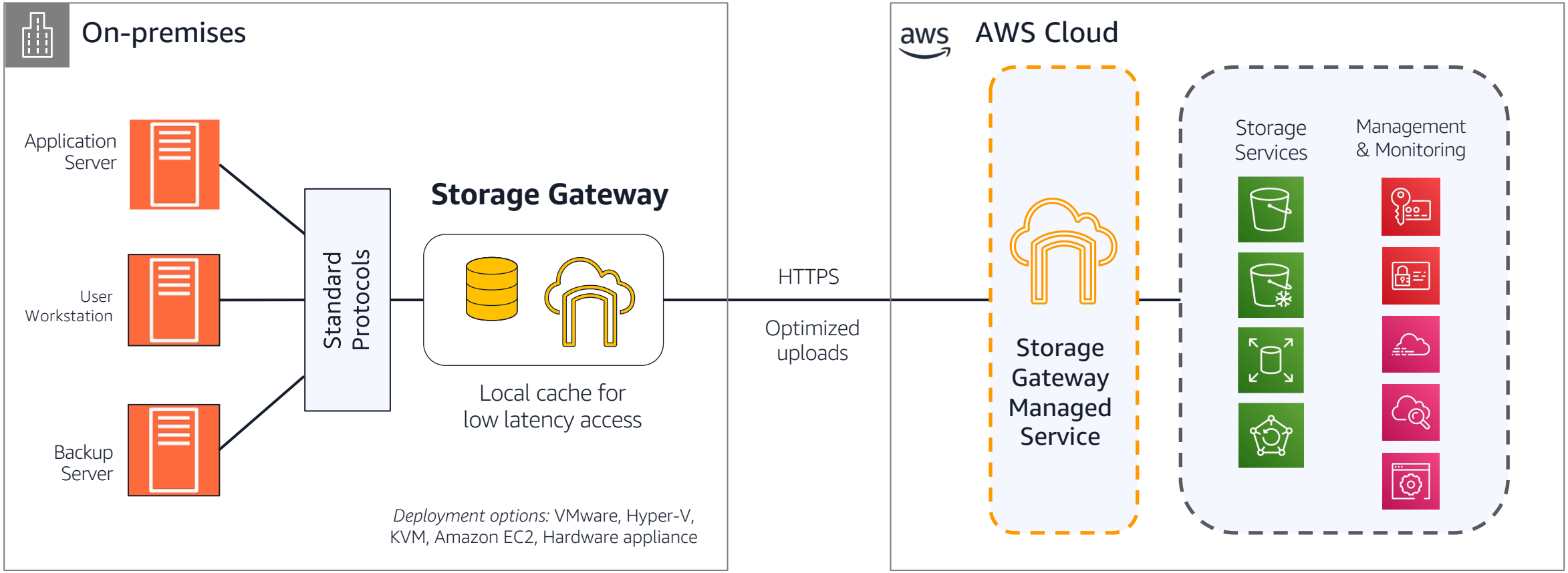
Supports concurrent access from hundreds of thousands of cores

TOYOTA
RESEARCH INSTITUTE

Toyota Research Institute uses FSx for Lustre for their large-scale machine learning workloads

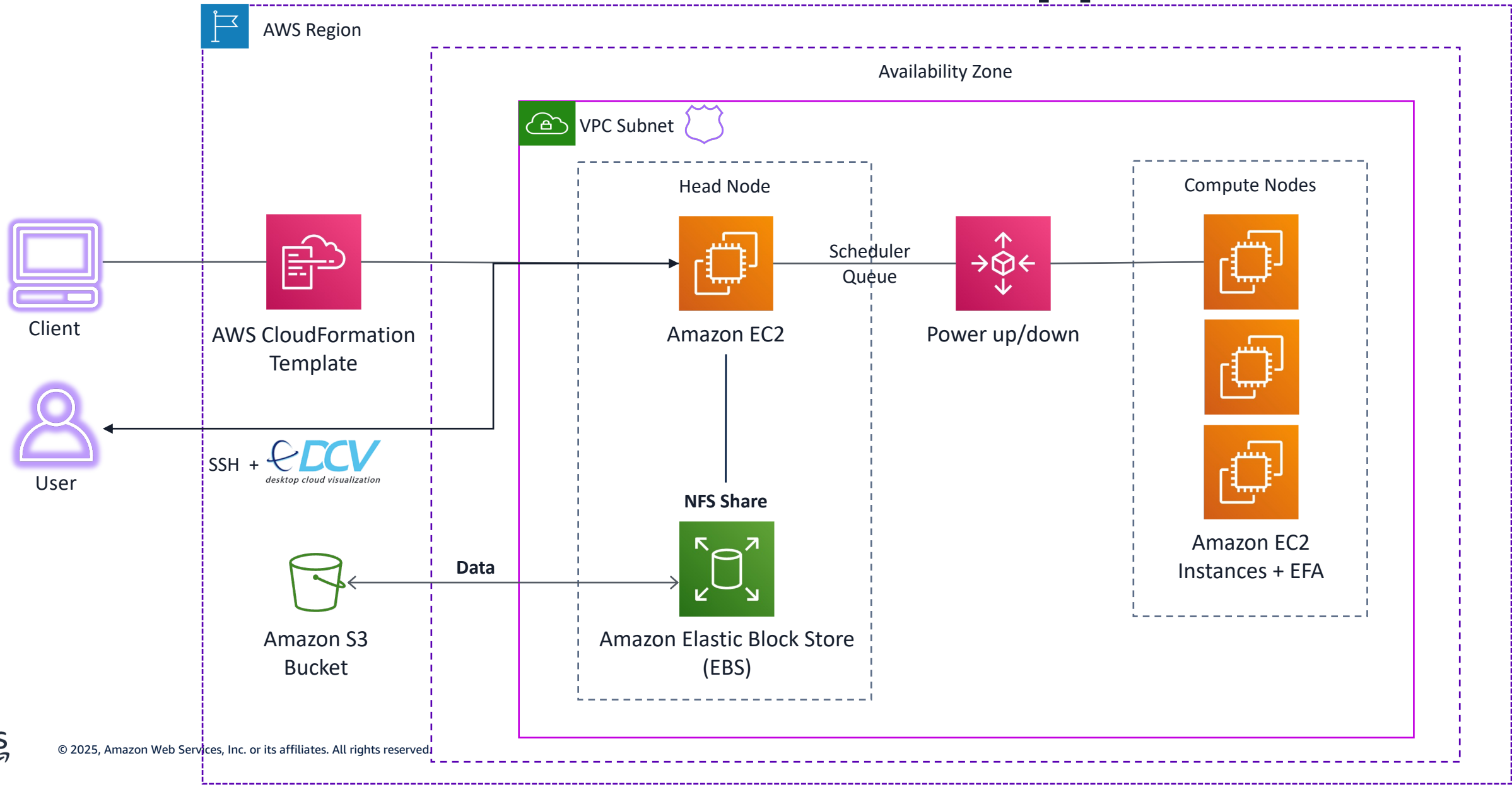
AWS Storage Gateway

On-premises access to virtually unlimited cloud storage



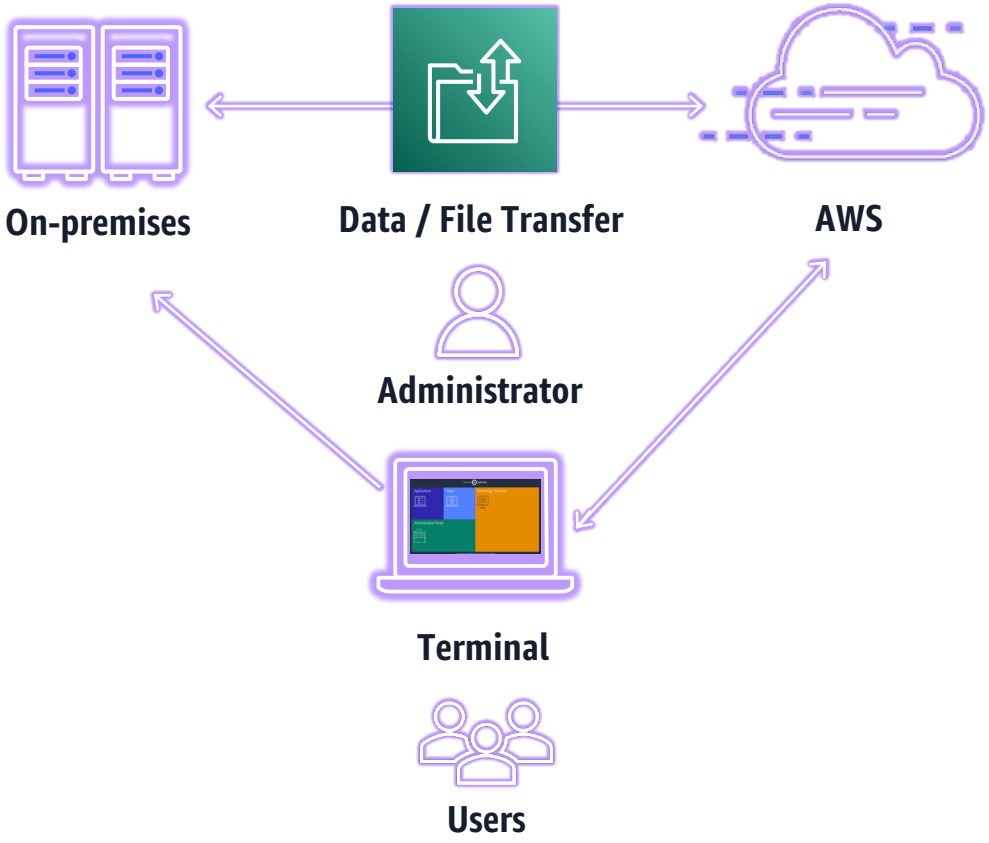
AWS ParallelCluster
AWS Batch

AWS ParallelCluster – run native applications

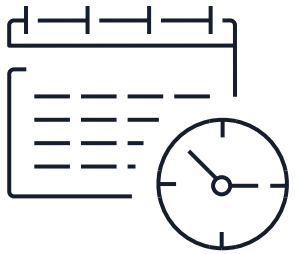


Hybrid HPC: Bursting

- Bursting with Slurm Federation
- Bursting with Slurm native plugin
- Bursting workload (Cloud native)

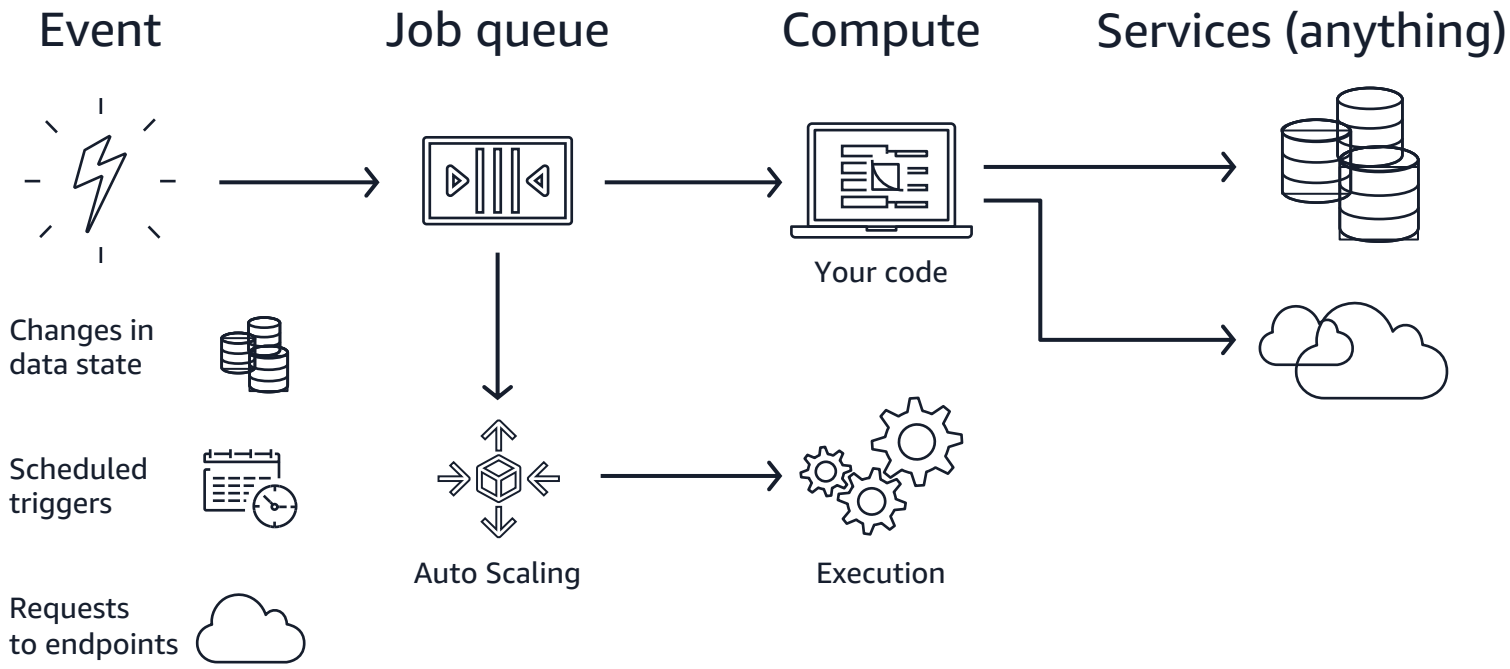


AWS Batch – run containerized applications



NEW!

AWS Fargate



Accelerated Computing

Accelerated Computing Workloads

Machine Learning

Computer vision, natural language understanding, recommendation engines, anomaly detection

Commonly used accelerator instances **include P5/P5e, P4d/P4de, G6/G6e, G5g/G5, G4dn, Tr1, DL1 for ML training and G4dn, G5, G5g, and Inf1/Inf2** for ML inference

High Performance Computing (HPC)

Seismic processing, reservoir simulation, cryogenic electron microscopy (cryo-EM), molecular dynamics (MD), computational fluid dynamics (CFD), database analytics

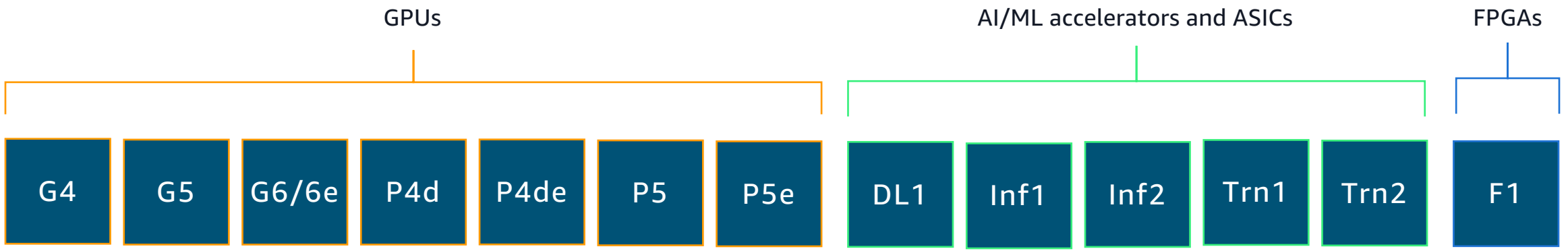
Commonly used accelerator instances include **G4dn, G5, G5g, P5/P5e, P4d/P4de,** and **F1**

Graphics

Rendering, transcoding, content streaming, product design, graphics workstations, game streaming

Commonly used accelerator instances include **G4dn, G4ad, G5, G5g,** and **VT1**

Broad and deep accelerated computing portfolio



Trainium accelerator
Inferentia accelerator
Graviton CPU

H100/200, A100, L40S
L4, T4G, A10G, T4

Gaudi accelerator

Radeon GPU
Xilinx accelerator
Xilinx FPGA

Artificial Intelligence/Machine Learning for research

The AWS AI/ML stack

Broadest and most complete set of machine learning capabilities

Gen AI SERVICES

- **Amazon Q** — Gen-AI Powered Assistant
- **Amazon Q for Business** — Gen-AI Powered Assistant
- **Amazon Q Developer** — Gen-AI Coding Assistant
- **Amazon Q Quicksight** — Gen-AI Powered Business Intelligence
- **Amazon Q Connect** — Gen-AI Powered Contact Center Assistant
- **Bedrock** — Gateway to foundation models
- **Foundation Models** — Amazon, Anthropic, etc.

AI SERVICES

SPECIALIZED

- **BUSINESS PROCESSES**: Amazon Personalize, Amazon Forecast, Amazon Fraud Detector, Amazon Lookout for Metrics
- **SEARCH**: Amazon Kendra
- **CONVERSATION**: Amazon Lex, Amazon Transcribe Call Analytics, Contact Lens, Voice ID
- **CODE + DEVOPS**: Amazon CodeGuru, Amazon CodeWhisperer, Amazon DevOps Guru
- **INDUSTRIAL**: Amazon Monitron, Amazon Lookout for Equipment, Amazon Lookout for Vision
- **HEALTH**: Amazon HealthLake, Amazon Comprehend Medical, Amazon Transcribe Medical, Amazon Omics

CORE

- **TEXT**: Amazon Translate, Amazon Comprehend
- **SPEECH**: Amazon Polly, Amazon Transcribe
- **VISION**: Amazon Textract, Amazon Rekognition, AWS Panorama

AMAZON SAGEMAKER

- **CANVAS** — No-code ML for business analysts
- **STUDIO LAB** — Learn ML
- **GROUND TRUTH** — Label data

STUDIO IDE

Prepare data · Geospatial ML · Build with notebooks · Train models · Deploy in production · Store features · Tune parameters · Manage and monitor

CI/CD | GOVERNANCE | RESPONSIBLE ML

- **EDGE MANAGER** — Manage edge devices

ML FRAMEWORKS & INFRASTRUCTURE

PyTorch, Apache MXNet, TensorFlow | Amazon EC2 | CPUs | GPUs | AWS Inferentia | AWS Trainium | Habana Gaudi | FPGA

The Infrastructure layer

AWS infrastructure for AI/ML

ML FRAMEWORKS AND INFRASTRUCTURE

ML Frameworks and open source	PyTorch	TensorFlow	Hugging Face	OpenXLA
Orchestration	HyperPod	Amazon EKS, ECS	AWS Batch	AWS ParallelCluster
Storage/Networking	Amazon EFS	Amazon S3	Amazon FSx for Lustre	EFA
EC2 Instances	Trn Inf	P5* P4*	G6 G5 G4	DL2q DL1

Trainium & Inferentia accelerators.
AWS Graviton CPU

H200, H100, A100, L40S, L4, A10G, T4 GPU

Gaudi accelerator

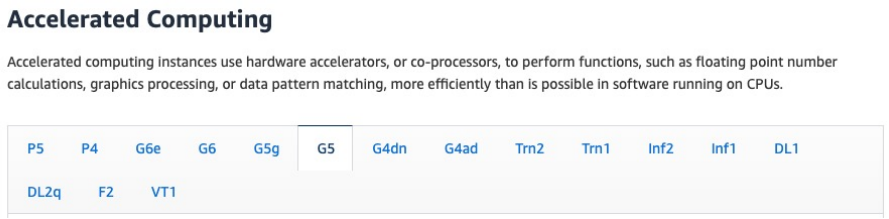
Radeon Pro V520 GPU
Xilinx accelerator
Xilinx FPGA

AI 100 accelerator

Amazon EC2 instances for AI/ML

https://aws.amazon.com/ec2/instance-types/

P5	P4	G6	G5	Trn2	Inf2
8 x H200, 8 x H100	8 x A100	Up to 8 L40S, L4 Tensor Core GPUs	T4G, A10G Tensor Core GPUs	16 AWS Trainium2 chips (20.8 FP8 petaflops) (30-40% better price performance then the current P5s)	Up to 12 AWS Inferentia2 chips (up to 2.3petaflops)
640 - 1128 GB HBM3 GPU memory	320 - 640 GB HBM2	Up to 384 GDDR6	Up to 768 GDDR6	1.5 TB	Up to 768 GB
3,200 Gbps EFA networking	400 Gbps EFA networking	Up to 400 Gbps EFA networking	Up to 100 Gbps of EFA networking	Up to 3.2 Tbps	Up to 100 Gbps



AWS Deep Learning AMIs (DLAMI) provides customized machine images that you can use for deep learning in the cloud

AWS Deep Learning Containers are pre-built Docker images that make it easier to run popular deep learning frameworks and tools on AWS.

Amazon Sagemaker AI

The next generation of Amazon SageMaker

THE CENTER FOR ALL YOUR DATA, ANALYTICS, AND AI

SageMaker

AI

Develop and scale AI use cases with the broadest set of tools

Unified Studio

Collaborate and build faster with a single data and AI dev. environment

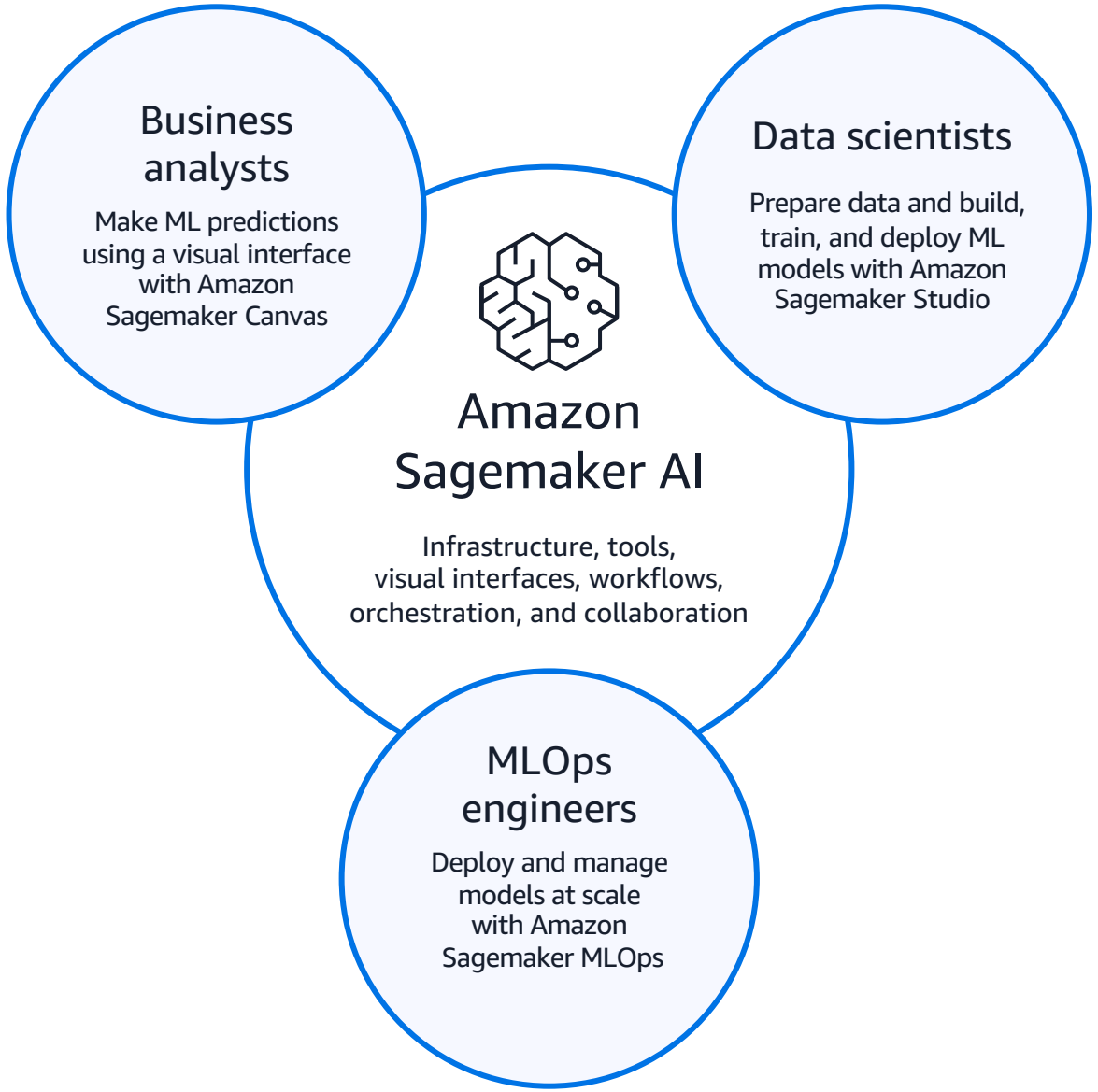
Lakehouse

Reduce data silos with an open lake house to unify all your data

Catalog

Meet your enterprise security needs with built-in data and AI governance

Amazon Sagemaker AI helps organizations harness ML



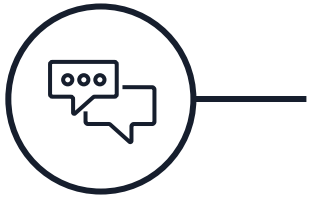
No-code ML tools

Amazon Sagemaker Canvas

Generate ML predictions – no code required

Quickly access and prepare data sources and data for ML with **SageMaker Data Wrangler**

AutoML built in to generate accurate predictions



Share ML models with data science teams

Amazon Sagemaker AI brings tools for every step of the ML lifecycle under one unified visual user interface

Amazon Sagemaker AI Notebooks

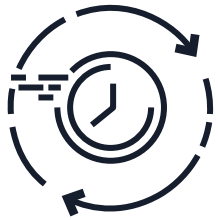
Fast-start sharable notebooks

Easy access with Single Sign-On (SSO)

Access your notebooks in seconds

Fully managed and secure

Administrators manage access and permissions



Fast setup

Start your notebooks without spinning up compute resources

Easy collaboration

Share notebooks with a single click

Flexible

Dial up or down compute resources

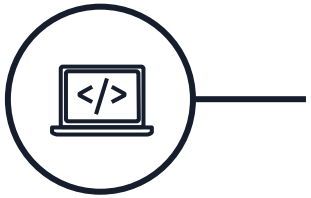
+ Integrated with Amazon Q Developer – your AI assistant

Sagemaker AI
Building ML models

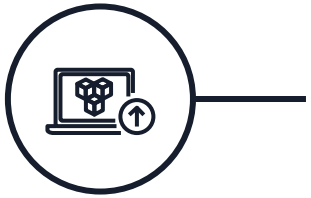
Build ML models

Fully managed shareable notebooks on Amazon EC2

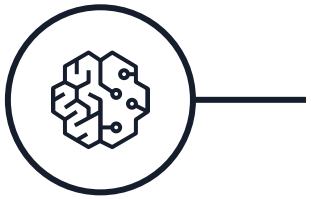
Fully managed, sharable Jupyter notebooks
Run notebooks on elastic compute resources



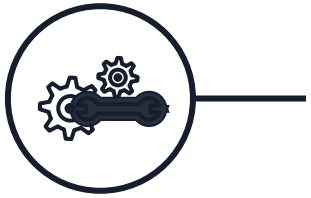
Built-in algorithms
15 built-in algorithms available in prebuilt container images



Prebuilt solutions and open-source models
Over 150 popular open-source models



AutoML
Automatically create ML models with full visibility



Support for major frameworks and toolkits
Optimized for popular deep learning (DL) frameworks such as TensorFlow, PyTorch, Apache MXNet, and Hugging Face

Amazon Sagemaker AI has built-in algorithms or bring your own

Classification
Linear Learner | XGBoost | KNN

Computer vision
Image classification | Object detection | Semantic segmentation

Topic modeling
LDA | NTM

Working with text
BlazingText | Supervised | Unsupervised

Recommendation
Factorization machines

Forecasting
DeepAR

Sequence translation
Seq2Seq

Regression
Linear Learner | XGBoost | KNN

Clustering
KMeans

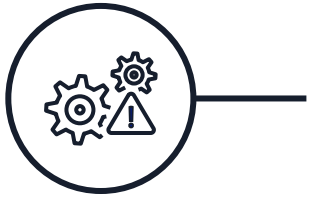
Anomaly detection
Random cut forests | IP Insights

Feature reduction
PCA

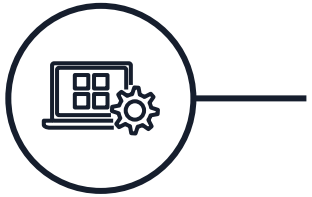
Sagemaker AI
Training ML models

Train ML models

Fast and cost-effective ML model training

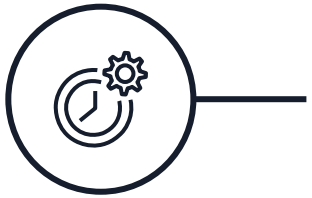


Experiment management and model tuning
Save weeks of effort by automatically tracking training runs and tuning hyperparameters



Debug and profile training runs
Use real-time metrics to correct performance problems

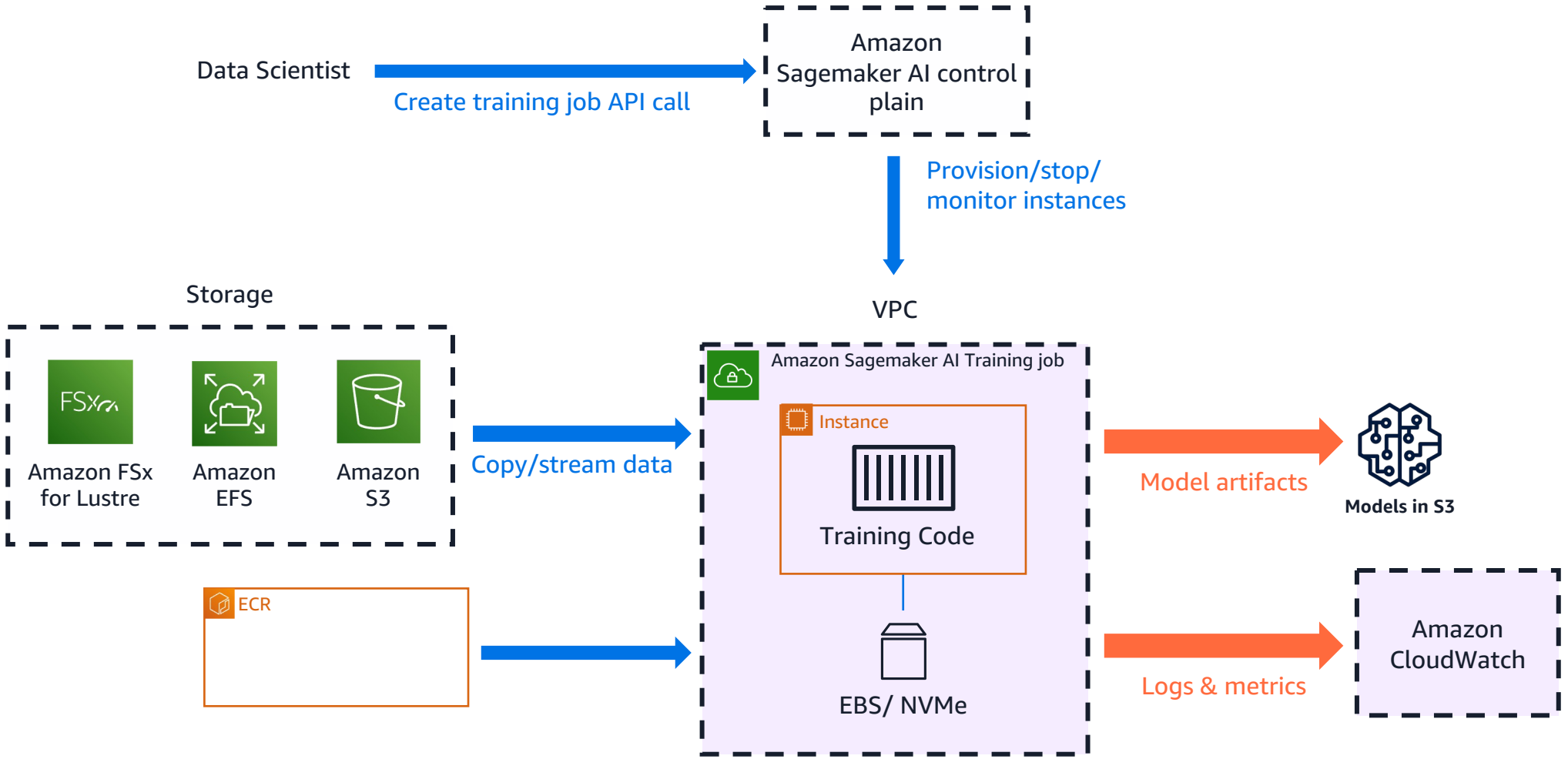
Distributed training
Complete distributed training up to 40% faster

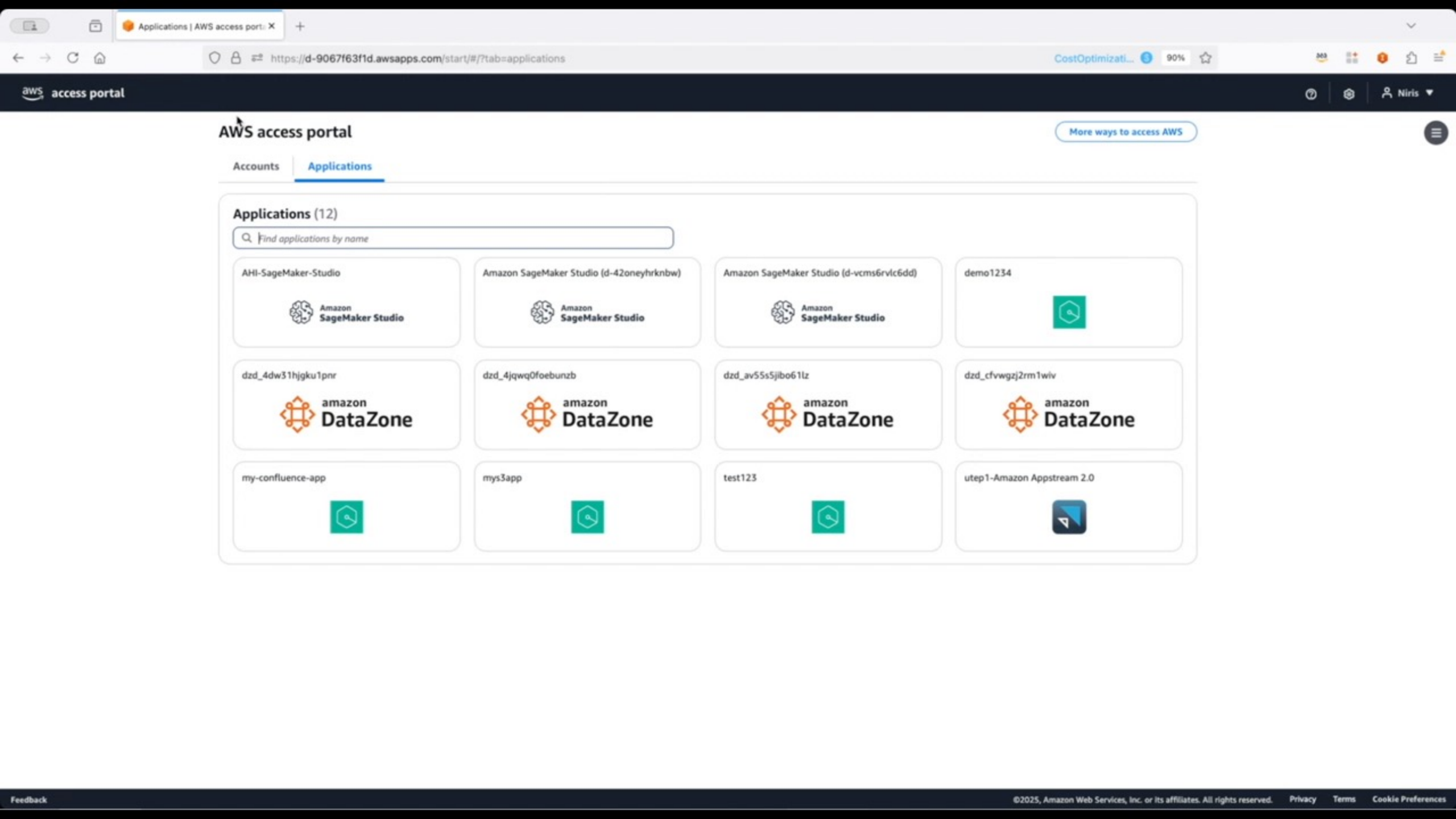


Training compiler
Accelerate training times by up to 50% through more efficient use of GPUs

Managed spot training
Reduce the costs of training by up to 90%

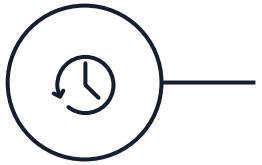
Training on Amazon Sagemaker AI



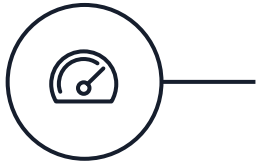

Applications | AWS access portal
https://d-9067f63f1d.awsapps.com/start/#/?tab=applications
access portal
Niris
AWS access portal
More ways to access AWS
Accounts
Applications
Applications (12)
Find applications by name
AHI-SageMaker-Studio
Amazon SageMaker Studio
Amazon SageMaker Studio (d-42oneyhrknbw)
Amazon SageMaker Studio
Amazon SageMaker Studio (d-vcms6rvlc6dd)
Amazon SageMaker Studio
demo1234
dzd_4dw31hjgku1pnr
amazon DataZone
dzd_4jqwq0foebunzb
amazon DataZone
dzd_av55s5jlbo61lz
amazon DataZone
dzd_cfvwgzj2rm1wiv
amazon DataZone
my-confluence-app
mys3app
test123
utep1-Amazon Appstream 2.0
Feedback
©2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy
Terms
Cookie Preferences

Distributed training

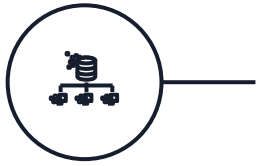
The fastest and easiest way to train large deep learning models



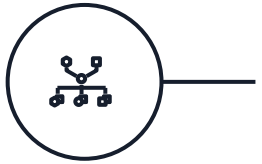
Reduced training time
Reduce training time by 25% with synchronization across GPUs



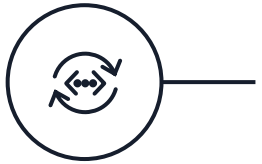
Optimized for AWS
Achieve near-linear scaling efficiency with data parallelism designed for AWS



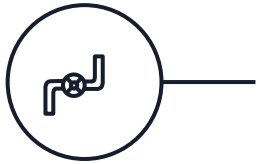
Support for popular ML framework APIs
Re-use existing APIs such as Horovod without custom training code



Automatic and efficient model partitioning
Avoid experimentation with automated model profiling and partitioning



Minimal code change
Implement model parallelism with fewer than 10 lines of code change



Efficient pipelining
Maximize resource usage with pipelining of micro-batches that keeps all GPUs active

Distributed training solutions on SageMaker AI

- SageMaker AI Distributed Data Parallelism library (SMDDP)
- SageMaker model parallel library (SMP)
- Open-Source distributed training frameworks

SageMaker AI Distributed Data Parallelism library (SMDDP)

For PyTorch DDP or FSDP

Initialize the process group as follows.

```
import torch.distributed as dist
import smdistributed.dataparallel.torch.torch_smddp

dist.init_process_group(backend="smddp")
```

For DeepSpeed or Megatron-DeepSpeed

Initialize the process group as follows.

```
import deepspeed
import smdistributed.dataparallel.torch.torch_smddp

deepspeed.init_distributed(dist_backend="smddp")
```

With the single line of backend specification, you can keep all the native PyTorch distributed modules and the entire training script unchanged

SageMaker model parallel library (SMP)

```
from sagemaker.framework import Framework

distribution={
    "smdistributed": {
        "modelparallel": {
            "enabled":True,
            "parameters": {
                ...   # enter parameter key-value pairs here
            }
        },
    },
    "mpi": {
        "enabled" : True,
        ...           # enter parameter key-value pairs here
    }
}

estimator = Framework(
    ...,
    instance_count=2,
    instance_type="ml.p4d.24xlarge",
    distribution=distribution
)
```

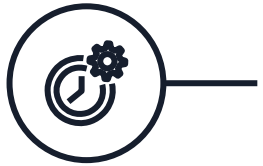
Example notebooks on GitHub repo

Sagemaker AI Hosting Deep Dive

Deploy ML models

Fully managed deployment for inference at scale

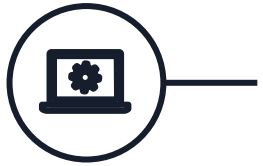
Wide selection of infrastructure
70+ instance types with varying levels of compute and memory to meet the needs of every use case



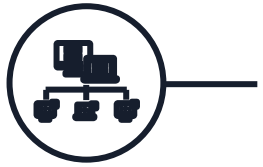
Single-digit millisecond overhead latency
For use cases requiring real-time responses

Asynchronous inference
Supports large models with long-running processing times

Cost-effective deployment
Multi-model/multi-container endpoints, serverless inference, and elastic scaling



Built-in integration for MLOps
ML workflows, CI/CD, lineage tracking, and catalog



Automatic deployment recommendations
Optimal instance type/count and container parameters, and fully managed load testing

Sagemaker AI inference options

Real-time inference

Low latency

Ultra high throughput

Multi-model endpoints

A/B testing

Batch transform

Process large datasets

Job-based system

Asynchronous inference

Near real-time

Large payloads (1 GB)

Long timeouts (15 mins)

Serverless inference

Serverless

Auto Scaling, no scaling policy

CPU based models

Amazon SageMaker JumpStart

ML Hub with foundation models

Access hundreds of foundation models including top open-weight and proprietary models that can be fine tuned easily for your use case

Ease of use

Easily use pre-trained models on SageMaker instances like Inf2 and optimized hosting configurations through presets

Evaluate and Customize

Evaluate, fine-tune, and optimize deployment using few clicks

Data security and access control

Keep inference and training data private and curate who can access and use models within your organization

Amazon Bedrock – Generative AI

Amazon Bedrock

The easiest way to build and scale generative AI applications with powerful tools and foundation models

Choice of leading FMs through a **single API**

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and data governance

Amazon Bedrock Security

Helps keep your data secure and private

None of the customer's data is used to train the underlying model

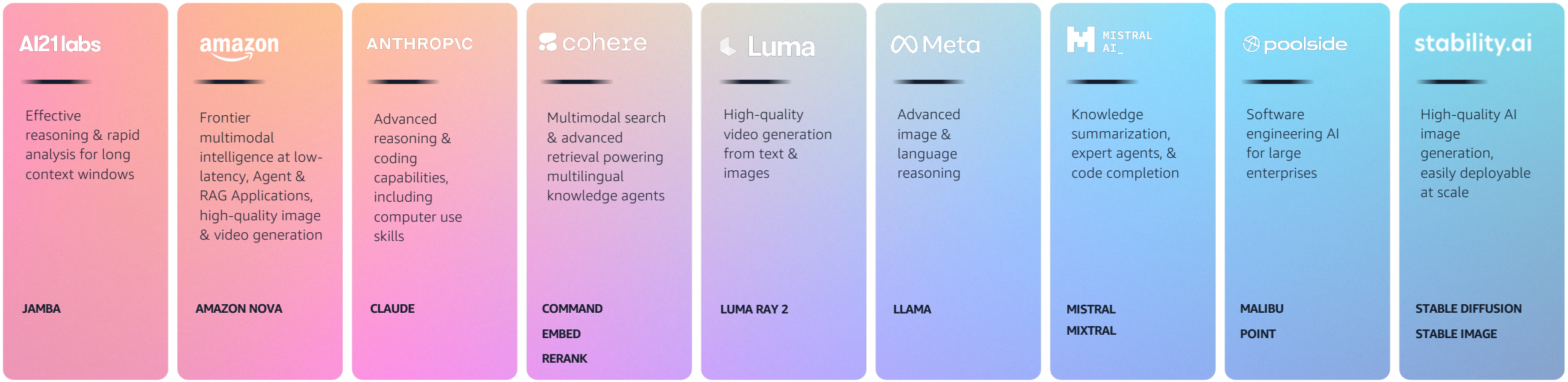
All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC

Data remains in the Region where the API is processed

Support for GDPR, SOC, ISO, CSA compliance, and HIPAA eligibility

Amazon Bedrock

BROAD CHOICE OF MODELS



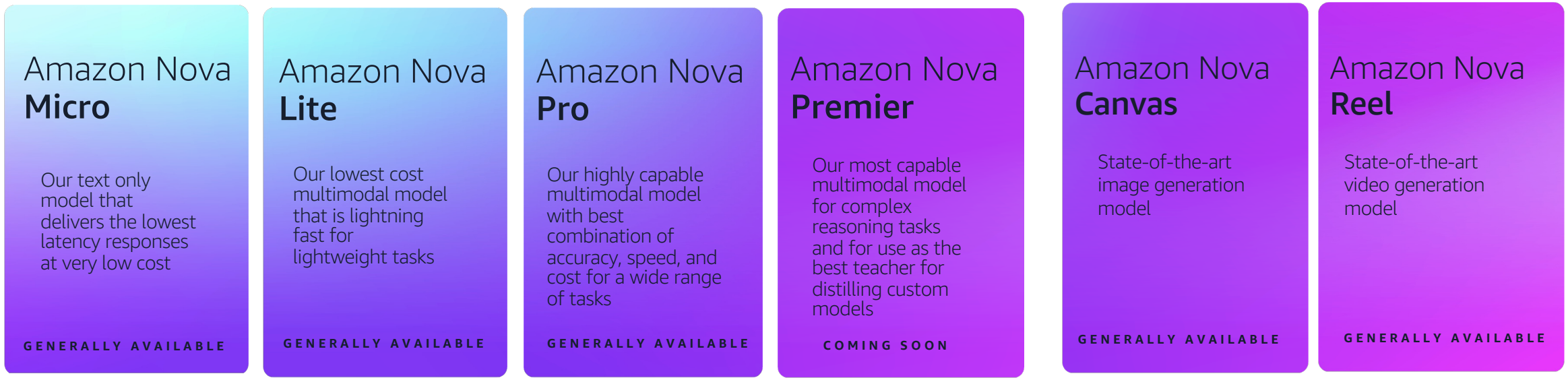
Coming soon

Amazon Nova Foundation Models

State-of-the-art foundation models that deliver frontier intelligence and industry leading price performance.

Understanding models

Creative content generation models



Lower Cost & Latency

Increasing Intelligence

Some features of Amazon Bedrock

Customization

Privately fine-tune FMs using your own labeled datasets in just a few quick steps

RAG

Knowledge Bases for Amazon Bedrock is a fully managed RAG capability that allows you to customize FM responses with contextual and relevant data

Agents

Enable generative AI applications to automate multistep tasks by seamlessly connecting with institutional systems, APIs, and data sources

Guardrails

- Filter harmful multimodal contents
- Redact sensitive data such as PII
- Guardrails supports contextual grounding checks to help detect and filter hallucinations if the responses are not factually accurate

Additional resources

aws-parallelcluster repository

Open Source cluster management tool to deploy and manage HPC clusters in the AWS cloud.

Sagemaker examples repo

Jupyter notebooks that demonstrate how to build, train, and deploy machine learning

Thank you!

Jianjun Xu

Principal Solutions Architect
Amazon Web Services
jianjx@amazon.com

Niris Okram

Sr. Solutions Architect
Amazon Web Services
niris@amazon.com

Please complete the survey for this session

Research Track

Run High Performance Computing (HPC) and Artificial Intelligence/Machine Learning for research

aws Learning Days
State, Local, and Education