AWS re:Invent

CMP208

EC2 Foundations

Raj Pai
Director of Product Management, EC2

Amazon EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Amazon EC2 Foundations

Resources

- **Instances**
- Storage
- Networking

Availability

- Regions and AZs
- Placement Groups
- Load Balancing
- Auto Scaling

Management

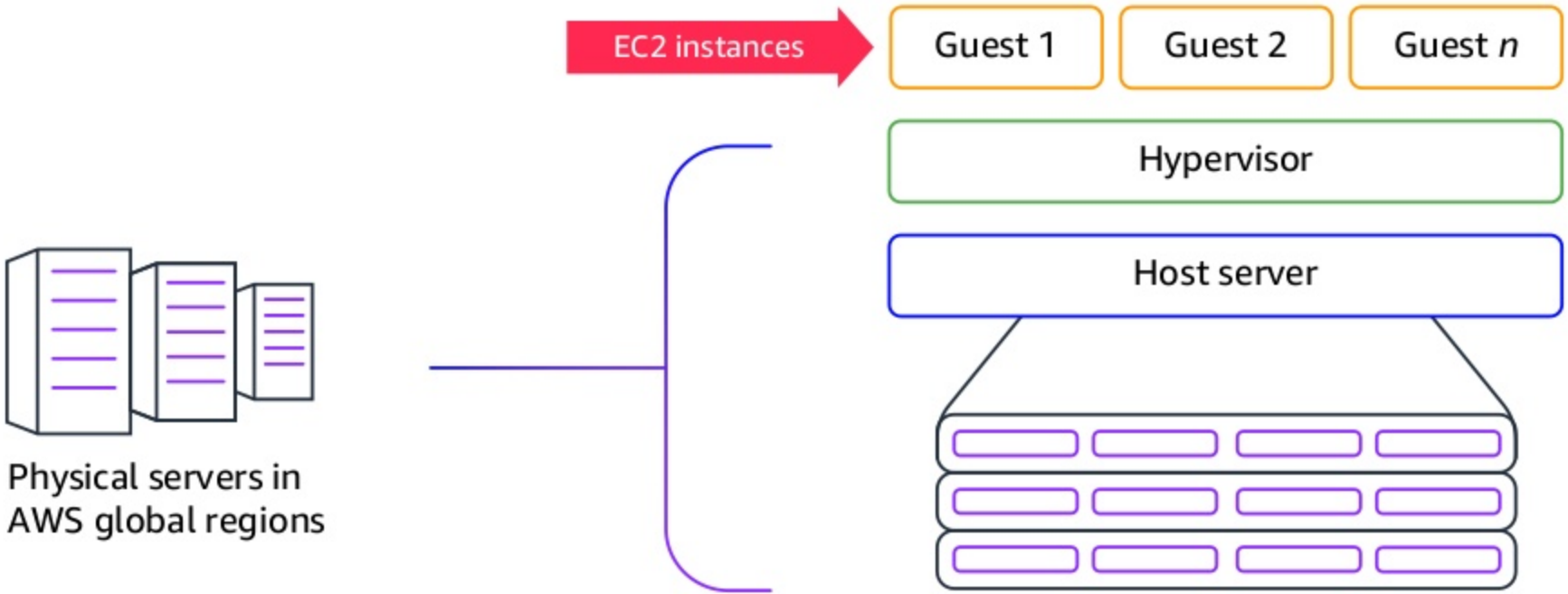
- Deployment
- Monitoring
- Administration

Purchase Options

- On Demand
- Reserved
- Spot

Amazon Elastic Compute Cloud (EC2)

Virtual servers in the cloud



Amazon EC2 12+ years ago…

Scale up or down quickly, as needed

Pay for what you use

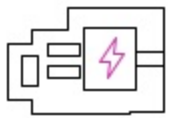
"One size fits all"

Continued rapid pace of innovation

Instance growth

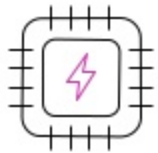
Innovation enabled by AWS Nitro System

Nitro Card



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

Nitro Security Chip



Integrated into motherboard
Protects hardware resources

Nitro Hypervisor

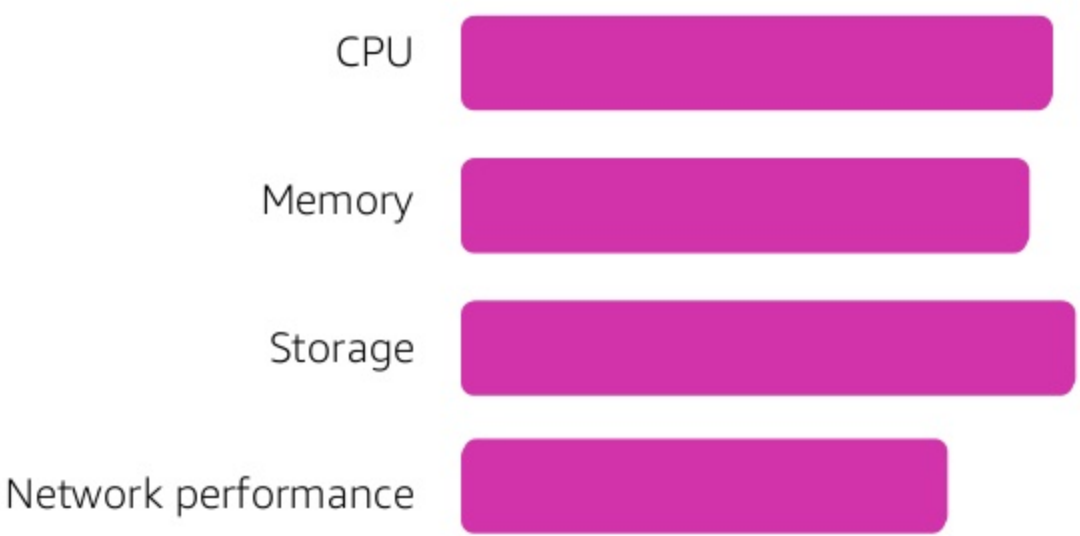
Lightweight hypervisor
Memory and CPU allocation
Bare Metal-like performance

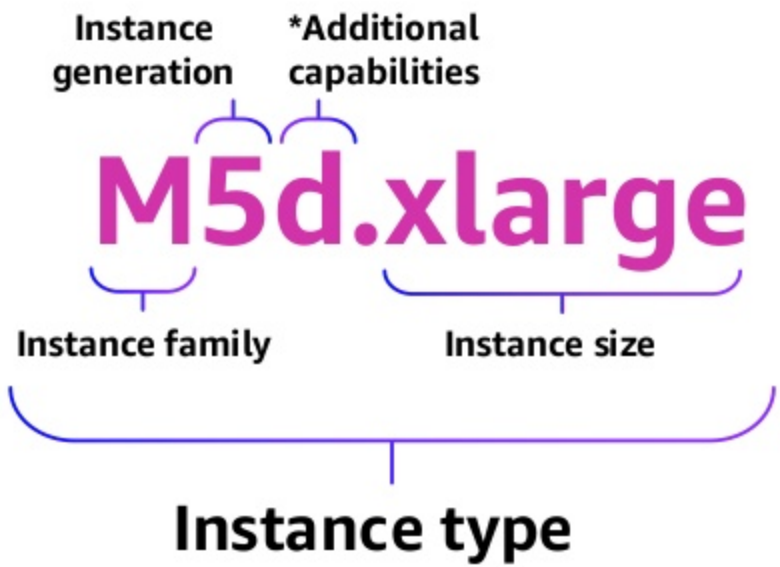
Modular building blocks for rapid design and delivery of Amazon EC2 instances

Run virtualized instances with perf indistinguishable from Bare Metal

Run bare metal workloads on Amazon EC2 with all the elasticity, security, scale, and services of AWS

Amazon EC2 instance characteristics





Amazon Machine Images (AMIs)

Amazon maintained

Broad set of Linux and Windows images

Kept up-to-date by Amazon in each region

Amazon Linux 2 with five years of long term support

Marketplace maintained

Managed and maintained by AWS Marketplace partners

Your machine images

AMIs you have created from Amazon EC2 instances

Can keep private, share with other accounts, or publish to the community

Choice of processors and architectures

Intel® Xeon® Scalable (Skylake) processor

AMD EPYC processor

AWS Graviton Processor based on 64-bit Arm arch

\+ Choice of GPUs and FPGAs for compute acceleration

Right compute for each application and workload

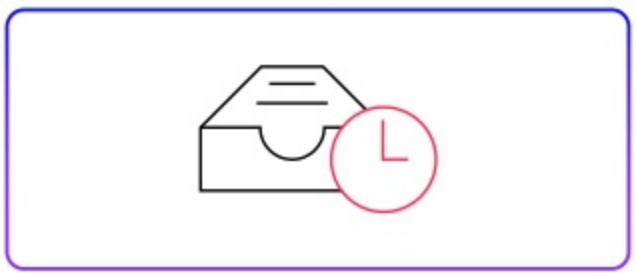
General purpose instance workloads

Web/app servers

Enterprise apps

Gaming servers

Caching fleets

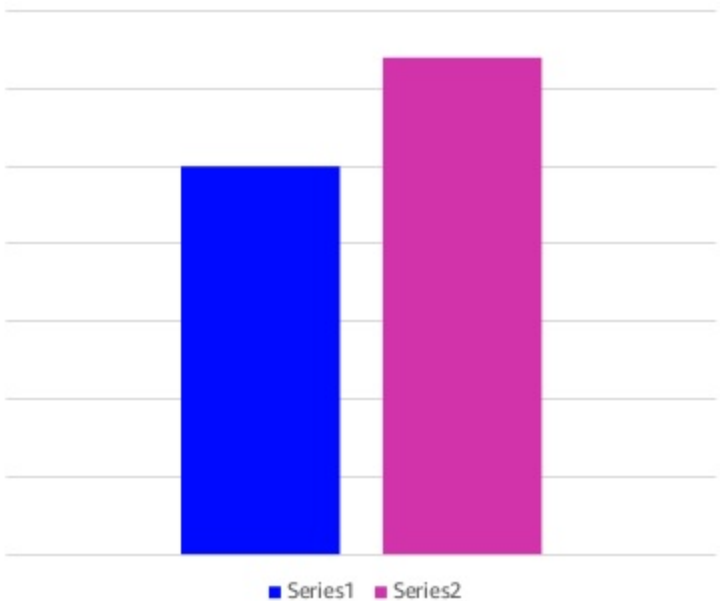


Analytics applications

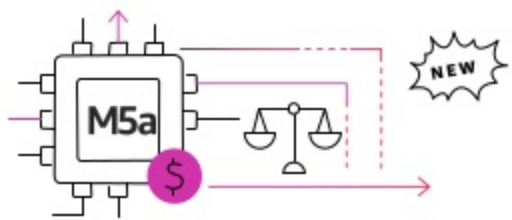
Dev/test environments

M5: General purpose instances

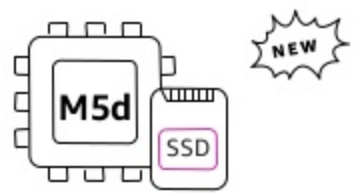
14% price/performance improvement with M5



- Balance of compute, memory, and networking resources
- Powered by 2.5 GHz Intel Xeon Scalable Processors **(Skylake)**
- Largest instance size, m5.24xlarge has **96 vCPUs** and **384 GiB of memory**
- Improved network and EBS performance on smaller sizes
- Support for Intel **AVX-512** offering up to twice the performance for vector and floating point workloads



M5a: Now available with AMD EPYC 7000 processor for 10% lower cost



M5d: Now available with high performance local NVMe SSD storage

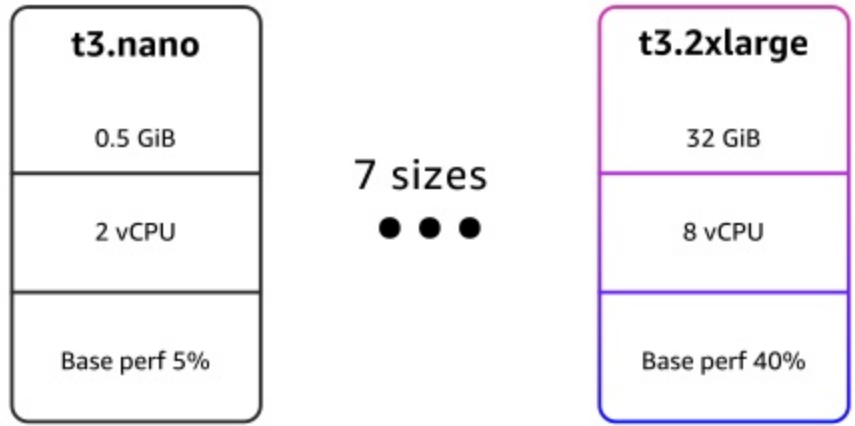
Opportunity: Most instances aren't very busy

T3: Burstable general-purpose instances

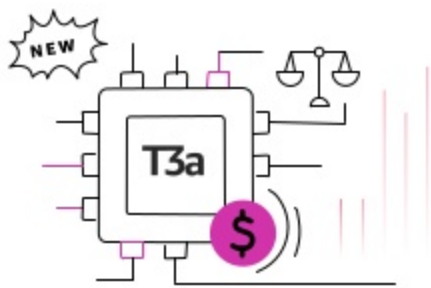
Balance of compute, memory, and network

Baseline level of CPU performance with the ability to burst CPU usage when needed at any time for as long as required

Lowest cost instance at $0.0052 per hour and up to 30% better price performance over T2



With T3 Unlimited bursting over baseline is only $0.05 per vCPU-hour, averaged over 24 hours



T3a: Coming soon with AMD EPYC 7000 processor for 10% lower cost

A1: First Arm instance in Amazon EC2

NEW

Optimized cost and performance for scale-out applications

a1.medium	•••	a1.4xlarge
2 GiB	5 sizes	32 GiB
1 vCPU		16 vCPU

Up to 45% cost savings

AWS Graviton Processor with Arm-based cores and custom silicon

Broad software and tooling support

Lower cost for scale-out workloads

Arm-based development platform

Choosing between Amazon EC2 General Purpose Instances

T3 Instances

Optimized for most workloads with occasional high CPU use

M5/M5a Instances

Balance of compute, memory, and network resources

A1 Instances

Workloads that can scale-out across multiple cores, fit within memory, run on Arm

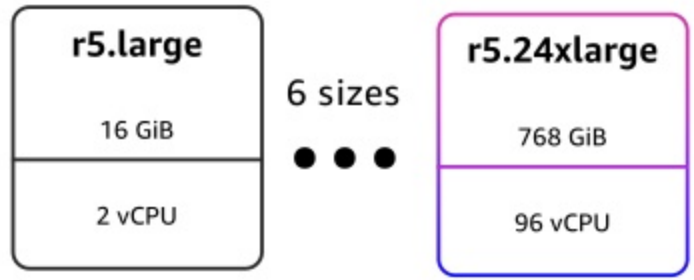
R5: Memory optimized instances

Memory-optimized instances with 8:1 GiB to vCPU

2.5 GHz Intel Xeon Scalable Processors (Skylake)

Up to 25 Gbps NW bandwidth

R5d instances include up to 3.6 TB of local NVMe SSD



In-memory caches

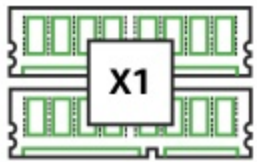
High performance databases

Big data analytics

R5a: Now available with AMD EPYC 7000 processor for 10% lower cost

R5.metal Bare Metal instances coming soon

X1 and X1e: Large-scale memory-optimized

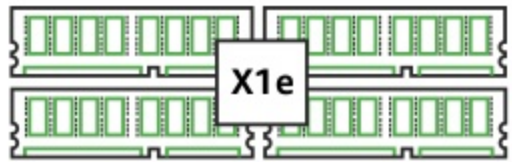


For large in-memory workloads

16:1 GiB to vCPU ratio

In-memory databases (e.g., SAP HANA), big data processing engines (Apache Spark, Presto), in-memory analytics

x1.16xlarge	x1.32xlarge
1 TB	2 TB
64 vCPU	128 vCPU



For memory-intensive workloads and very large in-memory workloads

32:1 GiB to vCPU ratio

High-performance databases, Large in-memory databases (e.g. SAP HANA), and DB workloads with vCPU based licensing (Oracle, SAP)

x1e.xlarge	6 sizes	x1e.32xlarge
122 GiB	● ● ●	4 TB
4 vCPU		128 vCPU

High Memory instances: Certified for SAP HANA

NEW

Up to 12TB Memory; SAP-Certified

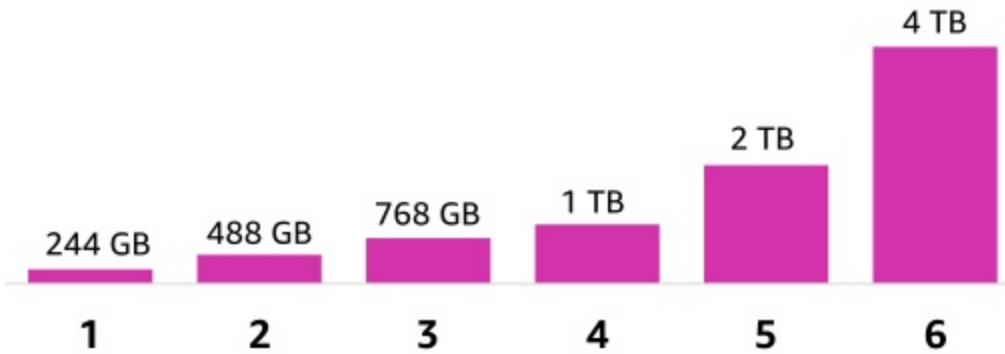
Custom Intel® Xeon® Scalable processor

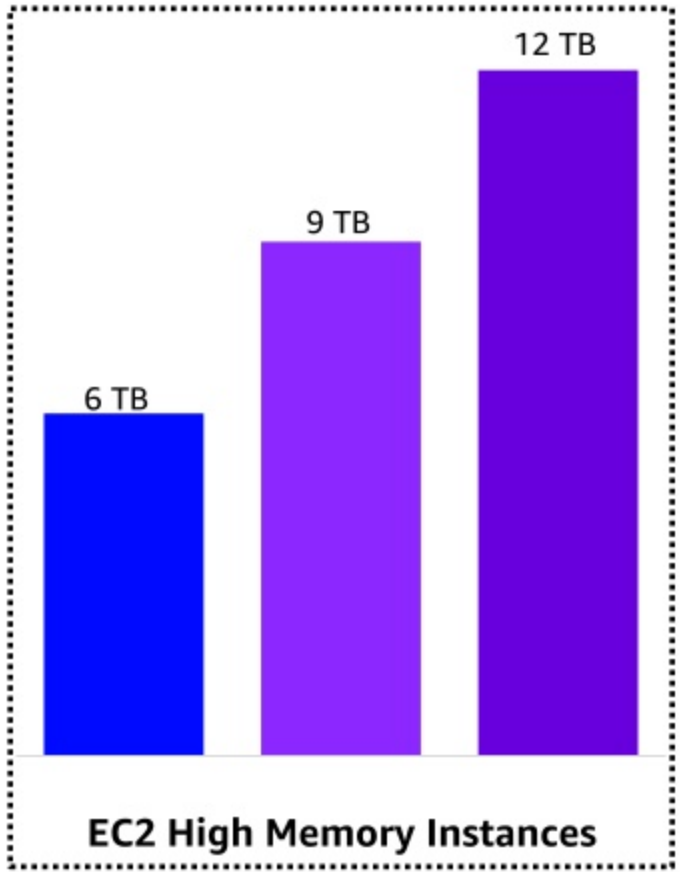
Native to AWS; Out-of-Box Integration

Simple Management: AWS CLI, Console, IAM

Flexibility to Scale; Resize in Minutes

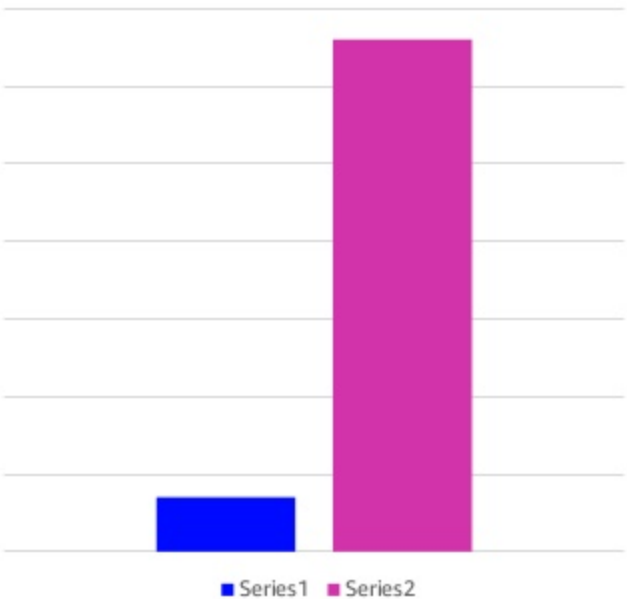
18 and 24 TB instance coming in 2019





I3: I/O optimized instances

9X as many IOPS
as I2



High-perf databases

Real-time analytics

Transactional workloads

No SQL databases

Intel Xeon E5 v4 (Broadwell) processors, with up to 15.2 TB of locally attached NVMe SSD storage, 64 vCPUs, and 488 GiB memory

Lowest cost per IOPS ($/IOPS)

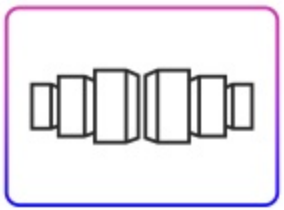
Offers very high Random I/O (up to 3.3 million IOPS) and disk throughput (up to 16 GB/s)

Up to 25 Gbps NW bandwidth

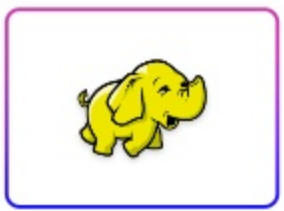
Available in Bare Metal, with i3.metal

D2 and H1: Dense storage workloads

Data warehousing



HDFS



Log processing

d2.8xlarge
244 GiB
36 vCPU
48 TB HDD

Lowest cost per storage ($/GB)

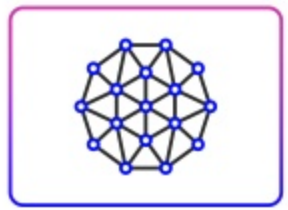
Supports high sequential disk throughput

h1.16xlarge
256 GiB
64 vCPU
16 TB HDD

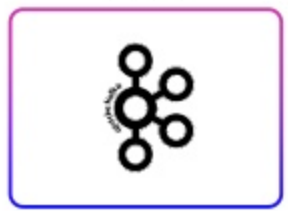
More vCPUs and memory per terabyte of disk

Lower costs for big data uses cases

Big data



Kafka

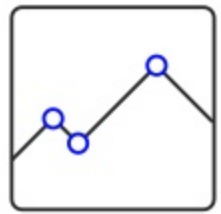


MapReduce

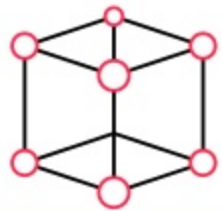
Compute-intensive workloads

Batch processing

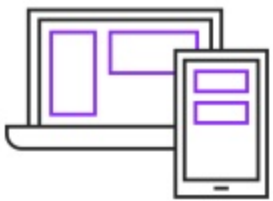
Distributed analytics



High-perf computing (HPC)

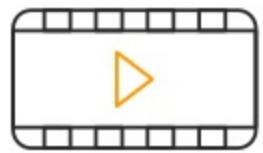


Ad serving



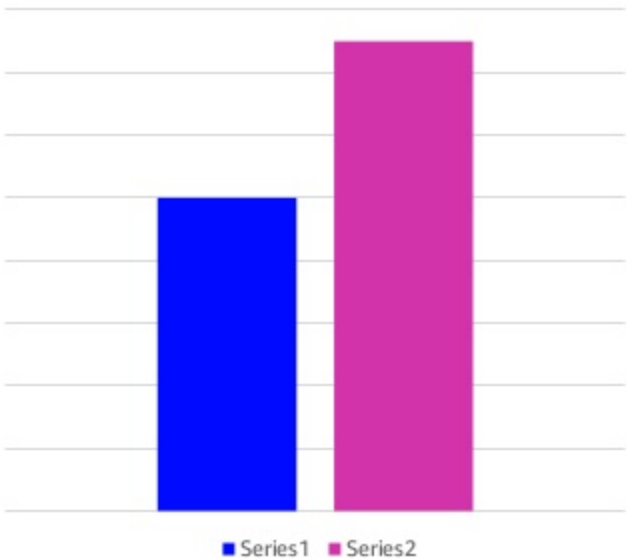
Multiplayer gaming

Video encoding



C5: Compute-optimized instances based on Intel Skylake

25% price/performance improvement over C4



Custom 3.0 GHz Intel Xeon Scalable Processors (Skylake)

Up to 72 vCPUs and 144 GiB of memory (2:1 Memory:vCPU ratio)

25 Gbps network bandwidth

Support for Intel AVX-512

C5d with local NVMe-based SSD storage

NETFLIX

"We saw significant performance improvement on Amazon EC2 C5, with up to a 140% performance improvement in industry standard CPU benchmarks over C4."

GRAIL

"We are eager to migrate onto the AVX-512 enabled c5.18xlarge instance size… We expect to decrease the processing time of some of our key workloads by more than 30%."

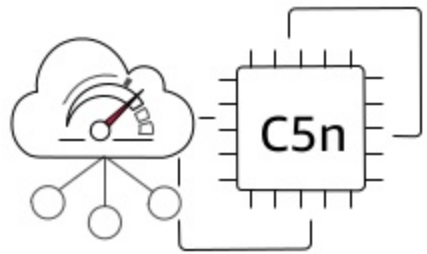
C5n: Fastest networking in the cloud

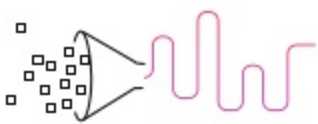
NEW

100 Gbps network bandwidth on largest instance sizes

25 Gbps peak bandwidth on smaller instance sizes

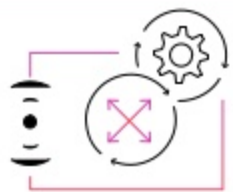
33% Increased memory footprint over C5 instances





Faster analytics and big data workloads

Lower costs for network-bound workloads



All of the elasticity, security, and scalability of AWS

z1d: High frequency for specialized workloads

NEW

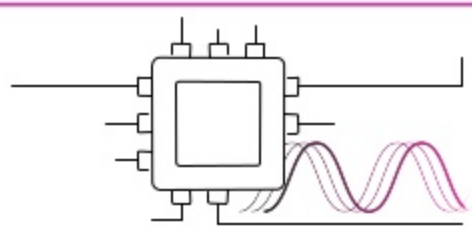
High Frequency instances with custom Intel® Xeon® Scalable Processors running at sustained 4 GHz all core turbo

8:1 GiB to vCPU ratio

Up to 25 Gbps network bandwidth and up to 1.8 TB of local NVMe storage

z1d.large	6 sizes	z1d.12xlarge
16 GiB	● ● ●	384 GiB
2 vCPU		48 vCPU

Electronic Design Automation



Relational databases

Gaming

NEW *z1d.metal Bare Metal instances coming soon*

Accelerated Computing Workloads

Applications that benefit from GPU and FPGA Acceleration

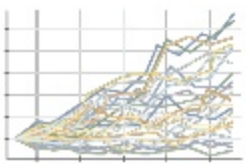
Machine Learning/AI

Natural Language Processing

Image and Video recognition

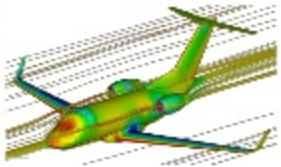
Autonomous vehicle systems

Recommendation Systems



High Performance Computing

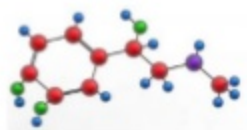
Computational Fluid Dynamics



Financial and Data Analytics

Genomics

Computational Chemistry



Graphics

Virtual Graphic Workstation

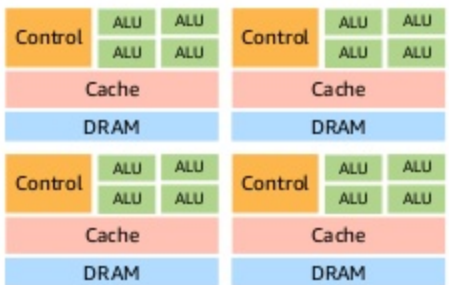
3D Modeling & Rendering

Video Encoding

AR/VR

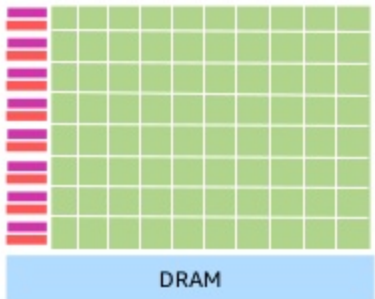
CPUs vs GPUs vs FPGAs – Architectural Comparison

CPU



- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and datapath widths
- Highly effective at parallel execution

FPGA

- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed parallel execution

Accelerated Computing Workloads such as training of machine learning models, running 3D fluid dynamics simulations, genomic sequencing and video encoding can take advantage of parallel compute architecture of GPUs and FPGAs

P3 instances: GPU Compute

Ideal for workloads needing massive parallel processing power

Training Machine Learning Model

Running HPC Simulations

Rendering 3D models

Video encoding

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*

Support all ML frameworks and model types

P3.2xlarge	P3.8xlarge	P3.16xlarge
1 V100 GPU	4 V100 GPU	8 V100 GPU
8 vCPU	32 vCPU	64 vCPU
61 GB Mem	244 GB Mem	488 GB Mem

P3dn - Most powerful GPU instance in the cloud

Efficiently scale ML model training and HPC simulations across multiple instances with 100Gbps of networking throughput

Fast access to training or simulation data via Amazon S3, network attached file systems or local instance storage

Train larger ML models or process more data via latest NVIDIA V100 GPU with 32GB of GPU memory

Optimize pre-processing of data with 96 vCPU using AWS Custom Skylake CPUs and 768GB of System Memory

NEW

p3dn.24xlarge
8 V100 GPU
96 vCPU
768 GB Mem
2 TB NVME SSD
100 Gbps Throughput

G3 instances: High Performance Graphics

Ideal for workloads needing massive parallel processing power

- Graphic Visualizations
- Cloud workstation
- Video encoding
- Virtual reality

4 GPUs, 64 vCPUs, 488 GiB of host memory, and 20 Gbps of network bandwidth

Tesla M60 GPU with included support for GRID Virtual Workstation features and licenses, and supports up to four monitors with 4096x2160 (4K) resolution

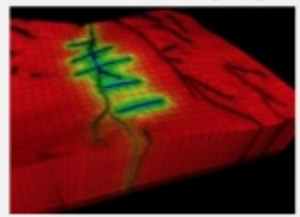
Seismic exploration and analytics for oil and gas

"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

HALLIBURTON

Seismic Imaging



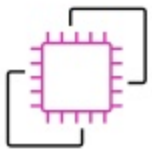
NEW

Available in 4 Sizes

G3s.xlarge	G3.4xlarge	G3.8xlarge	G3.16xlarge
1 M60 GPU	1 M60 GPU	2 M60 GPU	4 M60 GPU
4 vCPU	8 vCPU	32 vCPU	64 vCPU
30.5 GB Mem	61 GB Mem	244 GB Mem	488 GB Mem

New smaller size - Cost and performance optimized for remote workstations

Choice of accelerators for specialized workloads



Elastic Graphics

Easily add graphics acceleration to your EC2 instance

Configure right amount of graphics acceleration for your workload

Accelerate application for fraction of cost of standalone graphics instances

Elastic Inference

Reduce deep learning inference costs by up to 75%

Easily attach fractional sizes of a full GPU instance to EC2 or SageMaker instances

Scale inference acceleration up or down as needed with EC2 Auto Scaling

F1 Instances: First Cloud Instance with FPGA Accelerators

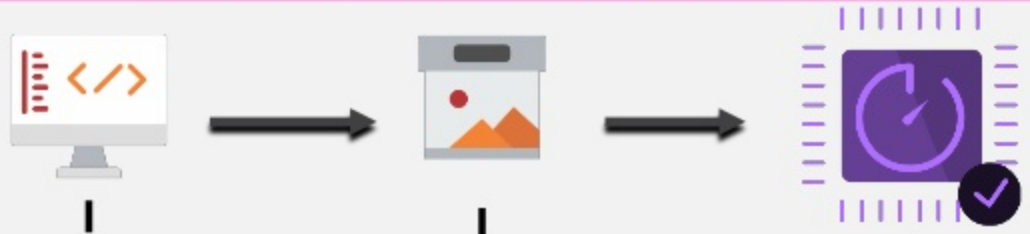
Speed up applications over 30x using hardware acceleration

- Genomics sequencing
- Financial computing
- Engineering simulations
- Image and video processing
- Big data and ML
- Security, compression

New – f1.4xlarge size to optimize price/performance

NEW

1. Develop Custom Accelerators or

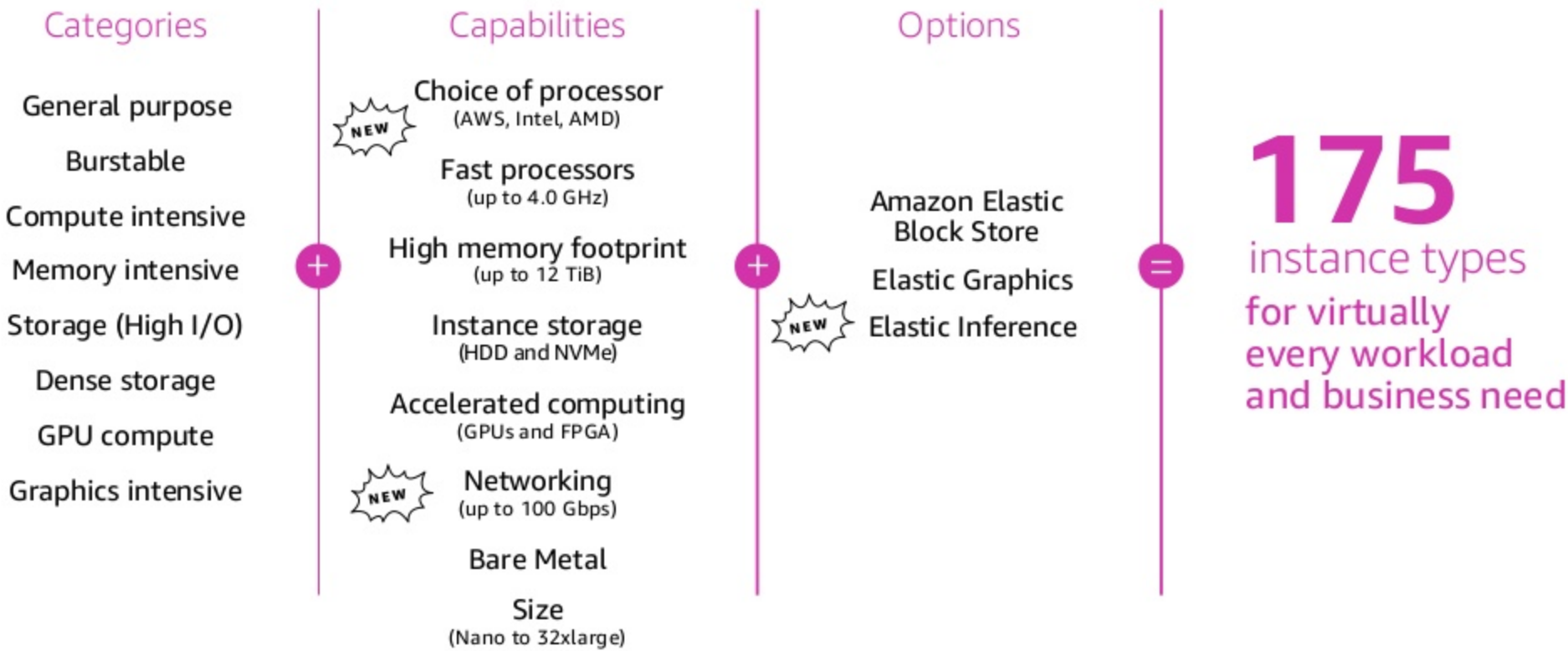


2. Use Pre-built Accelerators from F1 Partners

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

aws re:Invent

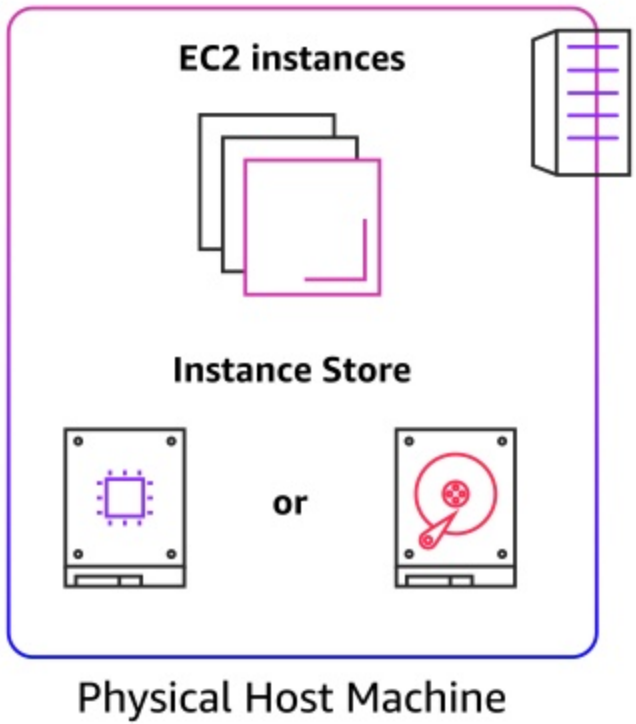
Broadest and deepest platform choice



Amazon EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Amazon EC2 instance store



Local to instance

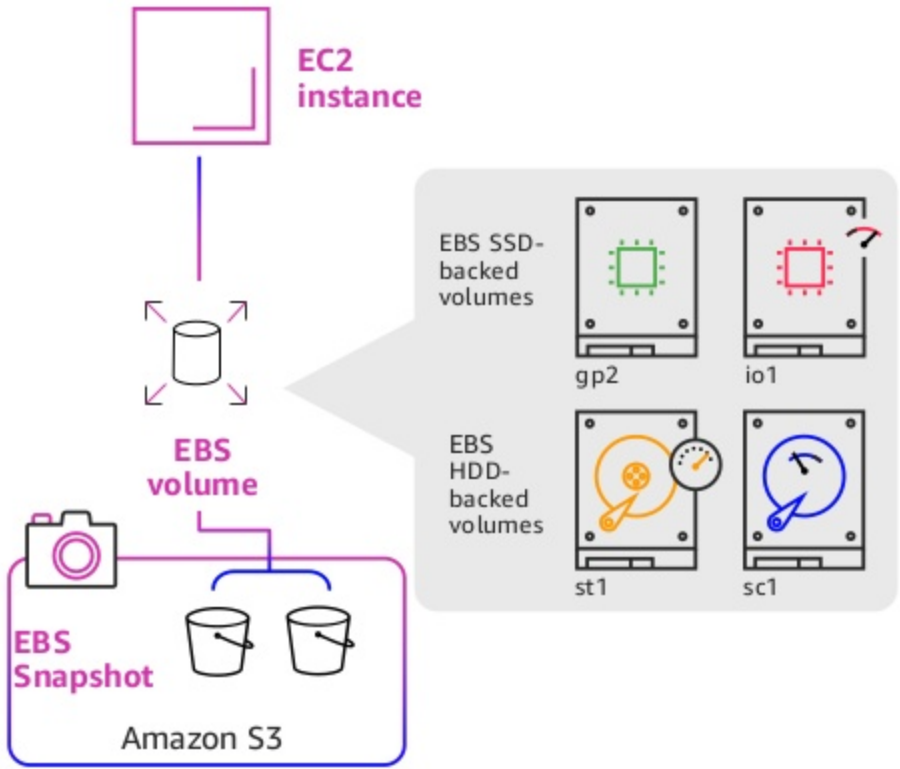
Non-persistent data store

Data not replicated
(by default)

No snapshot support

SSD or HDD

Amazon Elastic Block Store (EBS)



Block storage as a service

Create, attach, modify through an API

Select storage and compute based on your workload

Detach and attach between instances

Choice of magnetic and SSD-based volume types

Supports Snapshots: Point-in-time backup of modified volume blocks

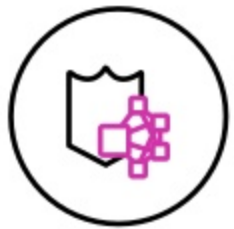
Amazon EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Amazon Virtual Private Cloud (VPC)

Virtual Private Cloud

Provision a logically isolated cloud where you can launch AWS resources into a virtual network



Security Groups & ACLs

NAT Gateway

Flow Logs

VPC Endpoints

Private and secure connectivity to Amazon S3 and Amazon DynamoDB

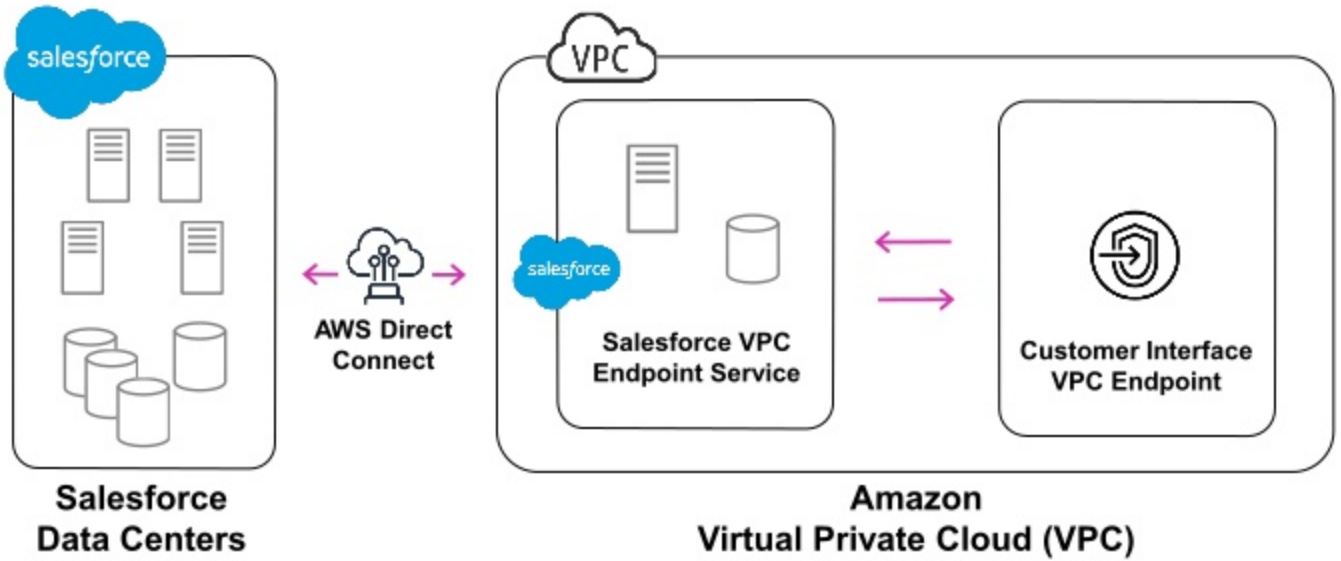
Shared VPC allows multiple accounts to launch their applications into a VPC

AWS PrivateLink

Share services privately between VPCs and on-premises networks

Secure. Scalable. Reliable.

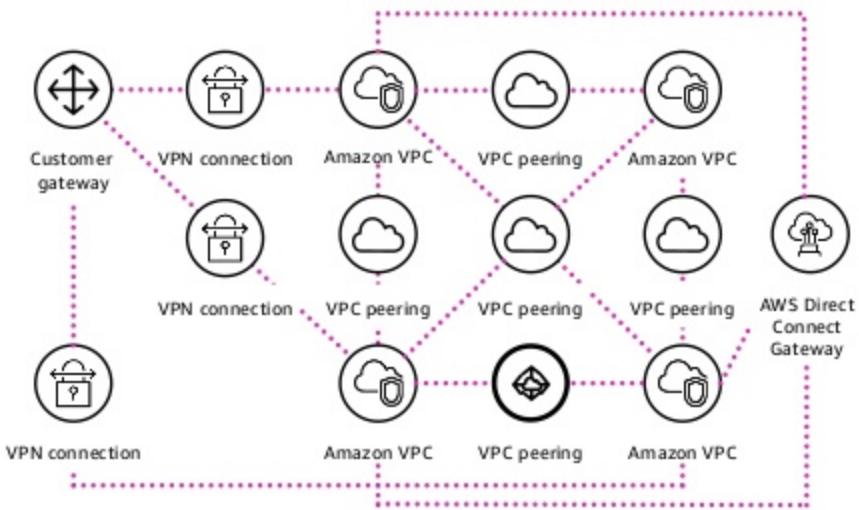
salesforce heroku



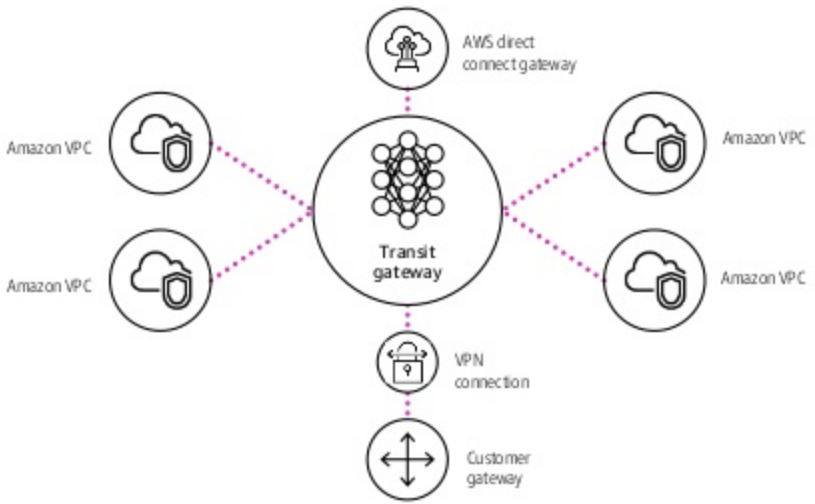
ca technologies · APPDYNAMICS · GE · Expedia · Vanguard · twilio · aqua · TIBCO · snowflake · cisco Stealthwatch Cloud · AUTODESK · dynatrace · Alfresco · SIGOPT

Simplifying the network with AWS Transit Gateway

Network topology today



NEW After Transit Gateway



Simple gateway to easily manage all connectivity

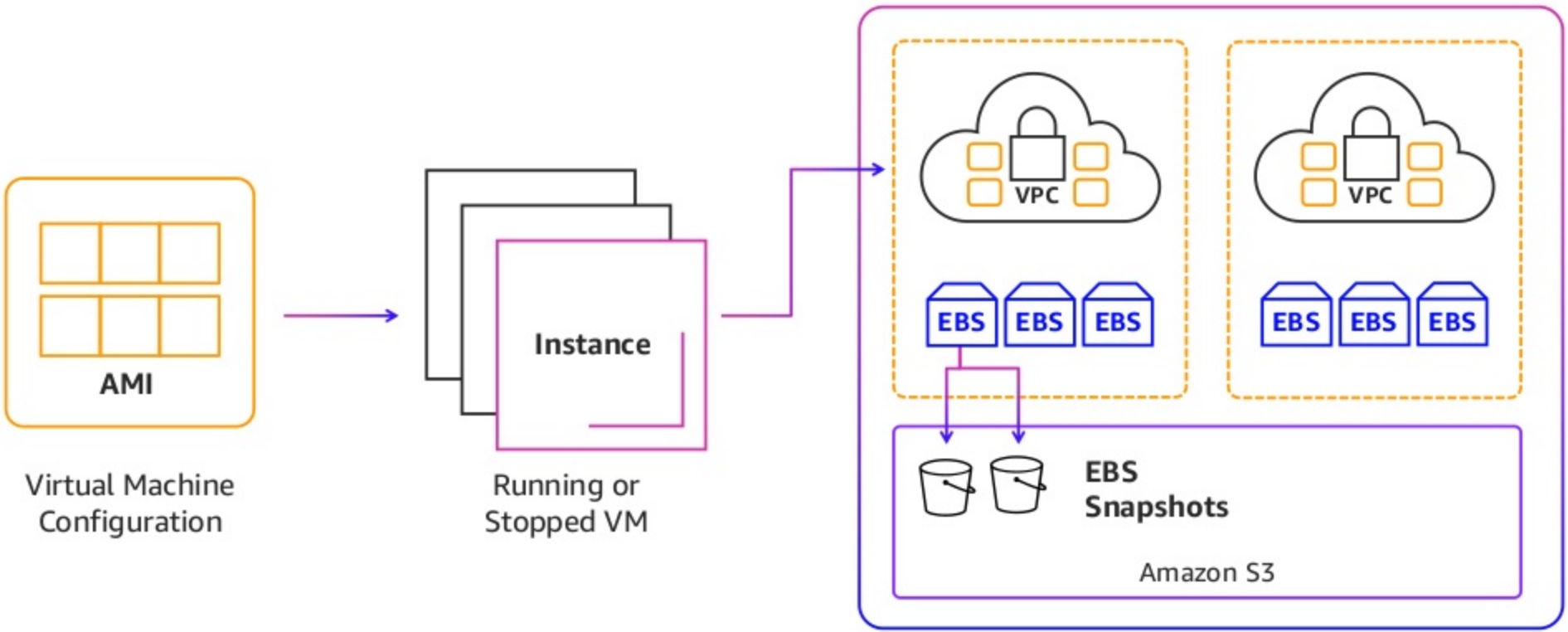
Attach 1000s of VPCs to create a peered network

Edge consolidation for Direct Connect* and VPN

Integrated with AWS Marketplace network partners

* Available in Q1 2019

Amazon EC2 Resources recap



Amazon EC2 Foundations

Resources	**Availability**	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

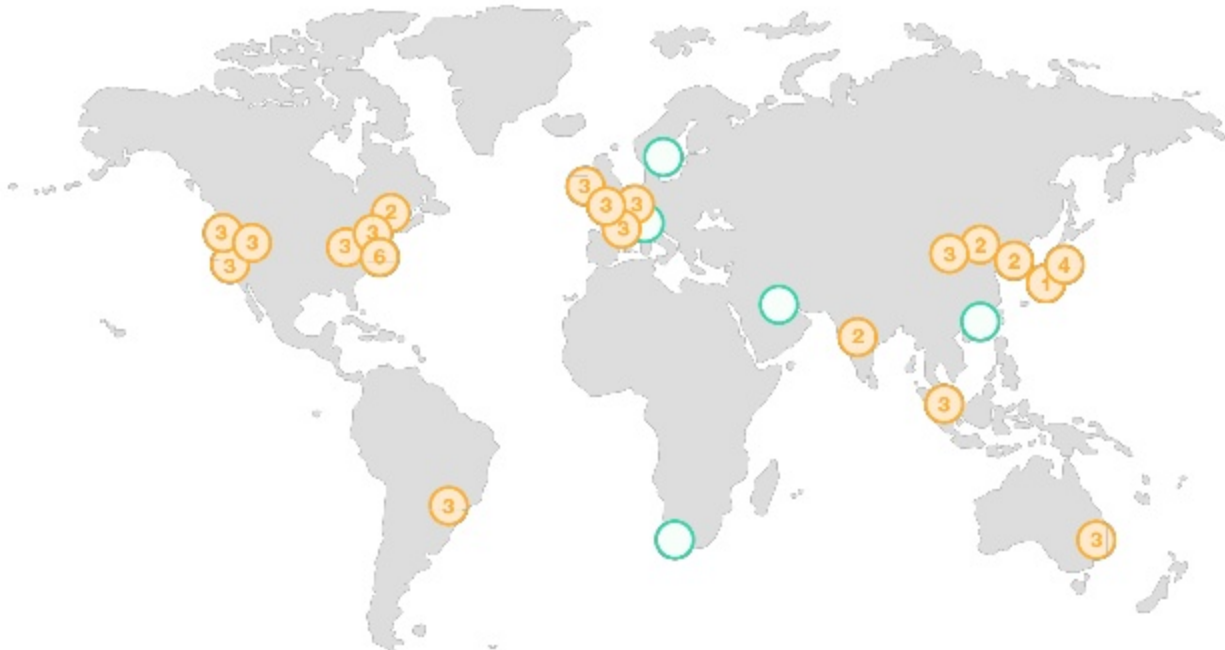
AWS global infrastructure

19 geographic regions

A region is a physical location in the world where we have multiple Availability Zones

57 Availability Zones

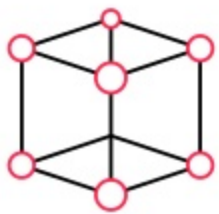
Distinct locations that are engineered to be insulated from failures in other Availability Zones



SLA of **99.99%** availability

Placement Groups

Placement Groups enable you to influence our selection of capacity for member instances, optimizing the experience for a workload



Cluster

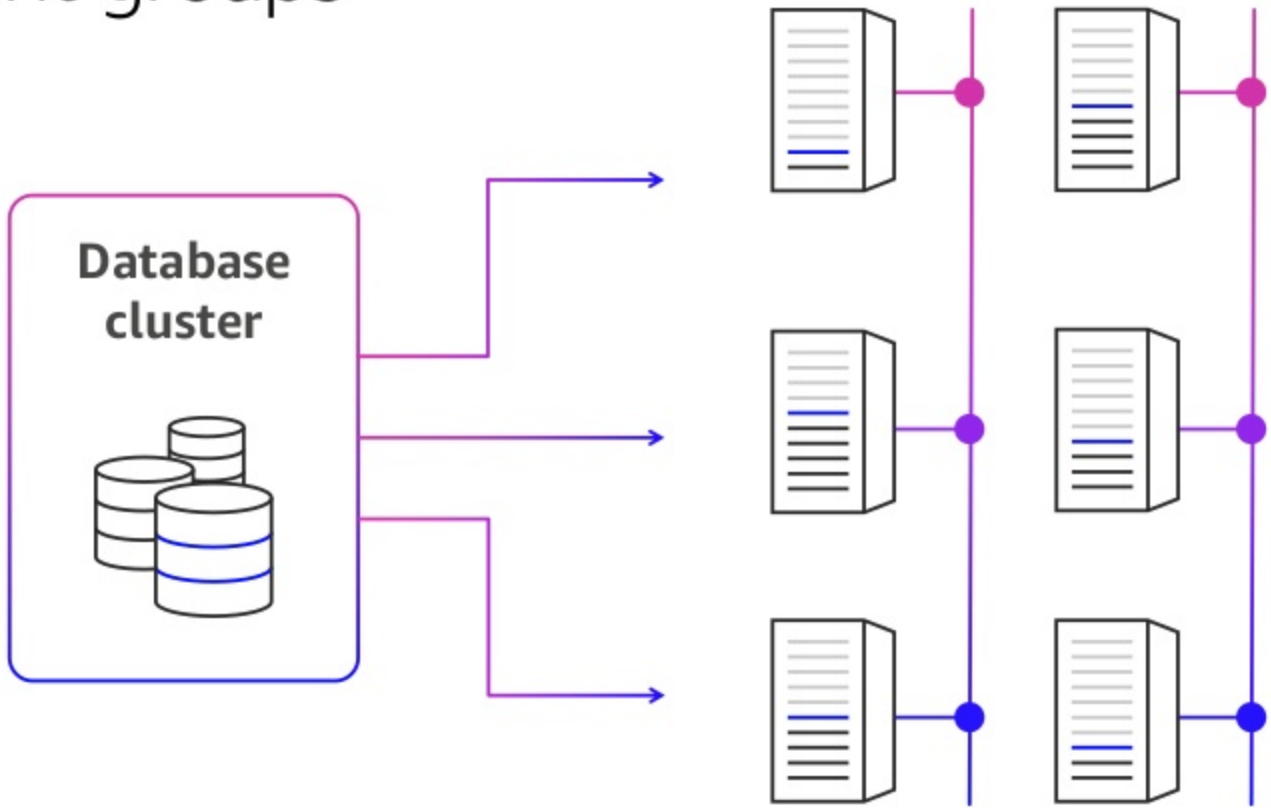
EC2 places instances closely together in order to optimize the performance of inter-instance network traffic

Spread

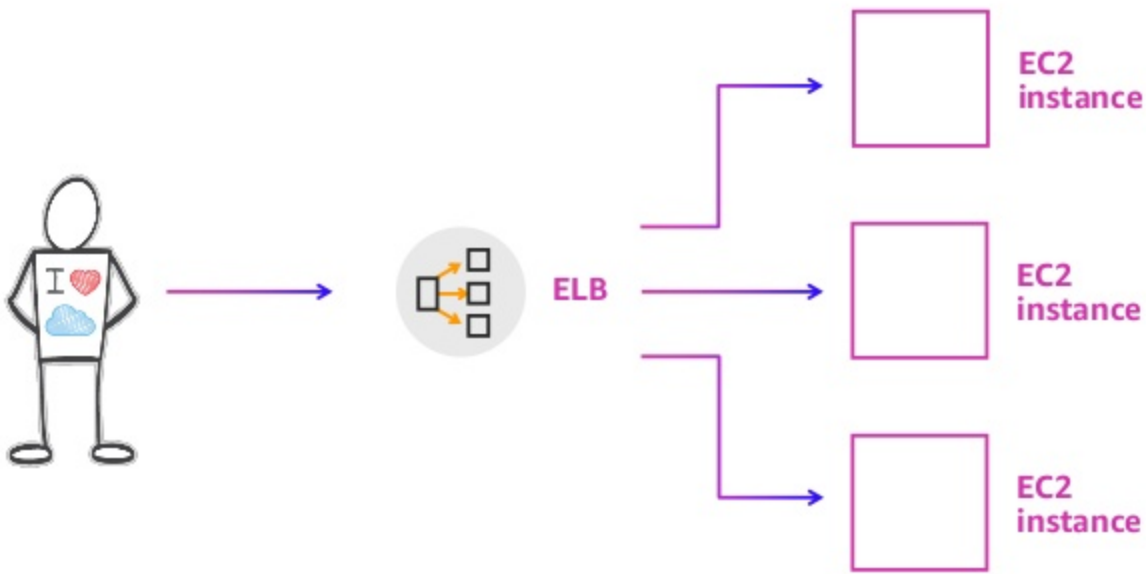
EC2 places instances on distinct hardware in order to help reduce correlated failures

Spread placement groups

When deploying a No SQL database cluster in Amazon EC2, Spread Placement will ensure the instances in your cluster are on distinct hardware, helping to insulate a single hardware failure to a single node



Elastic Load Balancing



Load balancer
used to route incoming requests to multiple Amazon EC2 instances, Containers, or IP addresses in your VPC

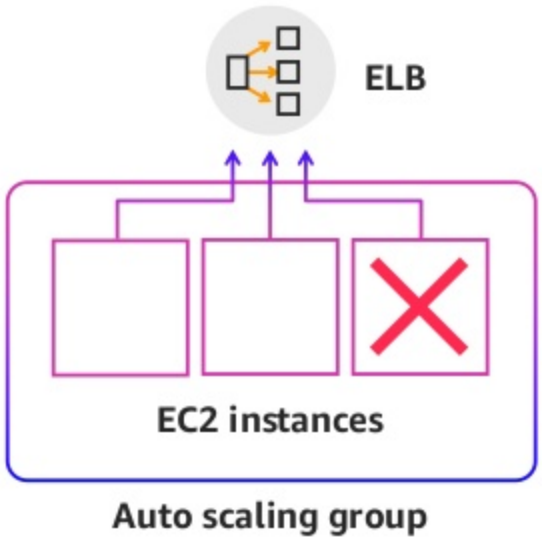
Elastic Load Balancing provides **high-availability** by utilizing multiple Availability Zones

Amazon EC2 Auto scaling

Dynamically react to changing demand, optimize cost

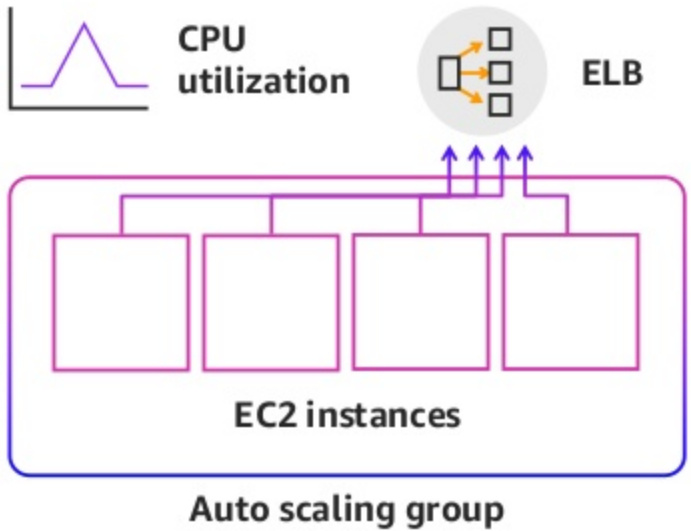
Fleet management

Replace unhealthy instances



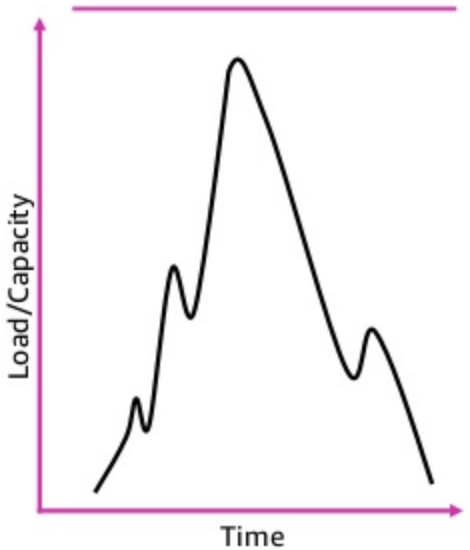
Dynamic scaling

Scale to demand

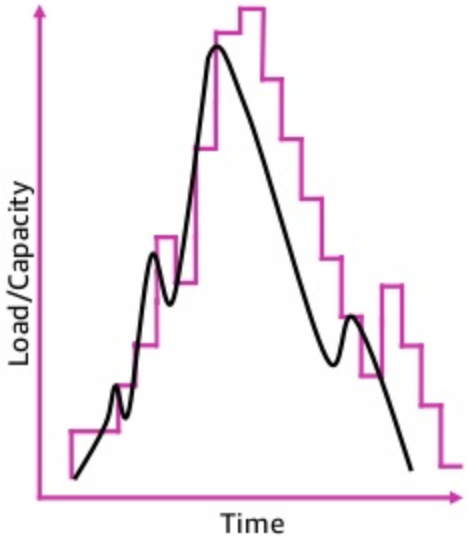


Ramp capacity before you need it with Predictive Scaling

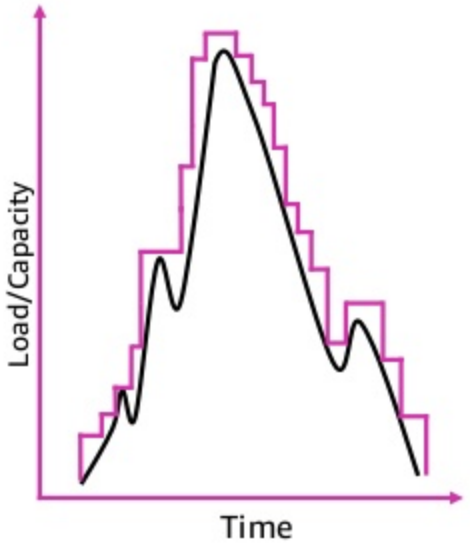
NEW



On-premises capacity provisioning



Capacity provisioning with target tracking



Capacity provisioning with predictive scaling and target tracking

Accommodates long warm-ups by bringing up instances ahead of spikes

Prevents downscaling during sudden untimely traffic dips

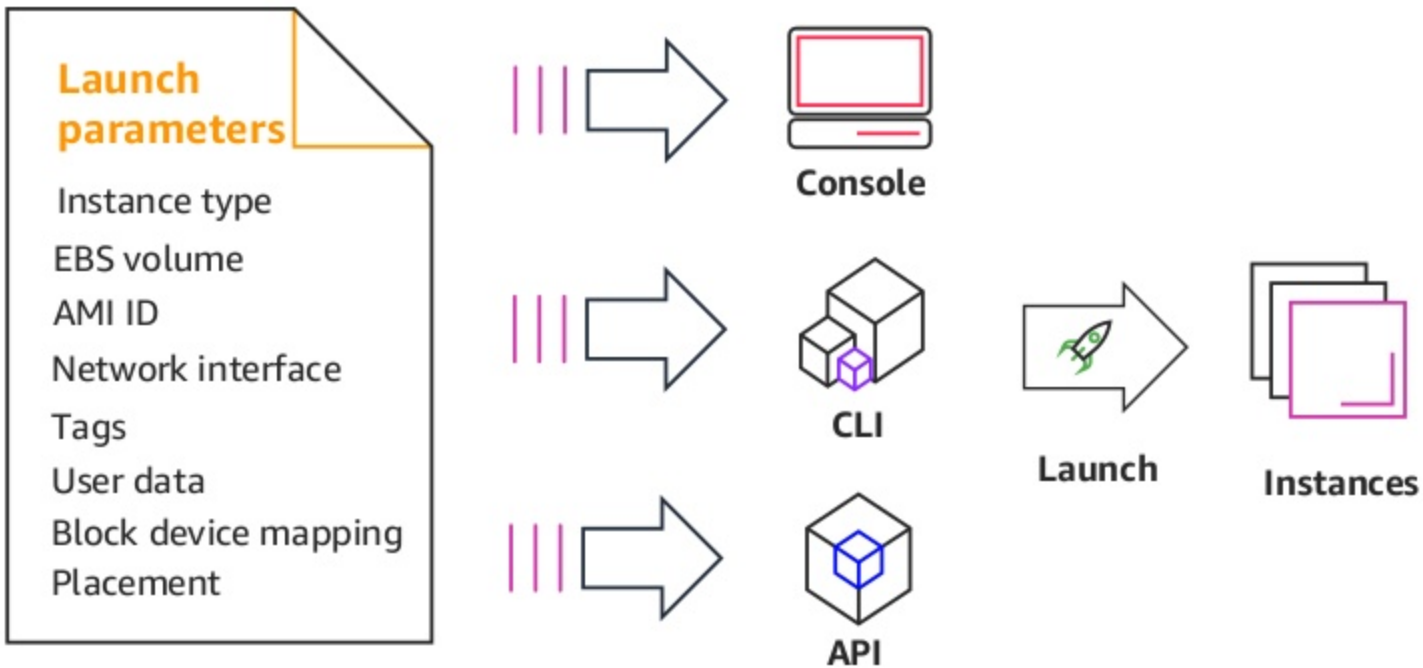
Automatically tracks changing traffic patterns with changing usage patterns

Amazon EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Launching instances with Launch Templates

Templatize launch requests in order to streamline and simplify future launches



Consistent experience

Simple Permissions

Governance and Best Practices

Increased Productivity

AWS Systems Manager: Operate Safely at Scale

Cloud
and
On Premises

Linux
and
Windows

Stay Patch and Configuration Compliant

Automate across accounts and regions

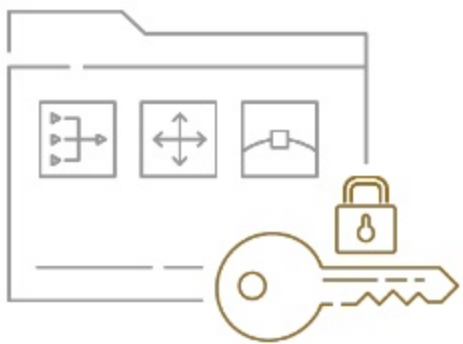
Connect to Amazon EC2 instances via browser and CLI

Track software inventory across accounts

Install agents safely across instances with rate control

AWS Resource Access Manager

NEW Securely share AWS resources with other accounts or AWS organizations



Reduces need to provision duplicate resources

Efficiently uses resources across different departments

AWS Identity and Access Management policies govern consumption of shared resources

Integration with Amazon CloudWatch and AWS CloudTrail

Supports resource sharing for License Manager Configs, Route 53 Resolver Rules, Subnets, and Transit Gateway

AWS License Manager

Simplified license management for on premises and cloud

More easily manage licenses from software vendors

Define licensing rules, discover usage, manage access

Gain single view of license across AWS and on-premises

Discover non-compliant software and help prevent misuse

Seamless integration with AWS Systems Manager and AWS Organizations

Free service for all customers

SAP

Windows

Microsoft SQL Server

ORACLE

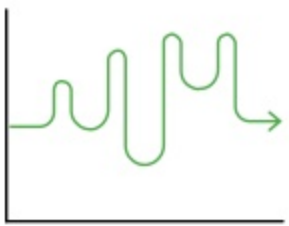
EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Amazon EC2 purchase options

On-Demand

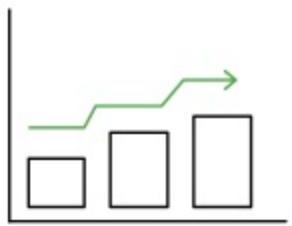
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads, to define needs

Reserved Instances

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



Committed and steady-state usage

Spot Instances

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

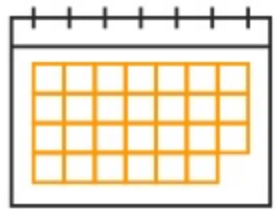
Fault-tolerant, flexible, stateless workloads

To optimize EC2, combine all three purchase options!

Amazon EC2 Reserved Instances pricing

Discount up to 75% off of the On-Demand price

Steady state and committed usage



1- and 3-year terms

Payment flexibility with 3 upfront payment options (all, partial, none)

Convertible RIs
Change instance family, OS, tenancy, and payment

Reserve capacity or opt for flexibility across AZs and instance sizes

On-Demand Capacity Reservations: Manage capacity and RI decisions independently

Amazon EC2 Spot pricing

Spare Amazon EC2 capacity at savings of up to 90% over On Demand

Faster results

Increase throughput up to 10x while staying in budget

Easy to use

Launch through AWS services (ex. Amazon ECS, Amazon EKS, AWS Batch, Amazon EMR) or integrated third-parties

Lean on Spot for these workloads!

Big data

CI/CD

Web services

HPC

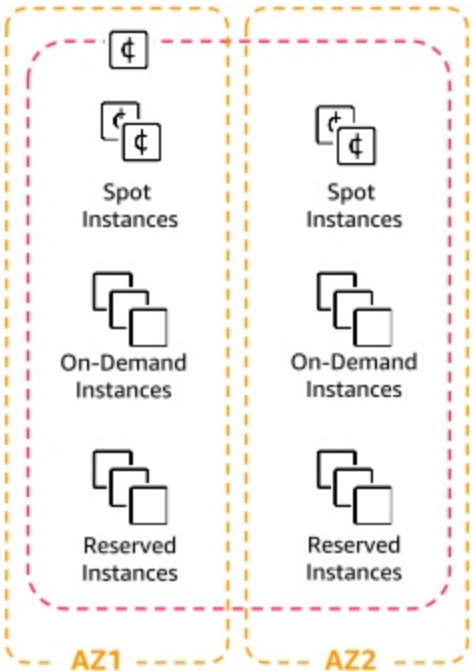
Or **containerized** workloads

Spot is ideal for:
- ☑Fault-tolerant
- ☑Flexible
- ☑Loosely coupled
- ☑Stateless workloads

Amazon EC2 Fleet

NEW

A single API that optimizes the provisioning of capacity across different instance types, AZs, and purchase options



Use all three purchase options to optimize costs

Integrated with Amazon EC2 Auto Scaling, Amazon ECS, Amazon EKS, and AWS Batch

Benefits

Reduce costs

Increase operational efficiency

Reduce development effort

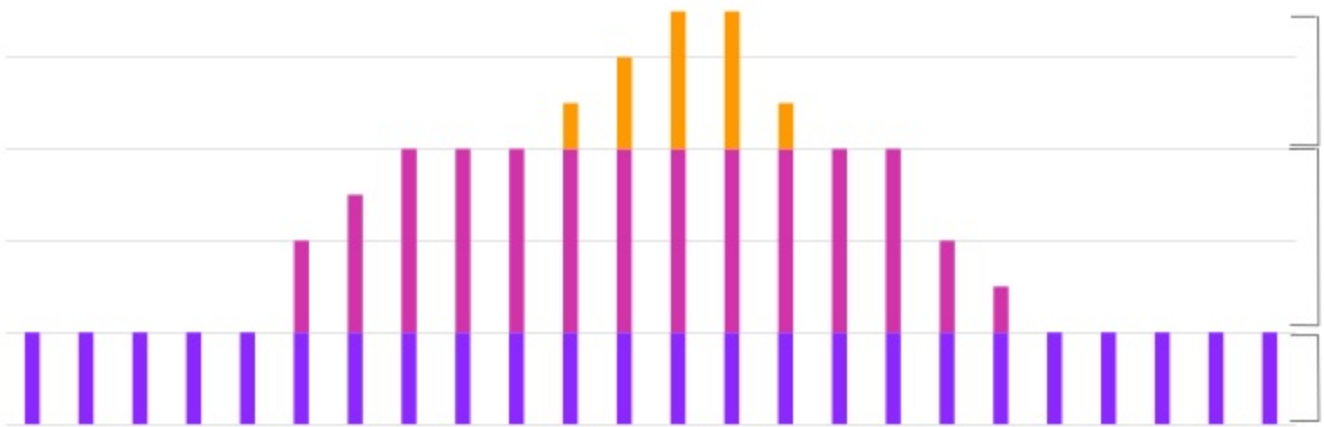
Key features

Flexible capacity allocation

Massive scale

Simplified provisioning

To optimize Amazon EC2, combine purchase options



Scale using **Spot** for fault-tolerant, flexible, stateless workloads

On-Demand, for unknown spiky workloads

Use **RIs** for known, steady-state workloads

NEW Now with Hibernate for Spot and On-demand

Amazon EC2 Foundations

Resources	Availability	Management	Purchase Options
Instances	Regions and AZs	Deployment	On Demand
Storage	Placement Groups	Monitoring	Reserved
Networking	Load Balancing	Administration	Spot
	Auto Scaling		

Thank you!

Thank you!

Please complete the session survey in the mobile app.