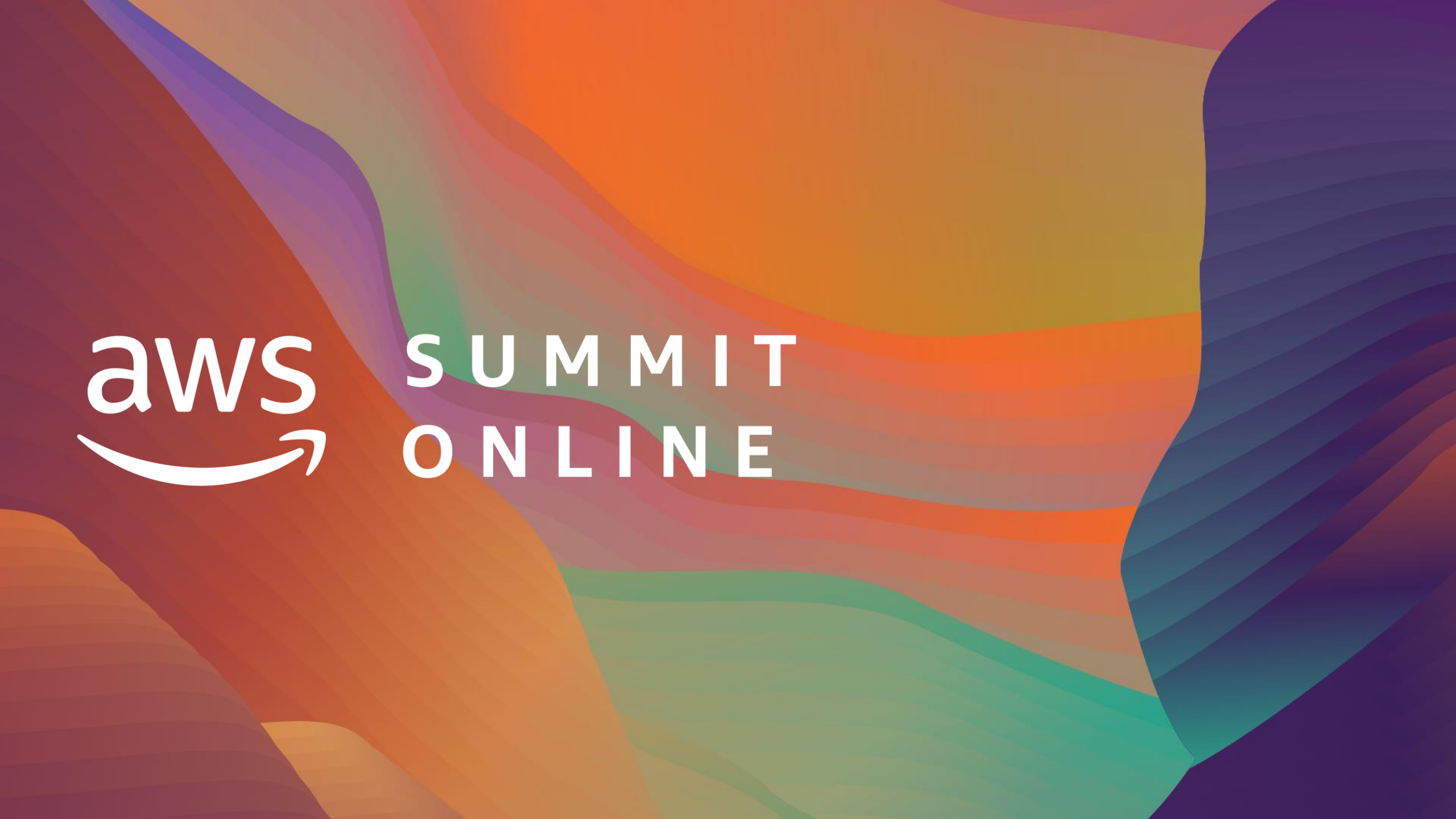

aws
SUMMIT
ONLINE

Getting more out of Amazon EC2

Matt Thomson
Global Head of EC2 Spot
Amazon Web Services

aws SUMMIT

Agenda

Amazon EC2 foundations

Compute for every workload

Pricing optimization

Capacity optimization

Guidance

Workload examples

Conclusion

Amazon EC2 13+ years ago

One size fits all

Pay for what you use

Scale up or down quickly, as needed

Broadest and deepest platform choice

Workloads

General-purpose

Burstable

Compute-intensive

Memory-intensive

Storage (high I/O)

Dense storage

GPU compute

Graphics-intensive

Inference

+

Capabilities

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 24 TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare metal

Size
(Nano to 32xlarge)

+

Options

Amazon Elastic Block Store
(Amazon EBS)

Elastic graphics

Amazon Elastic Inference

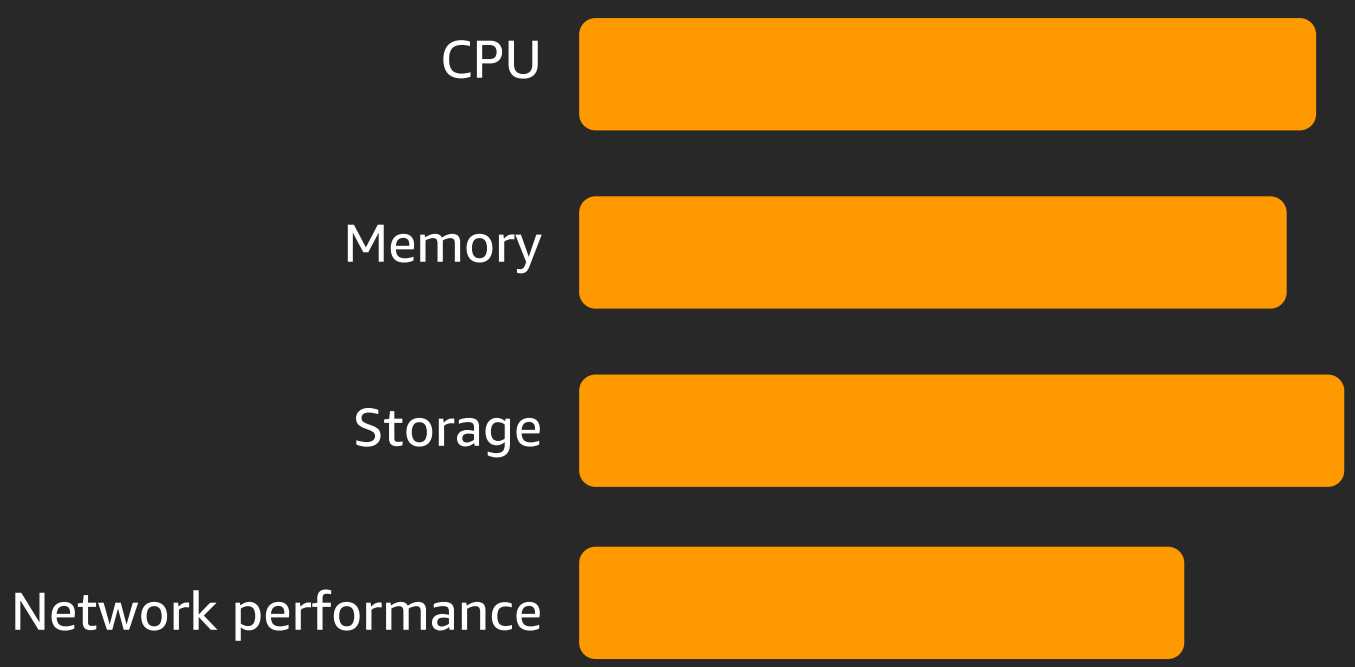
=

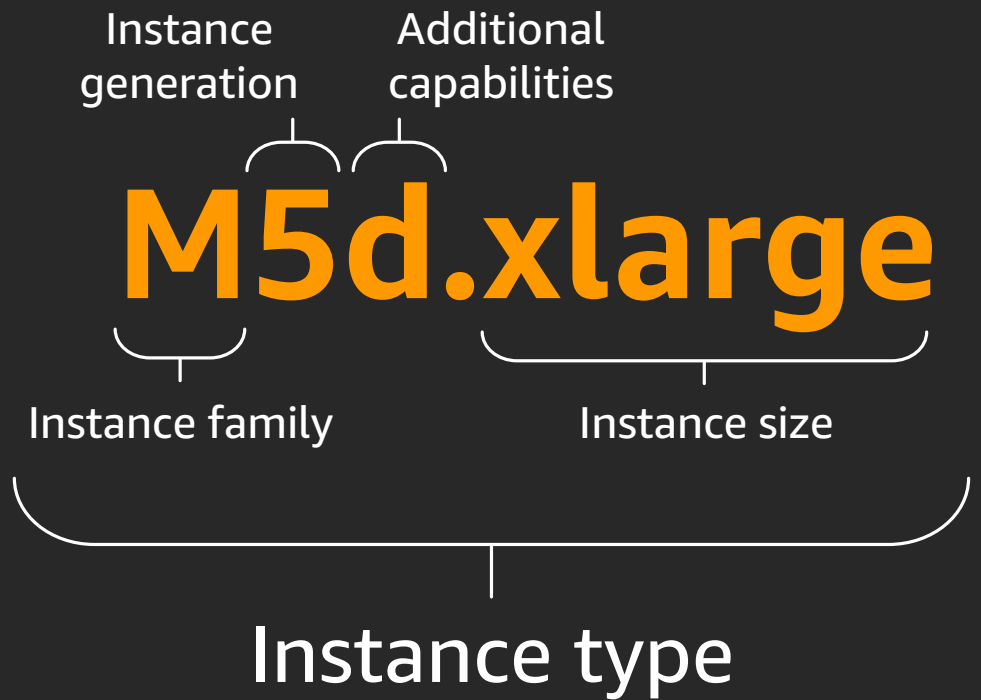
275+

instance types

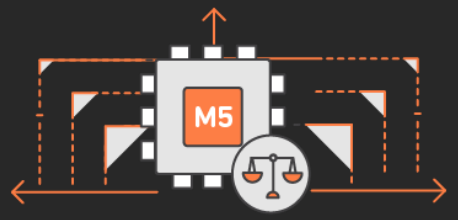
for virtually
every workload
and business need

Amazon EC2 instance characteristics





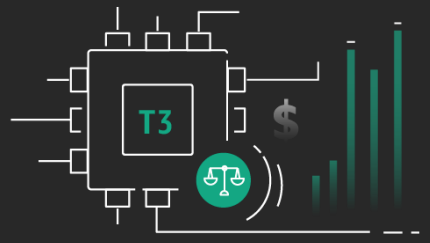
Amazon EC2 general-purpose instances



M5 instances

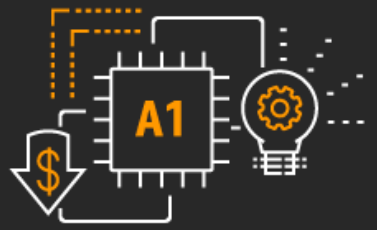
Balance of compute, memory, and network resources

4:1 memory-to-vCPU ratio



T3 instances

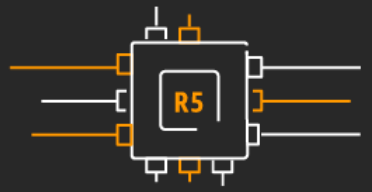
Baseline level of CPU performance with the ability to burst above the baseline for workloads that don't require sustained performance



A1 instances

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

Amazon EC2 memory-optimized instances



R5 instances

Accelerate performance for workloads that process large datasets in memory

8:1 memory-to-vCPU ratio

X1 and X1e instances

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

16:1 and 32:1 memory-to-vCPU ratios

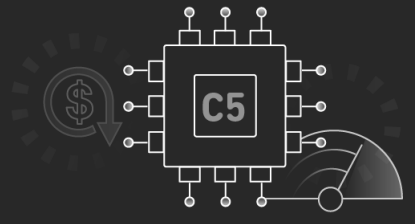
High-memory instances

Extreme memory needs

Certified to run SAP HANA

From 6 TB to 24 TB of memory

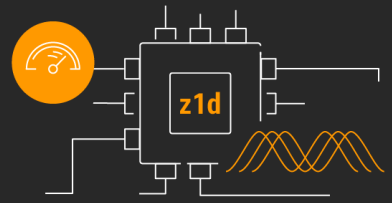
Amazon EC2 compute-optimized instances



C5 instances

High performance at a low price per vCPU ratio

2:1 memory-to-vCPU ratio



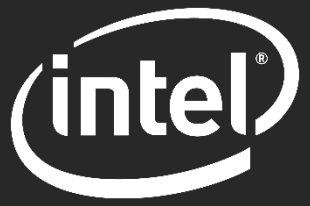
z1d instances

High single-thread performance

Fastest processor in the cloud at 4.0 GHz

8:1 memory-to-vCPU ratio

Broadest choice of processors



Intel Xeon
Scalable processors

AMD EPYC
processors

Graviton
processors

Announcing AWS Graviton2 processors

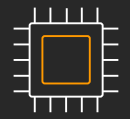
Graviton1 processor

First Arm-based processor in major cloud

Built on 64-bit Arm Neoverse cores with AWS-designed 16-nm silicon

Up to 16 vCPUs,10 Gbps enhanced networking, 3.5 Gbps Amazon EBS bandwidth

Graviton2 processor



Built with 64-bit Arm Neoverse cores; AWS-designed 7-nm silicon process

Up to 64 vCPUs, 20 Gbps enhanced networking, 14 Gbps Amazon EBS bandwidth

7x performance, 4x compute cores, and 5x faster memory

6 new instances powered by AWS Graviton2 processors

General purpose

4 GB DRAM/vCPU

M6g

M6gd

For early access, contact us

Coming in 2020

Compute-optimized

2 GB DRAM/vCPU

C6g

C6gd

Coming in 2020

Memory-optimized

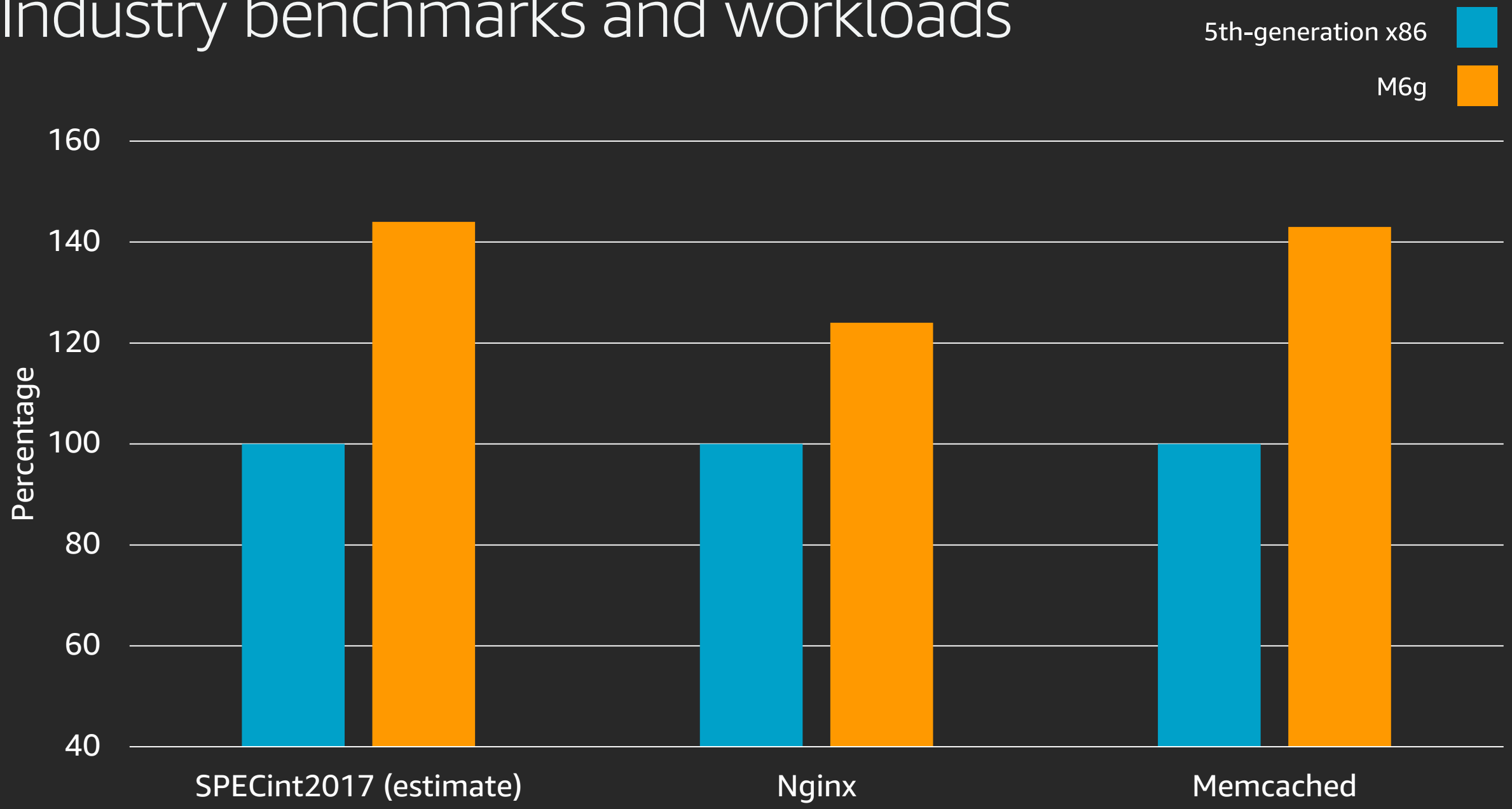
8 GB DRAM/vCPU

R6g

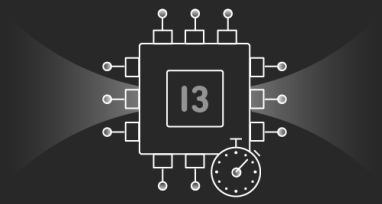
R6gd

All with enhanced networking, Amazon EBS, and 3 with local NVMe SSDs

Industry benchmarks and workloads

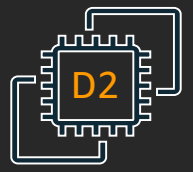


Amazon EC2 storage-optimized instances



I3/I3en instances

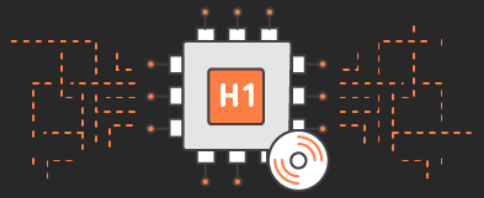
I/O optimized for high-transaction workloads and low-latency workloads



D2 instances

Lowest cost per storage ($/GB)

Supports high sequential disk throughput



H1 instances

Designed for applications that require low cost, high disk throughput, and high sequential disk I/O access to very large datasets

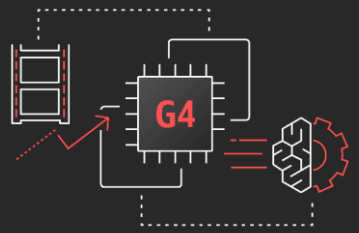
More vCPUs and memory per TB of disk than D2

Amazon EC2 accelerated computing instances

P series
P2/P3 instances

GPU **compute** instances for use cases including deep learning training, HPC simulations, financial computing, and batch rendering

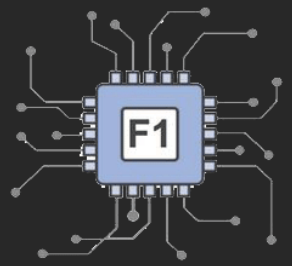
Feature latest NVIDIA high-end GPUs, including Volta V100



G series
G3/G4 instances

GPU **graphics** instances designed for workloads such as 3D rendering, remote graphics workstations, video encoding, and AR/VR

Feature NVIDIA midrange GPUs, such as Turing T4 GPUs, with GRID Virtual Workstation features and license



FPGA instances
F1 instances

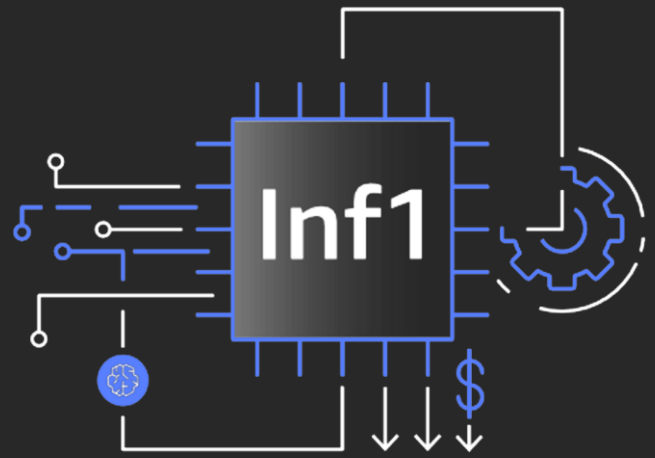
Customer-**programmable FPGAs** that provide dramatic performance improvements for applications such as financial computing, genomics, accelerated search, and image processing

Feature Xilinx Virtex UltraScale+ VU9P FPGAs in a single instance

Programmable via VHDL, Verilog, or OpenCL

Announcing Inf1 instances

Announcing Inf1 instances



High performance and the lowest-cost machine learning inference in the cloud

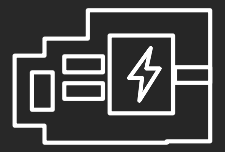
Up to 40% lower cost per inference than any Amazon EC2 GPU instance

Up to 2x higher inference throughput with up to 2,000 TOPS at sub-millisecond latency

Integration with popular machine learning frameworks, including TensorFlow, PyTorch, and MXNet

It starts with our investments in the AWS Nitro System platform

Nitro card

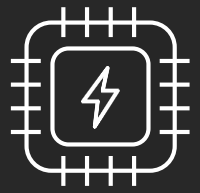


Local NVMe storage

Amazon EBS

Networking, monitoring, and security

Nitro security chip



Integrated into motherboard

Protects hardware resources

Nitro hypervisor

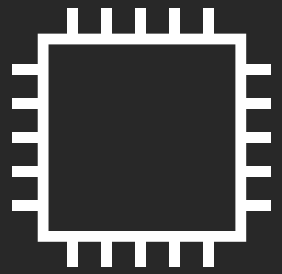
Lightweight hypervisor

Memory and CPU allocation

Bare metal–like performance

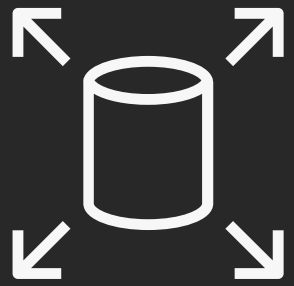
Modular building blocks for rapid design and delivery of EC2 instances

Block storage portfolio



Instance storage

Temporary block-level storage attached to host hardware that is ideal for storage of information that frequently changes or is replicated across multiple instances



Amazon EBS

Easy-to-use, high-performance block storage service designed for use with Amazon EC2 for both throughput- and transaction-intensive workloads

Snapshots

Incremental, point-in-time copies of your Amazon EBS data that can be used to restore new volumes, expand the size of a volume, or move volumes across Availability Zones

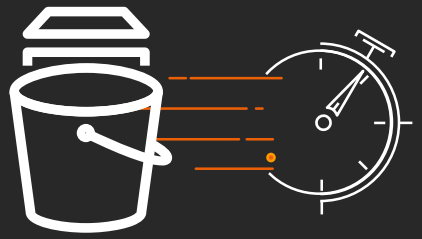
New EBS performance and security improvements

Encryption by default for EBS volumes with opt-in setting

Encrypt all newly created EBS volumes for an account in a Region

Easy to ensure compliance without change to workflows

Fast snapshot restore (FSR)

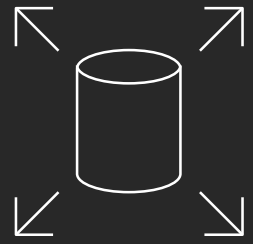


6x lower recovery time objective (RTO)

Skip pre-warming – instant access to data in snapshot and full performance upon volume creation

Restore up to 10 volumes simultaneously

36% higher EBS-optimized bandwidth on C5/C5d, M5/M5d, and R5/R5d instance types



Dedicated bandwidth to Amazon EBS

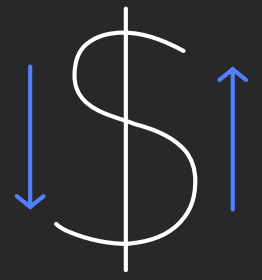
19 Gbps maximum bandwidth, the highest across EC2 instances

Optimizing Amazon EC2 cost and capacity

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

Pricing

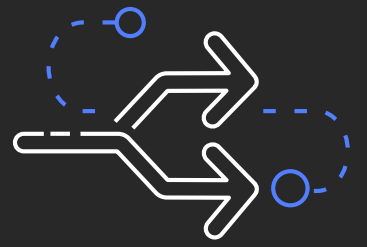


Achieve optimal price and performance with different purchase models

Capacity

Capacity management made easy on the broadest and deepest compute platform

Guidance

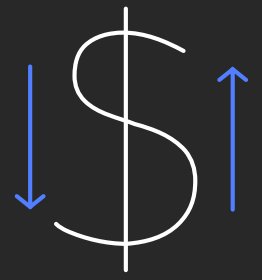


Cost and capacity recommendations enable ease of use and save time

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

Pricing

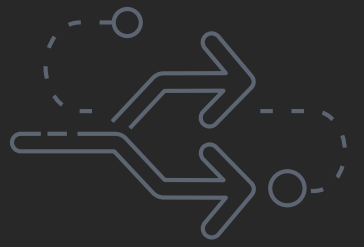


Achieve optimal price and performance with different purchase models

Capacity

Capacity management made easy on the broadest and deepest compute platform

Guidance

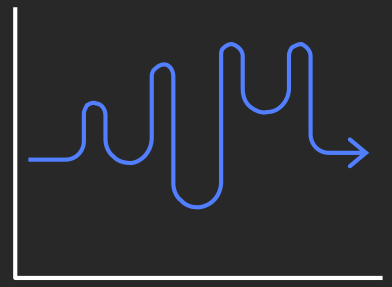


Cost and capacity recommendations enable ease of use and save time

Amazon EC2 purchase options

On-Demand

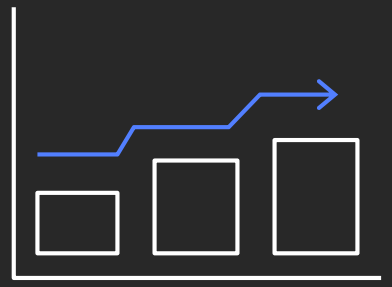
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads to define needs

Reserved Instances (RIs)

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



Committed and steady-state usage

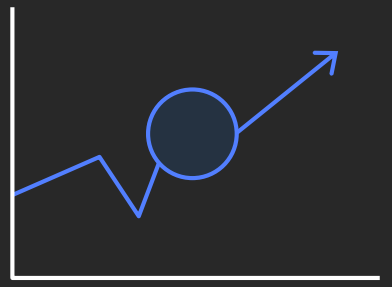
Savings Plans

Same great discounts as Amazon EC2 RIs with **more flexibility**

Flexible access to compute

Spot Instances

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

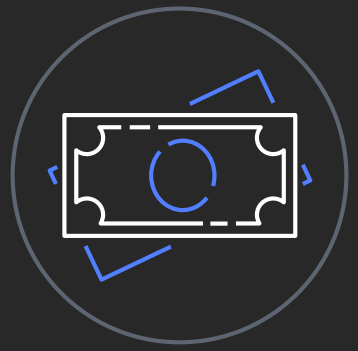


Fault-tolerant, flexible, stateless workloads

Introducing Savings Plans

Easy to use

Receive discounted rates automatically in exchange for a monetary commitment



Significant discounts

Select from two types of Savings Plans to receive discounts of up to 72% on EC2 Instance Savings Plans and 66% on Compute Savings Plans

Flexible

Make a single commitment that applies across multiple AWS compute services, even as your requirements change

Flexible purchase option that offers up to 72% discounts on Amazon EC2 and AWS Fargate usage

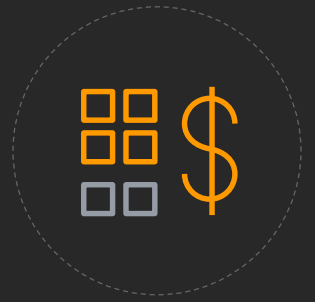
Types of Savings Plans

Compute Savings Plans

Offer the greatest flexibility, up to 66% off (same prices as Convertible RIs)

Flexible across

- ✓ Instance family: e.g., Move from C5 to M5
- ✓ Region: e.g., Change from EU (Ireland) to EU (London)
- ✓ OS: e.g., Windows to Linux
- ✓ Tenancy: e.g., Switch Dedicated tenancy to Default tenancy
- ✓ Compute options: e.g., Move from EC2 to Fargate



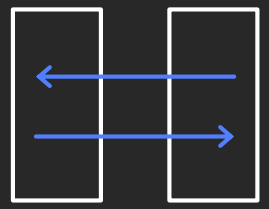
EC2 Instance Savings Plans

Provide the lowest prices, up to 72% off (same as Standard RIs) on the selected instance family (e.g., C5 or M5), in a specific AWS Region

Flexible across

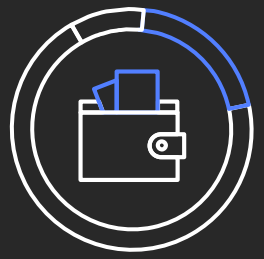
- ✓ Size: e.g., Move from m5.xl to m5.4xl
- ✓ OS: e.g., Change from m5.xl Windows to m5.xl Linux
- ✓ Tenancy: e.g., Modify m5.xl Dedicated to m5.xl Default tenancy

Save up to 90% using EC2 Spot Instances



Instances

Same infrastructure as On-Demand and RIs

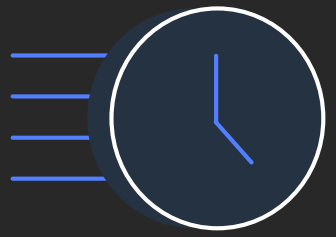


Pricing

Smooth, infrequent changes; more predictable

Usage

Choose different instance types, sizes, and AZs in a single fleet or EC2 Auto Scaling group

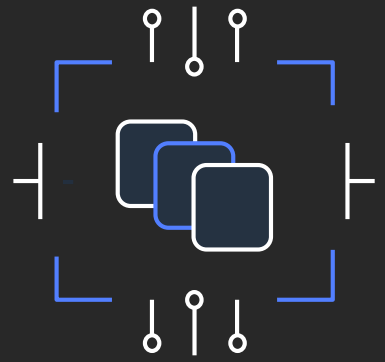


Capacity

AWS can reclaim with 2-minute notice; interruptions only happen if OD needs capacity

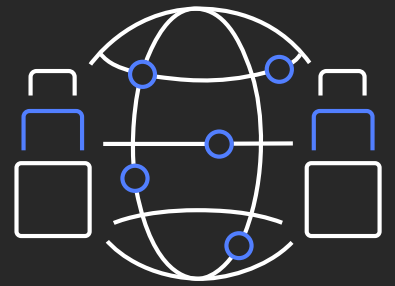
Pricing is based on long-term supply and demand trends; **no bidding!**

Flexibility is key to successful Spot usage



Instance flexible

Time flexible



Region flexible

Handling Spot interruptions

Less than 5% of Spot Instances were interrupted in the last 3 months

Minimal interruptions

Check for 2-minute interruption notification via instance metadata or Amazon CloudWatch events, and automate by

- ✓ Checkpointing
- ✓ Draining from ELB
- ✓ Using stop-start and hibernate to restart faster

Interruption handlers for Amazon ECS and Amazon EKS

Amazon EKS

Amazon ECS

- ✓ Connection between termination requests from AWS infrastructure to nodes
- ✓ Tasks running on Spot Instances will automatically be triggered for shutdown before the instance terminates, and replacement tasks will be scheduled elsewhere on the cluster

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

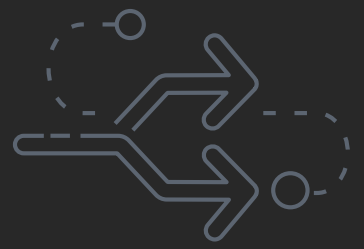
Pricing

Achieve optimal price and performance with different purchase models

Capacity

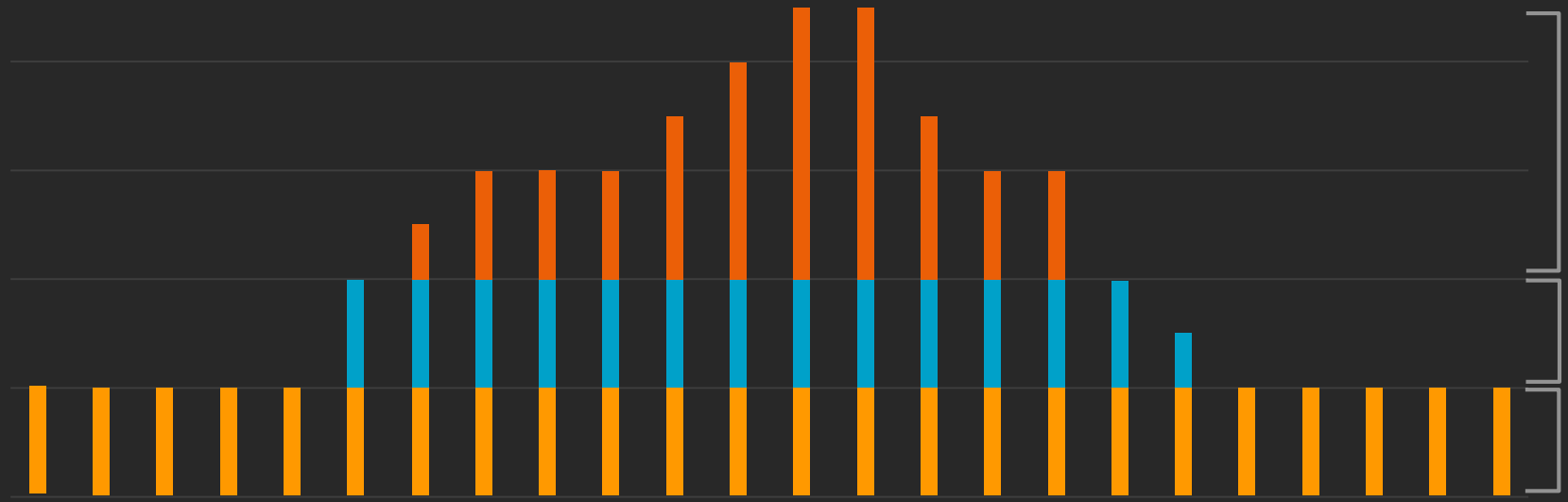
Capacity management made easy on the broadest and deepest compute platform

Guidance



Cost and capacity recommendations enable ease of use and save time

To optimize Amazon EC2, combine purchase options



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

Use **On-Demand** for new or stateful spiky workloads

Use **RIs or Savings Plans** for known, steady-state workloads

Using Amazon EC2 Auto Scaling

Automatically scale instances across instance families and purchase options in a single Auto Scaling group (ASG) to optimize cost

Capacity-optimized

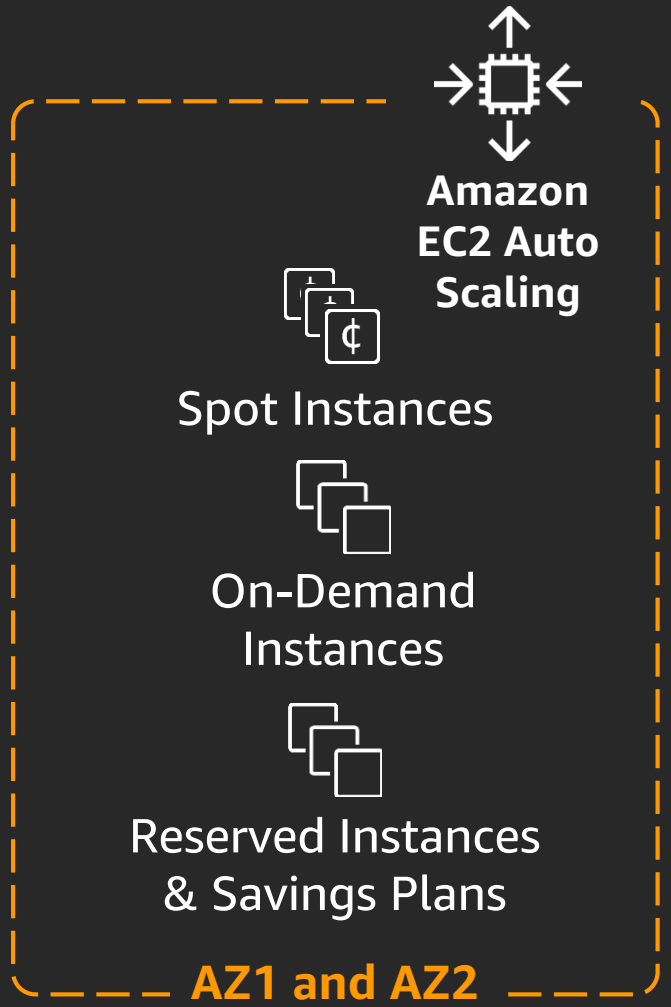
Prioritize deploying Spot Instances into greater Spot pool capacity order to lower the chance of interruptions

Lowest cost

Prioritize cost by selecting a mix of On-Demand and Spot Instances to launch based on the lowest available price

Prioritized list

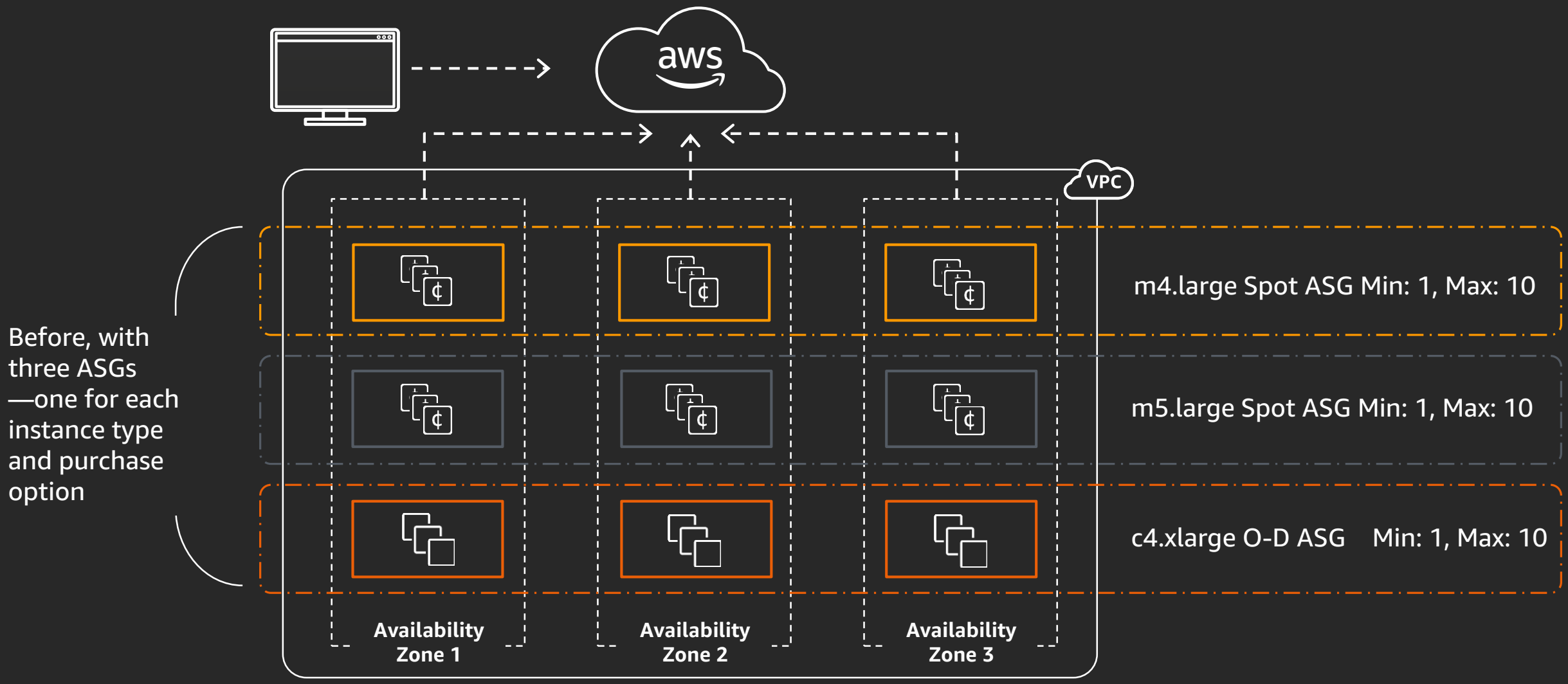
Use a prioritized list for On-Demand instance types to scale capacity during an urgent, unpredictable event to optimize performance



Guided workshop: https://ec2spotworkshops.com/running-amazon-ec2-workloads-at-scale.html

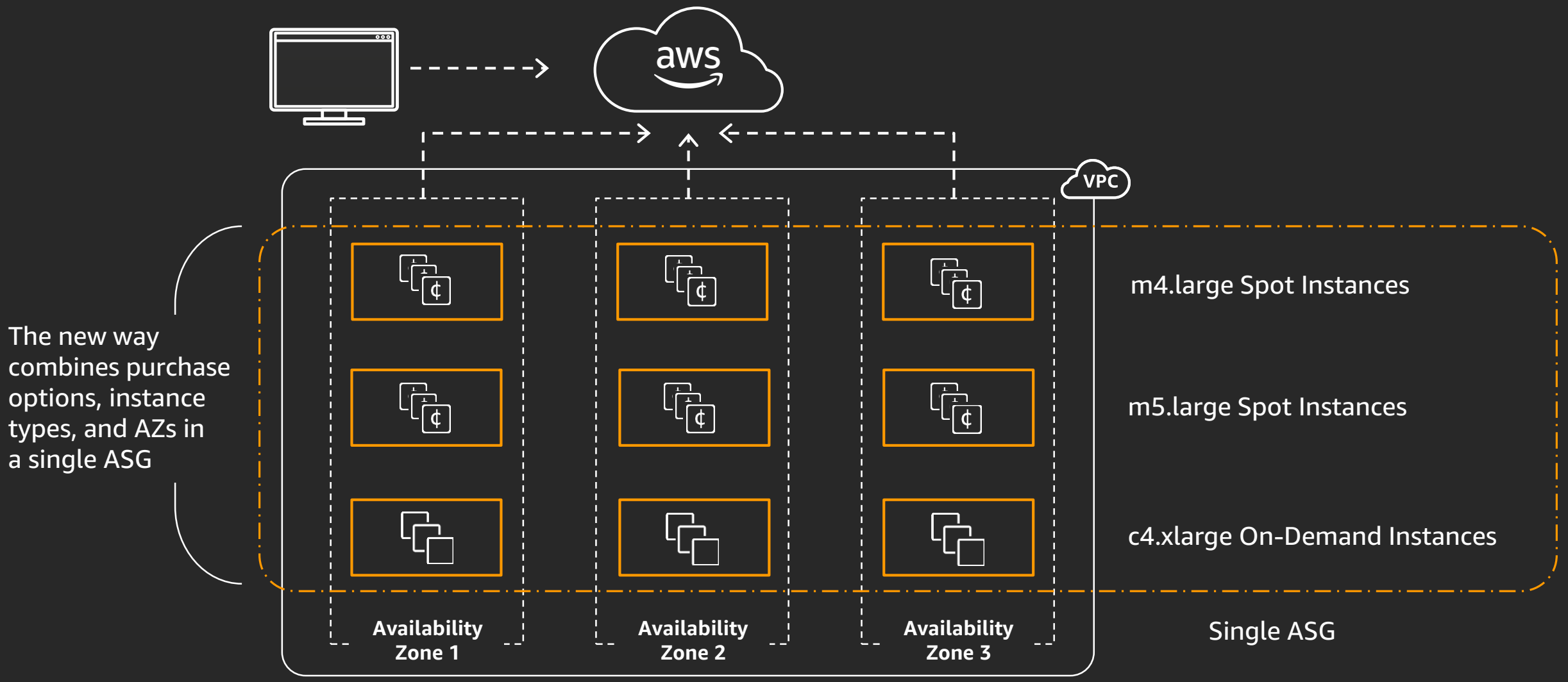
Reduce cost **Optimize performance** **Eliminate operational overhead**

Before: Multiple ASGs to use Spot, On-Demand, and RIs together

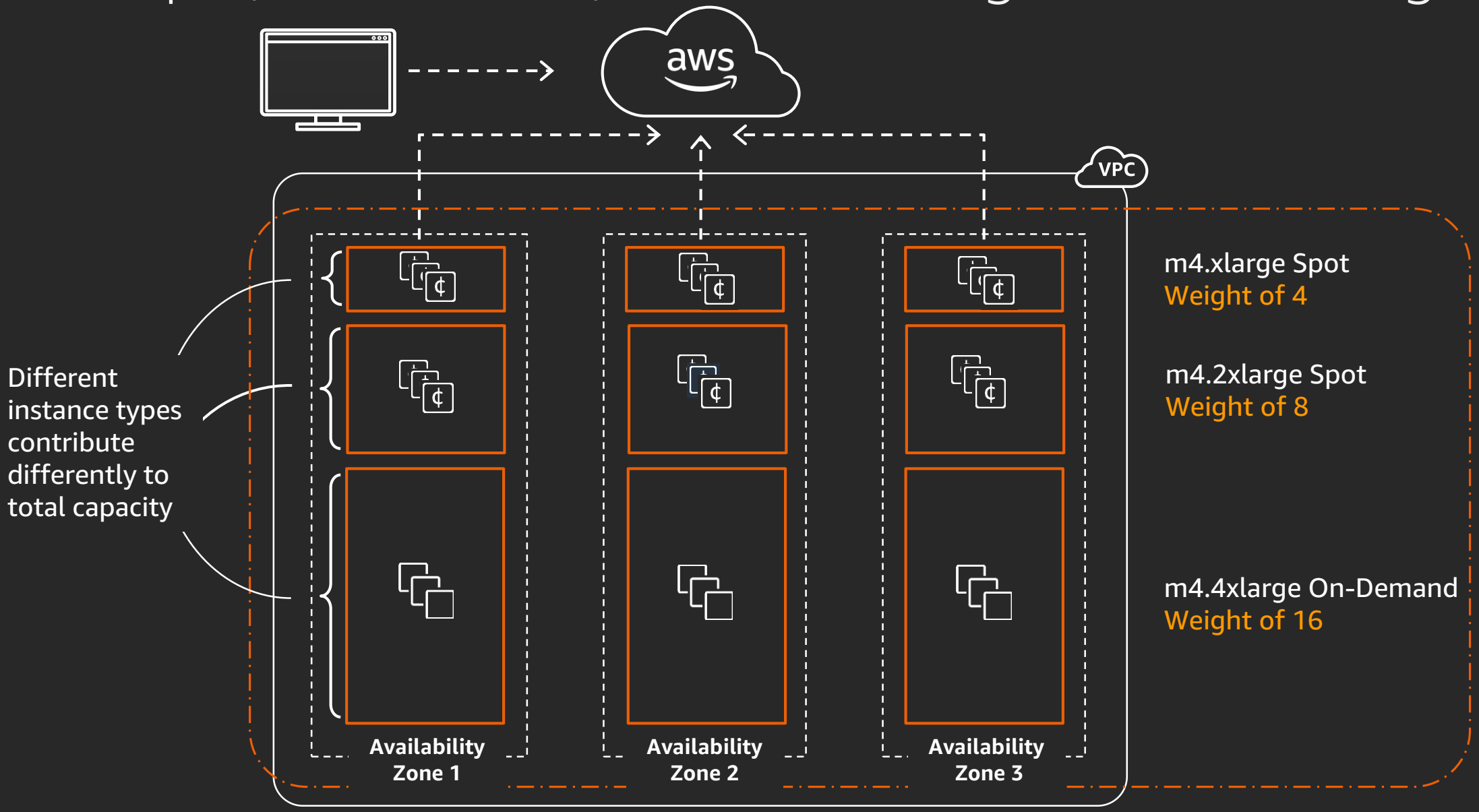


Before, with three ASGs —one for each instance type and purchase option

Then: Spot, On-Demand, and RIs in a single ASG



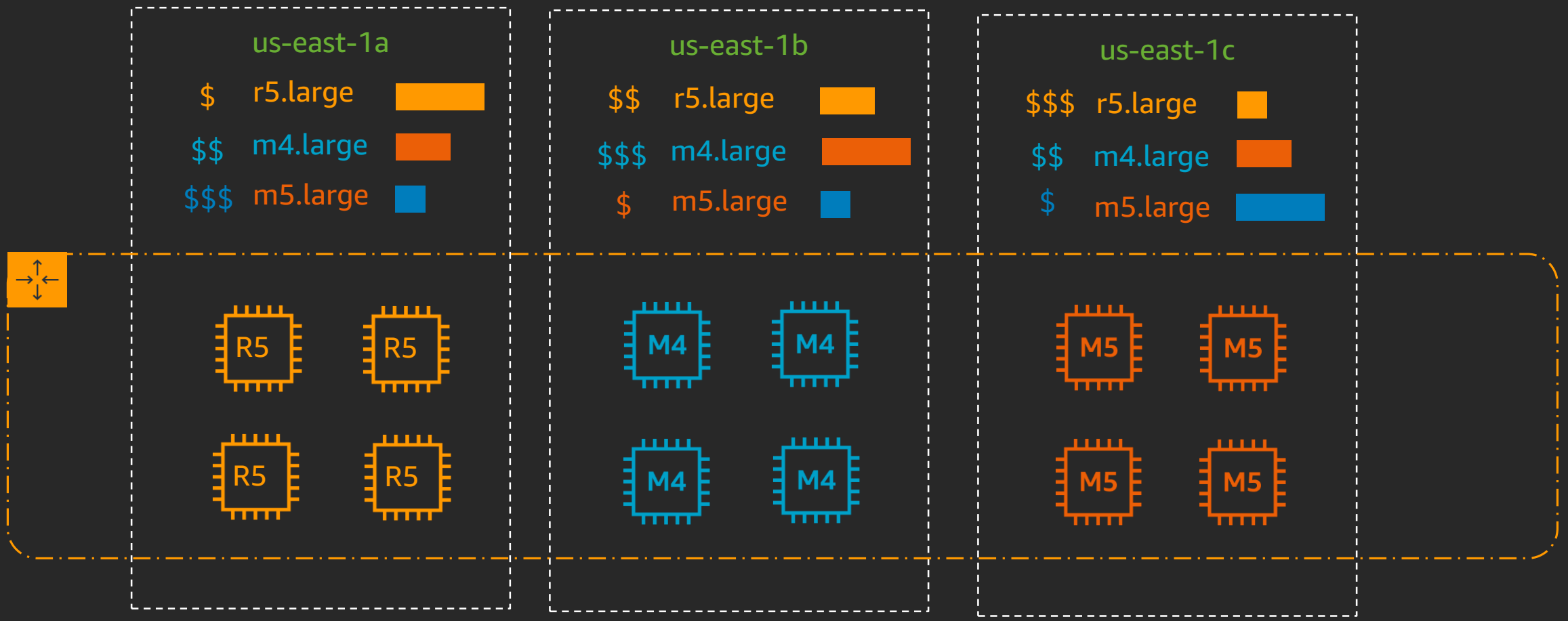
Now: Spot, On-Demand, and RIs in a single ASG with weights



ASG capacity-optimized allocation strategy

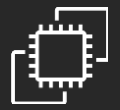
`Desired capacity: 12` `OnDemandBaseCapacity: 0` `OnDemandPercentageAboveCapacity: 0`

`Overrides: ["r5.large", "m4.large", "m5.large"]` `SpotAllocationStrategy: capacity-optimized`



AWS and third-party integrations with Spot Instances and EC2 Auto Scaling

Amazon EC2
Auto Scaling



Amazon EC2
fleet

AWS
Thinkbox

Amazon
EMR

AWS
CloudFormation

AWS
Batch

Amazon
ECS

Amazon
EKS

Amazon
SageMaker

AWS
Fargate

AWS Elastic
Beanstalk

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

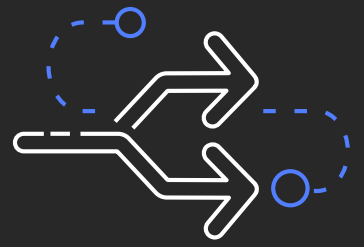
Pricing

Achieve optimal price and performance with different purchase models

Capacity

Capacity management made easy on the broadest and deepest compute platform

Guidance

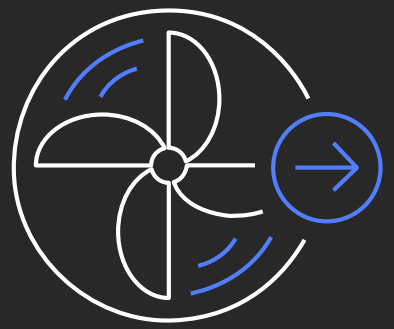


Cost and capacity recommendations enable ease of use and save time

AWS Compute Optimizer

Recommends optimal instances for Amazon EC2 and Amazon EC2 Auto Scaling groups from 140+ instances from M, C, R, T, and X families

Lowers **costs** and improves workload **performance**

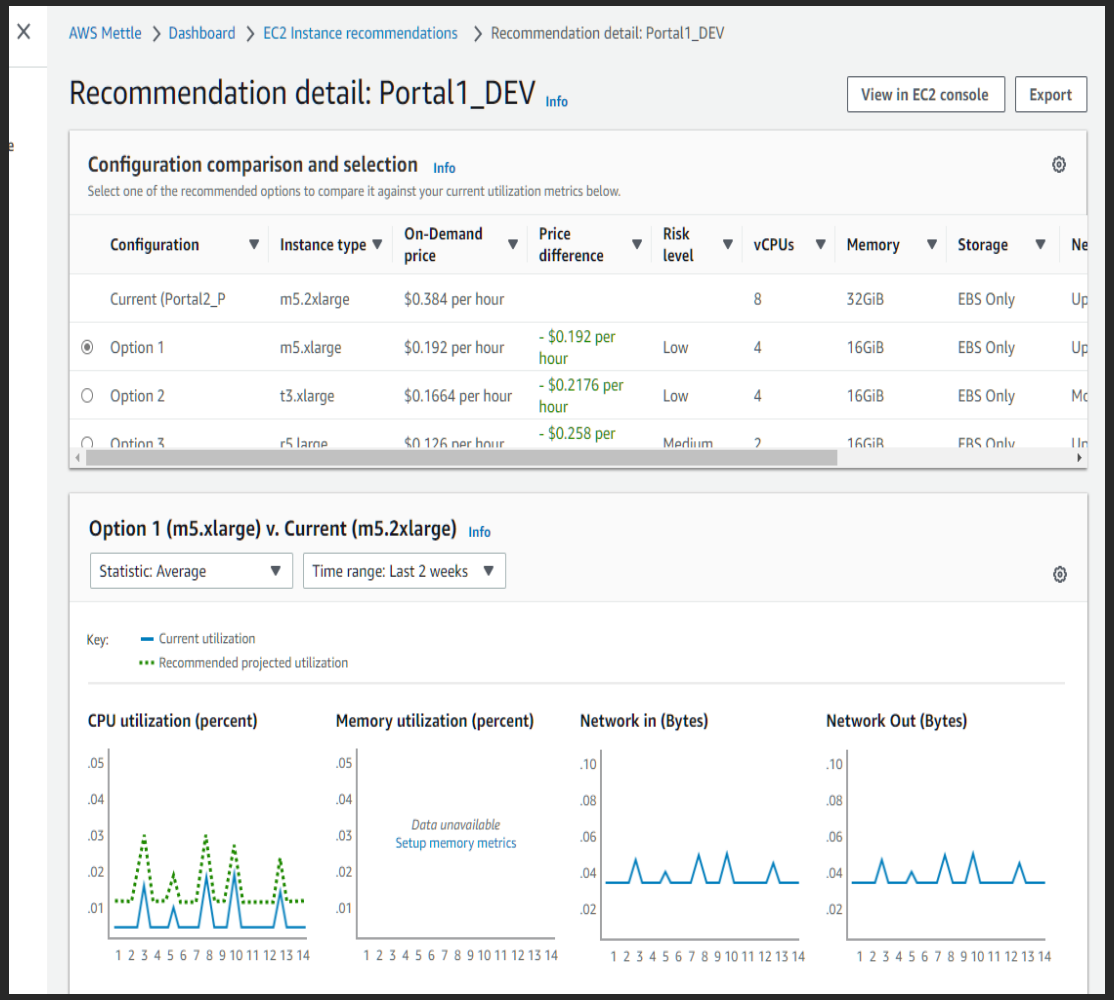


Applies insights from millions of workloads to make recommendations

Saves time comparing and selecting optimal resources for your workload

Easy to choose with AWS Compute Optimizer

New services that recommend optimal AWS compute resources to reduce costs up to 25%



- Recommends optimal EC2 instances
- Optimizes performance and reduces costs by making recommendations to help you right-size compute to your workloads
- Analyzes Amazon CloudWatch metrics and considers Auto Scaling group configuration for intuitive and actionable recommendations
- Up to three recommendations per workload
- Available at no additional charge

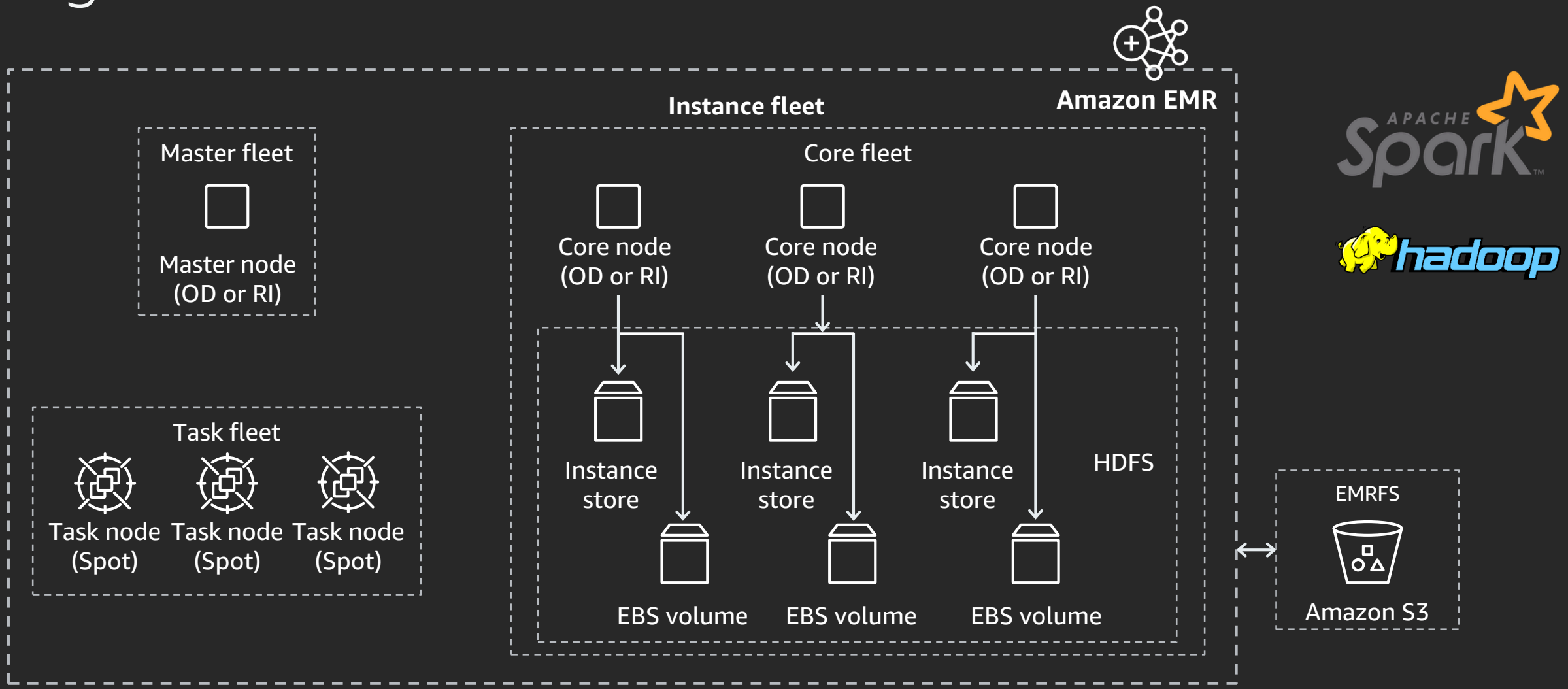
Workloads on AWS

Analytics and big data

DevOps—CI/CD

Websites and web applications

Big data reference architecture



Guided workshop: https://ec2spotworkshops.com/running_spark_apps_with_emr_on_spot_instances.html

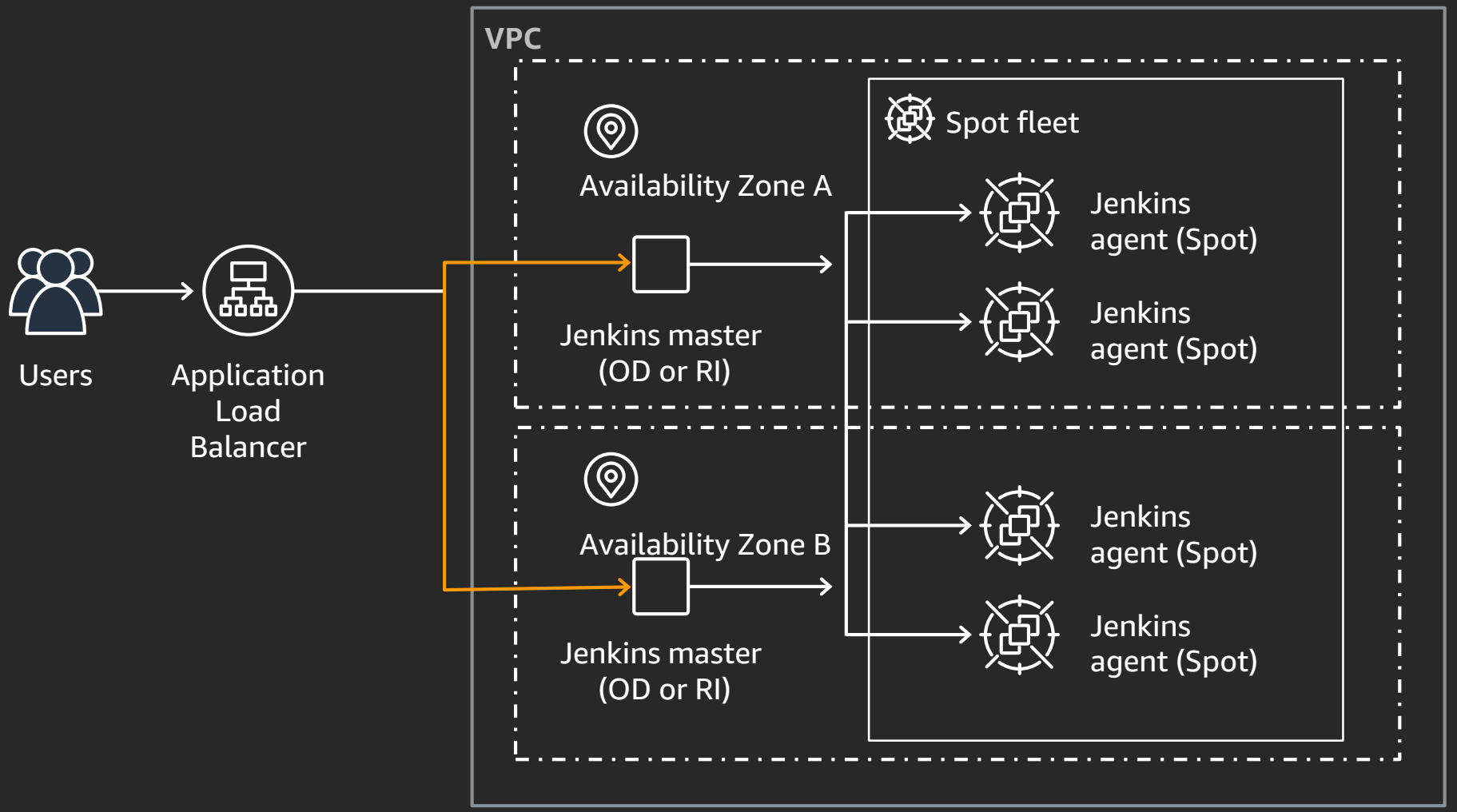
Workloads on AWS

Analytics and big data

DevOps—CI/CD

Websites and web applications

CI/CD reference architecture



Jenkins

Bamboo

https://github.com/awslabs/ec2-spot-jenkins-plugin/

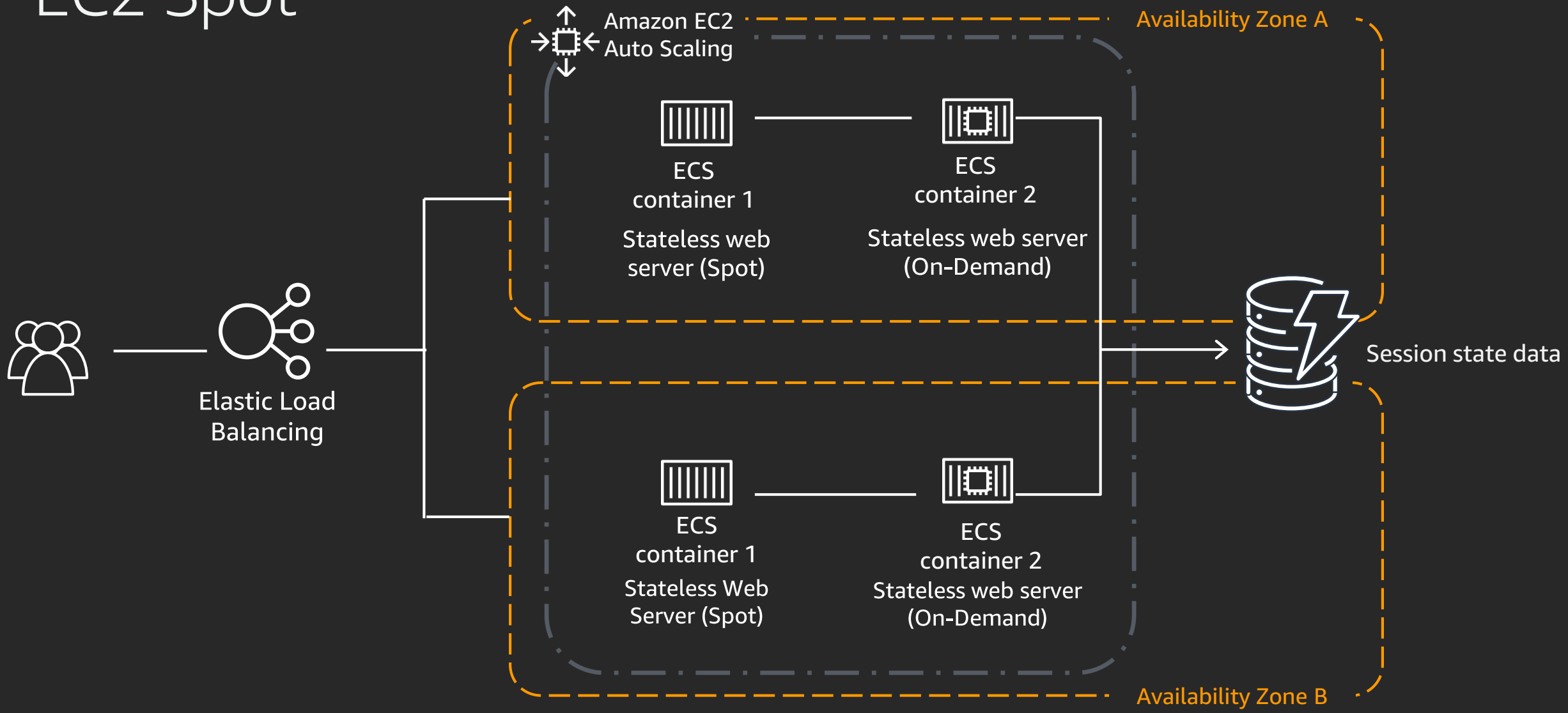
Workloads on AWS

Analytics and big data

DevOps—CI/CD

Websites and web applications

Running web applications with Amazon ECS on EC2 Spot



Key takeaways

1. EC2 has the right compute for every workload

Workload-optimized EC2 instances, AWS Nitro System, elastic block storage

2. Access compute at a lower cost to innovate faster

Spot Instances & Savings Plans

3. How to automate cost and capacity optimization

EC2 Auto Scaling

4. Optimize your workloads by using best practices

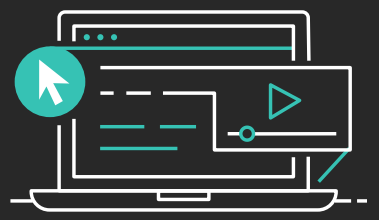
AWS Compute Optimizer

5. Get technical guidance in an AWS Immersion Day

CI/CD, analytics, big data, machine learning, and web services

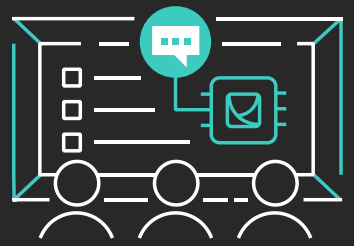
Learn compute with AWS Training and Certification

Resources created by the experts at AWS to help you build cloud compute skills



20+ free digital courses cover topics related to cloud compute, including introductions to the following services

- Amazon EC2
- Amazon EC2 Auto Scaling
- AWS Systems Manager
- AWS Inferentia and Amazon EC2 Inf1 instances



Compute is also covered in the classroom offering, **Architecting on AWS**, which features AWS expert instructors and hands-on activities

Visit the learning library at https://aws.training

Thank you!

aws SUMMIT