# AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

CMP210

# Optimize compute for cost and capacity

Chris Foote (he/him)
Sr. GTM Specialist, Compute
Amazon Web Services

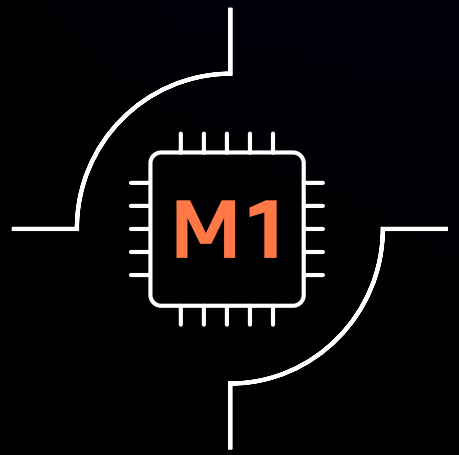
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

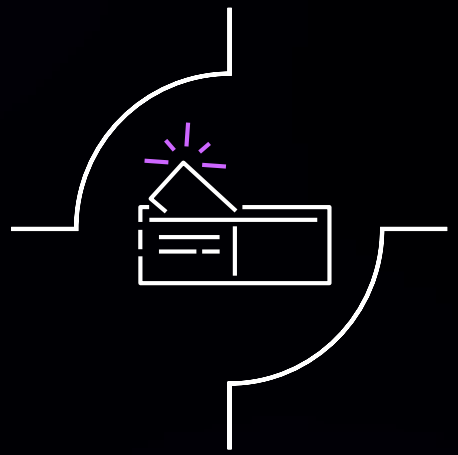
- Compute platform for virtually every workload
- Selecting the right Amazon EC2 instance types for you
- Amazon EC2 instance selection tools
- Price, capacity, and performance optimization
- Conclusion

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

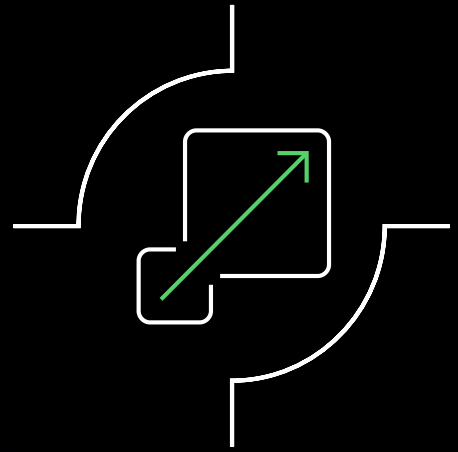
# Amazon EC2 15 years ago




One size fits all



Pay for what you use



Scale up or down quickly as needed

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Broadest and deepest platform choice

## CATEGORIES

General purpose

Burstable

Compute-intensive

Memory-intensive

Storage (high I/O)

Dense storage

GPU compute

Graphics-intensive

## CAPABILITIES

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High footprint
(up to 24 memory TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare metal

Size
(nano to 32xlarge)

## OPTIONS

Amazon EBS

Amazon Elastic Inference

Elastic Fabric Adapter

**400+**

**INSTANCE TYPES**

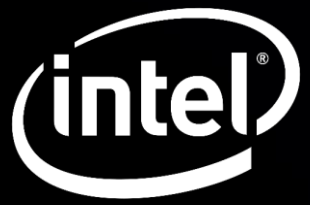
for virtually every workload and business need

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Broadest choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD




Intel Xeon Scalable (Skylake and Cascade Lake) processors




AMD EPYC processor




AWS Graviton2 Processor

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Broadest choice of accelerators

RIGHT COMPUTE FOR THE WORKLOAD




A100 Tensor Core and T4 Tensor Core GPUs



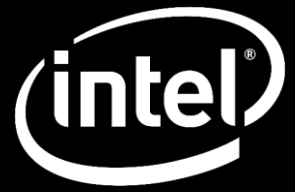

FPGAs for custom hardware acceleration




AWS Inferentia




G4ad instances




Habana-based instances

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Graviton2–based instances

UP TO 40% BETTER PRICE PERFORMANCE OVER COMPARABLE, CURRENT-GENERATION X86-BASED INSTANCES

**M6g**
General-purpose workloads

FREE TRIAL
**T4g**
Burstable general-purpose workloads

**C6g**
Compute-intensive workloads

**R6g**
Memory-intensive workloads

NEW
**X2gd**
Memory-optimized

Lowest price/GiB of RAM

## Also available

Local NVMe-based SSD storage options – general purpose **(M6gd)**, compute-optimized **(C6gd)**, and memory-optimized **(R6gd & X2gd)**

**C6gn** instances with 100 Gbps networking

Bare-metal options for M, C, R & X instance types – **M6g.metal, M6gd.metal, C6g.metal, C6gd.metal, R6g.metal, R6gd.metal, X2gd.metal**

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How do you select the right Amazon EC2 instance types for your workload?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Instance type selection is a process, but don’t overthink it

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Familiarize yourself with instance type categories

General purpose
Burstable

Compute-intensive
Memory-intensive

Storage (high I/O)
Dense storage

GPU compute
Graphics-intensive

## But no need to memorize everything

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Focus on using Amazon EC2 efficiently

Don’t try to completely optimize for cost on initial instance type selection

Optimize costs with purchasing options and additional compute optimization tools later

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# You know your application the best

There is a good chance multiple instance types can power your application/workload

Think flexible!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Take an attribute-based approach to instance type selection

## Work backwards from your application/workload performance requirements

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# What attributes do you need?

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High footprint
(up to 24 memory TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare metal

Size
(Nano to 32xlarge)

Amazon Elastic Inference

Elastic Fabric Adapter

Focus on the capabilities and options that are required for your application/workload

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Does AWS provide an easy way to discover and locate Amazon EC2 instances?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 instance discovery

SEARCH AND DISCOVERY EXPERIENCE FOR LOCATING INSTANCE TYPES



Single source of truth for the latest instance types, attributes, regional and zonal offerings, and pricing



Compare the hardware attributes, pricing, and availability of each instance type to select the best option



Programmatic access to instance type information with two APIs

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Attribute-based instance type selection NEW

## Selects and launches instance types for EC2 Fleet and Amazon EC2 Auto Scaling group based on your required attributes

- Express instance requirements as a set of attributes, such as vCPU, memory, storage type, and much more!
- Ideal for instance-type-flexible workloads and frameworks, such as containers, big data, and CI/CD
- Future-proofs fleet configuration by automatically adding newly released instance types

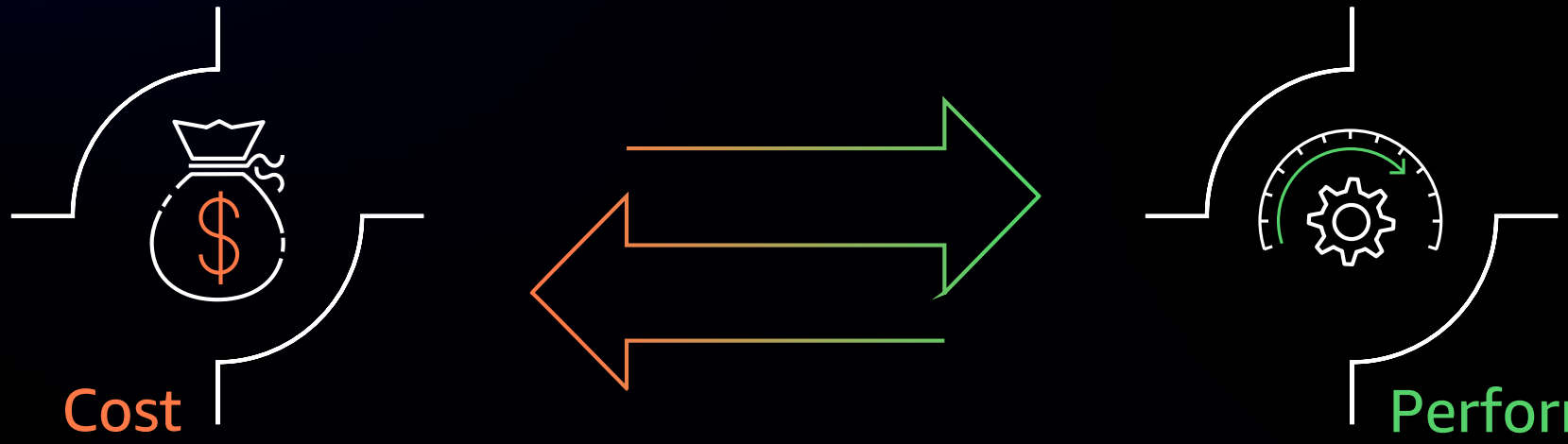
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Now, how can you optimize price, capacity, and performance?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Cost vs. performance



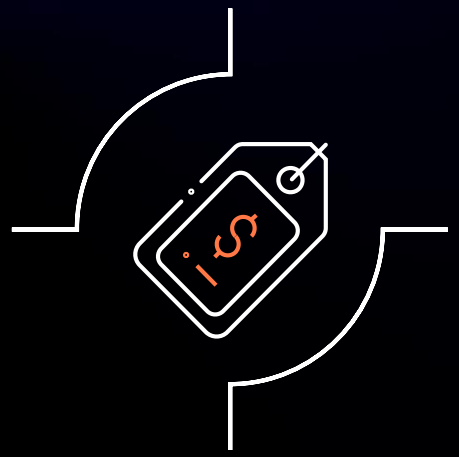

Is cost optimization always a trade-off for performance and availability?

No – it is a matter of understanding your pricing options, using the right tools for managing capacity, and taking advantage of guidance and best practices

aws

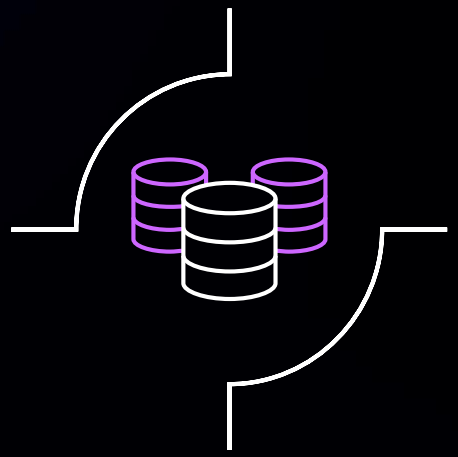
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Optimizing Amazon EC2 cost and capacity



## Pricing

Achieve optimal price and performance with different purchase models



## Capacity

Capacity management made easy on the broadest and deepest compute platform

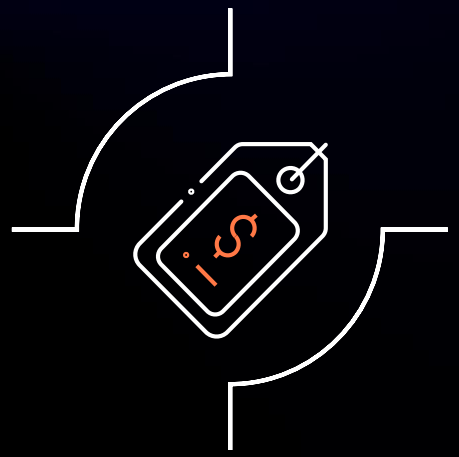


## Performance

Enable AWS resource recommendations to reduce cost and improve performance

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Optimizing Amazon EC2 cost and capacity



## Pricing

Achieve optimal price and performance with different purchase models



## Capacity

Capacity management made easy on the broadest and deepest compute platform



## Performance

Enable AWS resource recommendations to reduce cost and improve performance

aws

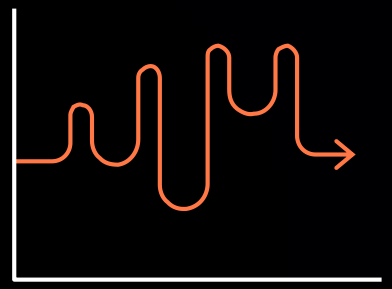
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 purchase options

Think flexible, and focus on these for optimizing performance and cost

## On-Demand Instances

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



Spiky workloads to define needs

Launches capacity

## Savings Plans

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

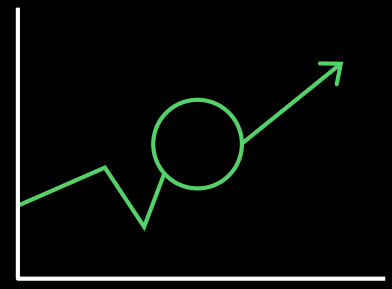


Flexible access to compute

Applies savings to On-Demand Instances

## Spot Instances

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



Fault-tolerant, flexible, stateless workloads

Launches capacity

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Types of Savings Plans

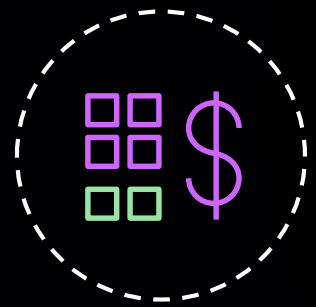


## Compute Savings Plans

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 US East (Ohio) to Europe (London)
- ✓ OS – e.g., Windows to Linux
- ✓ Tenancy – e.g., switch Dedicated tenancy to Default tenancy
- ✓ Compute options – e.g., move from Amazon EC2 to AWS Fargate



## EC2 Instance Savings Plans

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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# On-Demand Capacity Reservations

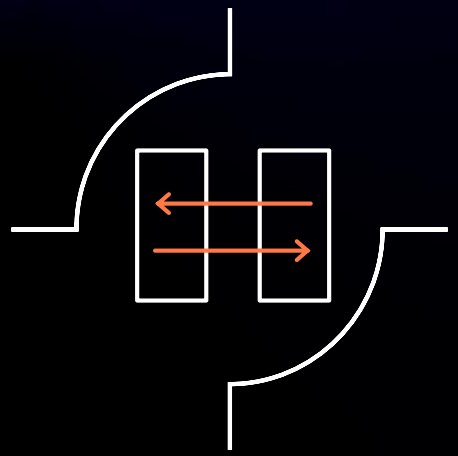
FOR STEADY-STATE WORKLOADS



- Manage capacity and discounts independently
- No commitment required; can be created and canceled as needed
- Reserve capacity by Availability Zone
- Capacity held whether or not you run instances
- Share reservations across accounts

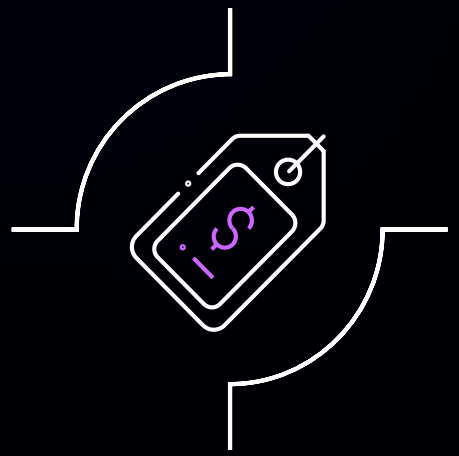
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deep savings using Amazon EC2 Spot Instances



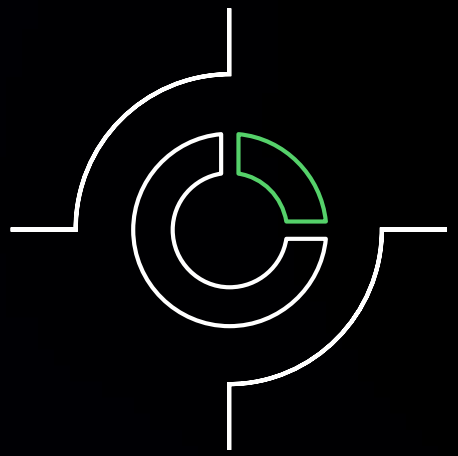
## EC2 Spot Instances

Spare capacity from the same infrastructure



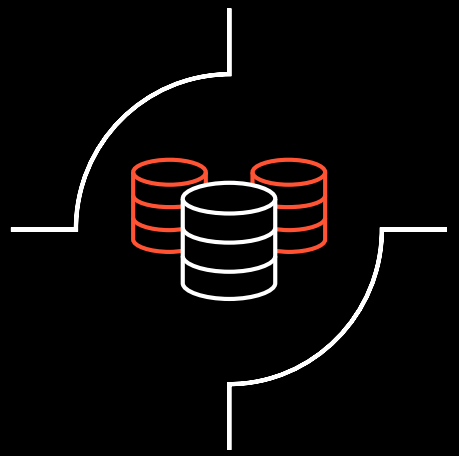
## Pricing

Smooth, infrequent changes; more predictable



## Usage

Choose different instance types, sizes, and Availability Zones in an Amazon EC2 Auto Scaling group



## Capacity

AWS can reclaim with 2-minute notice; interruptions happen when EC2 needs the spare capacity back

EC2 Spot Instances are integrated in many AWS services, AWS Partner offerings, and open-source software

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# 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 Elastic Kubernetes Service (Amazon EKS)

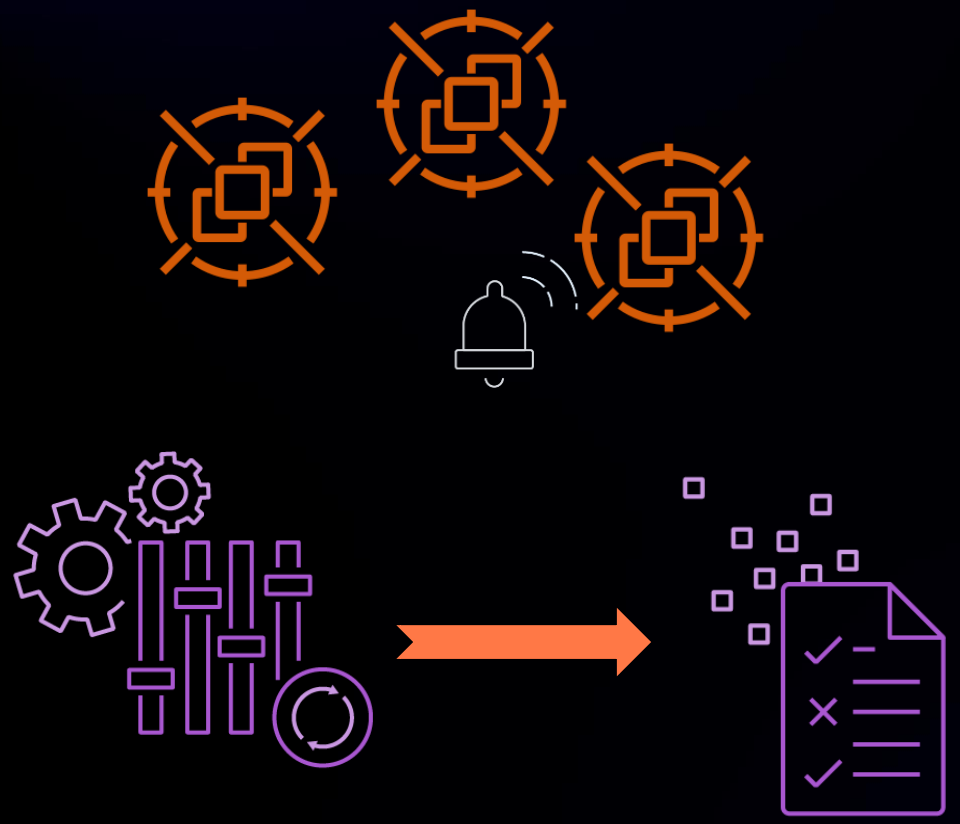
Amazon Elastic Container Service (Amazon ECS)

- ✓ 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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Rebalance recommendation for Spot Instances



- Rebalance recommendation signal
  - Notifies you when a Spot Instance is at elevated risk of interruption
  - Can arrive sooner than the 2-minute Spot Instance interruption notice
  - Helps to proactively rebalance your workload to new or existing Spot Instances that are not at elevated risk of interruption
- What do you do?
  - Start checkpointing your work early to save as much state as possible
  - Prevent scheduling new work on instances at elevated risk of interruption, thus increasing chance of completing work

Learn more: https://amzn.to/36pHhWM

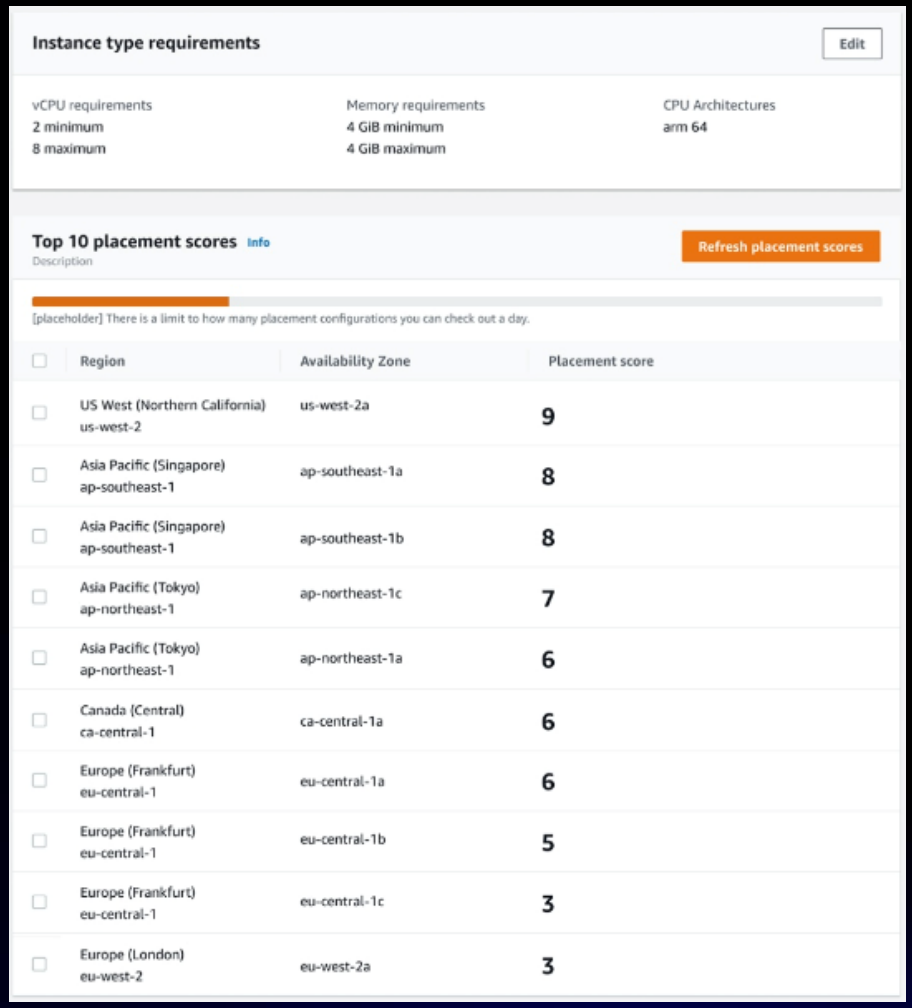
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Spot placement score

NEW

- Helps Amazon EC2 Spot customers select optimal locations
- Input the desired instance types and target capacity
- Outputs a scored list of Regions and/or Availability Zones



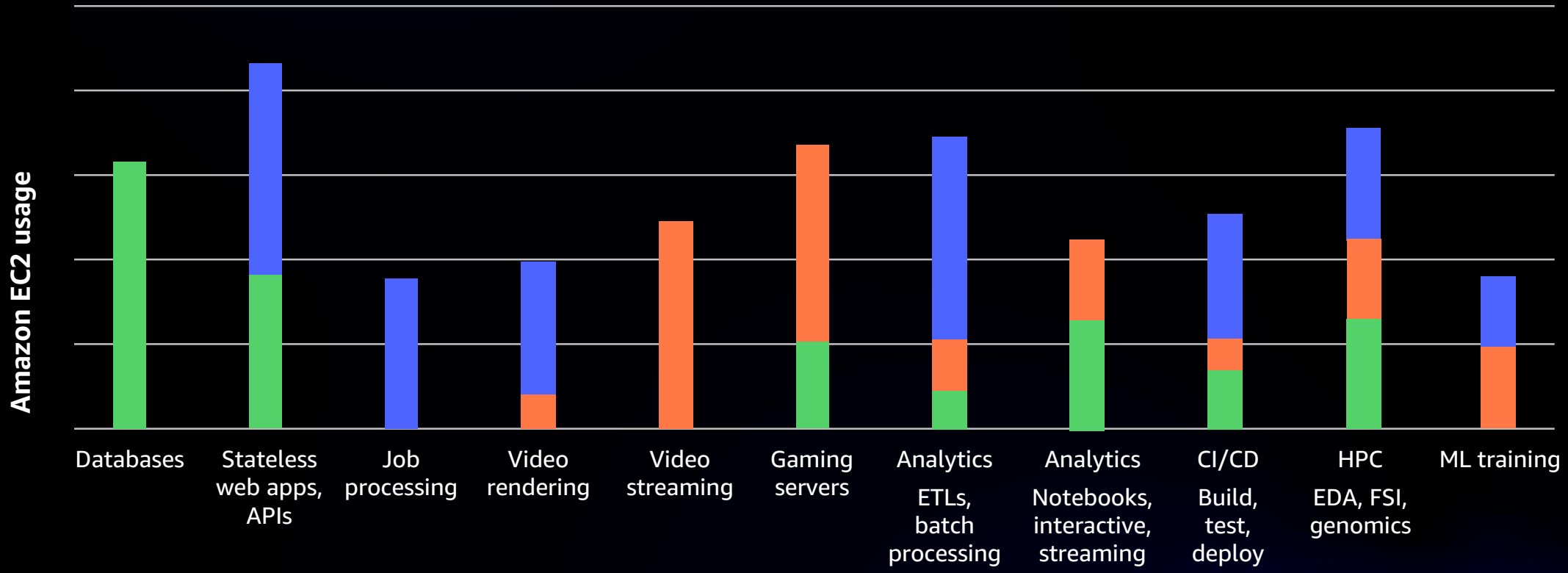

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Workload types and purchase options

Use **Savings Plans** for known/steady-state workloads

Scale using **On-Demand** for new or stateful spiky workloads

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




Example purchase option distributions

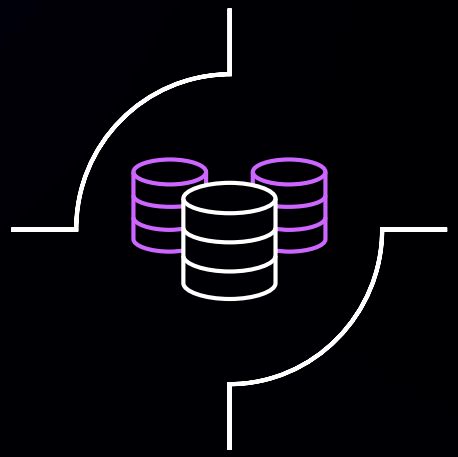
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Optimizing Amazon EC2 cost and capacity



## Pricing

Achieve optimal price and performance with different purchase models



## Capacity

Capacity management made easy on the broadest and deepest compute platform



## Performance

Enable AWS resource recommendations to reduce cost and improve performance

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Using Amazon EC2 Auto Scaling

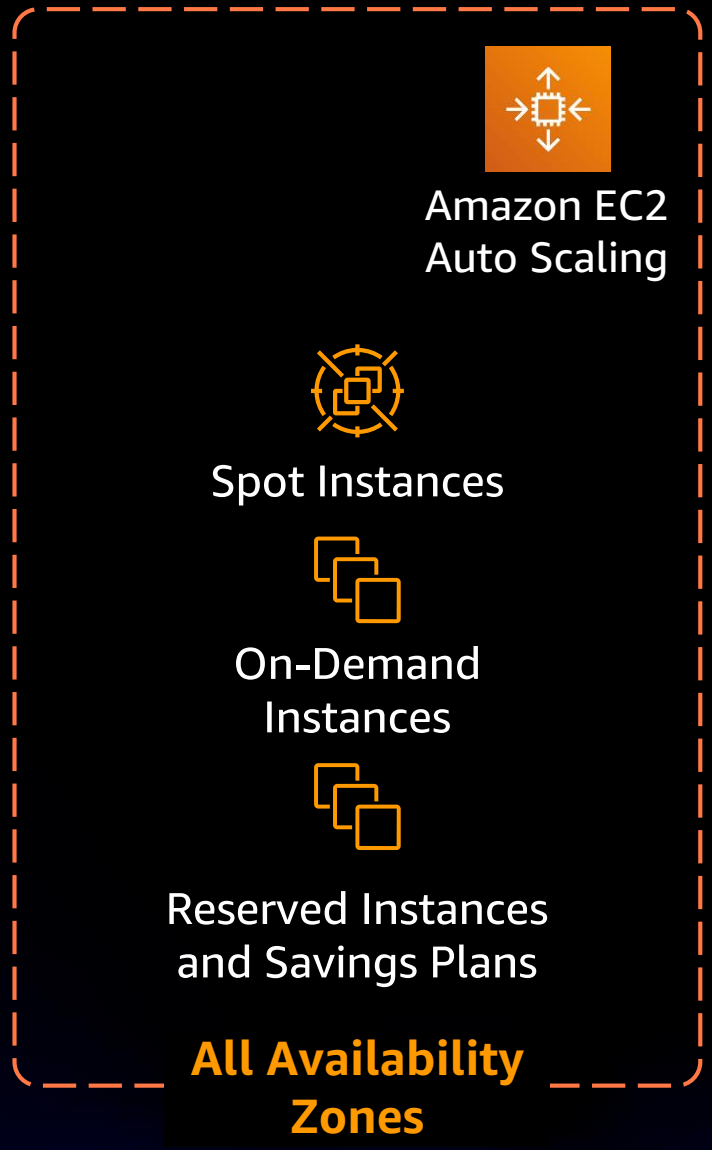
AUTOMATICALLY SCALE INSTANCE TYPES AND PURCHASE OPTIONS IN A SINGLE AUTO SCALING GROUP TO OPTIMIZE COST

**Reduce cost**
Select a mix of On-Demand and Spot Instances to reduce overall cost

**Optimize capacity**
Prioritize deploying Spot Instances into greater Spot Instance pool capacity to lower the chance of interruptions

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

**Maintain application availability**
Ensures that your application always has the right amount of compute and proactively provisions capacity with predictive scaling




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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Let EC2 Auto Scaling do the heavy lifting

JUST TELL AMAZON EC2 AUTO SCALING WHAT YOU NEED AND HOW FLEXIBLE YOU CAN BE

```
AutoScalingGroupName: myAutoScalingGroup
CapacityRebalance: true
DesiredCapacity: 12
MaxSize: 15
MinSize: 12
MixedInstancesPolicy:
  InstancesDistribution:
    OnDemandBaseCapacity: 3
    OnDemandPercentageAboveBaseCapacity: 25
    SpotAllocationStrategy: capacity-optimized
  LaunchTemplate:
    LaunchTemplateSpecification:
      LaunchTemplateName: myLaunchTemplate
      Version: $Default
    Overrides:
    - InstanceType: c5.large
    - InstanceType: c5a.large
    - InstanceType: m5.large
    - InstanceType: m5a.large
    - InstanceType: c4.large
    - InstanceType: m4.large
    - InstanceType: c3.large
    - InstanceType: m3.large
VPCZoneIdentifier: mySubnet1,mySubnet2,mySubnet3
```

5. Allow Amazon EC2 Auto Scaling to proactively rebalance Spot Instances with elevated risk of interruption to increase application availability

1. Allow Amazon EC2 Auto Scaling to provision a combination of On-Demand Instances and Spot Instances across multiple instance types

3. Set a base capacity for On-Demand Instances covered by a Savings Plan

4. Use the capacity-optimized allocation strategy to find the optimal spare capacity to reduce cost of interruptions

2. Provide a flexible list of instance types and Availability Zones to allow Amazon EC2 Auto Scaling to optimize for performance, cost, and availability

Learn more: https://amzn.to/3lihsiF

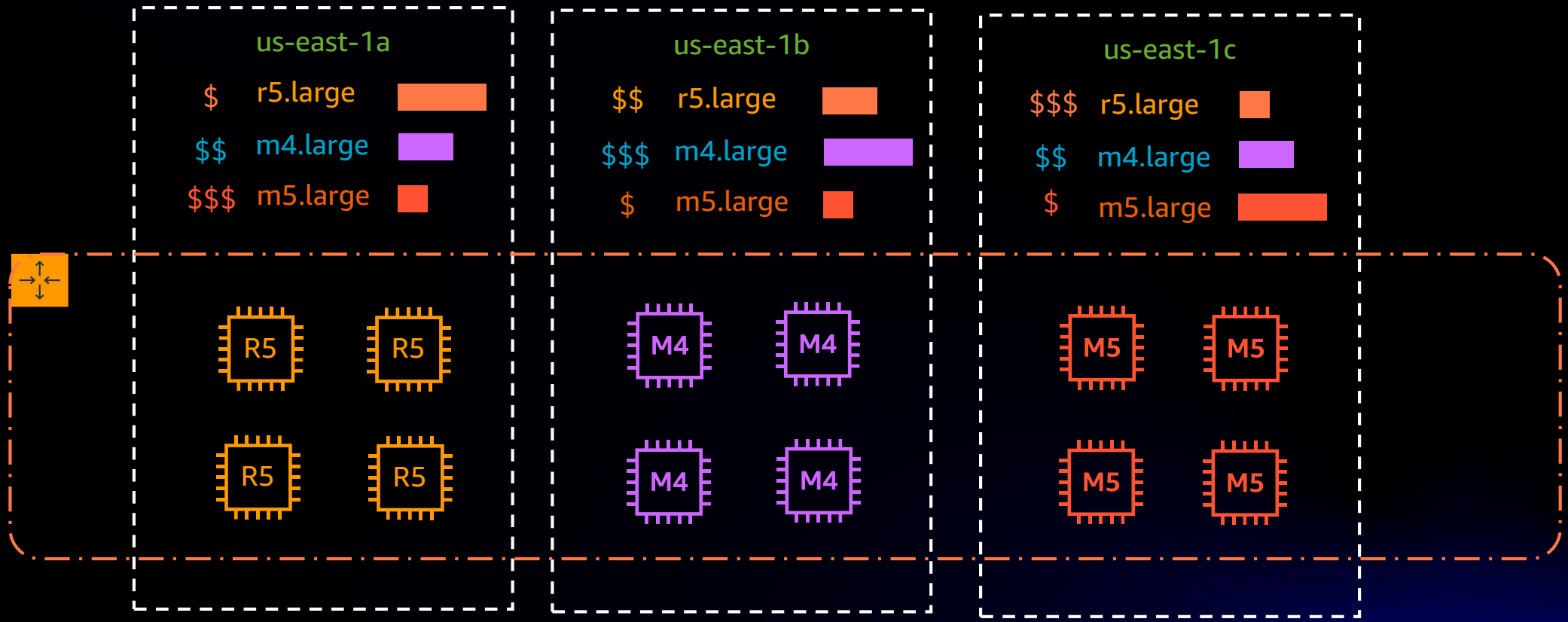
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Auto Scaling group capacity-optimized allocation strategy

Desired capacity: 12  OnDemandBaseCapacity: 0  OnDemandPercentageAboveCapacity: 0

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




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon EC2 Spot integrations


Amazon EC2
Auto Scaling


Amazon ECS


Amazon EKS


Fargate


AWS
Elastic Beanstalk


AWS
Batch


Amazon
SageMaker


Amazon
EMR


AWS
CloudFormation


AWS Thinkbox
Deadline


Jenkins







cloudera

Kubernetes










MESOS

Docker

Jenkins logo by Frontside / CC BY-SA 3.0

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Optimizing Amazon EC2 cost and capacity



## Pricing

Achieve optimal price and performance with different purchase models



## Capacity

Capacity management made easy on the broadest and deepest compute platform



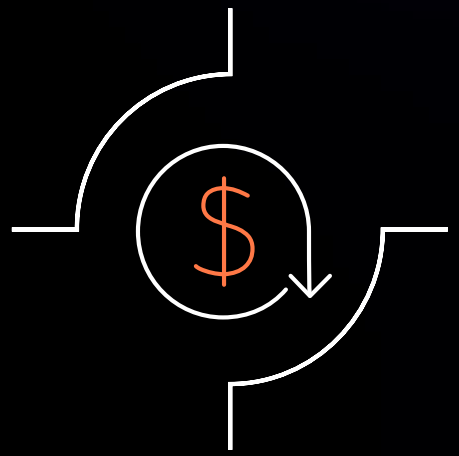
## Performance

Enable AWS resource recommendations to reduce cost and improve performance

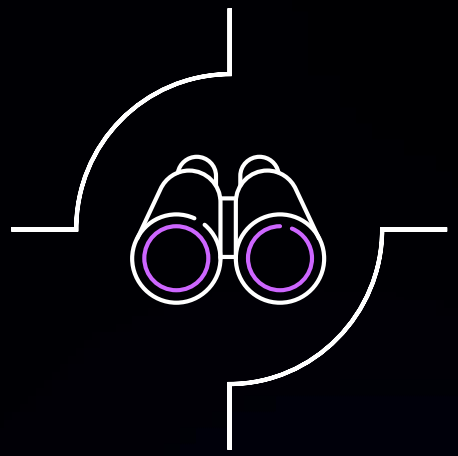
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Compute Optimizer

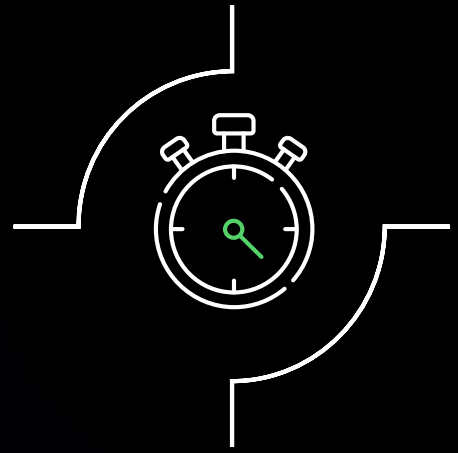
Recommends optimal instances for AWS Lambda, Amazon EC2, and Amazon EC2 Auto Scaling groups from 145+ instances from M, C, R, T, and X families



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

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Customers optimizing




"The number of posts the platform has to handle can vary from 5 to 50 per second. With Amazon EC2 Auto Scaling, we can scale our infrastructure by 10 times without any problem . . . As a result, we don't pay for resources that aren't used – and neither do our customers."

**Piotr Surma**
Associate Director, Managing Director

---




"Running our microbiome research on Amazon EC2 Spot Instances, we spend less money and less time to get scientific answers from the analysis – by accelerating our research processes on AWS, we can get closer to developing therapeutics to fight cancer."

**Sam Minot**
PhD and Staff Scientist

---



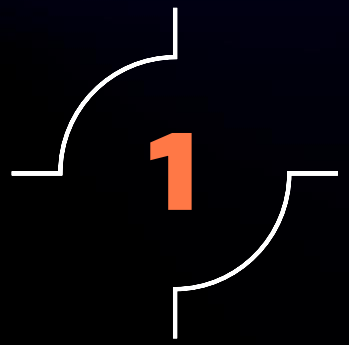

"With AWS Compute Optimizer, we will have a single pane of glass to identify optimal EC2 instance types for our workloads across our AWS environment. The ability to visualize projected resource utilization on various instance types will allow us to make more informed decisions before we update our fleet."

**Renan Capaverde**
Director of Engineering

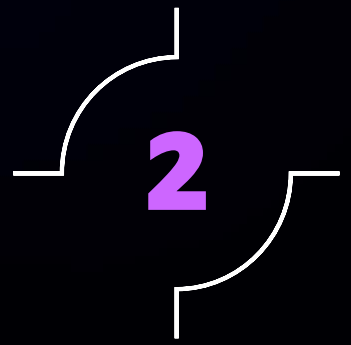
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Cost and capacity takeaways



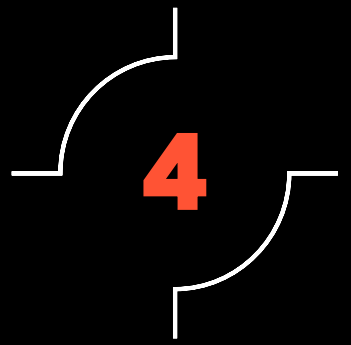

Familiarize yourself with instance type categories, but don’t stress over it




Take an attribute-based approach to instance type selection




Use instance type selection tools, such as Amazon EC2 Instance Discovery




Optimize price, capacity, and performance using purchasing options such as Savings Plans and Spot Instances, and tools such as Amazon EC2 Auto Scaling and AWS Compute Optimizer

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.