
AWS
re:Invent

CMP203

Optimize compute for performance and cost

Chad Schmutzer

Principal Developer Advocate, Amazon EC2

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Compute for virtually every workload

Pricing optimization

Capacity optimization

Tools for optimizing performance and cost

Guidance

Conclusion

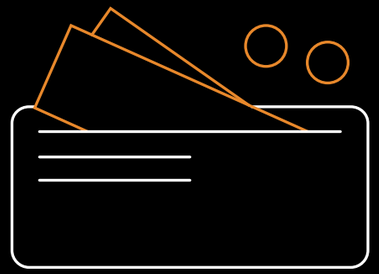


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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)

Amazon Elastic Graphics

Amazon Elastic Inference

=

350+

instance types

for virtually
every workload
and business need

Cost vs. performance



Cost

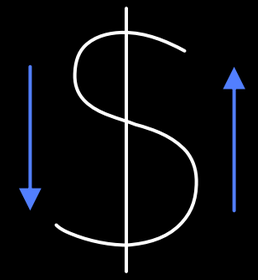
Performance

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

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

Optimizing Amazon EC2 cost and capacity

Pricing



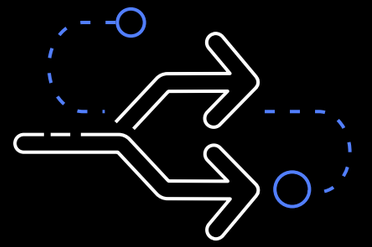
Achieve optimal price and performance with different purchase models

Capacity



Capacity management made easy on the broadest and deepest compute platform

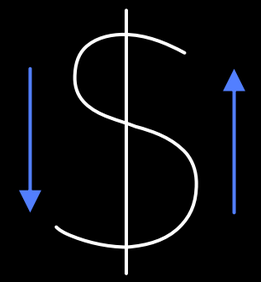
Guidance



Cost and capacity recommendations enable ease of use and save time

Optimizing Amazon EC2 cost and capacity

Pricing



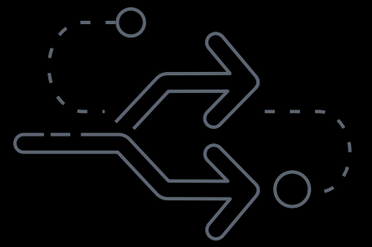
Achieve optimal price and performance with different purchase models

Capacity



Capacity management made easy on the broadest and deepest compute platform

Guidance



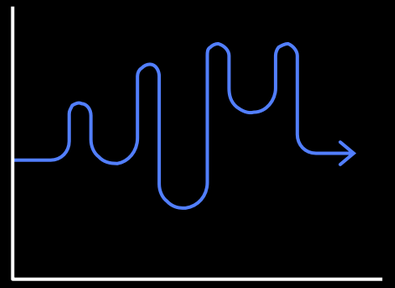
Cost and capacity recommendations enable ease of use and save time

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

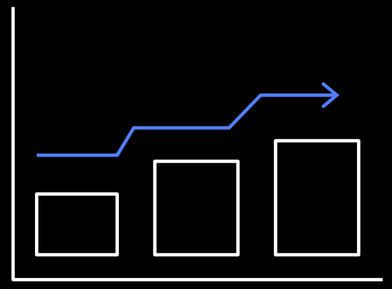


Spiky workloads to define needs

Launches capacity

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**

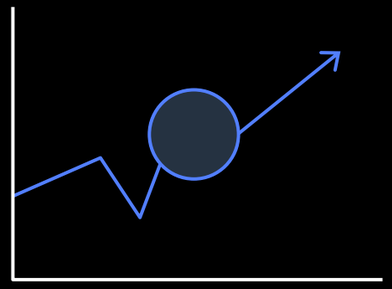


Flexible access to compute

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

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 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

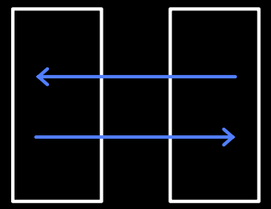
Provide the lowest prices, up to 72% off (same as Standard RIs) on the selected instance family (e.g., C5 or M5), in a specific AWS Region

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

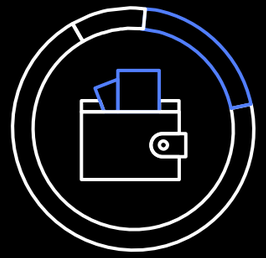
Learn more: https://amzn.to/3kftPtI

Deep savings using EC2 Spot Instances



EC2 Spot Instances

Spare capacity from the same infrastructure



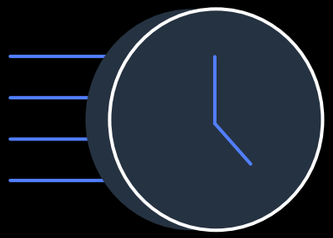
Pricing

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

https://github.com/aws/ec2-spot-instances-integrations-roadmap/

Optimizing Amazon EC2 cost and capacity

Pricing



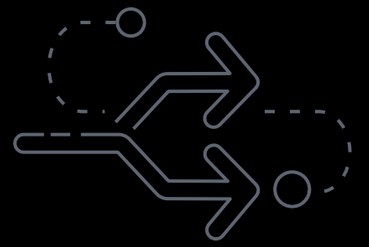
Achieve optimal price and performance with different purchase models

Capacity



Capacity management made easy on the broadest and deepest compute platform

Guidance



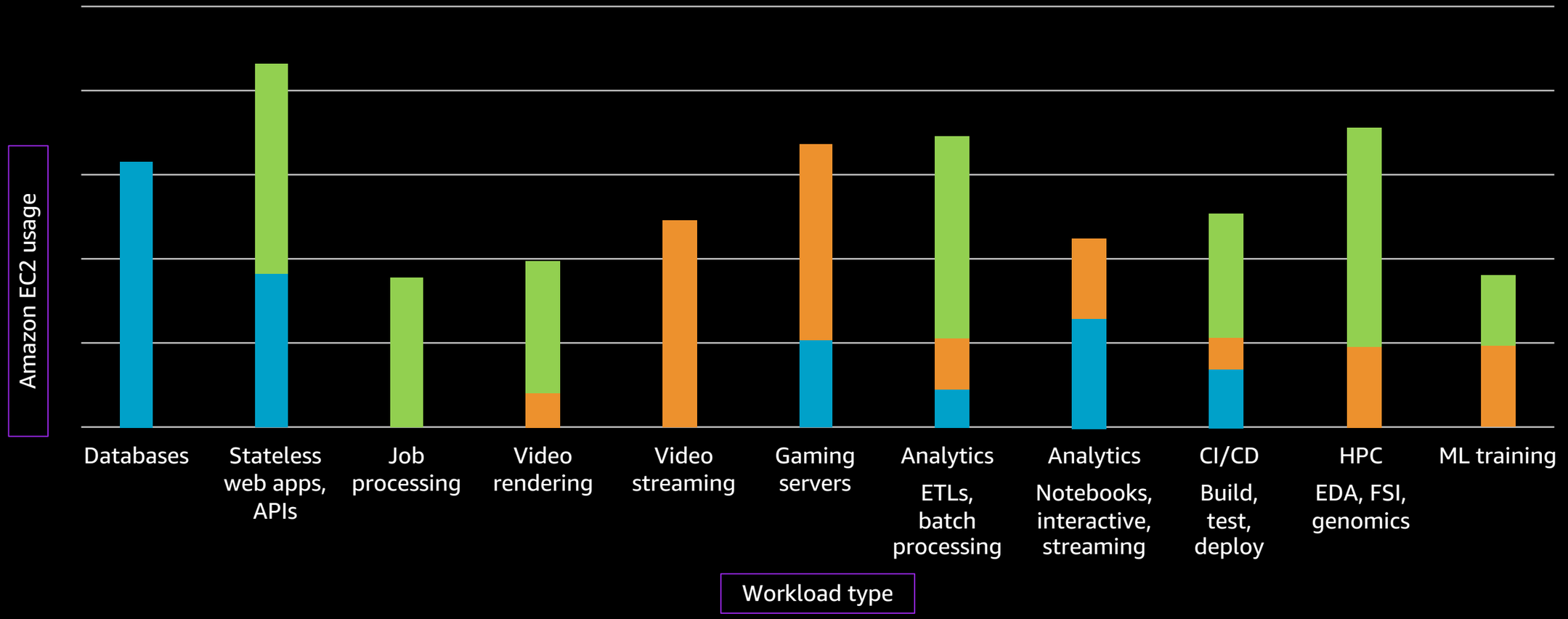
Cost and capacity recommendations enable ease of use and save time

Workload types and purchase options

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

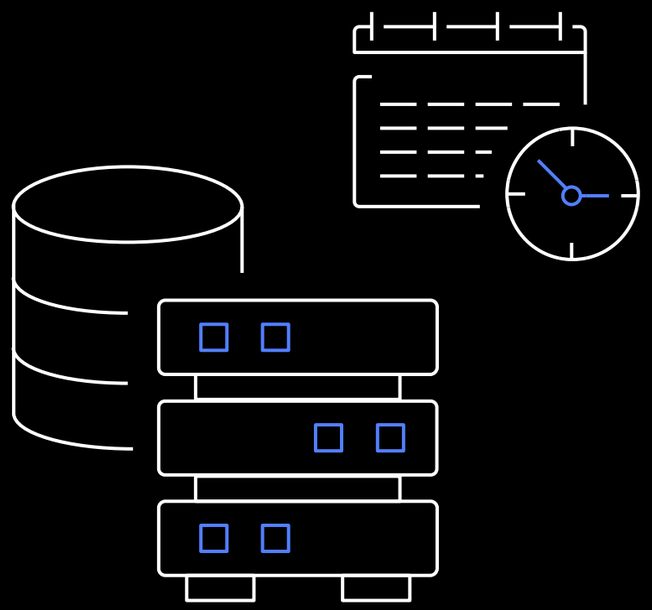
Scale using **On-Demand** for new or stateful spiky workloads

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



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 you run instances or not
- Share reservations across accounts
- Use AWS Resource Groups to organize Capacity Reservations into logical groups for different workloads, environments, teams, or special events
- Highly suggest combining Capacity Reservations with Savings Plans

Learn more: https://amzn.to/3eV7Bw5

Using Amazon EC2 Auto Scaling

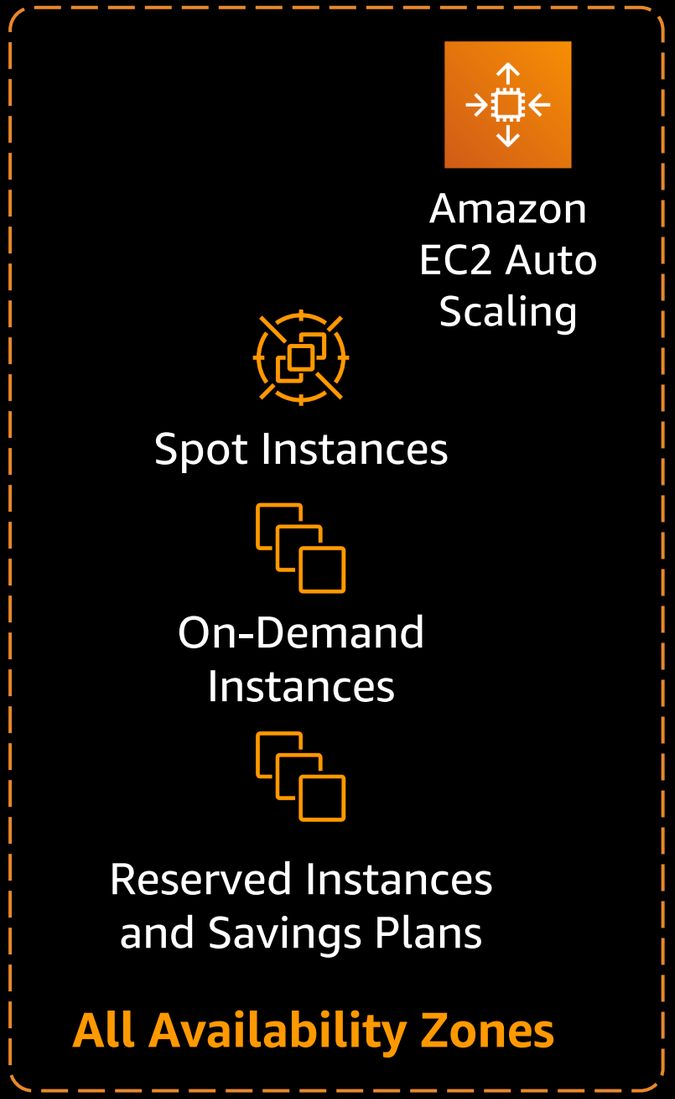
AUTOMATICALLY SCALE INSTANCE TYPES AND PURCHASE OPTIONS IN A SINGLE AUTO SCALING GROUP TO OPTIMIZE COST

Capacity-optimized
Prioritize deploying Spot Instances into greater Spot Instance pool capacity in 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

Support for On-Demand Capacity Reservations
Target EC2 Capacity Reservations within an AWS Resource Group using the Resource Group ARN



Self-guided workshop: https://bit.ly/3eH8lET

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

Let EC2 Auto Scaling do the heavy lifting

JUST TELL 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 EC2 Auto Scaling to proactively rebalance Spot Instances with elevated risk of interruption to increase application availability

1. Allow 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 EC2 Auto Scaling to optimize for performance, cost, and availability

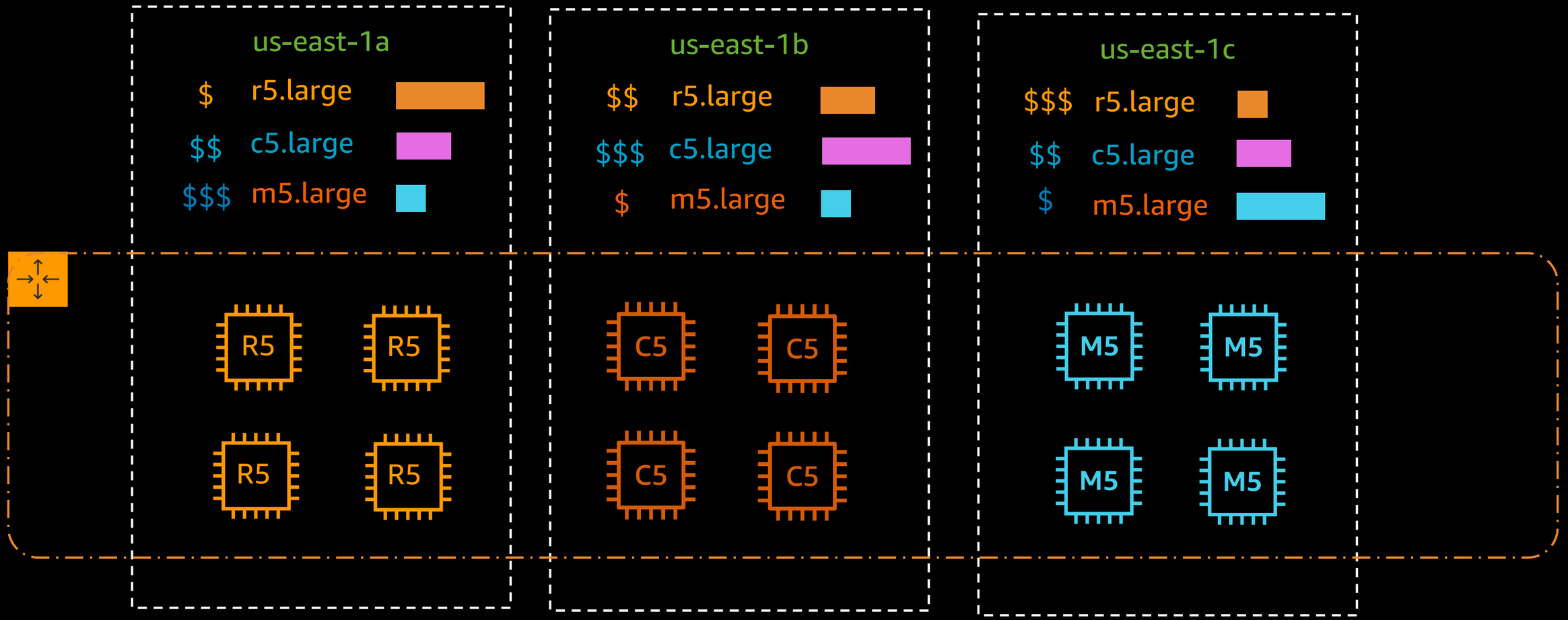
Learn more: https://amzn.to/3lhBPvl

Capacity-optimized allocation strategy

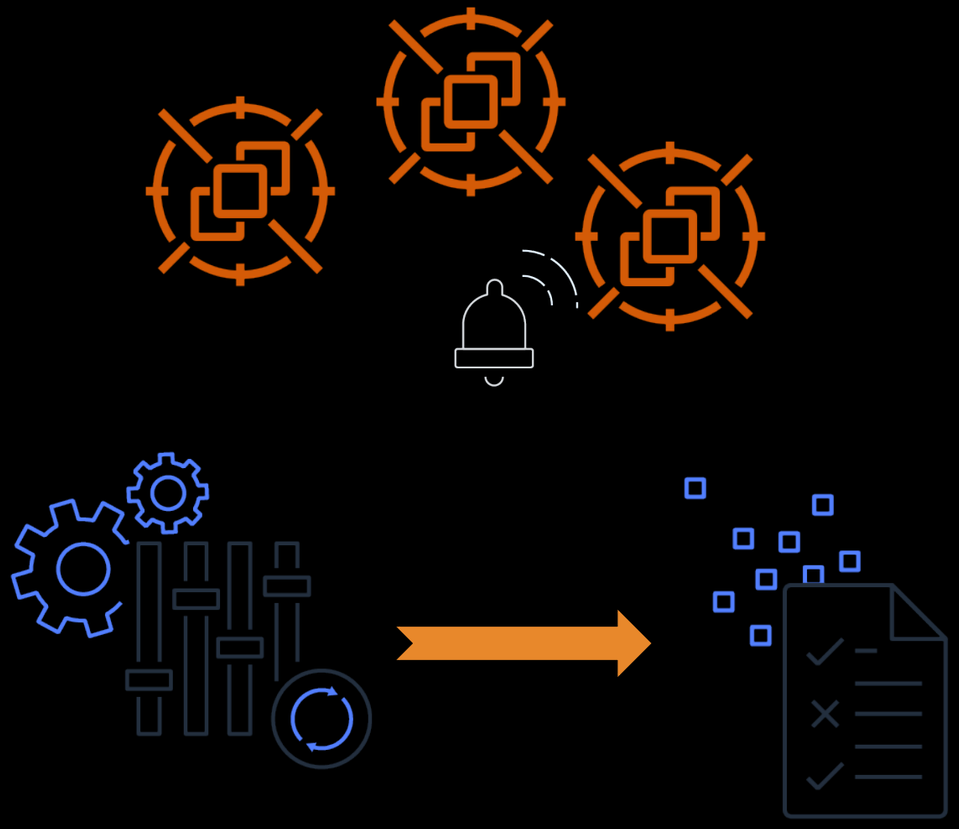
FOR EC2 AUTO SCALING SPOT-BACKED WORKLOADS

`DesiredCapacity: 12` `OnDemandBaseCapacity: 3` `OnDemandPercentageAboveCapacity: 25`

`Overrides: ["c5.large", "m5.large", "r5.large"]` `SpotAllocationStrategy: capacity-optimized`



Introducing the EC2 instance rebalance recommendation for Spot Instances

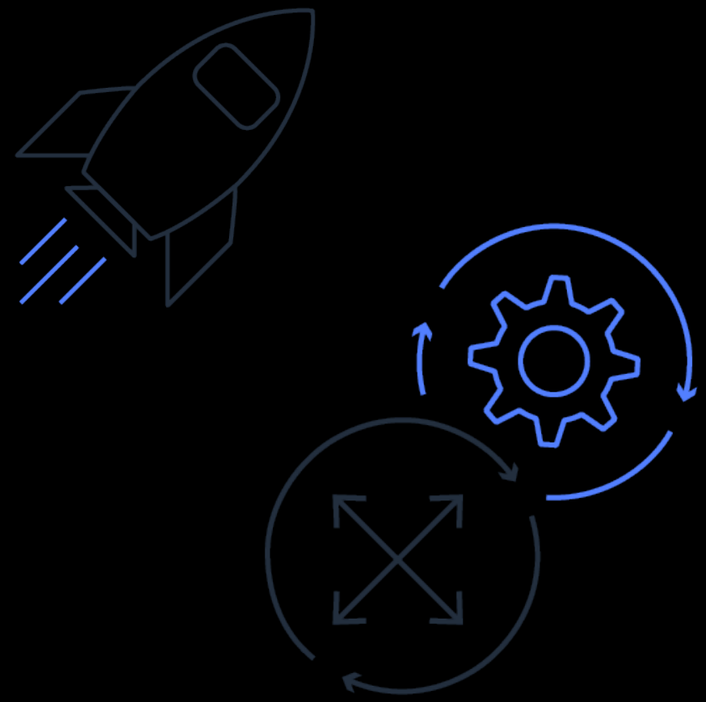


- A new signal that notifies you when a Spot Instance is at elevated risk of interruption
- The signal can arrive sooner than the 2-minute Spot Instance interruption notice
- Gives you the opportunity to proactively rebalance your workload to new or existing Spot Instances that are not at elevated risk of interruption
- 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

Introducing Capacity Rebalancing for EC2 Auto Scaling

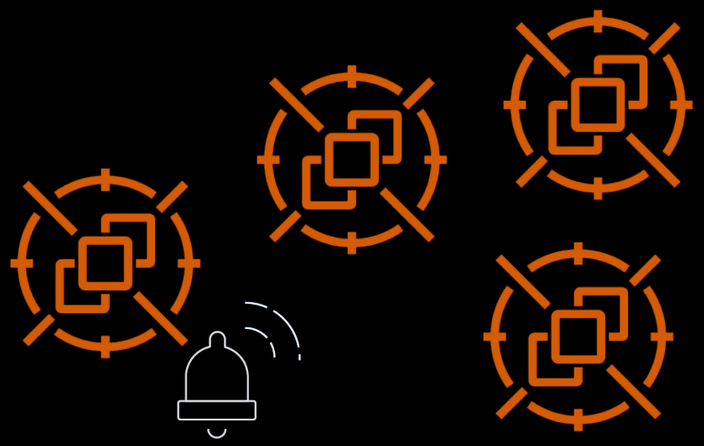
- A new feature for proactively managing the Amazon EC2 Spot Instance lifecycle in an Auto Scaling group
- Increases the emphasis on availability
- Provides a seamless and automated experience for maintaining desired capacity
- Works best when combined with mixed instances policy and capacity-optimized allocation strategy

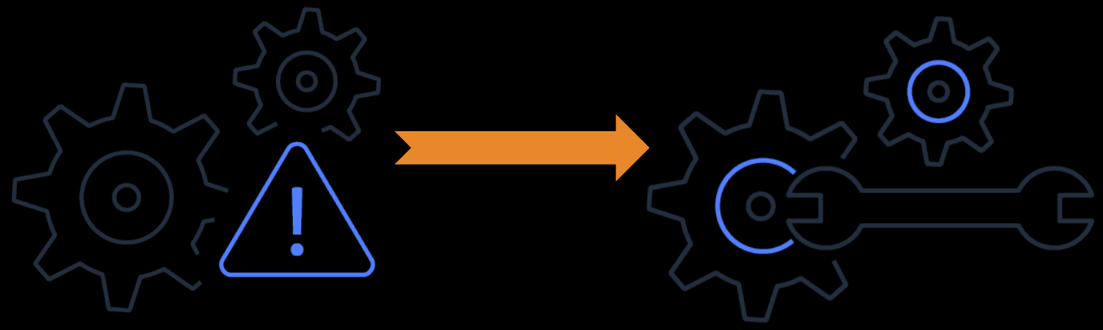


Learn more: https://amzn.to/3eCheQ2

How does Capacity Rebalancing work?

- EC2 Auto Scaling monitors for rebalance recommendations
- Proactively launches replacement capacity for existing Spot Instances when they are at elevated risk of interruption
- Detaches from Elastic Load Balancing if necessary and runs lifecycle hooks as configured
- Waits until the new instance passes its health check before proceeding with terminating the old instance



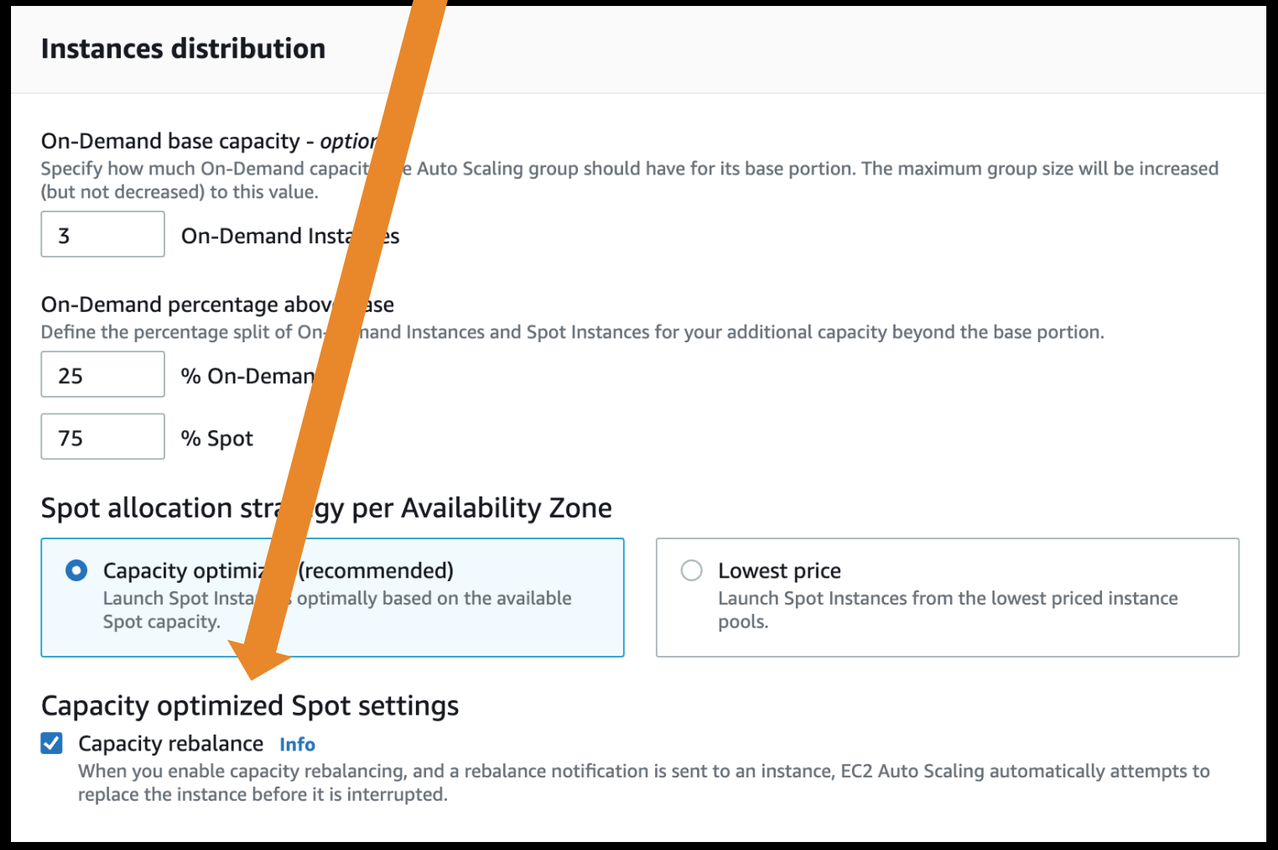


Enabling Capacity Rebalancing

Also available in AWS CloudFormation

```
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
```

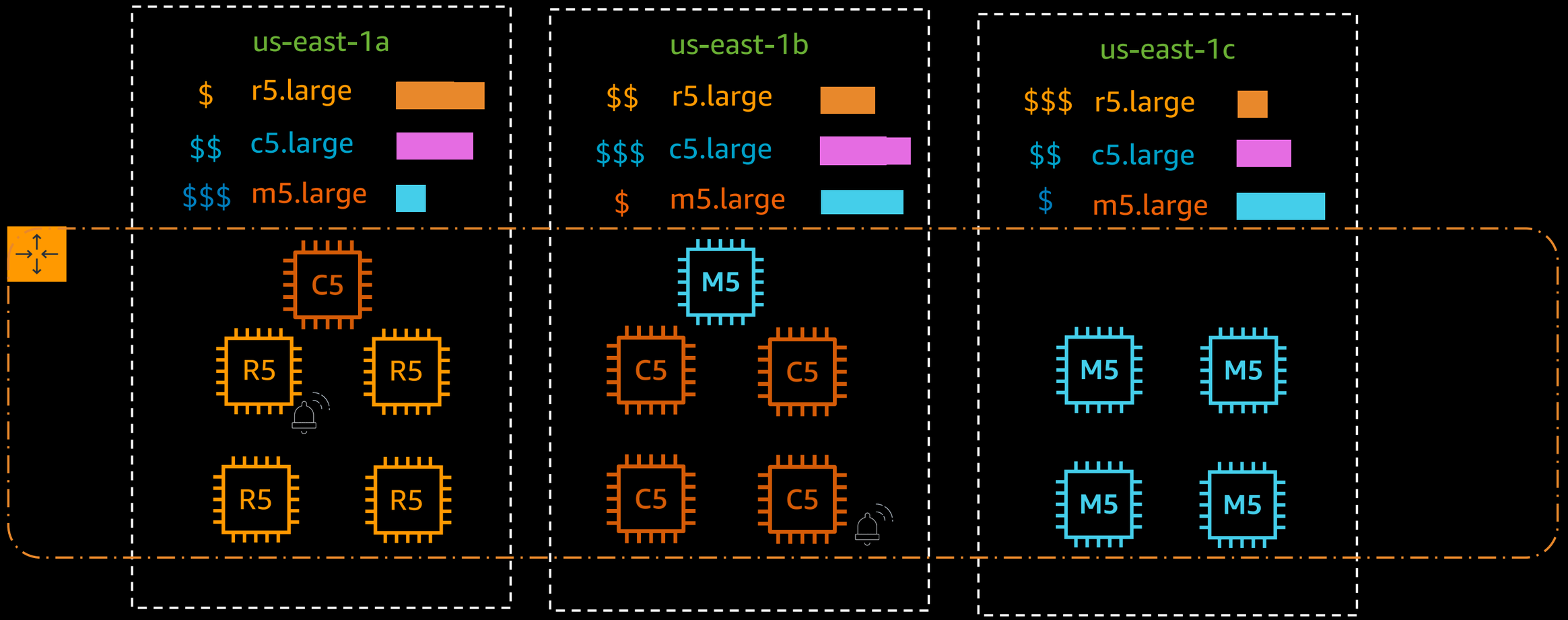
Only a single config



Capacity Rebalancing in action

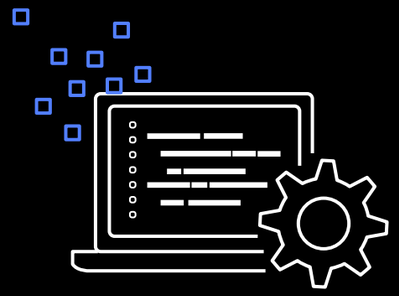
Desired capacity: 12 CapacityRebalance: true

Overrides: [“c5.large”, “m5.large”, “r5.large”] SpotAllocationStrategy: capacity-optimized

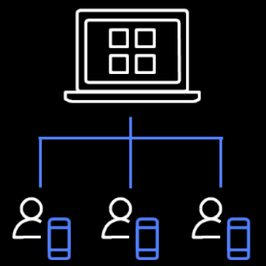


Amazon Elastic Container Service (Amazon ECS)

BUILT-IN SUPPORT FOR EC2 AUTO SCALING AND SPOT INSTANCES



Automated draining for Spot Instances running Amazon ECS services

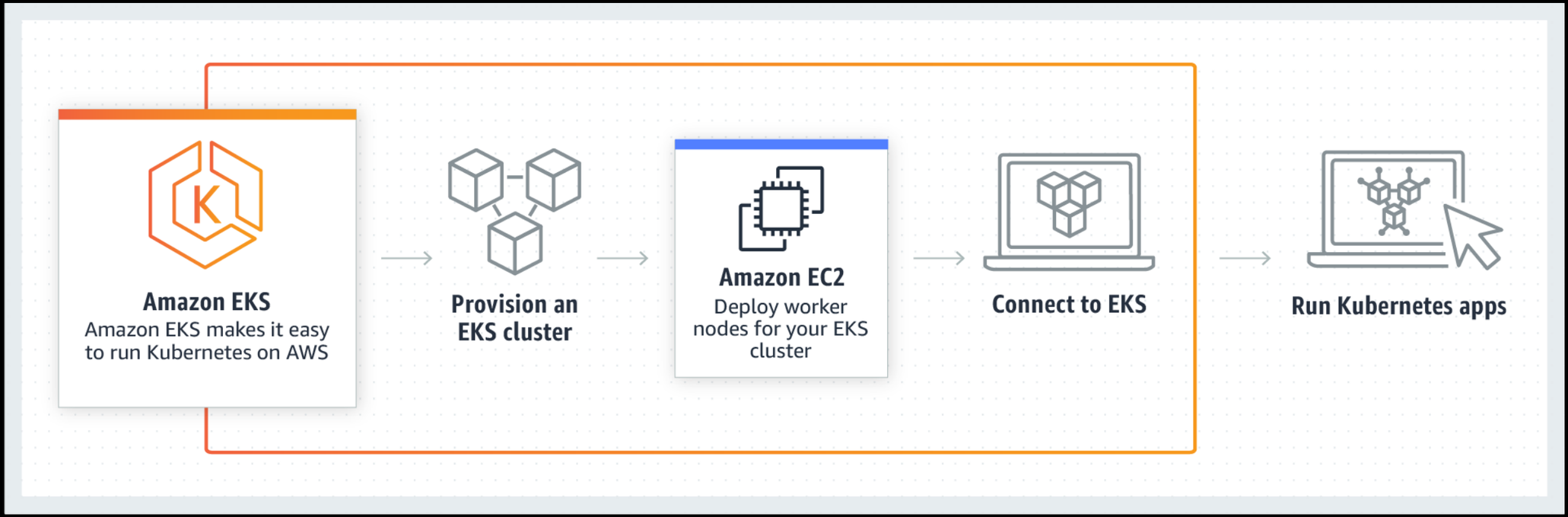


Fully manages Spot clusters

Learn more: https://amzn.to/2Dzo9HV

Amazon EKS

EASILY UTILIZE SPOT INSTANCES WITH K8S VIA EC2 AUTO SCALING



AWS Node Termination Handler supports automated draining for Spot Instance nodes on Kubernetes

https://github.com/aws/aws-node-termination-handler

Self-guided workshop: https://bit.ly/3kcf7DL

Amazon EMR

BUILT-IN SUPPORT FOR SPOT INSTANCES AND CAPACITY-OPTIMIZED ALLOCATION STRATEGY



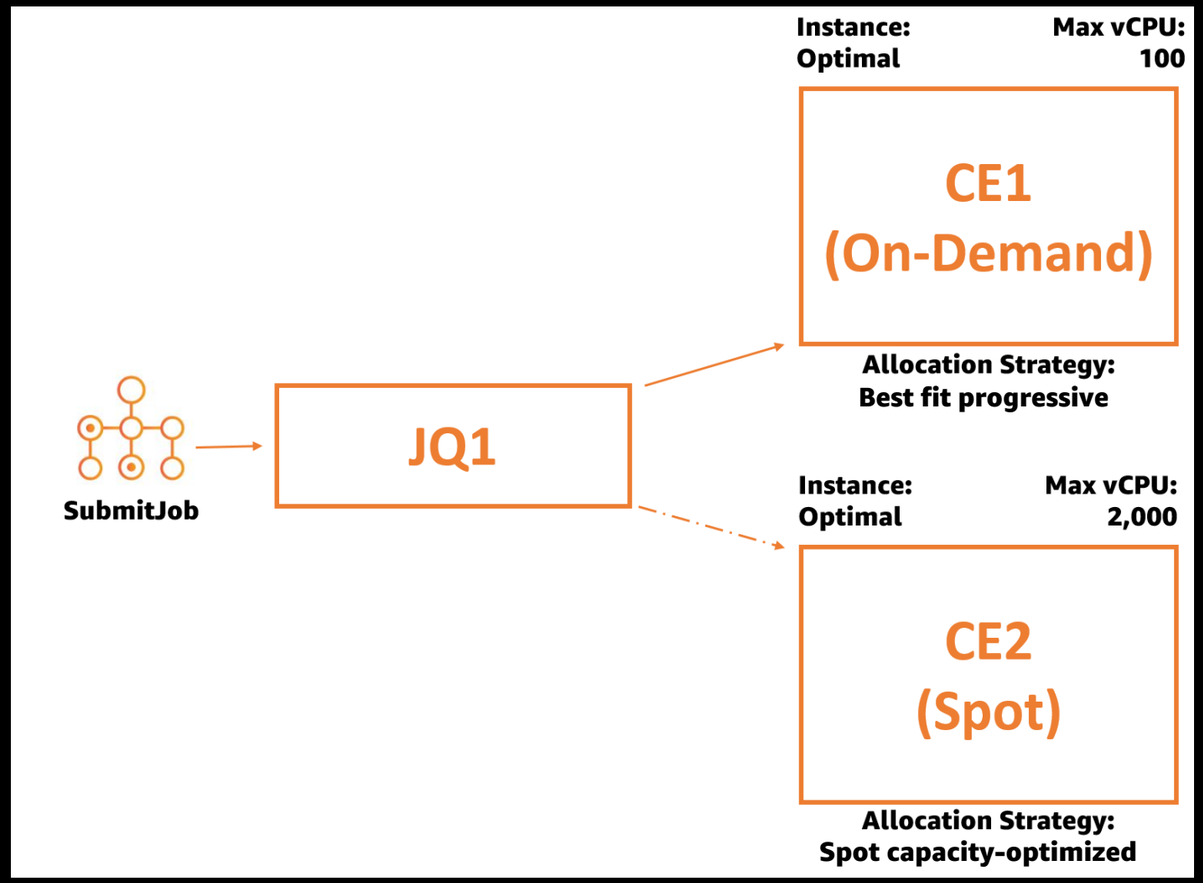
- Amazon EMR now offers allocation strategies for provisioning EC2 instances in an Amazon EMR cluster
- On-Demand Instances use a lowest-price strategy, which launches the lowest-priced instances first
- Spot Instances use a capacity-optimized strategy, which launches Amazon EC2 Spot Instances from Spot Instance pools that have optimal capacity for the number of instances that are launching
- Up to 15 instance types per task instance fleet when using allocation strategy option

Self-guided workshop: https://bit.ly/2InseEC

AWS Batch

BUILT-IN SUPPORT FOR EC2 AUTO SCALING AND SPOT INSTANCES

- Fully managed
- Cost-optimized resource provisioning
- Allocation strategies
  - Spot capacity-optimized
  - Best fit progressive



Learn more: https://amzn.to/33Bn7ph

Self-guided workshop: https://bit.ly/2IqRMRh

Optimizing Amazon EC2 cost and capacity

Pricing



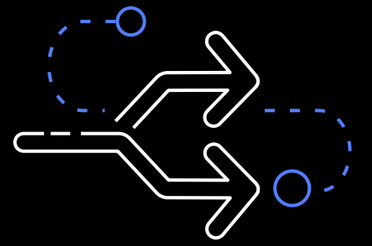
Achieve optimal price and performance with different purchase models

Capacity



Capacity management made easy on the broadest and deepest compute platform

Guidance



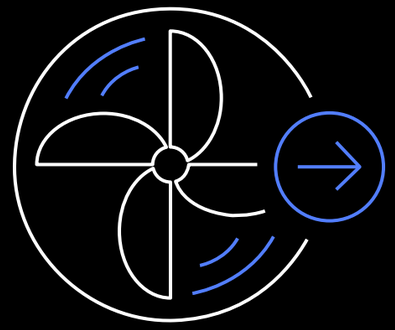
Cost and capacity recommendations enable ease of use and save time

AWS Compute Optimizer

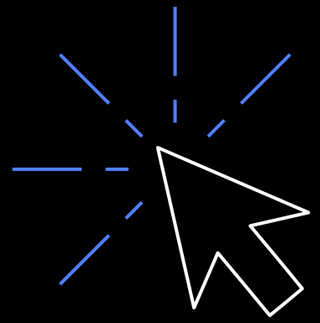
Recommends optimal instances for Amazon EC2 and Amazon EC2 Auto Scaling groups from 145+ instances from M, C, R, T, and X families



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



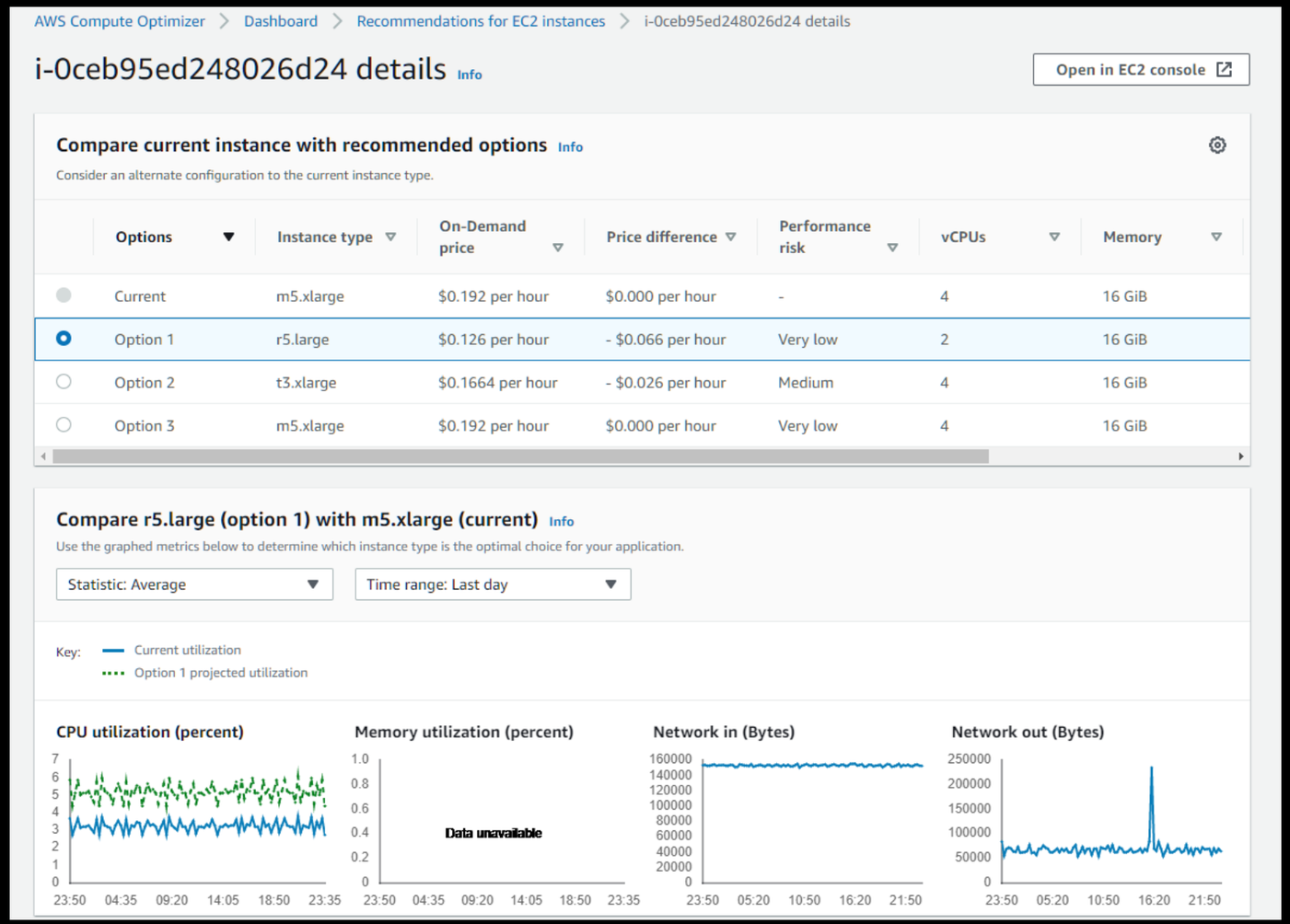
Applies insights from millions of workloads to make recommendations



Saves time comparing and selecting optimal resources for your workload

Learn more: https://amzn.to/3ki5xyW

Easy to choose with AWS Compute Optimizer



- Analyzes Amazon CloudWatch metrics and considers Auto Scaling group configuration for intuitive and actionable recommendations
- Recommends optimal EC2 instance types
- Optimizes performance and reduces costs by making recommendations to help right-size your workloads
- Up to 3 recommendations per workload
- Now analyzes additional Amazon EBS metrics
- Available at no additional charge

Key takeaways

1

Amazon EC2 is designed to provide the right compute for every workload

Workload-optimized EC2 instances

2

Access compute at a lower cost to innovate faster

Spot Instances and Savings Plans

3

How to automate cost and capacity optimization

Amazon EC2 Auto Scaling

4

Optimize your workloads by using best practices

AWS Compute Optimizer

Thank you!

Chad Schmutzer

@schmutze
schmutze@amazon.com

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.