AWS re:Invent

CMP202

Better, faster, cheaper—cost optimizing EC2

Jeanine Banks

General Manager
AWS Compute Services
Amazon Web Services

Alex Estrovitz

Director Platform Engineering
Salesforce

Key takeaways from this session…

1

Experiment and test at a lower cost to innovate faster

2

How to automate cost and capacity optimization

3

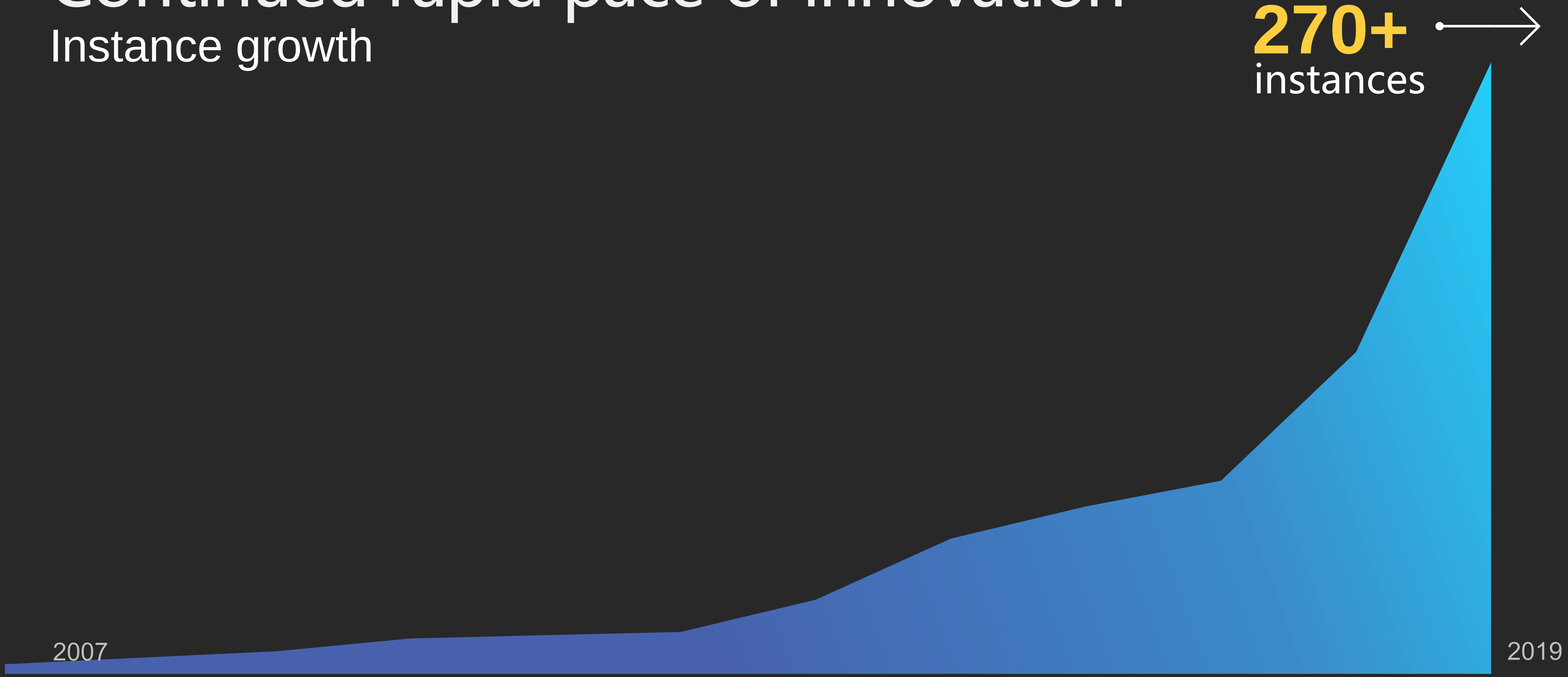
Optimize your workloads by using best practices

4

Get technical guidance in an Immersion Day

Continued rapid pace of innovation

Instance growth



Broadest and deepest platform choice

Workloads

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

NEW!

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

Elastic Graphics

Amazon Elastic Inference

=

270+
instance types

for virtually every workload and business need

Customer obsessed

90%

of roadmap originates with customer requests
and are designed to meet specific needs

AUTODESK

Uses Spot Instances and AWS Auto Scaling for it's Rendering-as-a-service workload to spend **less and scale more**

FRED HUTCH
CURES START HERE™

Decreased the time it took to analyze 10,000 biological samples from **7 years to 7 days**

Standard Chartered

Reduced grid infrastructure **costs by 60%**

Western Digital.

Completed **2.5 million** tasks in 8 hours by spinning up an Amazon EC2 cluster with over **1 million vCPUs**

skyscanner

Was able to **save 74%** on their K8s cluster

matterport

Processes tens of thousands of 3D models daily. Reduced compute costs by **70%**, savings **$1 million** yearly

NOVARTIS

What was originally estimated to take 39 years and $40 million took **9 hours and $4,232.**

lyft

Saved **75% a month** by changing four lines of code

MOBILEYE®

A job that took **weeks** in their data center, due to limited resources, takes **hours,** thanks to the great parallelism, at a very cost-efficient price

AdRoll

Processes over **100 billion** requests per day with an average response time of 90ms, saving over **$3M per year**

illumina®

Reduced monthly compute **costs by 75%** while gaining more compute power

QBiC

Reduced **queue time by 50%** by using Spot Instance

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

Pricing

Achieve optimal price/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/performance with different purchase models

Capacity

Capacity management made easy on the broadest and deepest compute platform

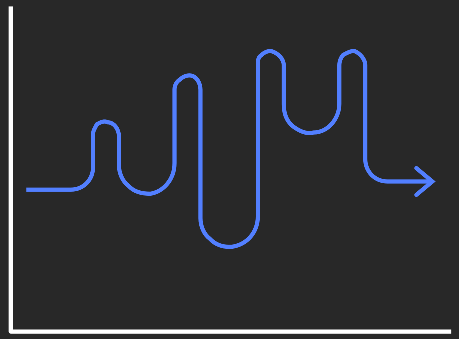
Guidance

Cost and capacity recommendations enable ease of use and save time

Amazon EC2 purchase options

On-Demand

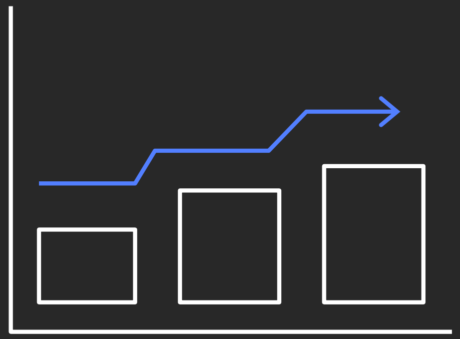
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads, to define needs

Reserved Instances

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



Committed and steady-state usage

Savings Plan

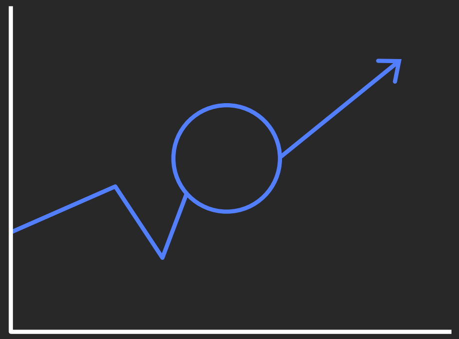
Same great discounts as Amazon EC2 RIs with **more flexibility**

Flexible access to compute

NEW!

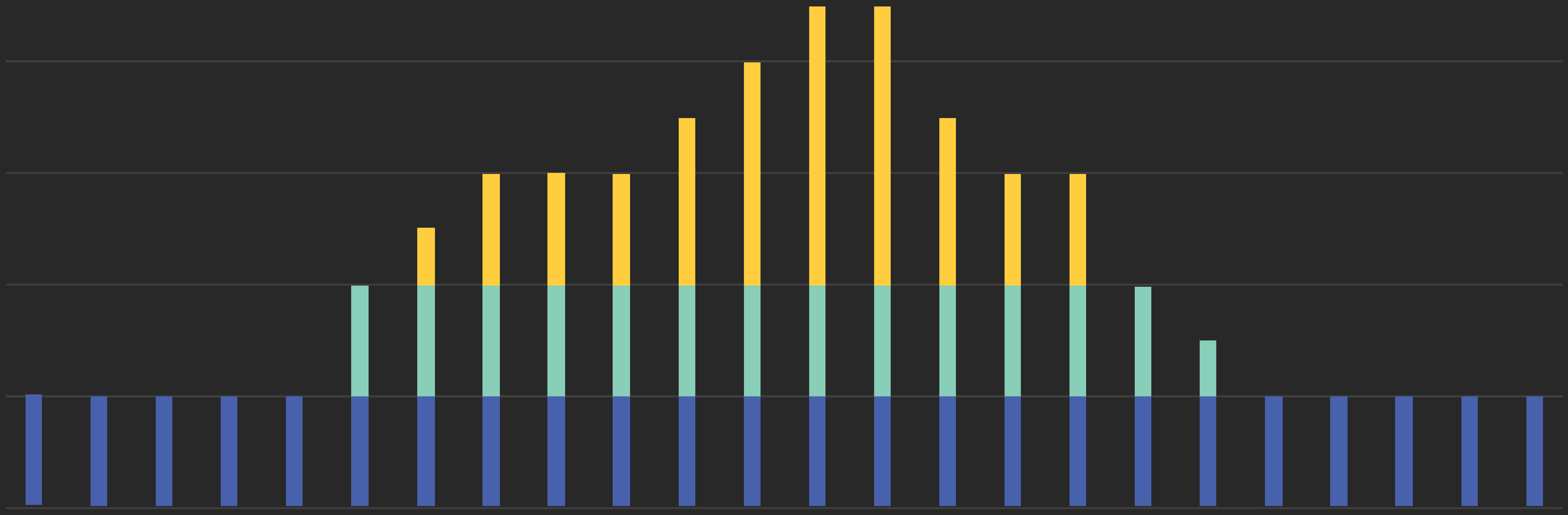
Spot Instances

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

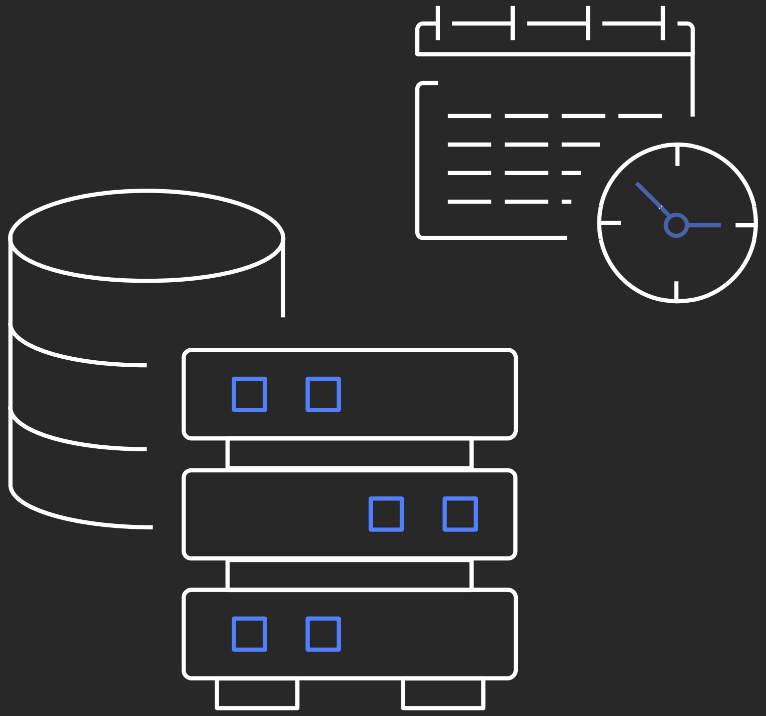


Fault-tolerant, flexible, stateless workloads

To optimize Amazon EC2, combine purchase options



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 NEW!

Introducing Savings Plans

NEW!

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 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

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

Capacity

Interruptions only
happen if OD needs
capacity

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

Why Spot Instances?

Low, predictable prices

Up to 90% discount over On-Demand prices

Faster results

Increase throughput up to 10x while staying in budget

Easy to use

Launch through AWS services (e.g., Amazon ECS, Amazon EKS, AWS Batch, Amazon SageMaker, Amazon EMR) or integrated third parties

Flexibility is key to successful Spot usage

Instance flexible

Time flexible

Region flexible

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

Pricing

Achieve optimal price/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

Using Amazon EC2 Auto Scaling

Automatically scale instances across instance families and purchase options in a single ASG to optimize cost

NEW!

Capacity optimized

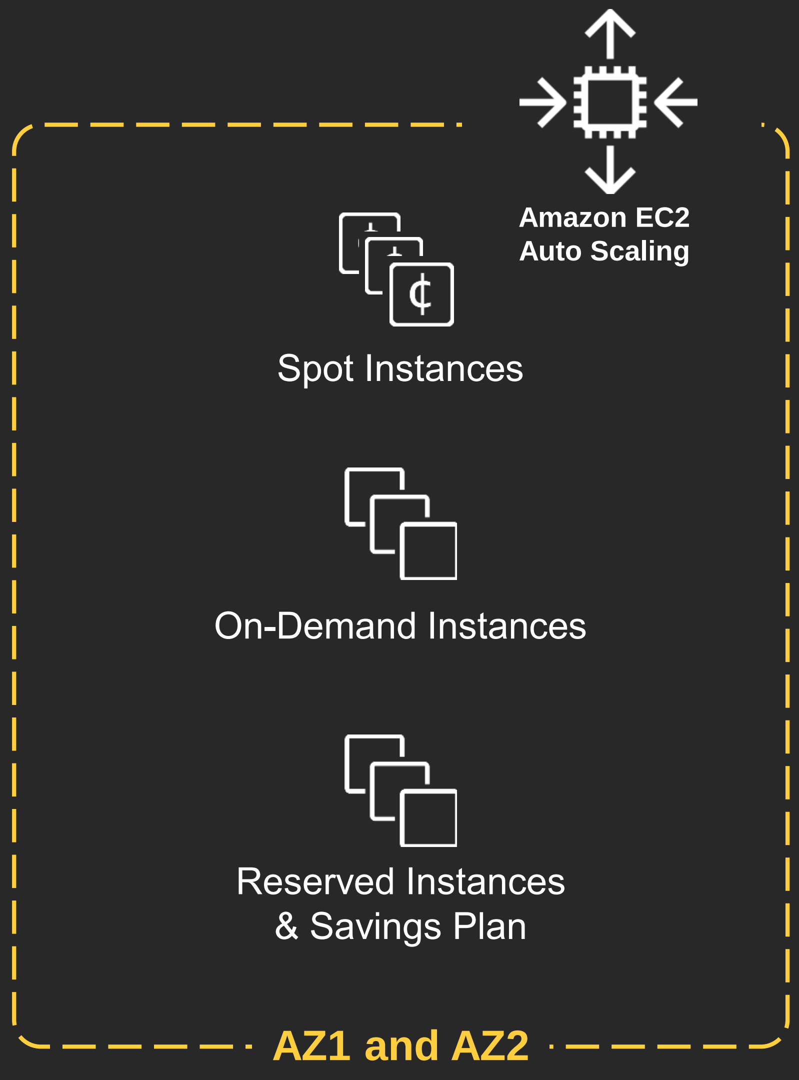
Prioritize deploying Spot Instances into greater Spot pool capacity order to lower the chance of interruptions

Lowest cost

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

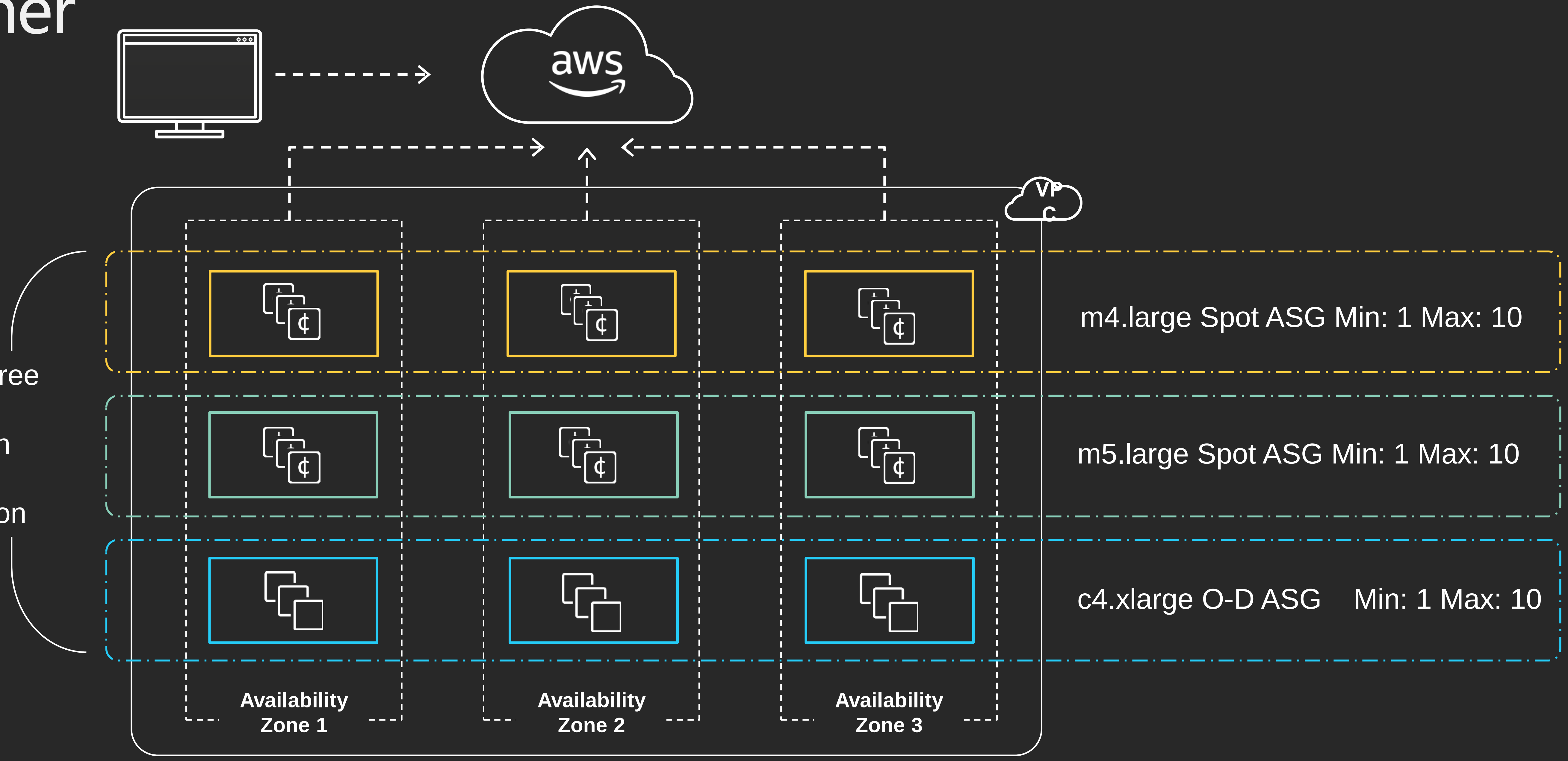
Prioritized list

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



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/ purchase option

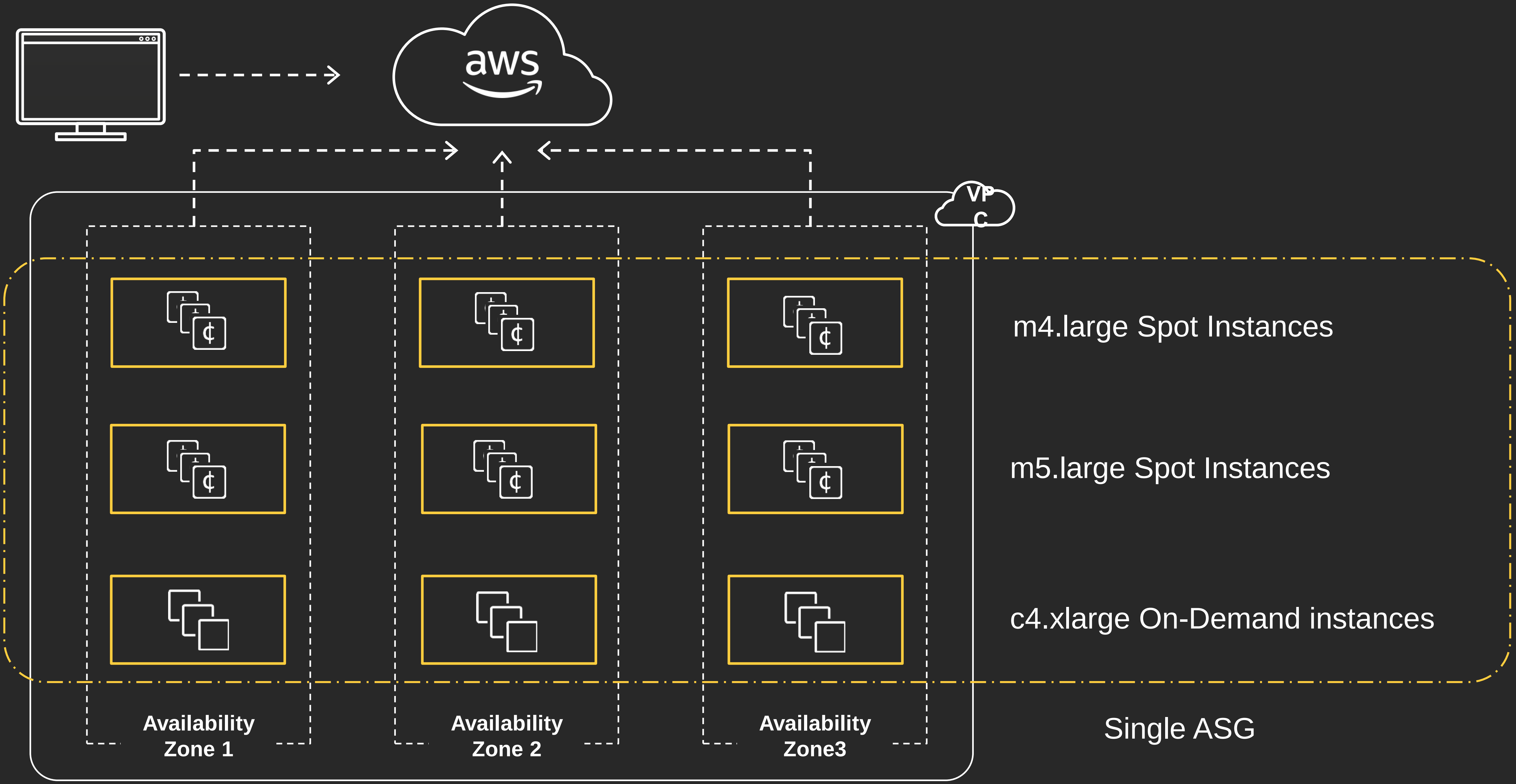
m4.large Spot ASG Min: 1 Max: 10

m5.large Spot ASG Min: 1 Max: 10

c4.xlarge O-D ASG Min: 1 Max: 10

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

The new way combines purchase options, instance types and AZs in a single ASG



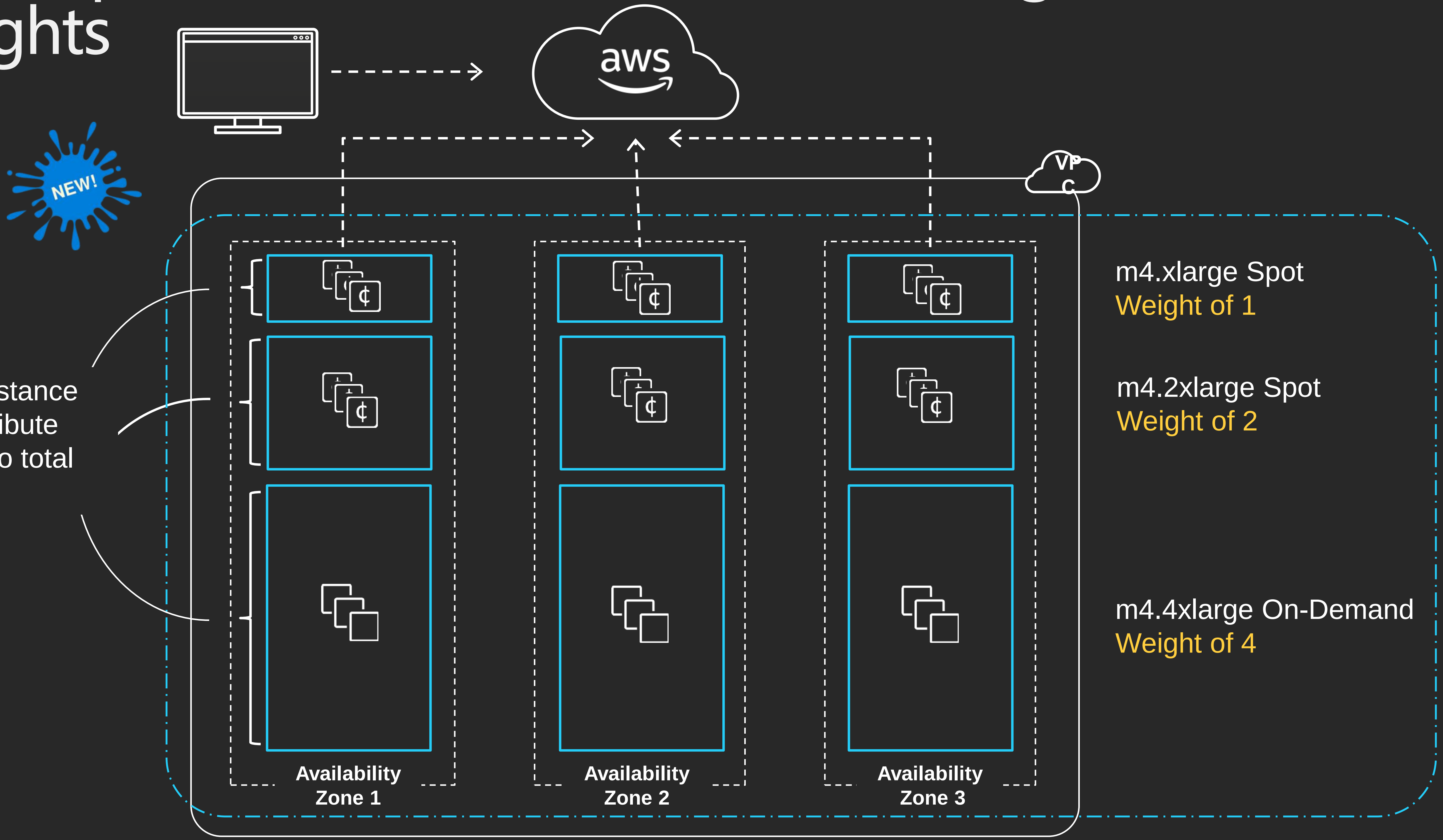
m4.large Spot Instances

m5.large Spot Instances

c4.xlarge On-Demand instances

Single ASG

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



Amazon EC2 Fleet

Consistent API across AWS services to launch a fleet of instances

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

AWS
CloudFormation

AWS services:
AWS Thinkbox Deadline,
Amazon GameLift

Third-party services:
Terraform, Jenkins, Qubole

Use Amazon EC2 Fleet for DIY control over instance management, otherwise let Auto Scaling Groups reduce the undifferentiated heavy lifting

AWS and third-party integrations

Amazon EC2 Auto-scaling

Amazon EC2 Fleet

AWS Thinkbox

Amazon EMR

AWS CloudFormation

AWS Batch

NEW!

Amazon Elastic Container Service

NEW!

Amazon Elastic Container Service for Kubernetes

NEW!

Amazon SageMaker

NEW!

AWS Fargate

NEW!

AWS Elastic Beanstalk

Qubole

Terraform

cloudbees

kubernetes

NEW!

IBM Spectrum Symphony

Schedule an immersion day

AWS experts are here to help and it’s FREE!

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

Pricing

Achieve optimal price/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

Workloads on AWS

- Analytics and big data
- Databases
- DevOps-CI/CD
- Enterprise apps
- IoT
- Machine learning
- Storage
- Websites and web apps

AWS Compute Optimizer

NEW!

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

Lower **costs** and improve workload **performance**

Applies insights from millions of workloads to make recommendations

Saves time comparing and selecting optimal resources for your workload

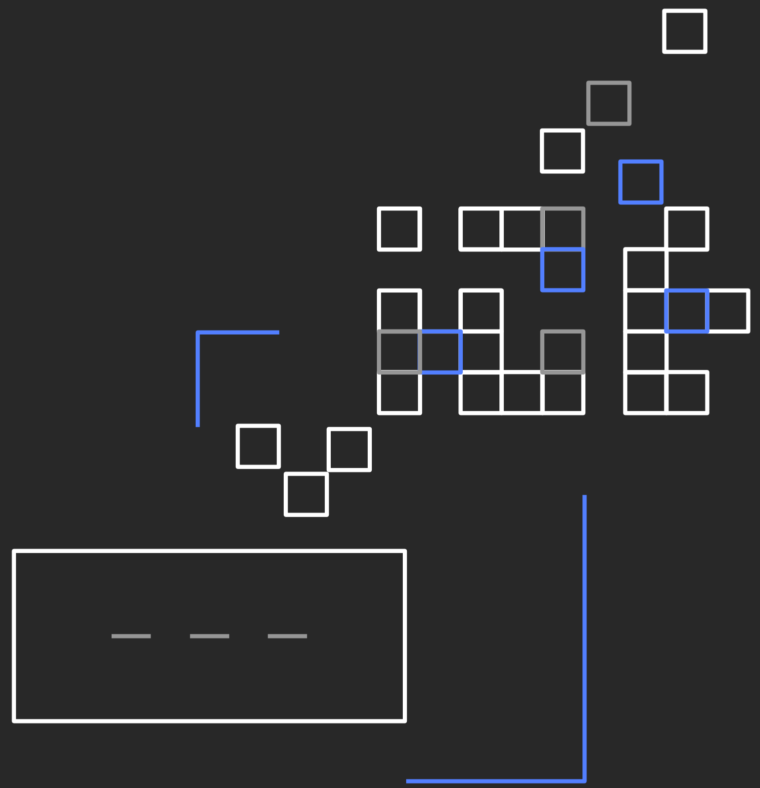
Workloads on AWS

Analytics, big data, and machine learning

DevOps – CI/CD

Websites and web apps

Big data



Amazon EMR

hadoop

APACHE Spark

Massive scale and cost savings to run hyper-scale workloads for data analysis

Unleash your talented data scientists in the age of data

Machine learning

Get ML solutions to market faster with access to built-in algorithms, ML frameworks, and custom models

NEW!

Amazon SageMaker
Managed Spot Training

Save up to 90% in training costs with Managed Spot Training

Automatically manages Spot capacity on your behalf

All instance types, training models, and configurations

aws

salesforce

Salesforce Audience Studio

Alex Estrovitz

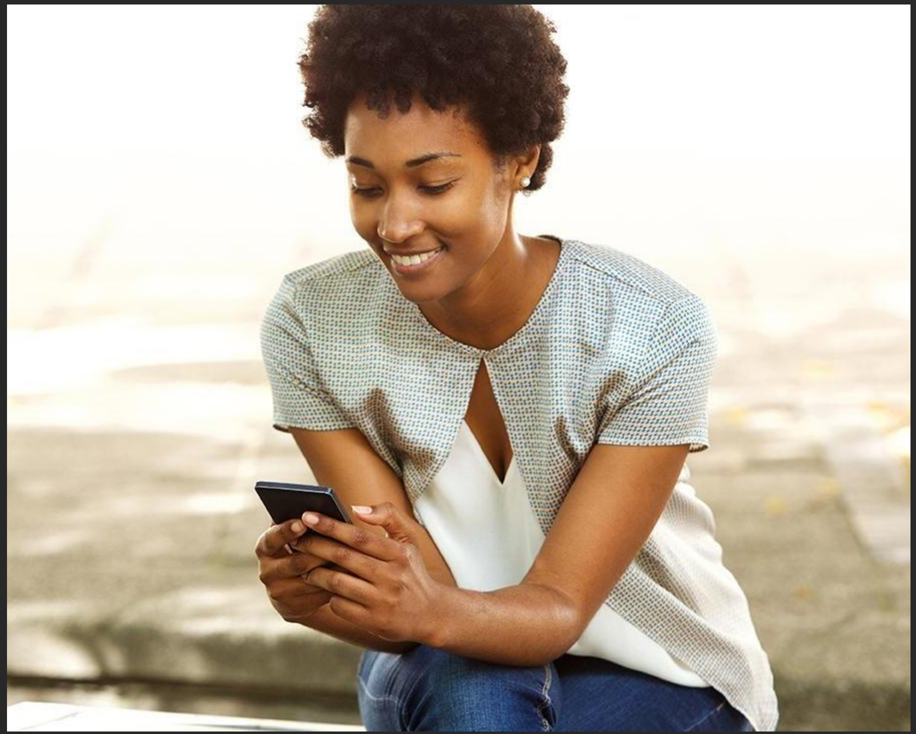
Director Platform Engineering
Salesforce

Audience Studio is a DMP; what's that?

Collect and store data

Online behavior, offline purchases, etc.

Unify data to single user



Rich consumer profiles

Segment into audiences

"Cereal moms"

Publisher challenged by scale and ability to prove audience value

Publisher

How do I scale my audience and offer demonstrable value to my advertisers?

Marketers

I need media buys that scale and perform and complement search and social platforms and other partners I utilize

Consumer rights—RTBF and portability request

As part of GDPR consumer rights management, we have to honor two very important requirements

Right to be forgotten

Delete all references for given user IDs/ organization from the entire system (across all captured and transformed user logs)

Portability

Export raw user-level logs captured for a given user/organization

Scale of Audience Studio

Users		Application
Real-time user activities **~ 200+k qps ~ 17B data points/day**	S3 Storage **~ 70+PB**	Application metrics **~ 2.0M/sec**
Offline log ingestion **~ 10s of TB/day**	EMR Clusters **~ 2500 clusters/day**	User activation **~ 10s of billions of user segments**
	EMR Instances > **200k instance hours per day**	
	85%+ on EC2 Spot	

Real-world example of 7.2-hour job

	Lead Count (EC2 On-Demand) 1 nodes	**Task Count (EC2 Spot) 38 nodes**	**Amazon EMR Cost**
EC2 Costs	$40.32	$182.53	$129.11
Per Instance Cost	$13.44	$4.80	–
Job Total			$351.96
Job Total if On-Demand			$680.15

48% total savings on EC2 Spot

aws

salesforce

Related sessions: Analytics, Big Data & AI/ML

Wednesday, 12/4

ANT226—Lower costs on Amazon EMR: AWS Auto Scaling and Spot pricing
4:45 PM–5:45 PM | Mirage, St. Thomas B

Friday, 12/6

ANT308-R1—[REPEAT 1] Deep dive into running Apache Spark on Amazon EMR
10:45 AM–11:45 AM | Venetian, Level 3, Lido 3005

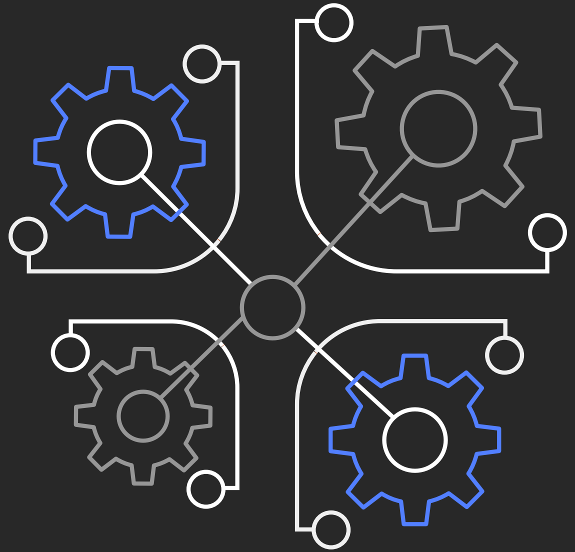
Workloads on AWS

Analytics, big data, and machine learning

DevOps – CI/CD

Websites and web apps

DevOps – CI/CD

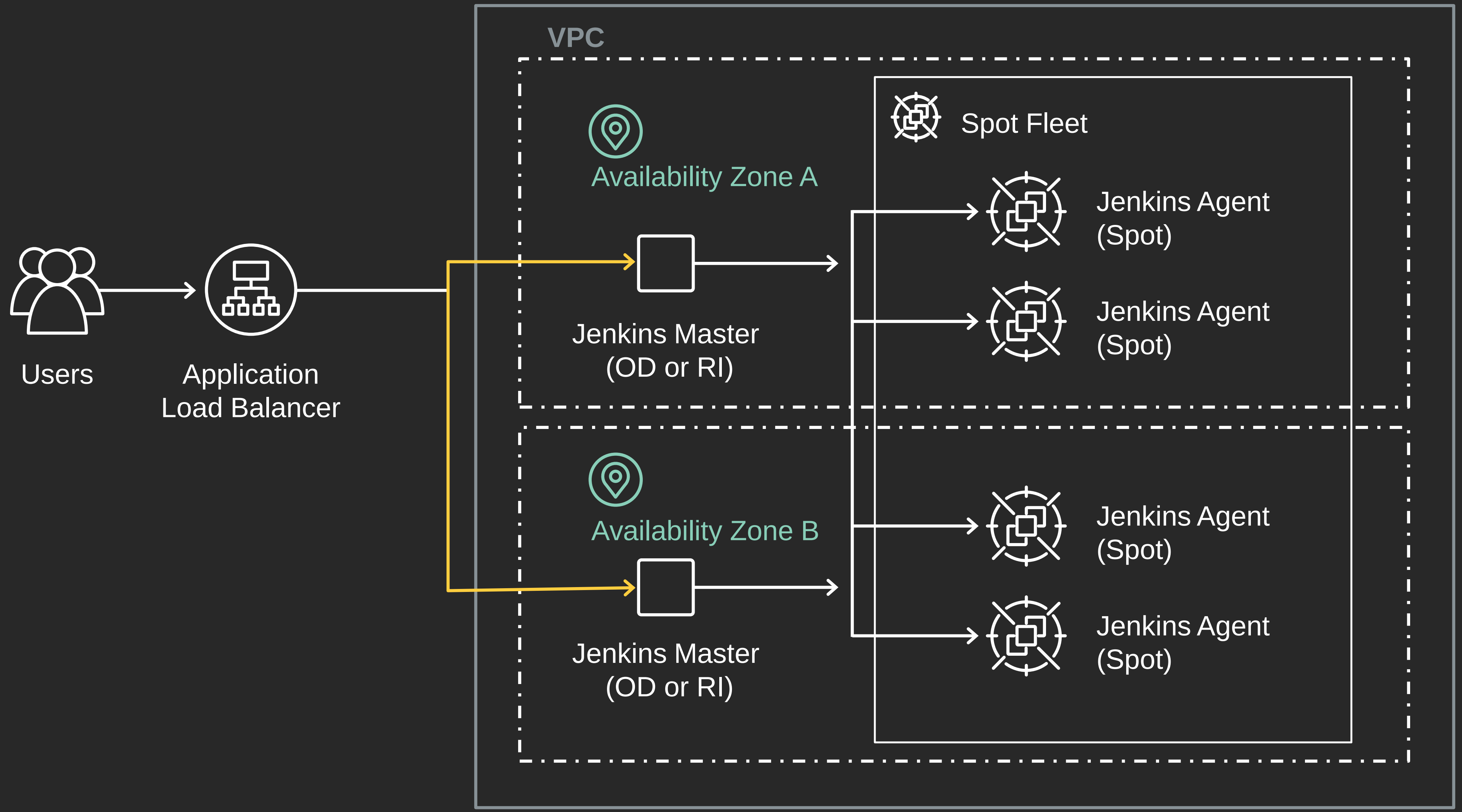


Jenkins

Configure Jenkins with the EC2 Fleet plug-in to automatically scale a Fleet of Spot instances based on the number of CI/CD jobs

Accelerate your integration and deployment pipelines, get to market faster

CI/CD reference architecture



Jenkins

Bamboo

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

Related sessions: CI/CD

Thursday, 12/5

CMP401-R1—Deploying Amazon EC2 Auto Scaling in your CI/CD pipeline

1:00 PM–2:00 PM | Mirage, Grand Ballroom B - Table 2

Friday, 12/6

CMP403-R3—Running enterprise test/dev on Amazon EC2 Spot Instances

10:00 AM–11:00 AM | Mirage, Events Center C1 - Table 3

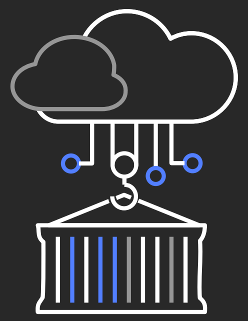
Workloads on AWS

Analytics, big data, and machine learning

DevOps – CI/CD

Websites and web apps

Websites and web apps



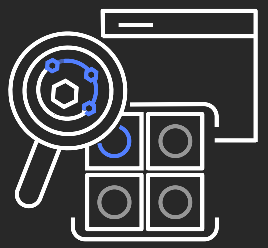
Amazon Elastic Container Service

Amazon Elastic Container Service for Kubernetes

AWS Fargate

Amazon EC2 Auto Scaling

Amazon EC2 Fleet



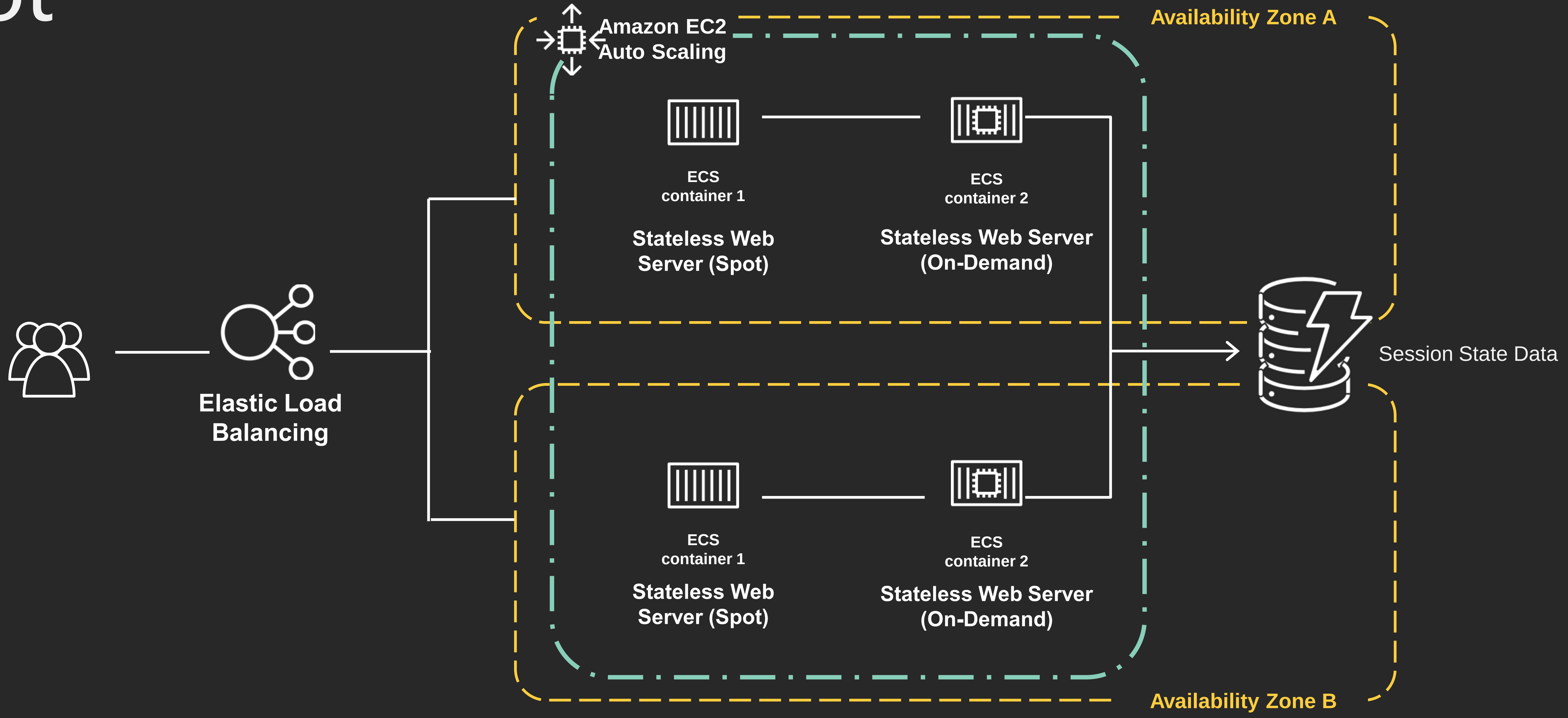
Run web services ranging from ad servers to real-time bidding servers

Deploy web apps or services on containers and scale clusters at a fraction of the cost

Use Auto Scaling with ECS or EKS to run any containerized workload, including a web app

Scale in real time, pay in seconds, save up to 90%

Running web apps with Amazon ECS on EC2 Spot



Running Kubernetes with Amazon EKS on EC2 Spot

Delivery Hero

Delivery Hero is among the largest food delivery networks in the world

Delivery Hero operates in 39 countries with 310,000 restaurant partners, and transports 1 million food orders daily

> “Our experience running Amazon EKS on Amazon EC2 Spot Instances was eye-opening. It has become a big cost saver and **freed our time and energy** to focus on business growth instead.”

—Vojtech Vondra
Senior Director of Engineering, Logistics

Delivery Hero uses Amazon EKS with Spot Instances to deploy 90% of their Kubernetes clusters

Introducing AWS Fargate with EC2 Spot

Run containers without managing servers or clusters

NEW!

AWS Fargate

Up to 70% off over regular Fargate tasks

Only pay for the resources you use by autoscaling based on tasks, vCPUs, and memory

VM-level isolation by design

Related sessions: Containers

Wednesday, 12/4

CON308-S—How Ticketmaster runs Kubernetes for 80% less without managing VMs

5:30 PM–6:30 PM | Aria, Level 1 East, Joshua 9

Thursday, 12/5

CMP318-R1—[REPEAT 1] Kubernetes on Spot Instances: Optimize for scale and cost

3:15 PM–5:30 PM | Mirage, Grand Ballroom G

Thursday, 12/5

CON324-R1—[REPEAT 1] Cost Optimization with Containers and Spot

1:00 PM–2:00 PM | MGM, Level 1, Grand Ballroom 119

To tie it all together…

Key takeaways from this session…

1

Experiment and test at a lower cost to innovate faster

Spot Instances

2

How to automate cost and capacity optimization

Auto Scaling

Savings Plan

3

Optimize your workloads by using best practices

Compute Optimizer

4

Get technical guidance in an Immersion Day
+
$50 EC2 Spot Credit

CI/CD, Analytics, Big Data, Machine Learning & Web Services

Collect $50 Spot Credits

Amazon EC2
Spot Instances

Save up to 90% on Compute

$50 service credit

aws

Thank you!

Jeanine Banks

@femtechie

AWS re:Invent

aws

!

Please complete the session survey in the mobile app.

AWS re:Invent

aws