
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CMP413-R

Set up adaptable, elastic, and efficient compute for ML

Kinnar Kumar Sen (he/him/his)

Sr. Solutions Architect – EC2 Spot
AWS

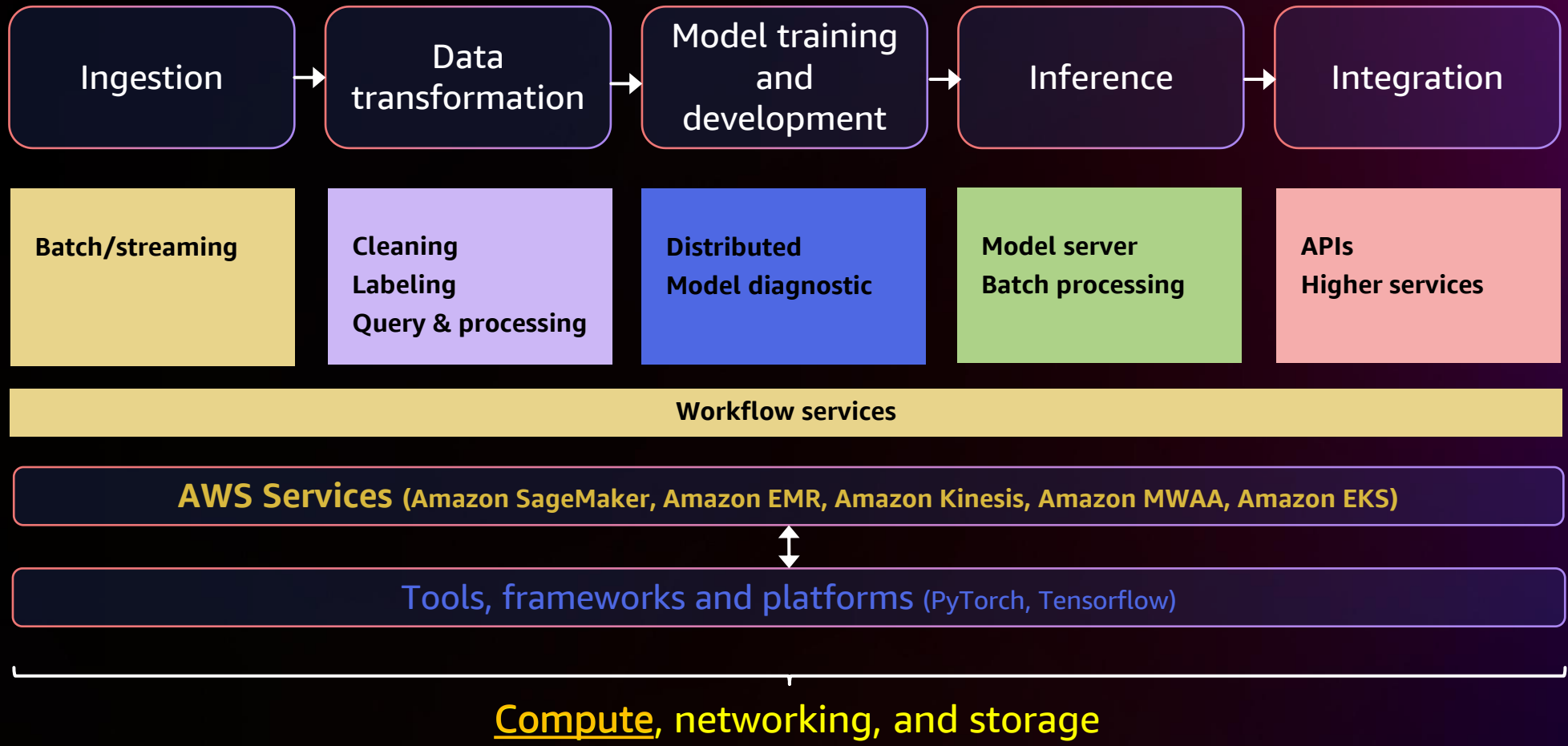
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- ML pipeline and key challenges
- Compute flexibility: Least common denominator
- Customer obsession: Tools
- Architecture: Distributed and flexible training

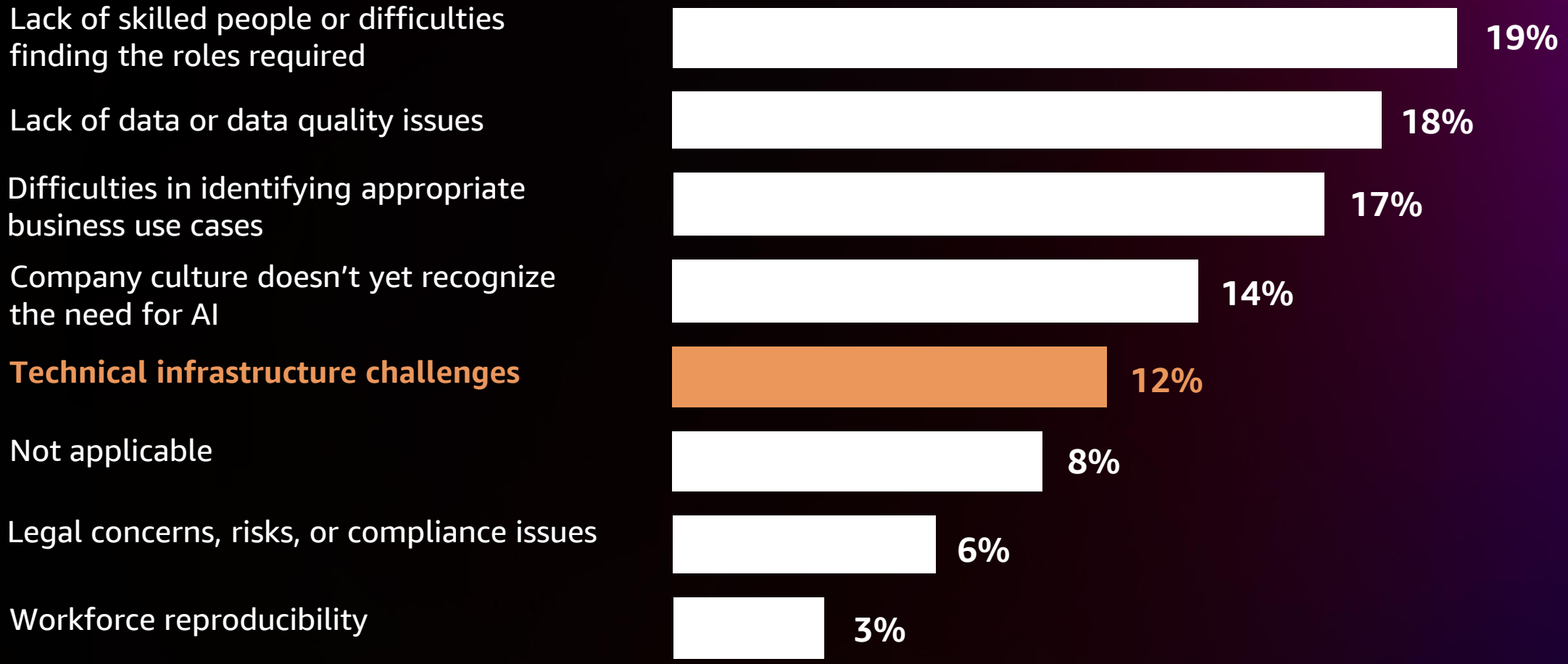
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML pipeline



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key challenges to AI/ML adoption



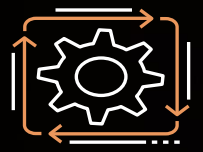
Source: O'Reilly, AI Adoption in the Enterprise Report 2021

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges in ML infrastructure

- Machine learning workloads are not one-size-fits-all
- Training and deploying models are expensive
- Procuring the right amount of infrastructure requires analysis
- Customizing the infrastructure to meet performance and budget
- Reducing barriers to launch capacity for data scientists/engineers



FLEXIBLE



SCALE



COST

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

"Using Spot provides us with the most cost-effective solution for the analysis of many terabytes of metagenomic data. We further benefit from the flexibility of Amazon Spot, as different analysis stages in the pipeline involve different workloads (data cleaning, DNA comparison, machine learning). Each task is best performed by different EC2 instance types, so in each stage we launch a different cluster. Clusters in large projects such as this typically involve several thousand instances that run for dozens of hours."

Elad Kehat

VP of R&D Software
BiomX

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compute: Integrations and core services



Amazon EMR



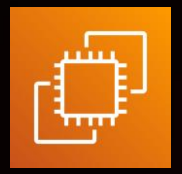
Amazon SageMaker



Amazon Elastic Kubernetes Service (Amazon EKS)



AWS Batch



Amazon EC2



Amazon EC2 Auto Scaling



Spot Instances



Inferentia Instances



Accelerated Instances



Graviton Instances

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Compute Flexibility

To help you adopt architectural best practices and specific AWS services that optimize and improve performance of your workloads while minimizing waste through sizing, scaling, and flexibility.

Amazon EC2 services that help you with flexibility


Amazon EC2


Amazon EC2 Auto Scaling


Amazon EC2 Spot



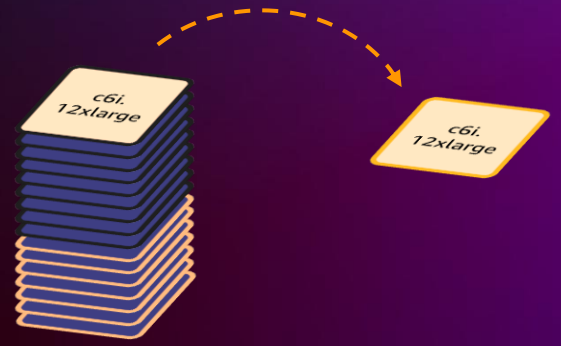
AWS Graviton


AWS Compute Optimizer

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Spot Instances

RUN FAULT-TOLERANT WORKLOADS FOR UP TO 90% OFF



Infrastructure
Spare EC2 capacity



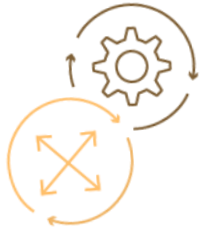
Same as
On-Demand and
Reserved Instances

Pricing
Up to 90% off compared to On-Demand



Pricing is based on long-term supply and demand (no bidding!)

Interruption
Interrupted only if OD needs capacity



Instances can be interrupted with a 2-minute warning if capacity is required

Diversification
and flexibility is key



Make use of different instance types, sizes, Availability Zones, and times

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Graviton processor

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



Most cost-effective instances in Amazon EC2 for Android game streaming with up to 30% lower cost per stream per hour

Get significantly lower cost per inference for machine learning inference over x86-based GPU instances

Powered by AWS Graviton2 and featuring NVIDIA T4G Tensor Core GPUs

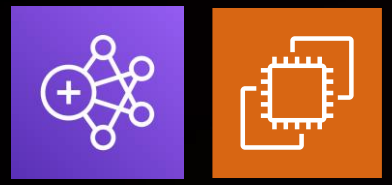
Amazon EMR

Amazon EMR provides up to 35% lower cost and up to 15% improved performance for Apache Spark workloads on Graviton2-based instances versus previous generation instances

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR: Optimize using Spot and Graviton

EASILY RUN AND SCALE APACHE SPARK, HIVE, PRESTO, AND OTHER BIG DATA WORKLOADS



Amazon EMR on Amazon EC2



EC2 Spot

performance, scale, and up to 90% savings



Amazon EMR on Amazon EKS



Graviton

35% lower cost and up to 15% improved performance

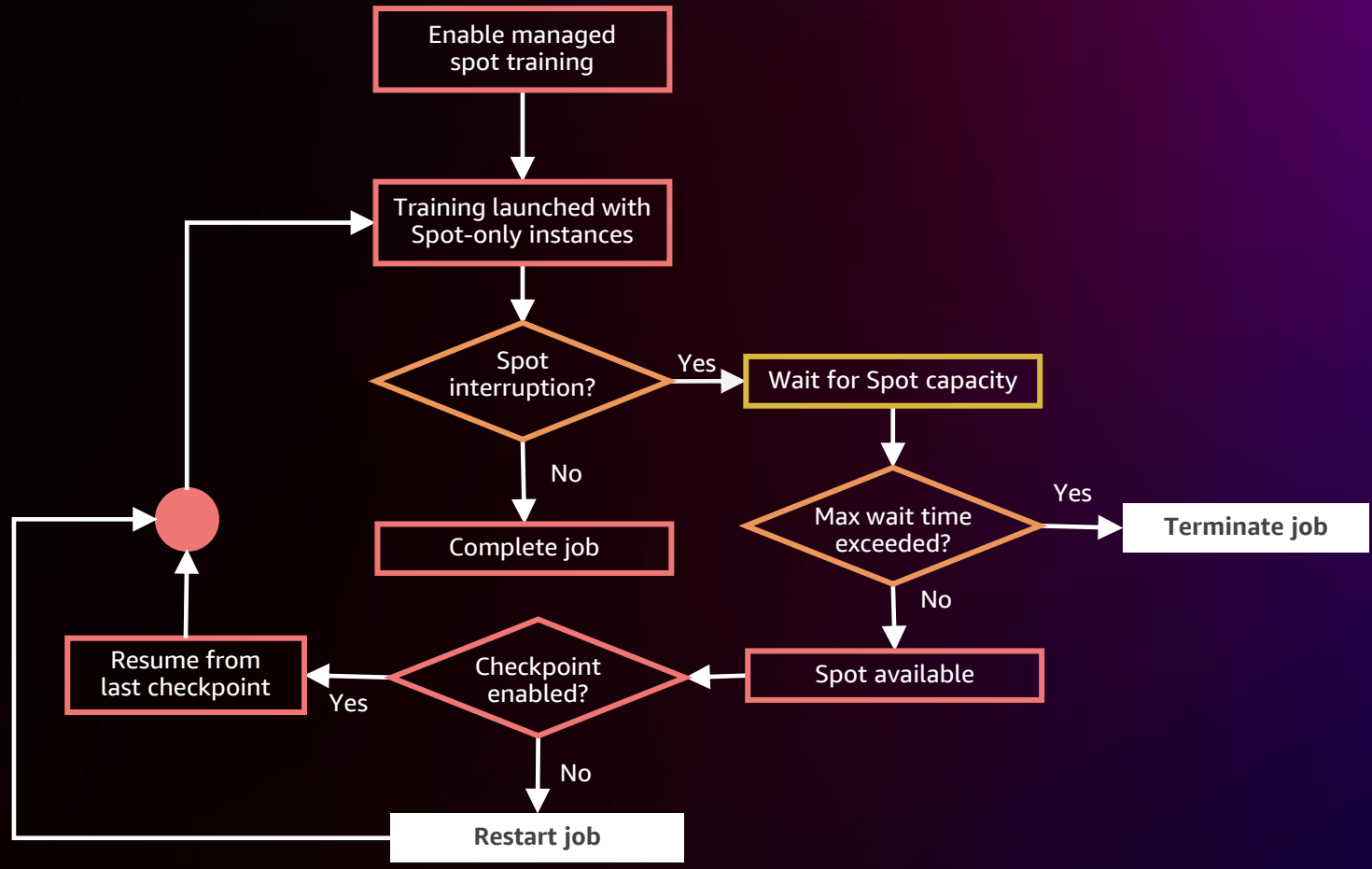
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker: Managed spot training

SAVE UP TO 90% ON MODEL TRAINING COSTS





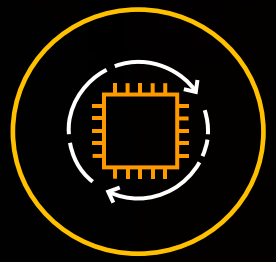
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

EC2 Auto Scaling

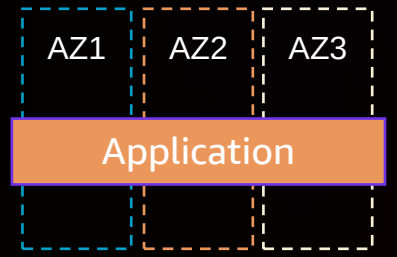
AUTOMATICALLY SCALE INSTANCE TYPES AND PURCHASE OPTIONS IN A SINGLE AUTO SCALING GROUP TO OPTIMIZE COST

Improves fault tolerance

- Instance lifecycle management
- Across Availability Zones



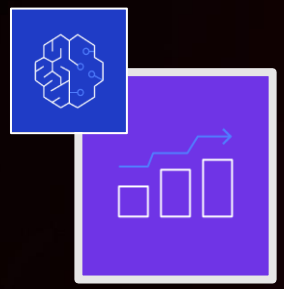
Lifecycle management



Fault tolerant application

Increases application availability

- Match capacity to demand
- Warm pools
- Proactive provisioning



Predictive scaling

Lowers costs

- Scale across instance types, purchase options, architectures
- Reduce over-provisioning



EC2 Spot Instances



Savings plan



AWS Graviton

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Flexible compute tools/features

BUILD ADAPTABLE, ELASTIC, AND EFFICIENT ML PIPELINES

Attribute-based instance selection/Amazon EC2 instance selector

- Provision capacity/select compatible instance types based on instance attributes (such as CPU/memory/GPU)

Karpenter

- Provisions and scales Kubernetes clusters and improves the efficiency and cost of running workloads

Spot placement score

- Finds the optimal location for your workload and recommends locations (Reg/AZ) and the likelihood of getting the Spot capacity

AWS Fault Injection Simulator (FIS)

- Enables chaos engineering and supports imitation of Spot interruptions

Blueprints: Spot and EKS

- Dynamic infrastructure code generator, providing best-practice reference architecture and add-ons

Warm pools/predictive scaling

- Proactively initialize/manage capacity using EC2 Auto Scaling

On-Demand capacity reservation

- Reserve compute capacity in an AZ for steady-state workloads

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

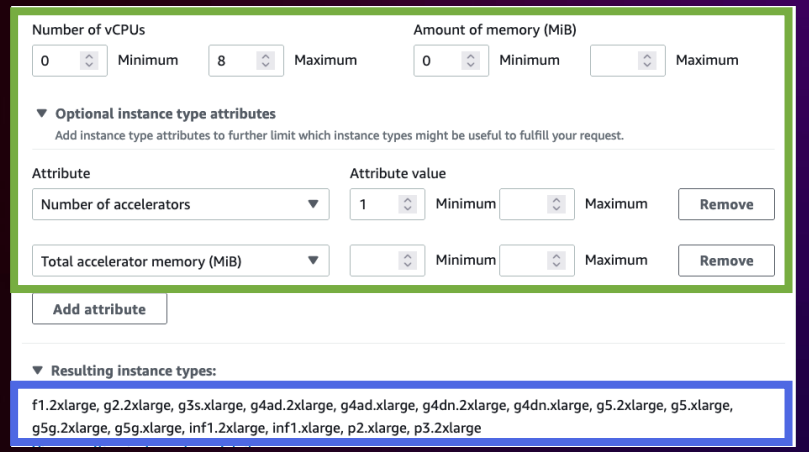
Diversify with attribute-based instance selection

PROVISION CAPACITY BASED ON INSTANCE ATTRIBUTES DEPENDING ON COMPUTE REQUIREMENTS, SUCH AS VCPU, RAM, STORAGE TYPE, AND MORE

Benefits

- Price protected
- Instance flexibility
- Simplifies provisioning
- Mixed instance groups
- Future proof
- AWS Management Console, AWS CLI, and Amazon Chime SDK
- Preview instances

- ABIS requires the use of launch templates
- Supports only **lowest-price on-demand allocation strategy**
- Supports only **capacity-optimized and lowest-price Spot allocation**

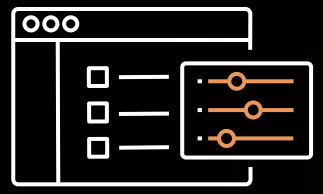


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Predictive scaling

PROACTIVELY LAUNCH CAPACITY BASED ON HISTORICAL TRENDS

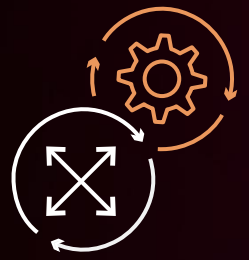
A **new** Amazon EC2 Auto Scaling policy that proactively launches Amazon EC2 instances in anticipation of predicted demand



Proactively manages capacity



Reduces overprovisioning



Works within defined boundaries



Adjusts automatically

CMP412 : Proactive auto scaling for optimal cost and availability

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Karpenter

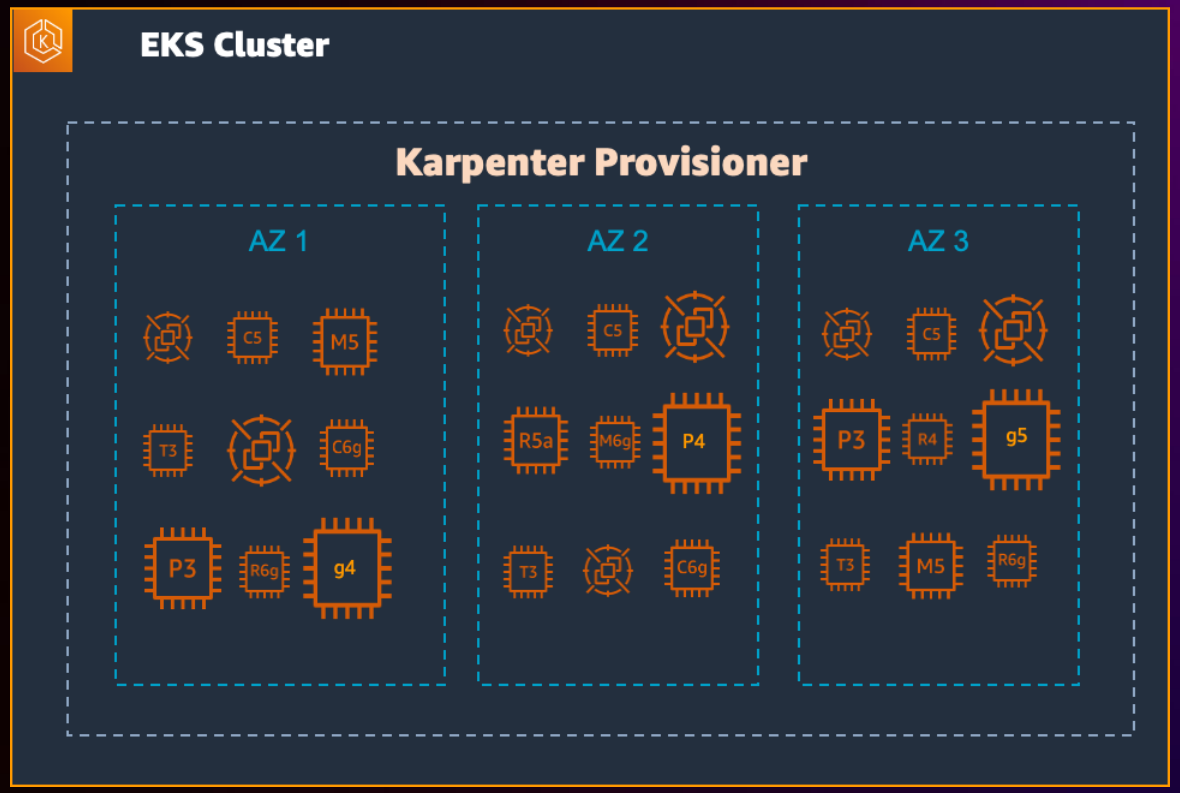
KARPENTER SIMPLIFIES KUBERNETES INFRASTRUCTURE WITH THE RIGHT NODES AT THE RIGHT TIME

Application-first infrastructure
Node provisioning based on pod requirements

Simplified configuration
Single configuration with On-Demand and Spot purchasing options and diverse instance types

Diversify with Spot and On-Demand
Karpenter helps simplify diversification across purchase options

Generally Available on AWS
Ready for Production workloads since November 2021

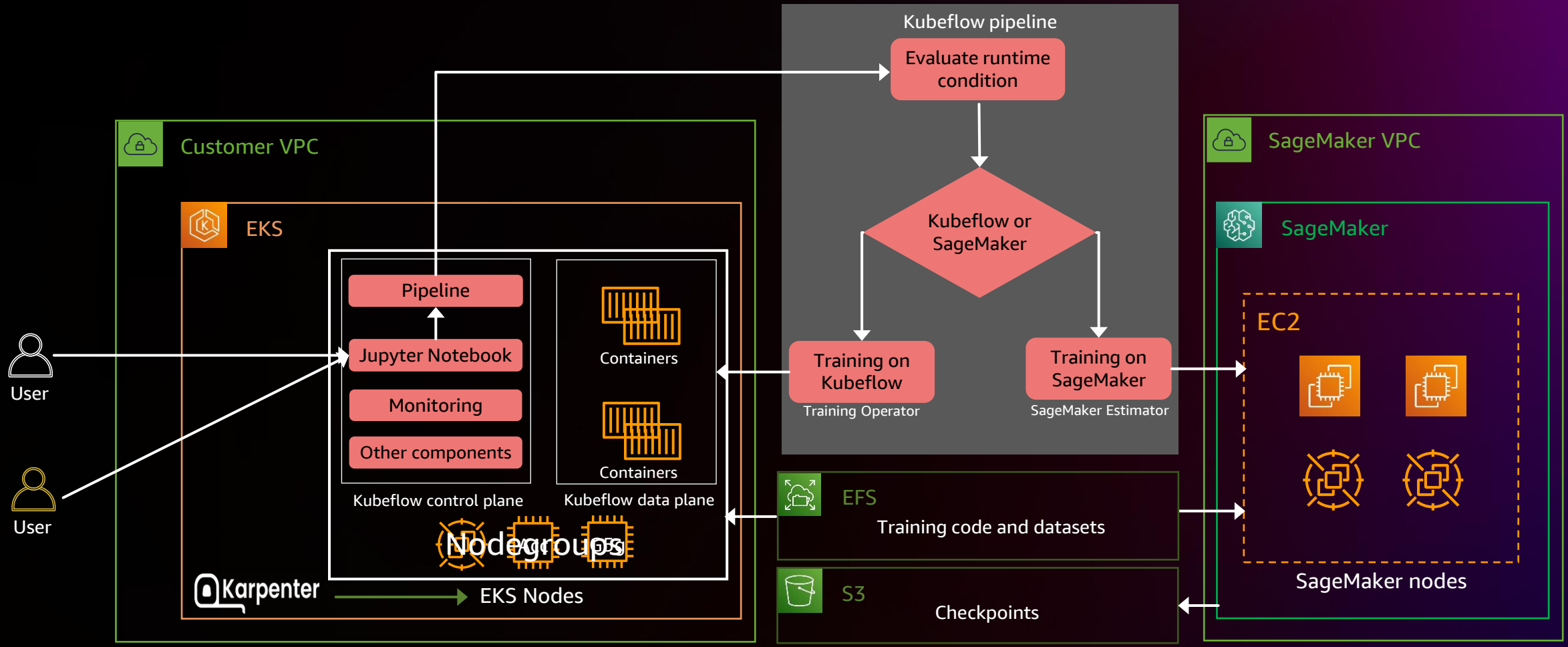


CMP405 : Running efficient Kubernetes clusters on Amazon EC2 with Karpenter

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kubeflow on AWS/Amazon SageMaker

FLEXIBLE, SCALABLE, AND DISTRIBUTED TRAINING ARCHITECTURE



Reference : https://aws.amazon.com/blogs/machine-learning/build-flexible-and-scalable-distributed-training-architectures-using-kubeflow-on-aws-and-amazon-sagemaker/

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Kinnar Kumar Sen
kinnarse@amazon.com

LinkedIn :





Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.