AWS re:Invent

CON415-R

Auto scale Kubernetes workload by GPU

Theodore Salvo

Container Specialist Solutions Architect

Amazon Web Services

AWS re:Invent

aws

Agenda

Introduction to AWS container services

Amazon Elastic Kubernetes Service (Amazon EKS) overview

Machine learning on Amazon EKS

GPU autoscaling on Amazon EKS

Demo

Q&A

Related breakouts

CON415-R1 Auto scale Kubernetes workload by GPU

Wednesday, Dec 4, 4:00 p.m.–5:00 p.m. – Mirage, Events Center C1 Table 3

CON415-R2 Auto scale Kubernetes workload by GPU

Thursday, Dec 5, 2:30 p.m.–3:30 p.m. – Mirage, Events Center C1 Table 3

CON415-R3 Auto scale Kubernetes workload by GPU

Friday, Dec 6, 11:30 a.m.–12:30 p.m. – Mirage, Grand Ballroom B Table 2

AWS container services landscape

Management

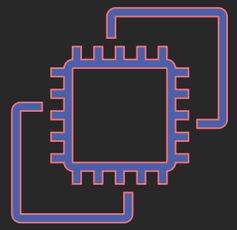
Deployment, scheduling, scaling, & management of containerized applications

Amazon Elastic Container Service (Amazon ECS)

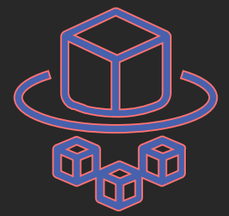
Amazon Elastic Kubernetes Service (Amazon EKS)

Hosting

Where the containers run



Amazon Elastic Compute Cloud (Amazon EC2)



AWS Fargate

Image registry

Container image repository

Amazon Elastic Container Registry (Amazon ECR)

Amazon EKS: A year in review

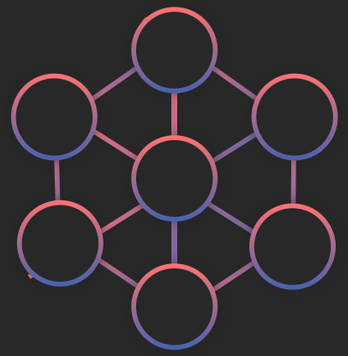
June – December 2018

Amazon EKS achieves K8s conformance, HIPAA eligibility, generally available
Amazon EKS AMI build scripts and AWS CloudFormation templates available in GitHub
Support for GPU-enabled EC2 instances; support for HPA with custom metrics
Amazon EKS launches in Dublin, Ireland
Amazon EKS simplifies cluster setup with `update-kubeconfig` CLI command
Amazon EKS adds support for Dynamic Admission Controllers (Istio), ALB support with the AWS ALB ingress controller
Amazon EKS launches in Ohio, Frankfurt, Singapore, Sydney, and Tokyo
Amazon EKS adds managed cluster updates and support for Kubernetes version 1.11, CSI driver for Amazon EBS

2019

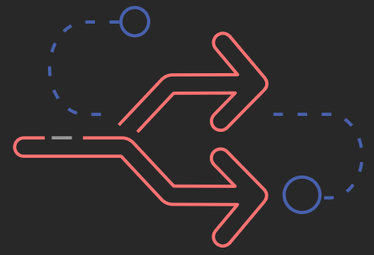
Amazon EKS launches in Seoul, Mumbai, London, and Paris
Amazon EKS achieves ISO and PCI compliance, announces 99.9% SLA, cluster creation limit raised to 50
API server endpoint access control, AWS App Mesh controller
Windows support (preview), Kubernetes version 1.12
CSI drivers for Amazon EFS, Amazon FSx for Lustre, control plane logs, A1 (ARM) instance support (preview)
Deep Learning Benchmark Utility, public IP address support
Simplified cluster authentication, SOC compliance, Kubernetes 1.13, pod security policies
Container Insights, CNI 1.5.0, Amazon ECR, AWS PrivateLink support

How are customers using Amazon EKS?

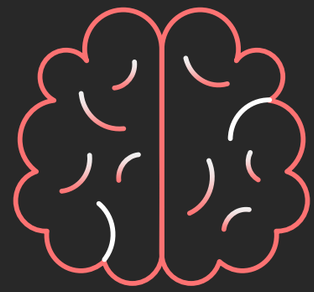


Microservices

Platform as a service



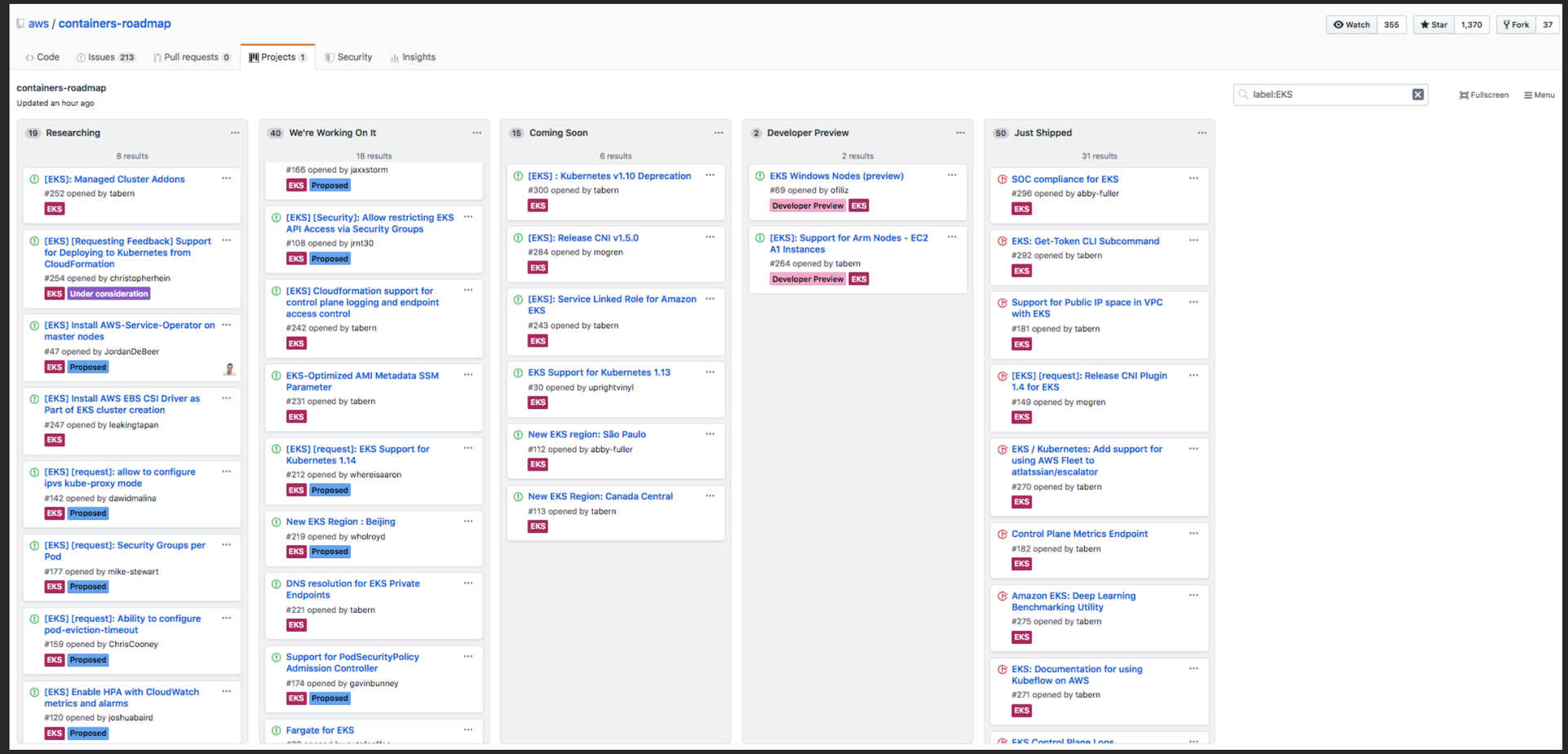
Enterprise app migration



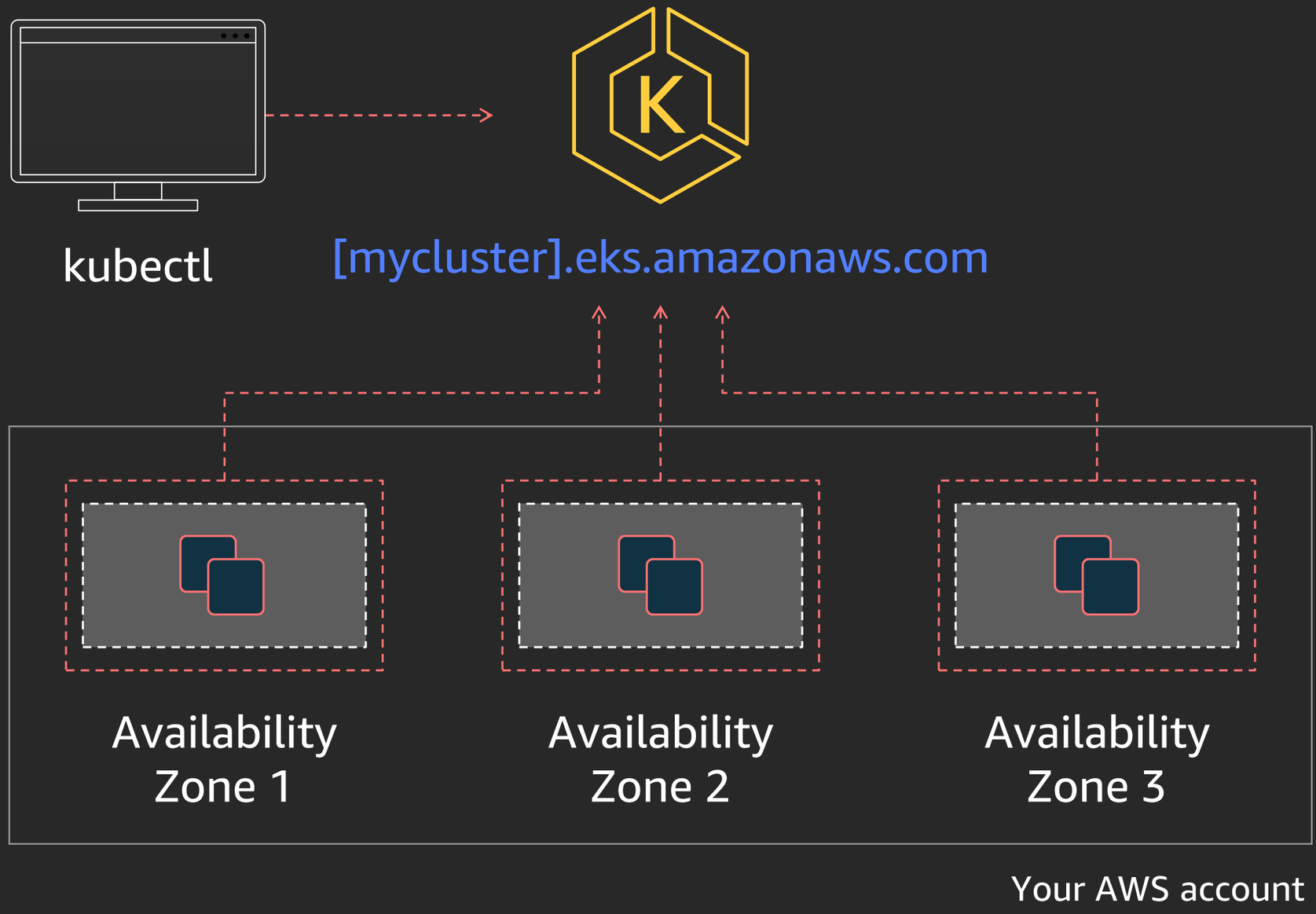
Machine learning

Open-source roadmap

https://github.com/aws/containers-roadmap/



Architecture


kubectl
[mycluster].eks.amazonaws.com
Availability Zone 1
Availability Zone 2
Availability Zone 3
Your AWS account

Configuration and setup

Provisioning worker nodes

AWS CloudFormation

eksctl

eksctl

Terraform
Pulumi
Rancher

And more

Partners

Amazon EKS-optimized GPU AMI

Includes NVIDIA packages to support Amazon P2 and P3 instances

Easily run TensorFlow on Amazon EKS

Now supporting P3dn.24xlarge instances

CUDA 10 with NVIDIA v410 coming soon

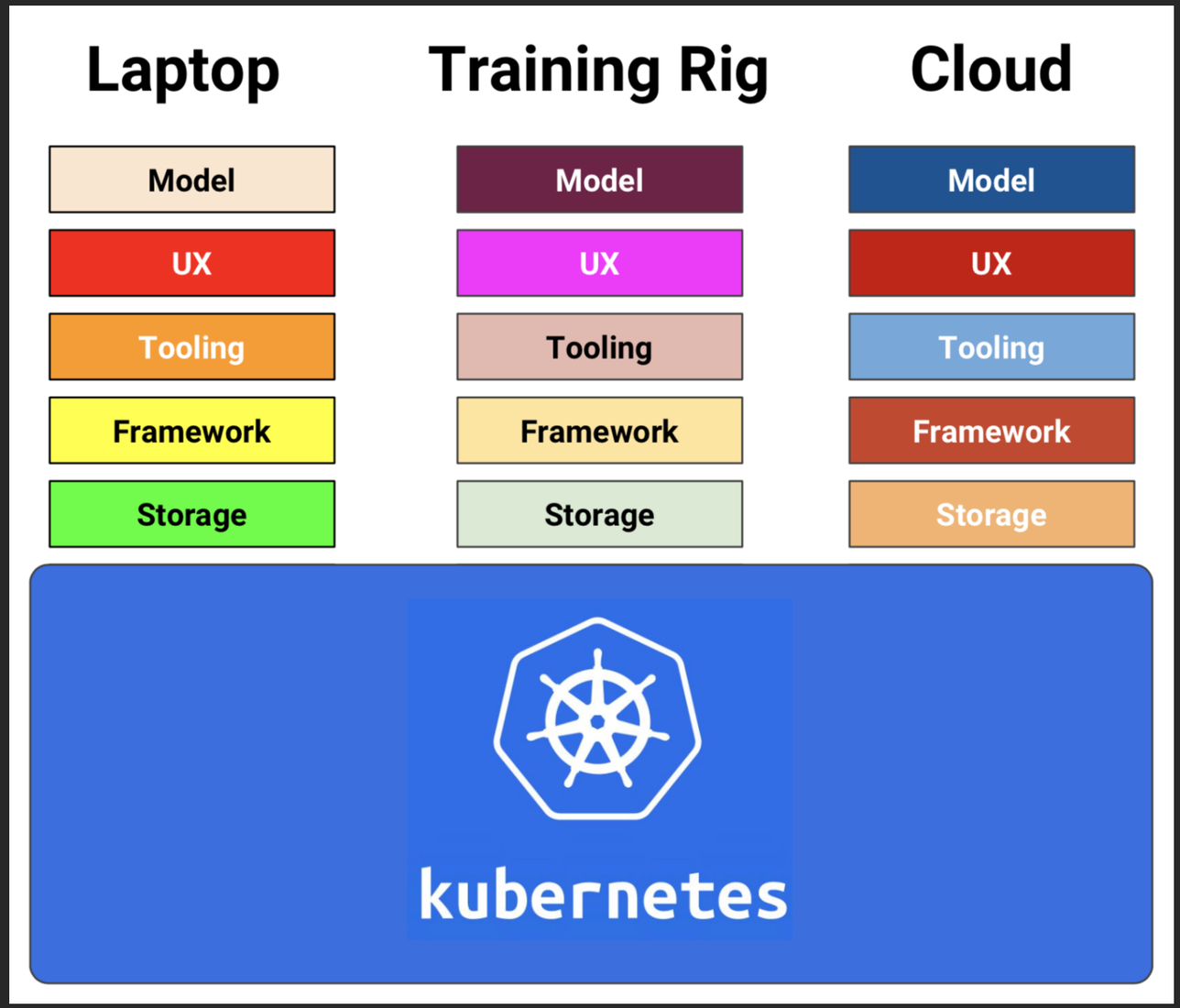
End-to-end ML framework

Why machine learning on Kubernetes?

- ✓ Composability
- ✓ Portability [On-Premises and Cloud]
- ✓ Scalability

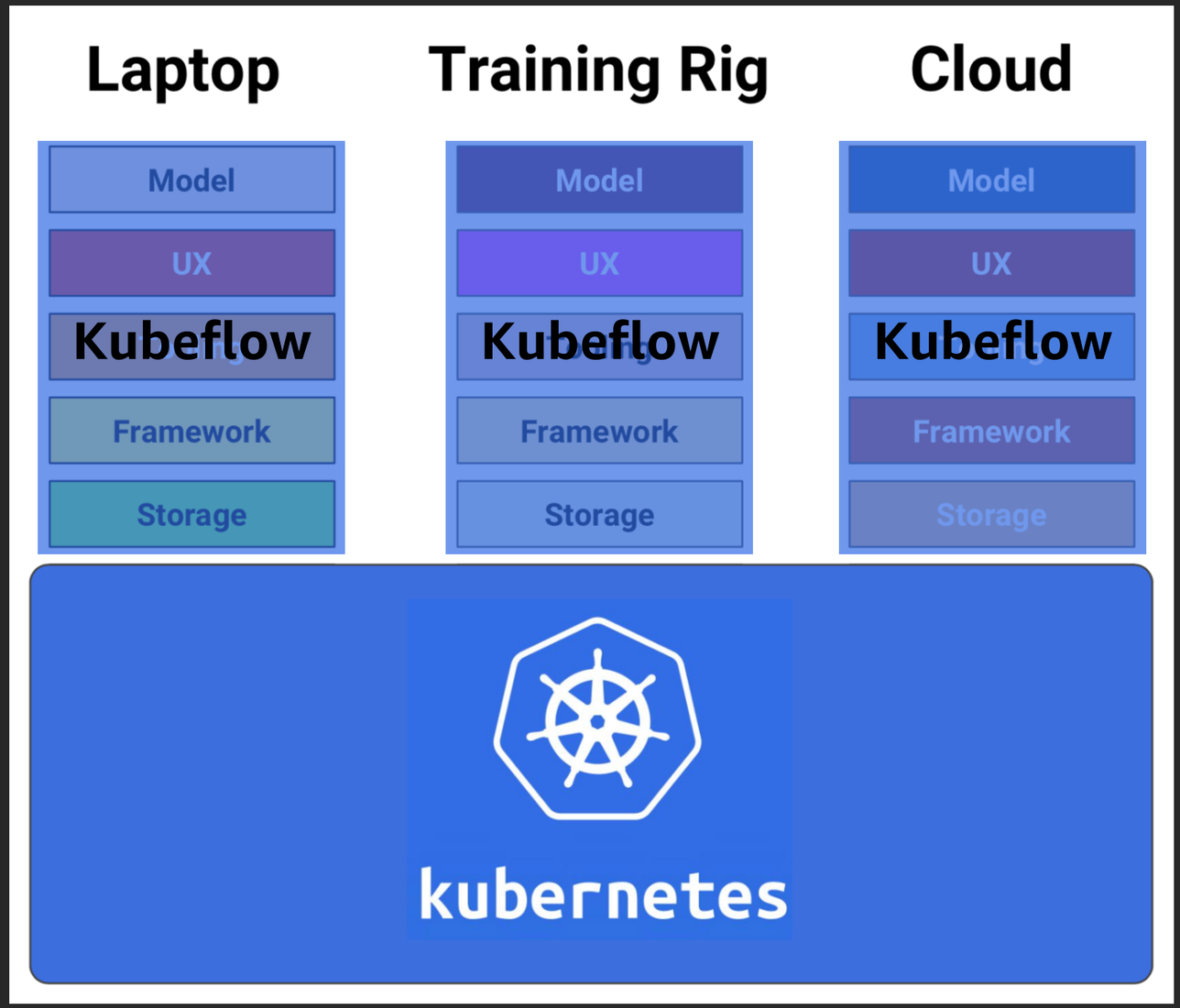
http://www.shutterstock.com/gallery-635827p1.html

ML on K8s—without Kubeflow



Credits: @aronchik

ML on K8s—with Kubeflow



Credits: @aronchik

What's in Kubeflow?

Notebook for collaborative & interactive training

Serving deployment & training controller

For workflows

For complex inference and non TF models

Framework operators

ReverseProxy (Ambassador)

Wiring to make it work on any K8s anywhere

Kubernetes autoscaling

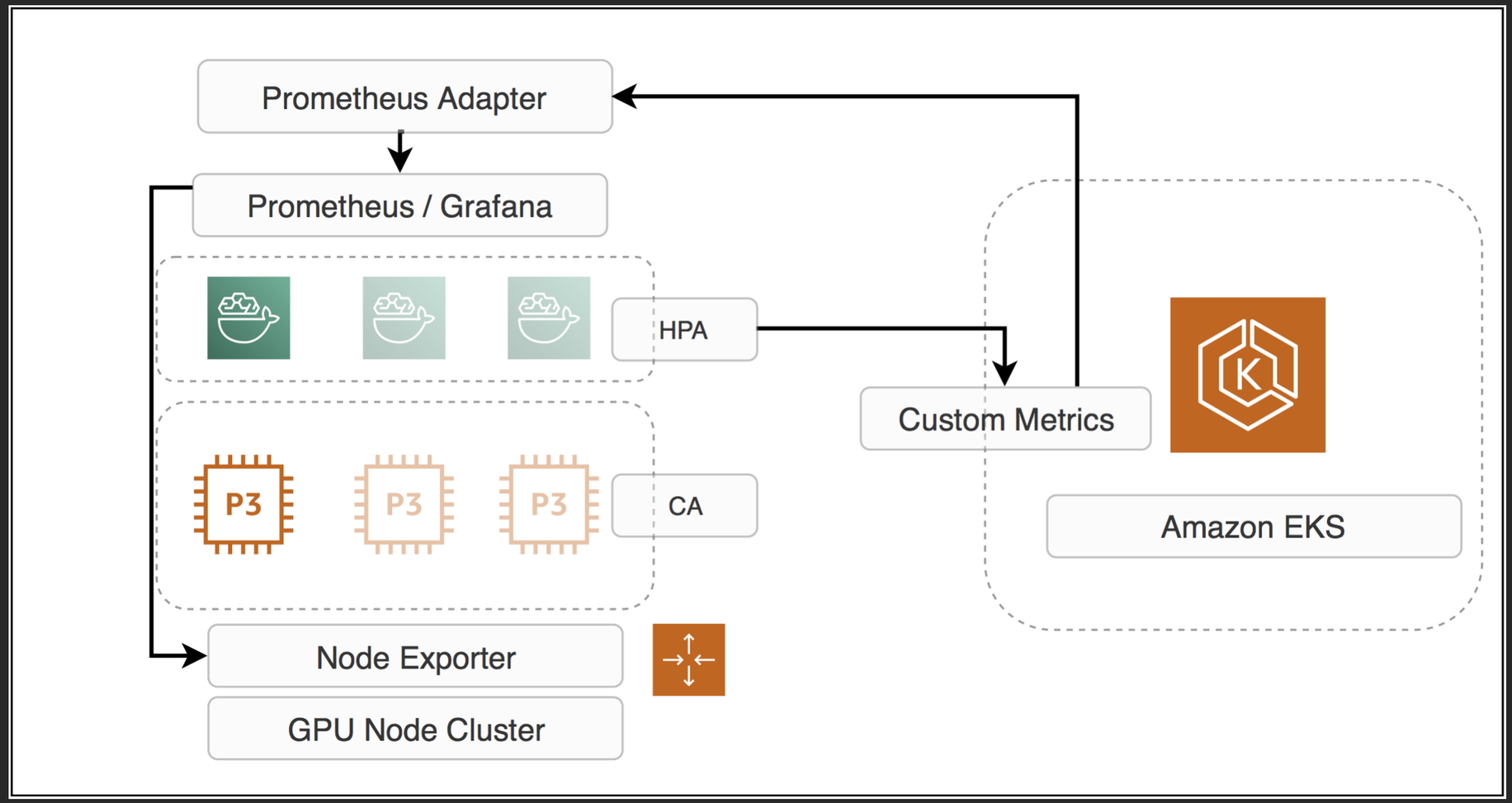
Horizontal Pod Autoscaler (HPA)

Scales the pods in a deployment or replica set. It is implemented as a K8s API resource and a controller. The controller manager queries the resource utilization against the metrics specified in each HorizontalPodAutoscaler definition. It obtains the metrics from either the resource metrics API (for per-pod resource metrics), or the custom metrics API (for all other metrics)

Cluster Autoscaler (CA)

This is the default K8s component that can be used to perform pod scaling as well as scaling nodes in a cluster. It automatically increases the size of an Amazon EC2 Auto Scaling group so that pods have a place to run. And it attempts to remove idle nodes—that is, nodes with no running pods

GPU autoscaling



Steps

- ✓ Ensure an Amazon EKS cluster has been created
- ✓ Label your GPU nodes
- ✓ Install the Nvidia device plugin
- ✓ Install Helm
- ✓ Install Prometheus and GPU Node Exporter
- ✓ Install the Prometheus adapter to generate custom metrics
- ✓ Deploy the GPU stress-testing application
- ✓ Configure Horizontal Pod Autoscaler (HPA)
- ✓ Test the scaling
- ✓ Configure Cluster Autoscaler (CA)
- ✓ Test the scaling

Demo

Thank you!

Pradyumna Dash

pradyd@amazon.co.uk

Please complete the session survey in the mobile app.