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

BOA323

Demystifying Kubernetes for machine learning practitioners

Randy DeFauw
Senior Principal Solutions Architect
AWS

Sina Sojoodi
Principal Solutions Architect
AWS

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

What we'll cover today

- Why containers are great for machine learning (ML) projects
- What's the big deal about Kubernetes (K8s)?
- Should you use Kubeflow?

10 minutes

- Demo

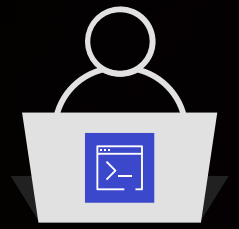
5 minutes

- Q&A

aws

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

A quick poll by show of hands



Who is here today?





Data scientist, data engineer, or ML engineer



Application developer, full-stack engineer, or enterprise architect



Product manager or business stakeholder

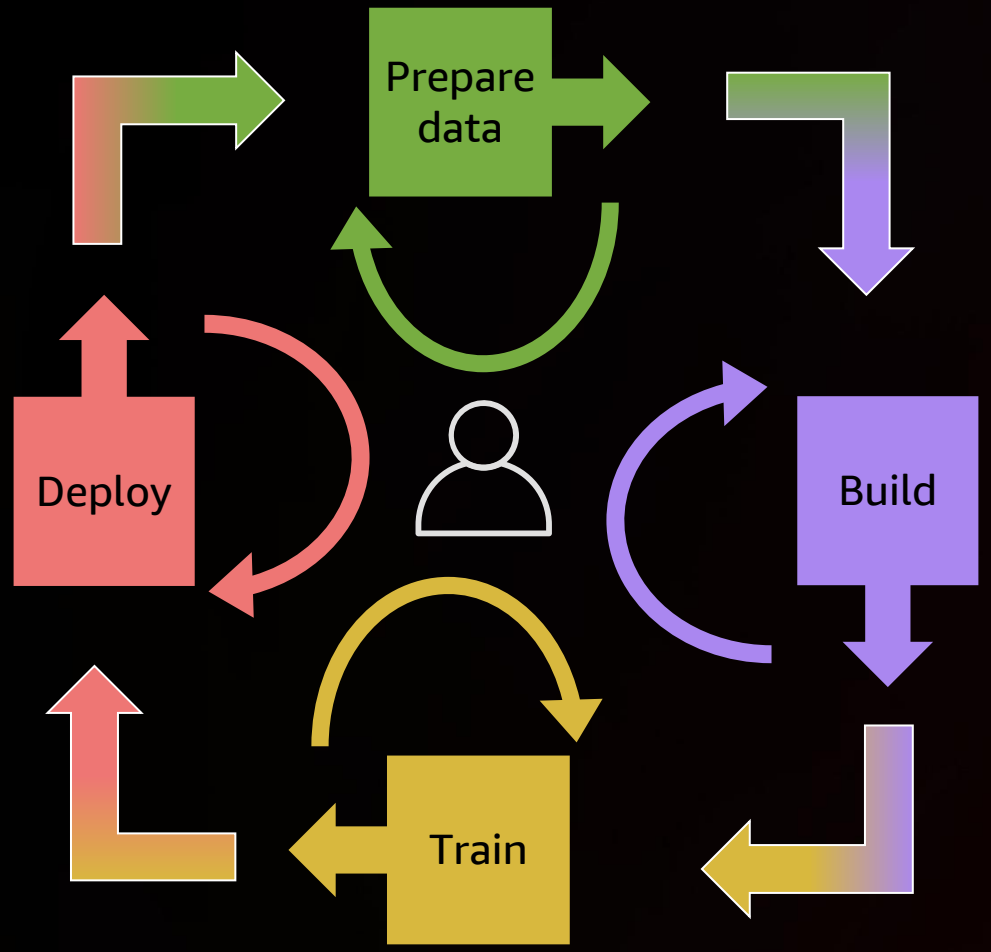


Other

aws

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

Key challenges in ML workflows



- ❑ Reproducibility
- ❑ Efficient utilization
- ❑ Scalability

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

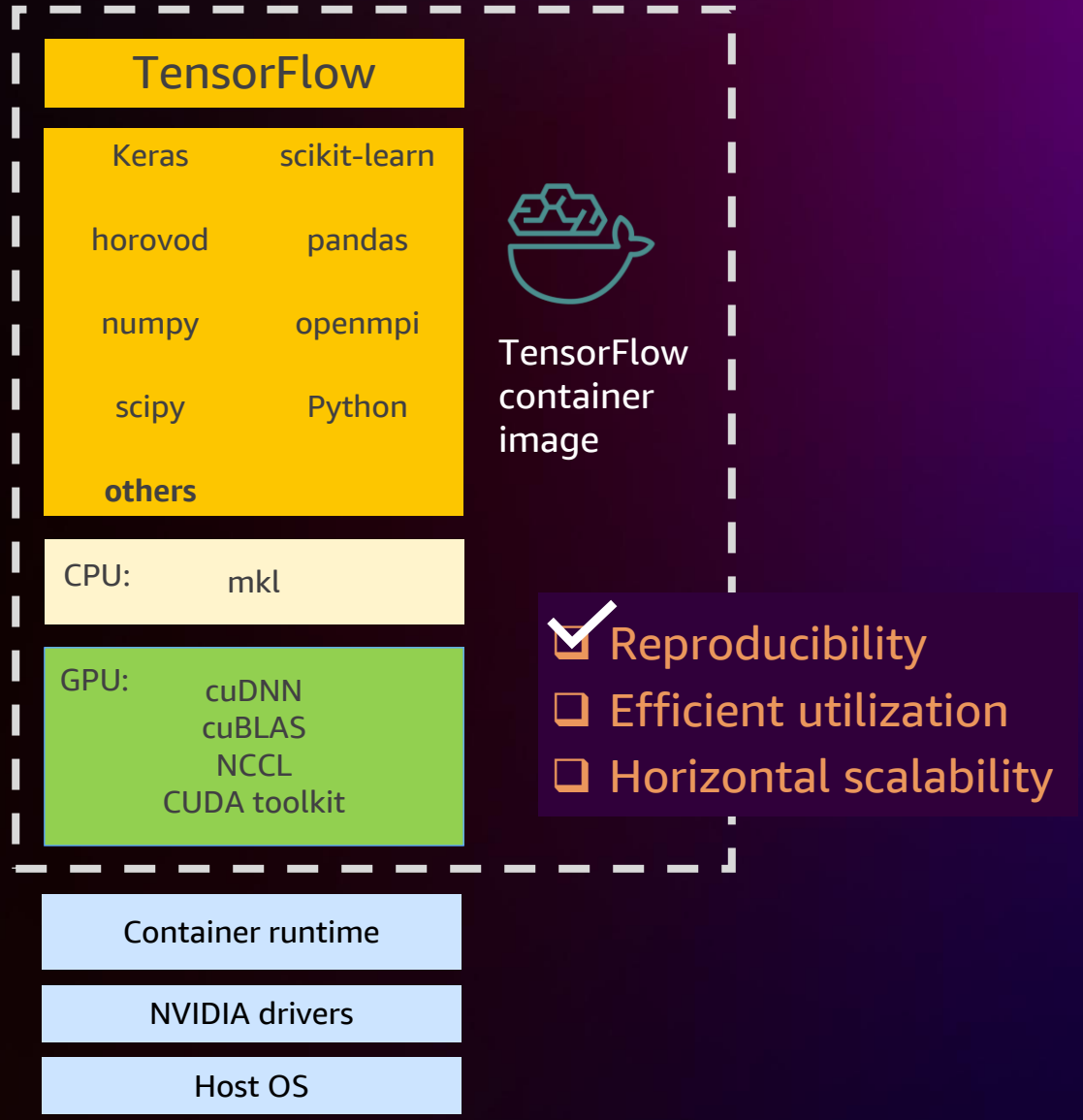
Containers for custom ML environments

Containers can package

- Training code
- Dependencies
- Configurations

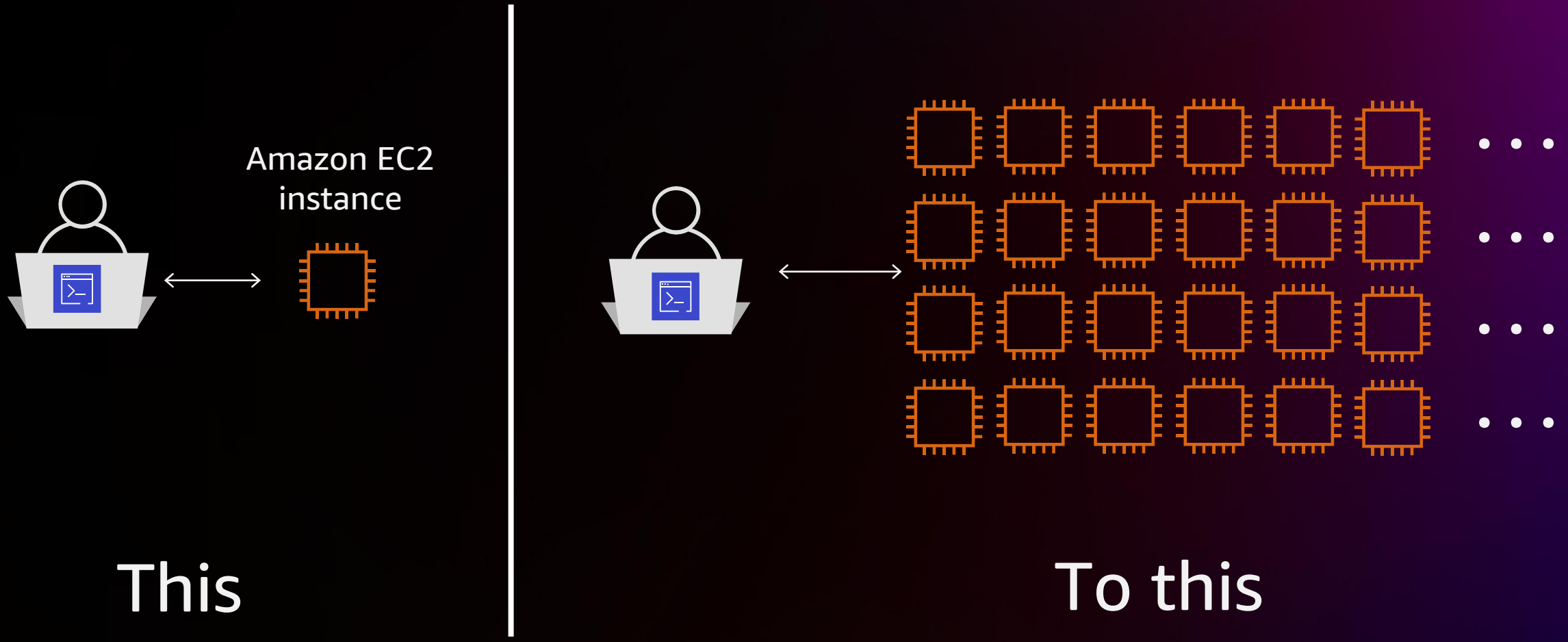
Benefits

- Dependencies packaged in
- Portable: train somewhere, deploy anywhere
- Inference pipelines



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

How do we scale ML?



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

ML infrastructure and cluster management

Management
Deployment, scheduling, scaling, and management of containerized applications



Amazon ECS



Amazon EKS

Compute
Where the containers run



Amazon EC2

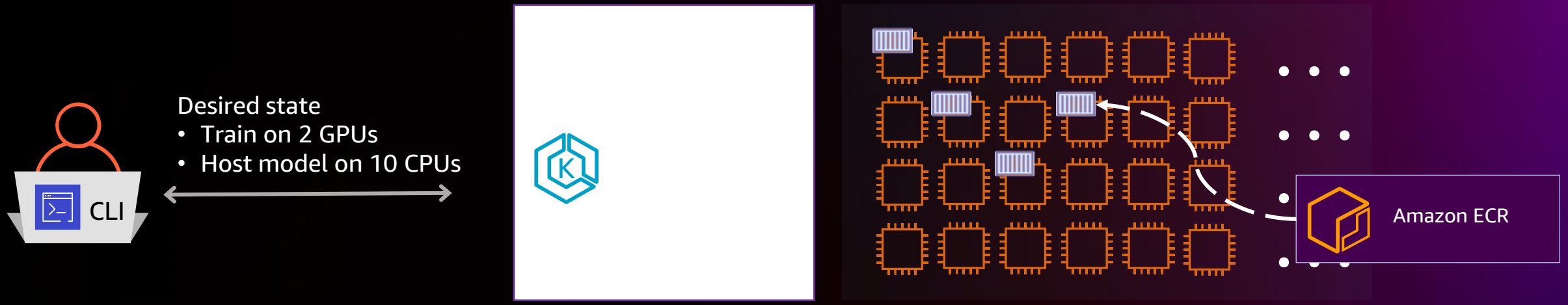
Image registry
Container image repository



Amazon ECR

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

Kubernetes for ML



Operating system for the cluster

- ✓ Better utilization of CPU and GPU resources on the cluster
- ✓ Automatic scaling to heterogeneous GPUs and multi-GPU nodes
- ✓ Integrate with applications deployed on K8s

- ☑ Reproducibility
- ☑ Efficient utilization
- ☑ Horizontal scalability

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

Kubernetes + Kubeflow for ML



Kubeflow

jupyter

Model development

APACHE Spark

Data prep

TensorFlow
PYTORCH
mxnet

Model training

Model tracking

TensorRT

Model deployment

Katib

Hyperparam tuning

Amazon EKS

https://awslabs.github.io/kubeflow-manifests/

PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc. TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

aws

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

Scaling ML on Kubernetes with Amazon SageMaker



Kubeflow

1

Amazon SageMaker components for Kubeflow Pipelines (KFP)

github.com/kubeflow/pipelines/tree/master/components/aws/sagemaker



Kubernetes

2

Amazon SageMaker Operators for Kubernetes

github.com/aws/amazon-sagemaker-operator-for-k8s

Webinar: “Scaling machine learning on Kubernetes and Kubeflow with SageMaker,” https://www.youtube.com/watch?v=LKmkiUdhV58

aws

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

Thank you!

Randy DeFauw
Linked In: @randalldefauw

Sina Sojoodi
Linked In & Twitter: @sinasojoodi



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

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