Generative AI on AWS

Self-managed and Managed ML workloads

Kanwaljit Khurmi

Principal AI/ML Solutions Architect

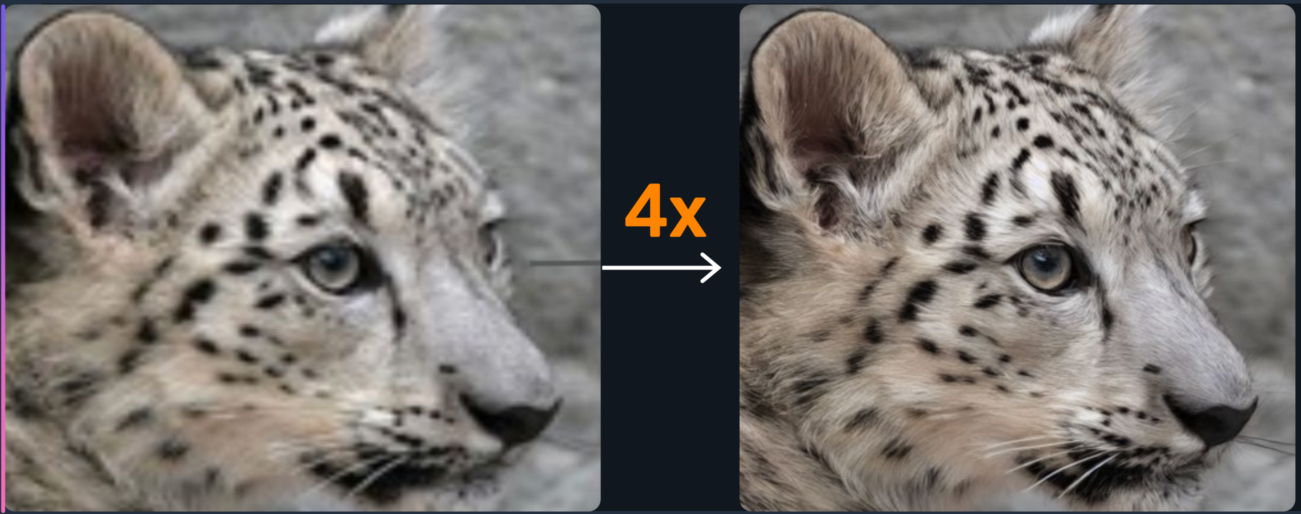
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

GenAI is transforming AI

Image generation, transformation, upscaling

Text to Image: Generated by Stable Diffusion 2.0 This interior does not exist

Seamless transformation



Upscaling

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges with training large-scale models

Hardware

Health checks

Orchestration

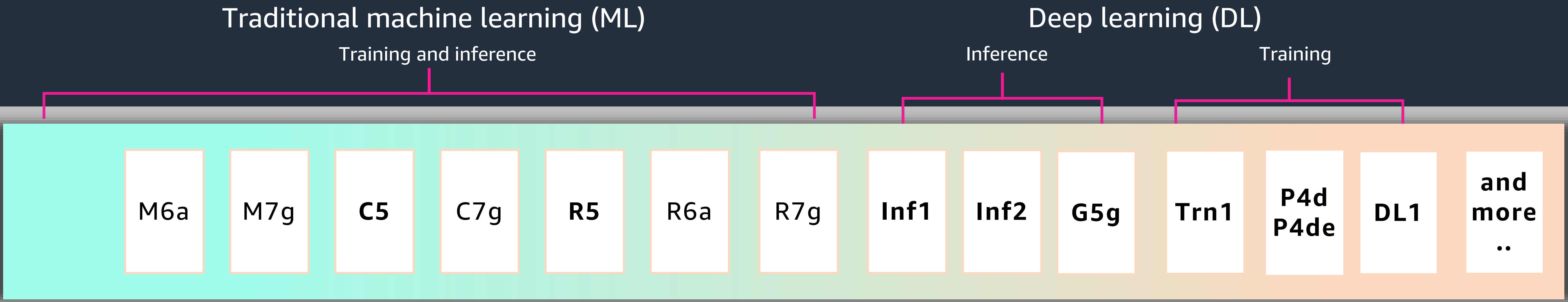
Data

Scaling up

Cost

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS AI/ML stack for Self-Managed and Managed Envs



Traditional machine learning (ML)

Training and inference

Deep learning (DL)

Inference

Training

M6a | M7g | C5 | C7g | R5 | R6a | R7g | Inf1 | Inf2 | G5g | Trn1 | P4d P4de | DL1 | and more ..

intel

AMD
EPYC CPU

annapurnalabs an amazon company
Graviton CPU
Inferentia and Trainium chip

NVIDIA
A100, V100, T4 GPUs

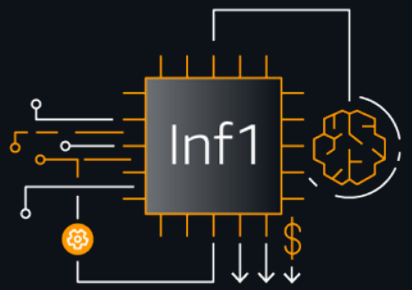
AWS Nitro Enclaves

Elastic Fabric Adapter

Amazon FSx for Lustre

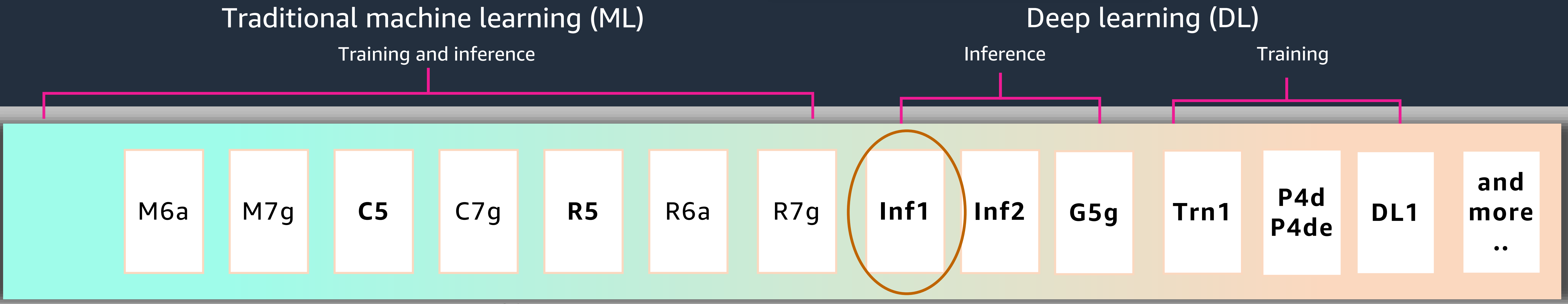
Broadest and deepest compute for AI/ML

AWS Inferentia1



Lowest cost inference in the cloud for running deep learning models

Up to 70% lower cost per inference


Traditional machine learning (ML)

Training and inference

Deep learning (DL)

Inference

Training

M6a

M7g

C5

C7g

R5

R6a

R7g

Inf1

Inf2

G5g

Trn1

P4d
P4de

DL1

and more
..

intel

AMD
EPYC CPU

annapurnalabs
an amazon company
Graviton CPU
Inferentia and Trainium chip

NVIDIA
A100, V100, T4 GPUs

AWS Nitro Enclaves

Elastic Fabric Adapter

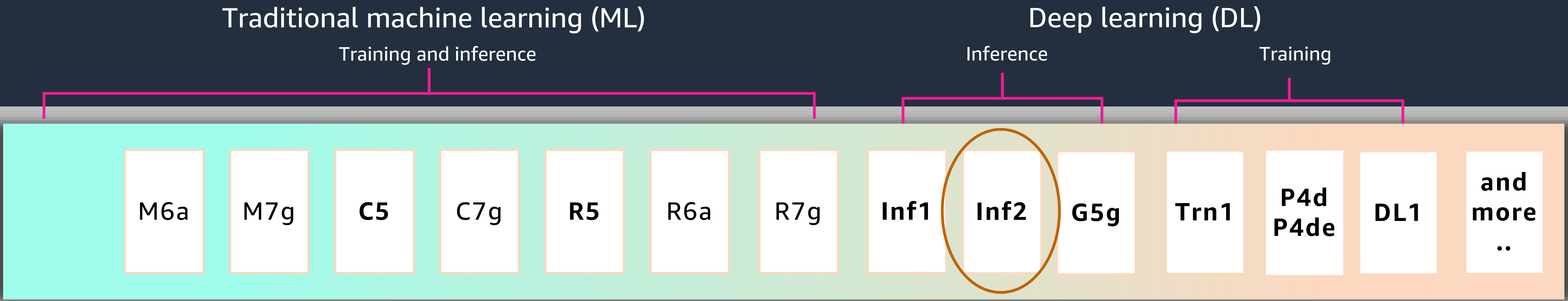
FSx
Amazon FSx for Lustre

Broadest and deepest compute for AI/ML

Up to 40% better price performance for Generative AI

AWS Inferentia2

High performance at the lowest cost per inference for LLMs and diffusion models



intel

AMD
EPYC CPU

annapurnalabs
an amazon company
Graviton CPU
Inferentia and Trainium chip

NVIDIA
A100, V100, T4 GPUs

AWS Nitro Enclaves

Elastic Fabric Adapter

Amazon FSx for Lustre

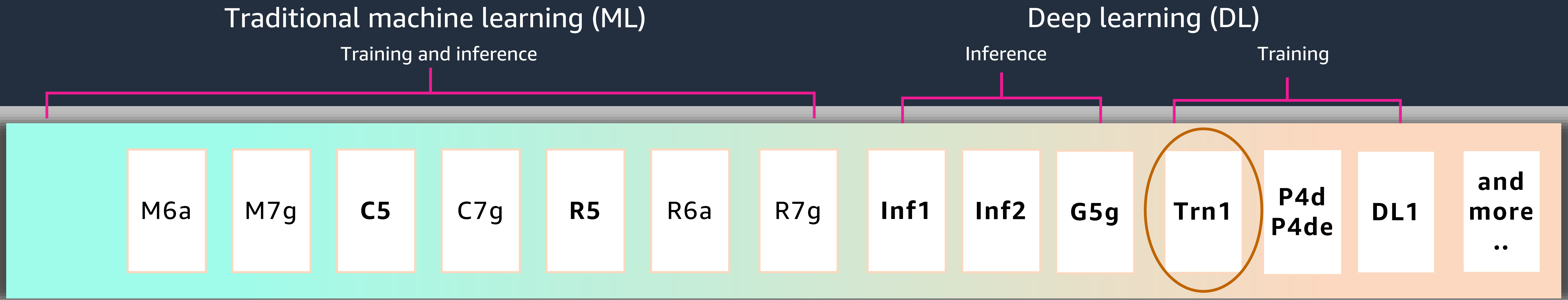
Broadest and deepest compute for AI/ML

AWS Trainium

Up to 50% cost-to-train savings

Trn1

The most cost efficient for high-performance training of LLMs and diffusion models



Traditional machine learning (ML)

Training and inference

Deep learning (DL)

Inference

Training

M6a | M7g | C5 | C7g | R5 | R6a | R7g | Inf1 | Inf2 | G5g | Trn1 | P4d P4de | DL1 | and more ..

EPYC CPU

annapurnalabs an amazon company

Graviton CPU

Inferentia and Trainium chip

A100, V100, T4 GPUs

AWS Nitro Enclaves

Elastic Fabric Adapter

Amazon FSx for Lustre

Broadest and deepest compute for AI/ML

Amazon EC2 | AWS ParallelCluster | Amazon EKS | Amazon ECS | AWS Fargate | more …. | Amazon SageMaker

Self-Managed → Managed

Broadest and deepest compute services for AI/ML

M6a | M7g | C5 | C7g | R5 | R6a | R7g | Inf1 | Inf2 | G5g | Trn1 | P4d P4de | DL1 | and more ..

intel

AMD

EPYC CPU

annapurnalabs
an amazon company

Graviton CPU
Inferentia and Trainium chip

NVIDIA

A100, V100, T4 GPUs

AWS Nitro Enclaves

Elastic Fabric Adapter

Amazon FSx for Lustre

Broadest and deepest compute for AI/ML

AWS Deep Learning AMIs

AWS Deep Learning Containers

Amazon EC2

AWS ParallelCluster

Amazon EKS

Amazon ECS

AWS Fargate

more ....

Amazon SageMaker

Self-Managed → Managed

Broadest and deepest compute services for AI/ML

M6a	M7g	**C5**	C7g	**R5**	R6a	R7g	**Inf1**	**Inf2**	**G5g**	**Trn1**	**P4d P4de**	**DL1**	**and more ..**

intel

AMD EPYC CPU

annapurnalabs an amazon company Graviton CPU Inferentia and Trainium chip

NVIDIA A100, V100, T4 GPUs

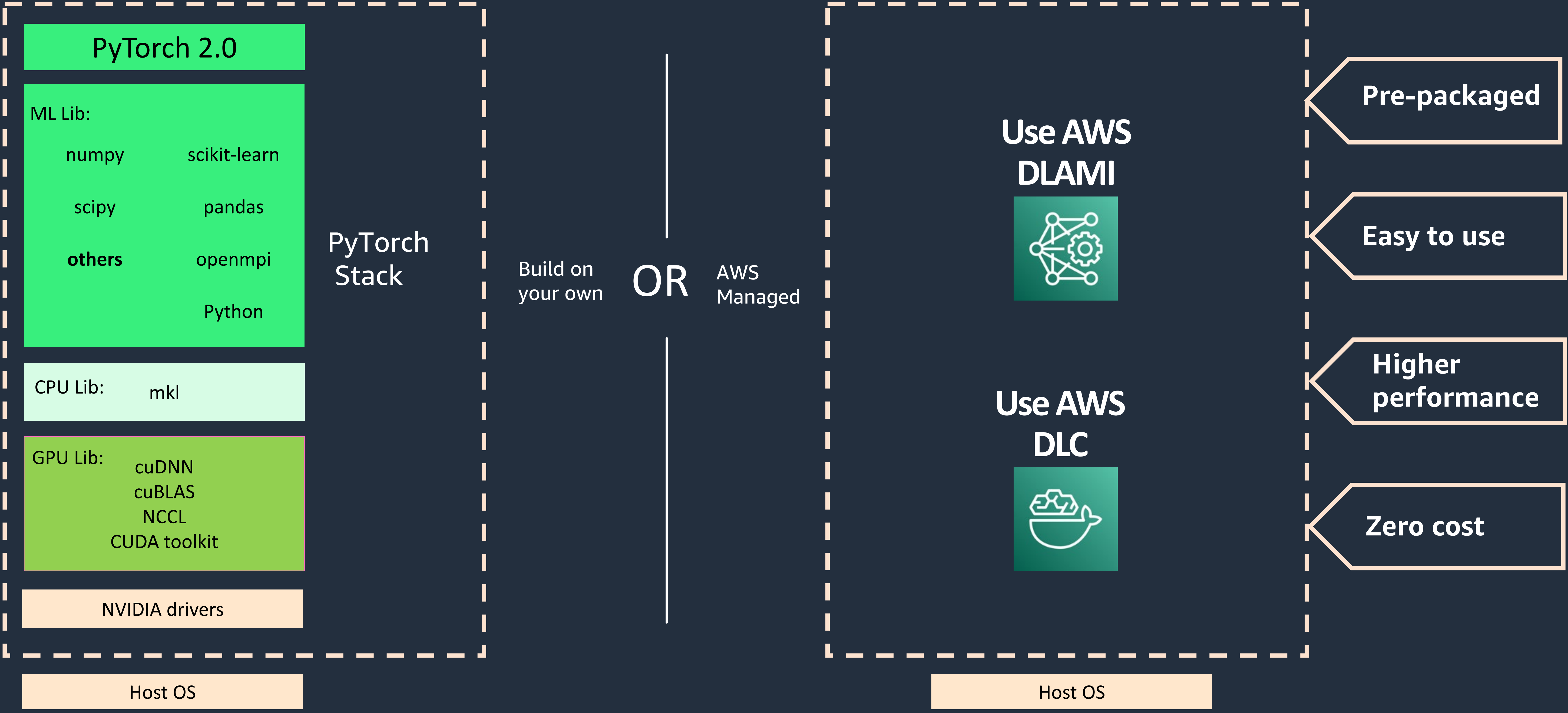
AWS Nitro Enclaves

Elastic Fabric Adapter

Amazon FSx for Lustre

Broadest and deepest compute for AI/ML

Why use DLAMI and DLC ?



PyTorch 2.0

ML Lib:

numpy

scikit-learn

scipy

pandas

others

openmpi

Python

CPU Lib: mkl

GPU Lib:

cuDNN

cuBLAS

NCCL

CUDA toolkit

NVIDIA drivers

PyTorch Stack

Host OS

Build on your own

OR

AWS Managed

Use AWS DLAMI

Use AWS DLC

Host OS

Pre-packaged

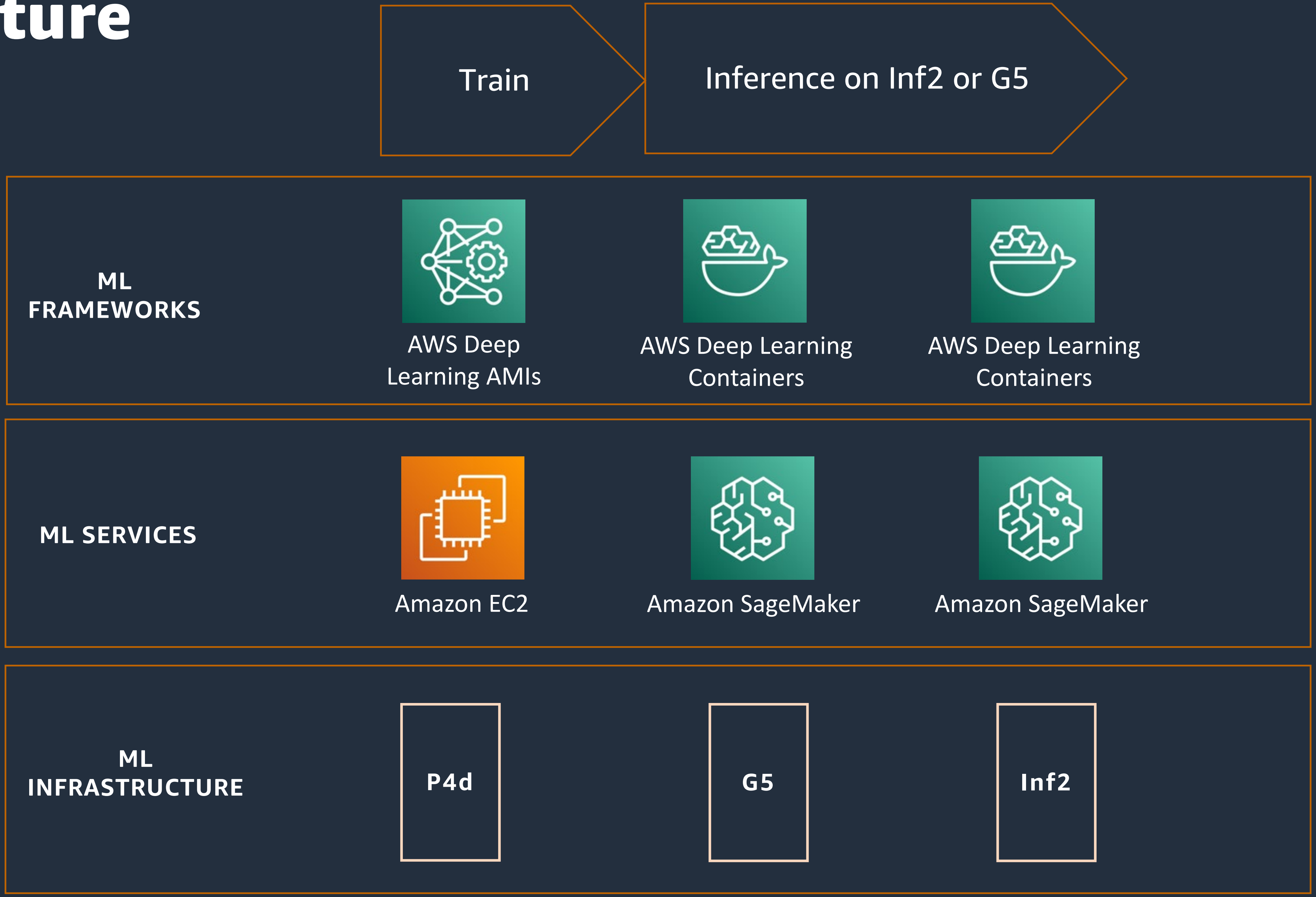
Easy to use

Higher performance

Zero cost

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Train and deploy Stable Diffusion model using Hybrid Archtecture



	Train	Inference on Inf2 or G5	
ML FRAMEWORKS	AWS Deep Learning AMIs	AWS Deep Learning Containers	AWS Deep Learning Containers
ML SERVICES	Amazon EC2	Amazon SageMaker	Amazon SageMaker
ML INFRASTRUCTURE	P4d	G5	Inf2

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

References

- AWS Neuron Documentation: https://awsdocs-neuron.readthedocs-hosted.com/en/latest/index.html
- Hugging Face Optimum Neuron: https://huggingface.co/docs/optimum-neuron/index
- AWS DLC: https://docs.aws.amazon.com/deep-learning-containers/
- AWS DLAMI: https://docs.aws.amazon.com/dlami/

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Kanwaljit Khurmi
kkhurmi@amazon.com

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Amazon Web Services, AWS, the Powered by AWS logo, and all AWS service names used in this slide deck are trademarks of Amazon.com, Inc. or its affiliates.