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

CMP330

AWS managed services to run ML training workloads at scale with AWS Trainium

Scott Perry

Sr. Solutions Architect – Annapurna Labs ML
AWS

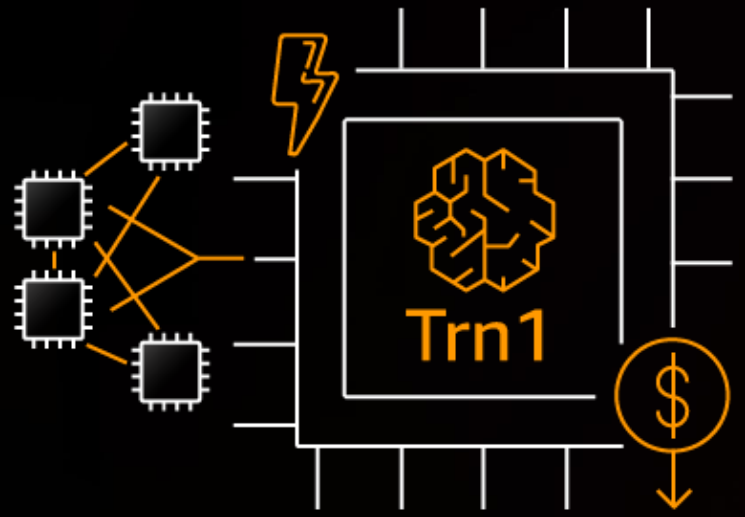
KC Tung

Sr. Solutions Architect – Annapurna Labs ML
AWS

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

Amazon EC2 Trn1 instances

COST-EFFECTIVE DL TRAINING IN THE CLOUD



60% higher accelerator memory (vs. P4d)

2x network bandwidth (vs. P4d)

Native support for PyTorch and TensorFlow

Train on Trn1 and deploy anywhere

Instance size	vCPUs	Trainium accelerators	Accelerator memory (GB)	NeuronLink	Instance memory (GiB)	Instance networking (Gbps)	Local instance storage
Trn1.2xlarge	8	1	32	N/A	32	Up to 10	500 GB NVMe
Trn1.32xlarge	128	16	512	Yes	512	800 Gbps (1.6 TBps in Q1 '23)	8 TB NVMe

aws

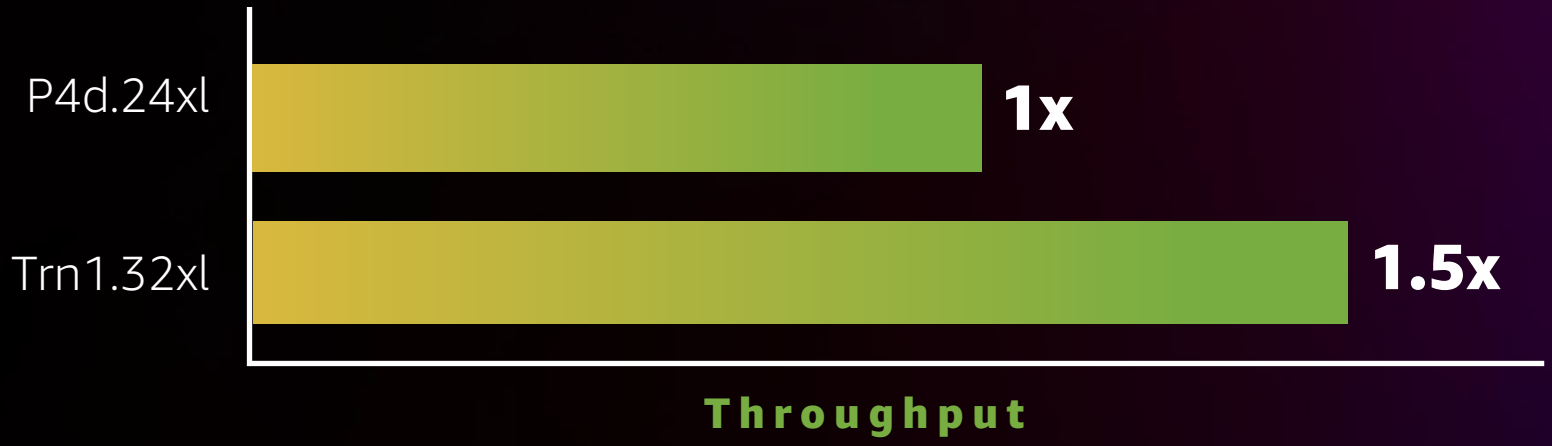
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best cost to train with Trn1

AMAZON EC2 TRN1 INSTANCES DELIVER THE BEST TRAINING PRICE-PERFORMANCE

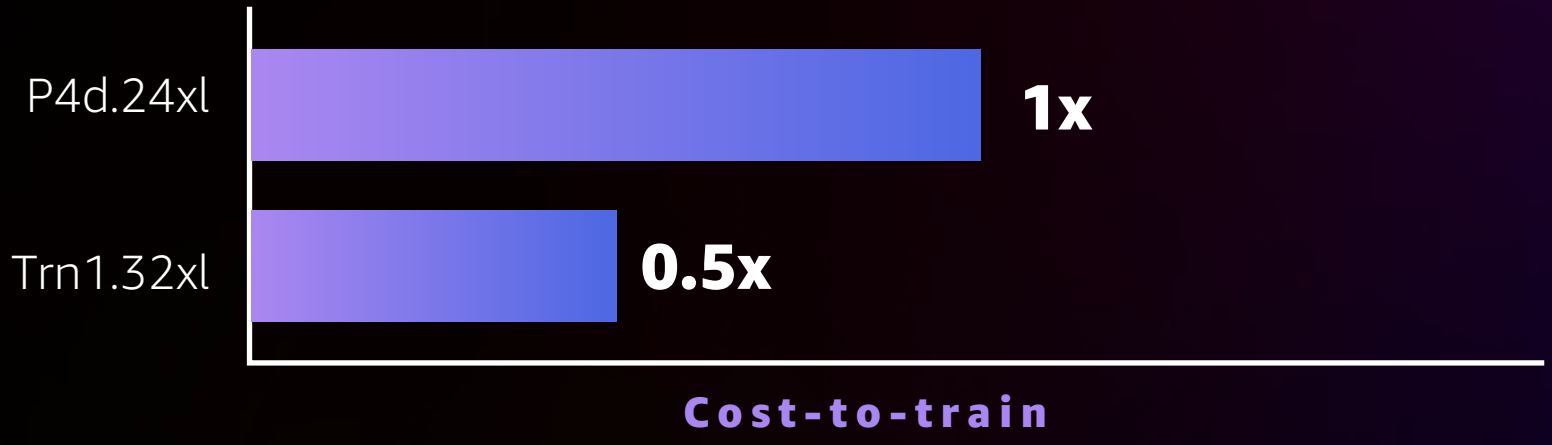
1.5x

Higher throughput



50%

Lower cost

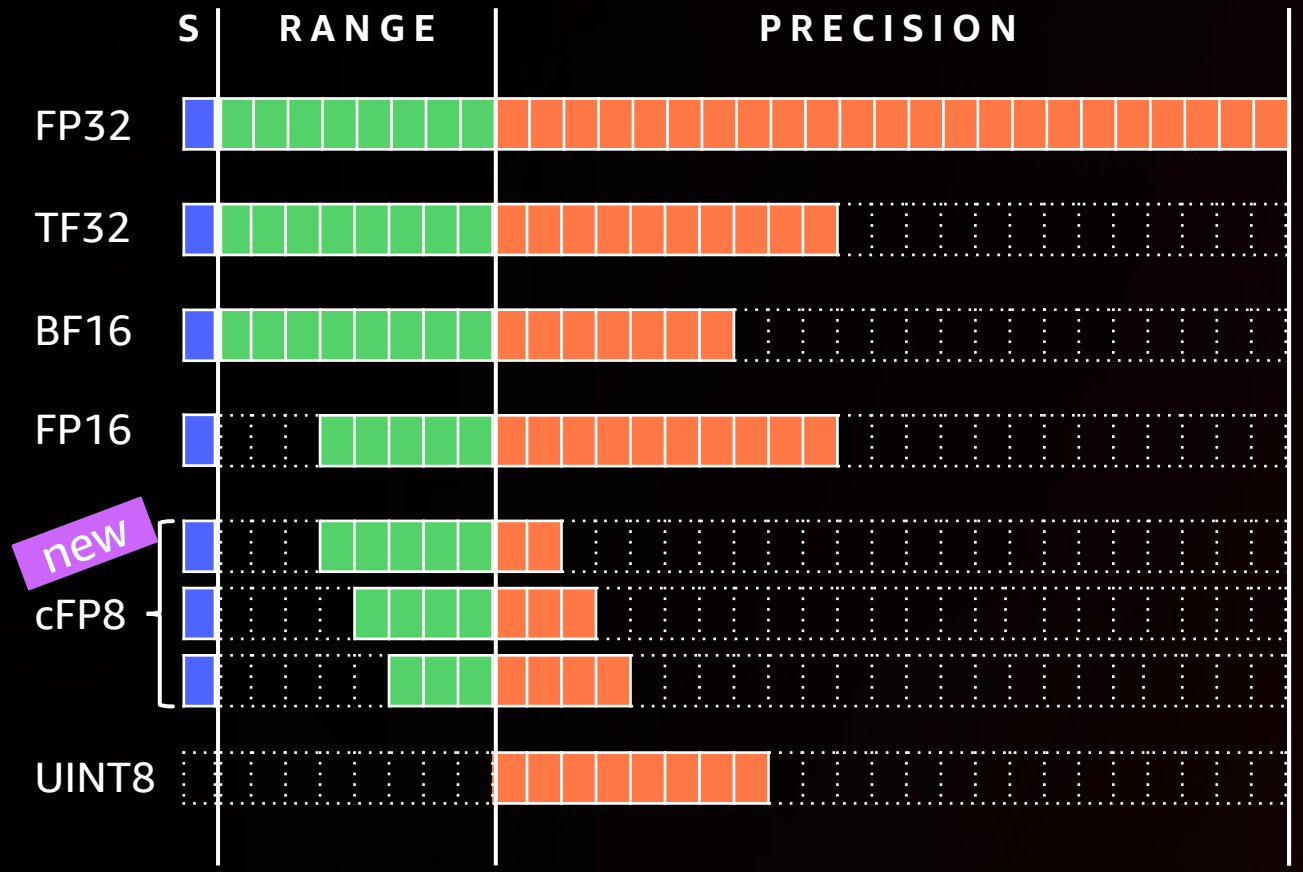


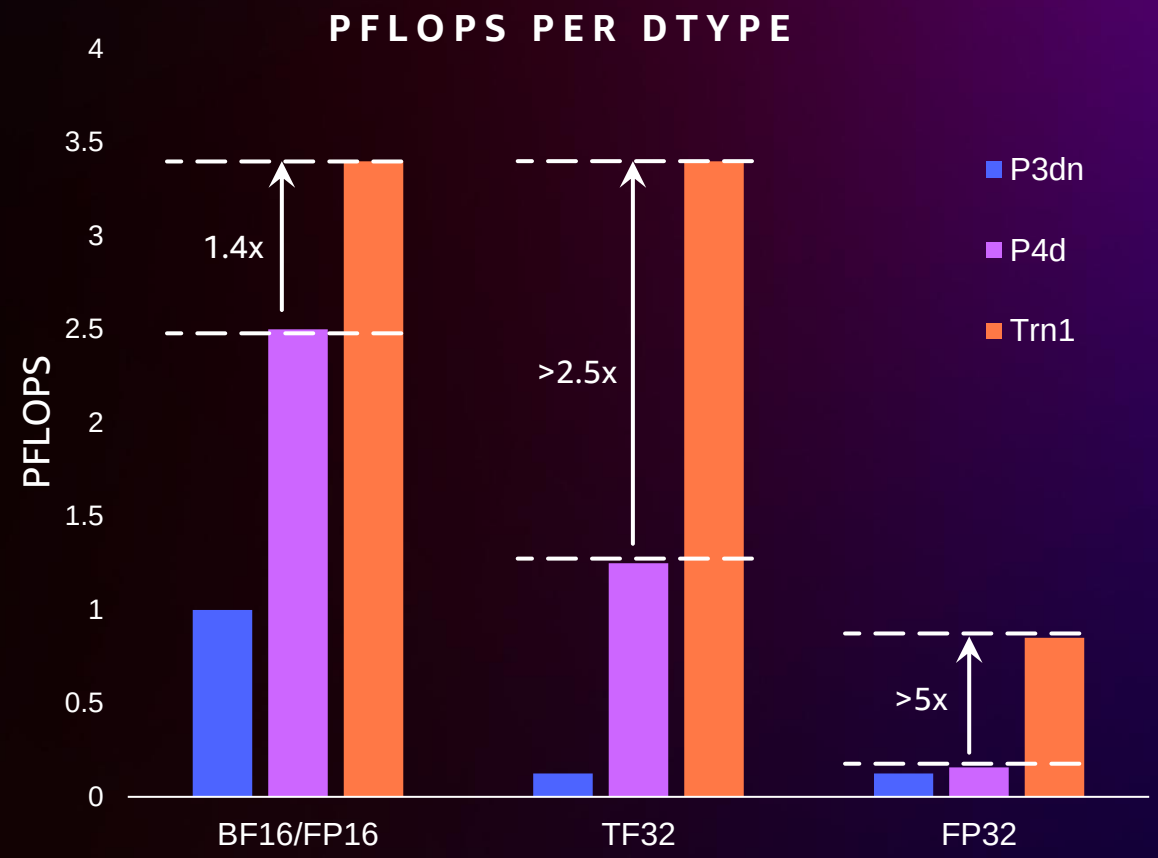
* Measured on PyTorch BERT-Large

aws

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

Rich data-type selection





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

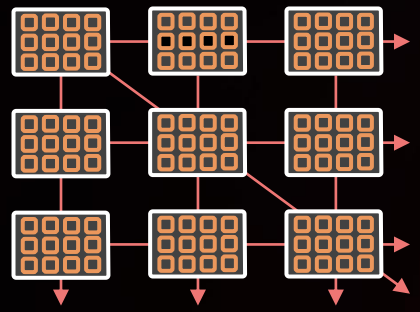
AWS Neuron for Inferentia and Trainium

EASY-TO-USE DEEP LEARNING SOFTWARE DEVELOPMENT KIT



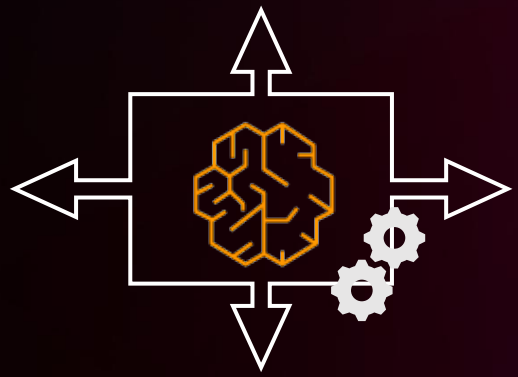
Framework integration

Full framework integration, JIT, eager debug mode, collective communication



Distributed training

Scale up to 30K+ devices, integration with distributed training libraries, and EFA



Flexible and extendable

Support for custom ops, dynamic shapes, new data types, and stochastic rounding



Cloud native

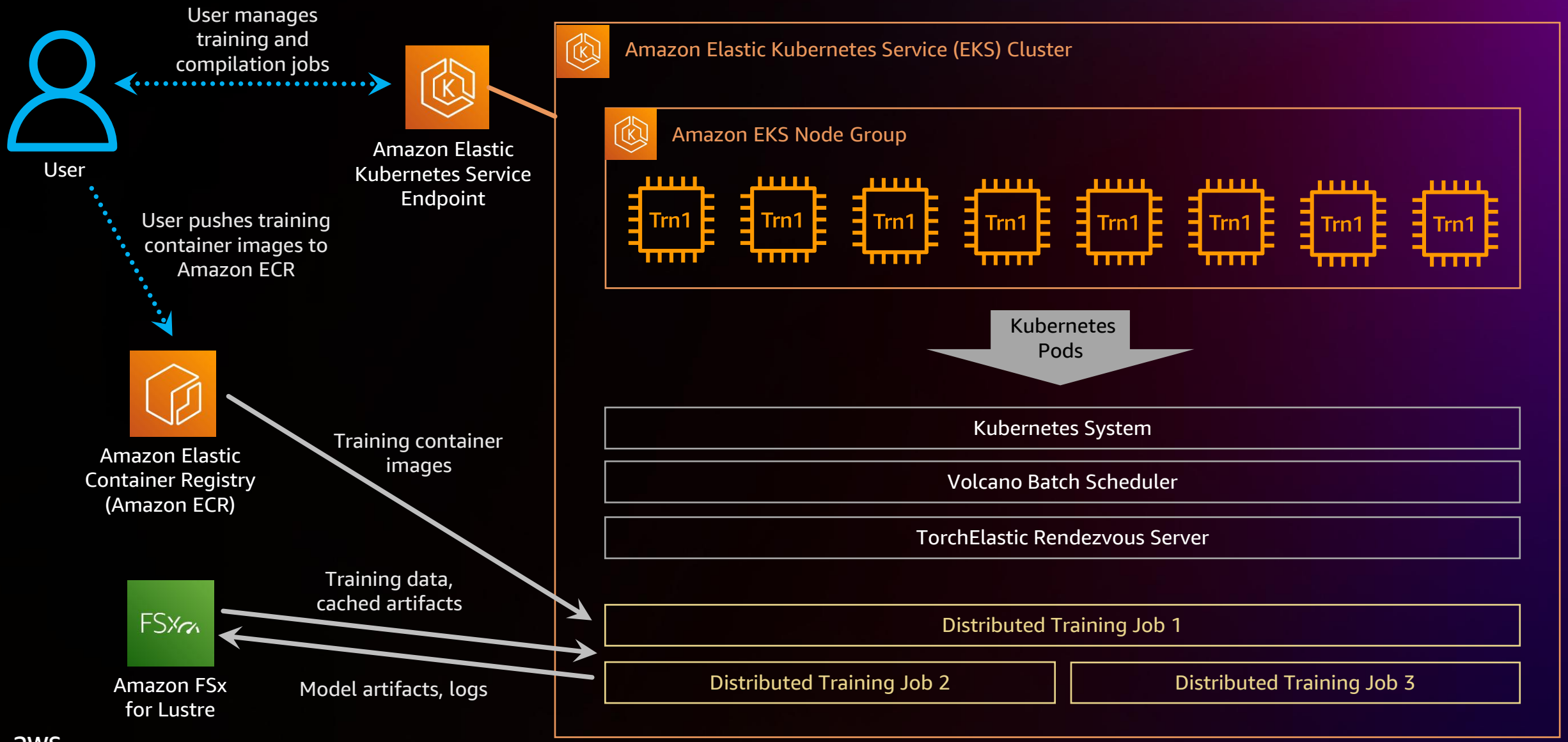
Amazon SageMaker, Amazon EKS, Amazon ECS, AWS ParallelCluster, AWS Deep Learning Containers, and AWS Deep Learning AMIs (DLAMI)

aws

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

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

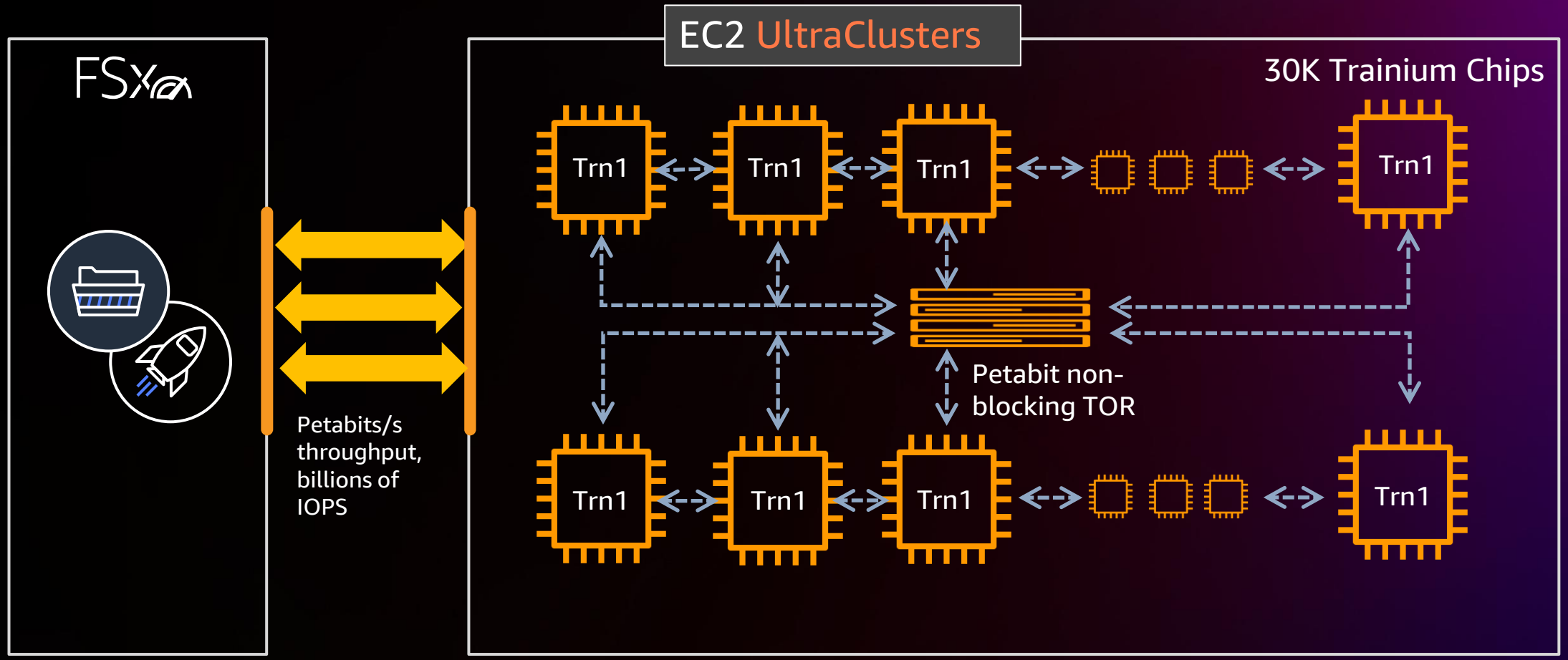
Distributed training with AWS Trainium and EKS



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

UltraCluster scale out for ultra-large models

30K TRAINIUM ACCELERATORS, 6 EXAFLOPS



On-demand access to a world-class supercomputer

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

Thank you!



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

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