
AWS
re:Invent

CMP304

AWS infrastructure for large-scale distributed ML training

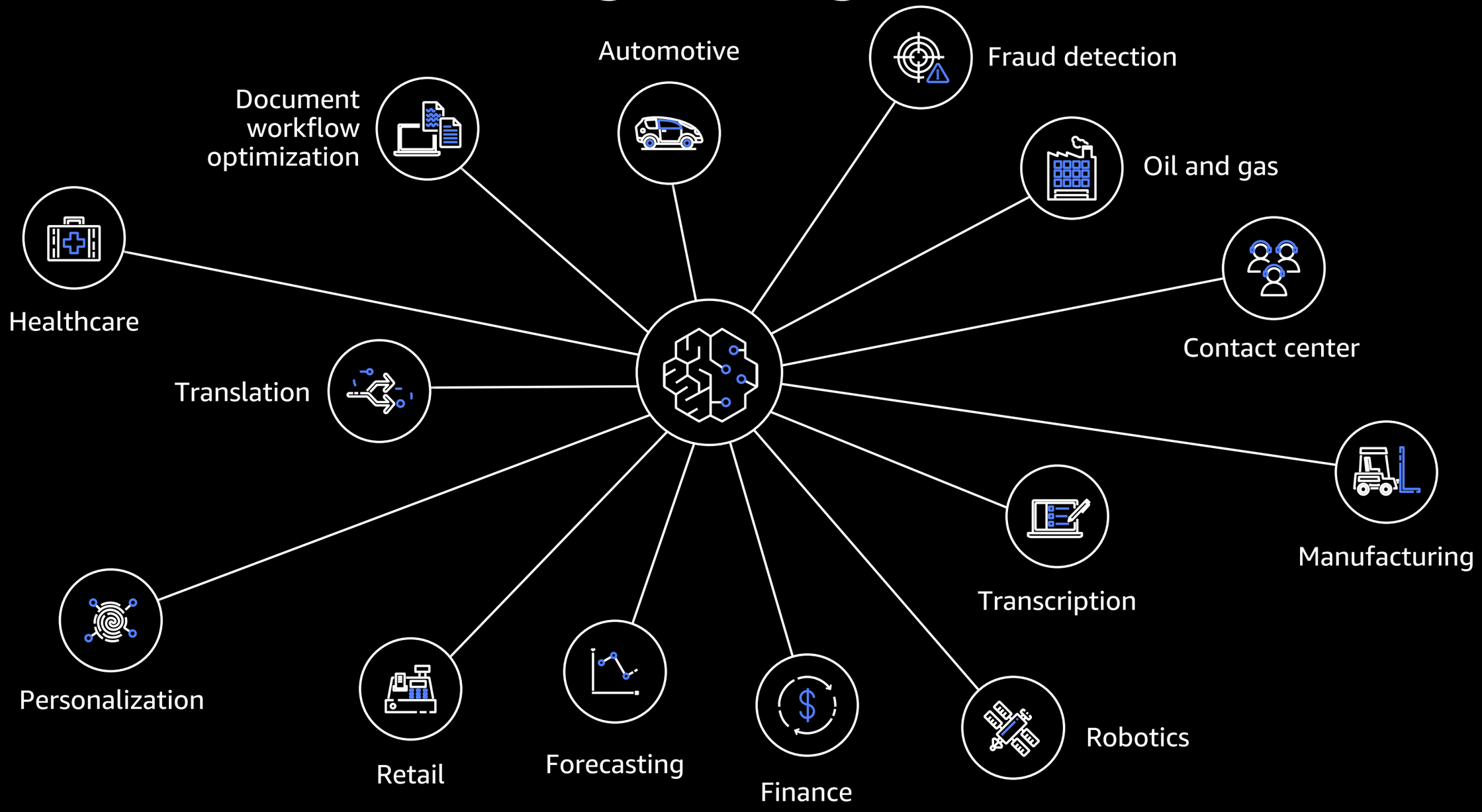
James Jeun
Senior Product Manager
AWS

Amr Ragab
Senior Solutions Architect
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The reach of ML is growing



Our mission at AWS

Put machine learning in the
hands of every developer

The AWS ML Stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI SERVICES

VISION
- Amazon Rekognition

SPEECH
- Amazon Polly
- Amazon Transcribe *+Medical*

TEXT
- Amazon Comprehend *+Medical*
- Amazon Translate
- Amazon Textract

SEARCH
- Amazon Kendra

CHATBOTS
- Amazon Lex

PERSONALIZATION
- Amazon Personalize

FORECASTING
- Amazon Forecast

FRAUD
- Amazon Fraud Detector

DEVELOPMENT
- Amazon CodeGuru

CONTACT CENTERS
- Contact Lens *For Amazon Connect*

ML SERVICES

- Amazon SageMaker
- Ground Truth
- AWS Marketplace for ML

SageMaker Studio IDE
- Built-in algorithms
- Notebooks
- Experiments
- Processing
- Model training & tuning
- Debugger
- Autopilot
- Model hosting
- Model Monitor

- Neo
- Augmented AI

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow · mxnet · PYTORCH

GLUON · Keras · learn · HOROVOD · DeepGraphLibrary

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Inferentia | FPGA

Why AWS for ML?

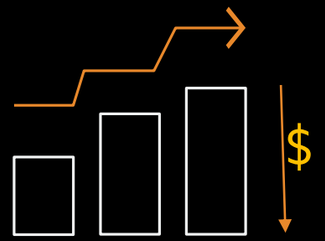


Broadest and deepest set of AI and ML services

Solutions for everyone from ML scientists to application developers

Freedom of choice between self-managed, managed, and ready-to-use API services

Supports broad open-source ecosystem



Many choices to optimize development costs for scaling ML workloads

Up to 70% cost reduction in data labeling

Custom accelerators for lowering inference costs by 54%

Significant cost reduction with Spot, Reserved Instances, and savings plans



Comprehensive cloud services offerings

Highly secure, reliable, fully featured data store

The strongest set of compute, storage, security, database, and analytics capabilities to build upon

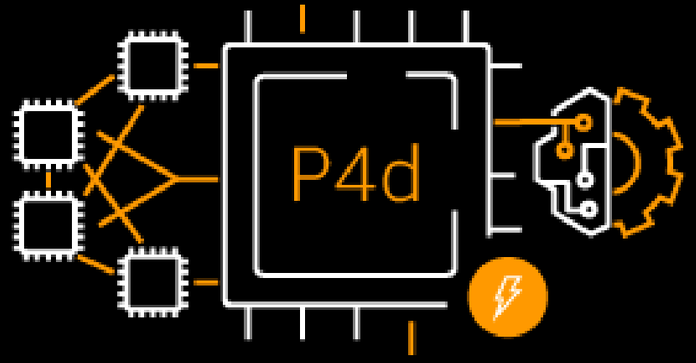
ML is happening in companies of every size and in every industry

10,000+ CUSTOMERS | MORE ML HAPPENS ON AWS THAN ANYWHERE ELSE



Introducing Amazon EC2 P4d instances

The most powerful GPU instances in the cloud



Up to 60% lower cost to train ML model

Up to 3x more deep learning performance

4x higher network bandwidth

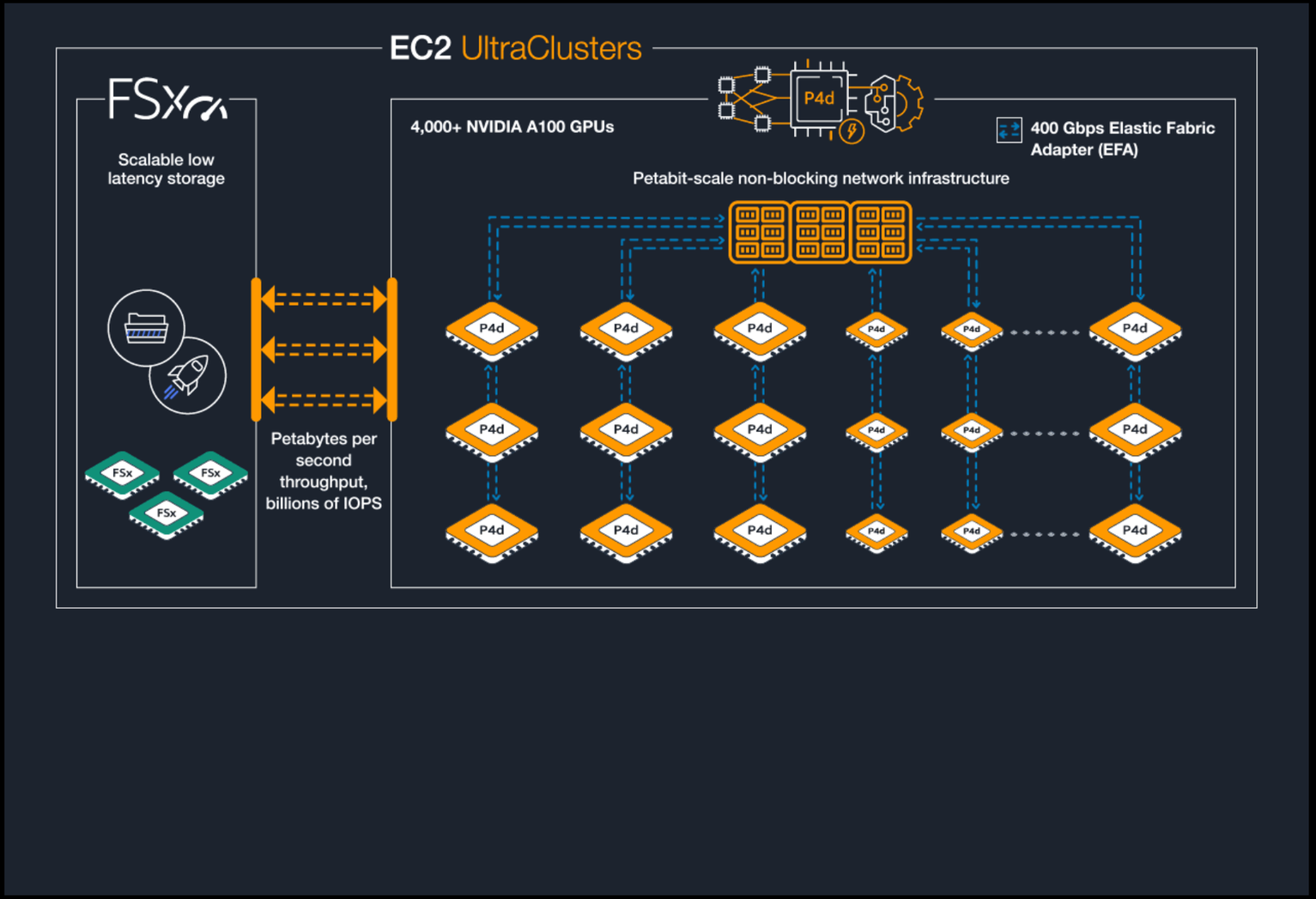
25% more GPU memory

4x more local storage

Instance size	vCPUs	Instance memory	GPUs	GPU memory	GPU memory BW	GPU interconnect	Instance networking	Local instance storage
P4d.24xl	96	1152 GB	8x A100	320 GB	1555 (GB/s)	600 GB/s NVSwitch	400 Gbps EFA + GPUDirect RDMA	8 TB NVMe

EC2 UltraClusters of P4d instances

On-demand access to a world-class supercomputer

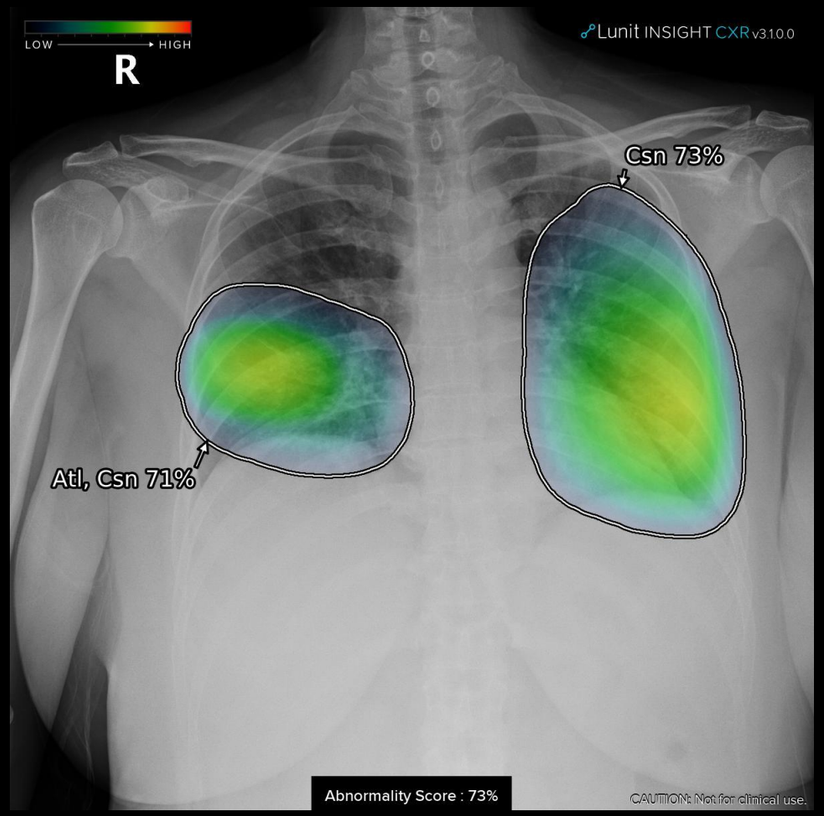


Over 4,000 A100 GPUs

Fully non-blocking petabit-scale network infrastructure

High-throughput, low-latency storage from Amazon FSx for Lustre

GE Healthcare uses AWS to accelerate medical imaging

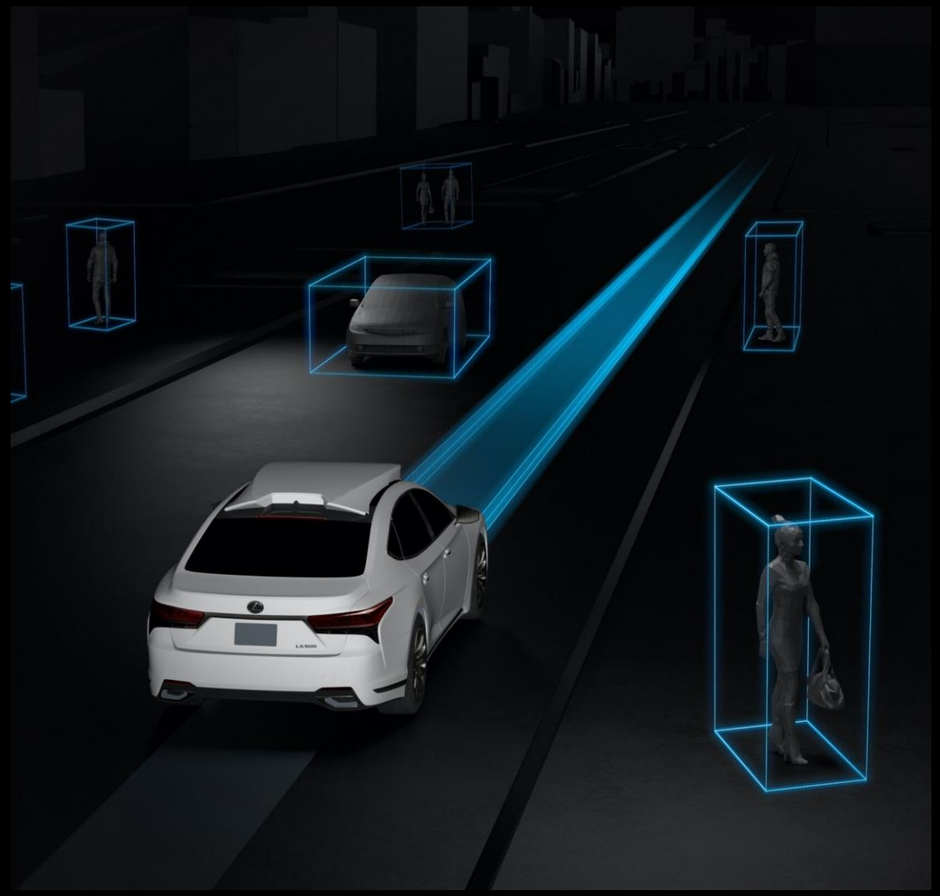


GE Healthcare

"Our medical imaging devices generate massive amounts of data that need to be processed by our data scientists. Using the new P4d instances reduced processing time from days to hours. We saw two- to three-times greater speed on training models with various image sizes, while achieving better performance with increased batch size and higher productivity with a faster model development cycle."

Karley Yoder
VP & GM, Artificial Intelligence
GE Healthcare

Toyota Research Institute uses AWS to accelerate autonomous driving

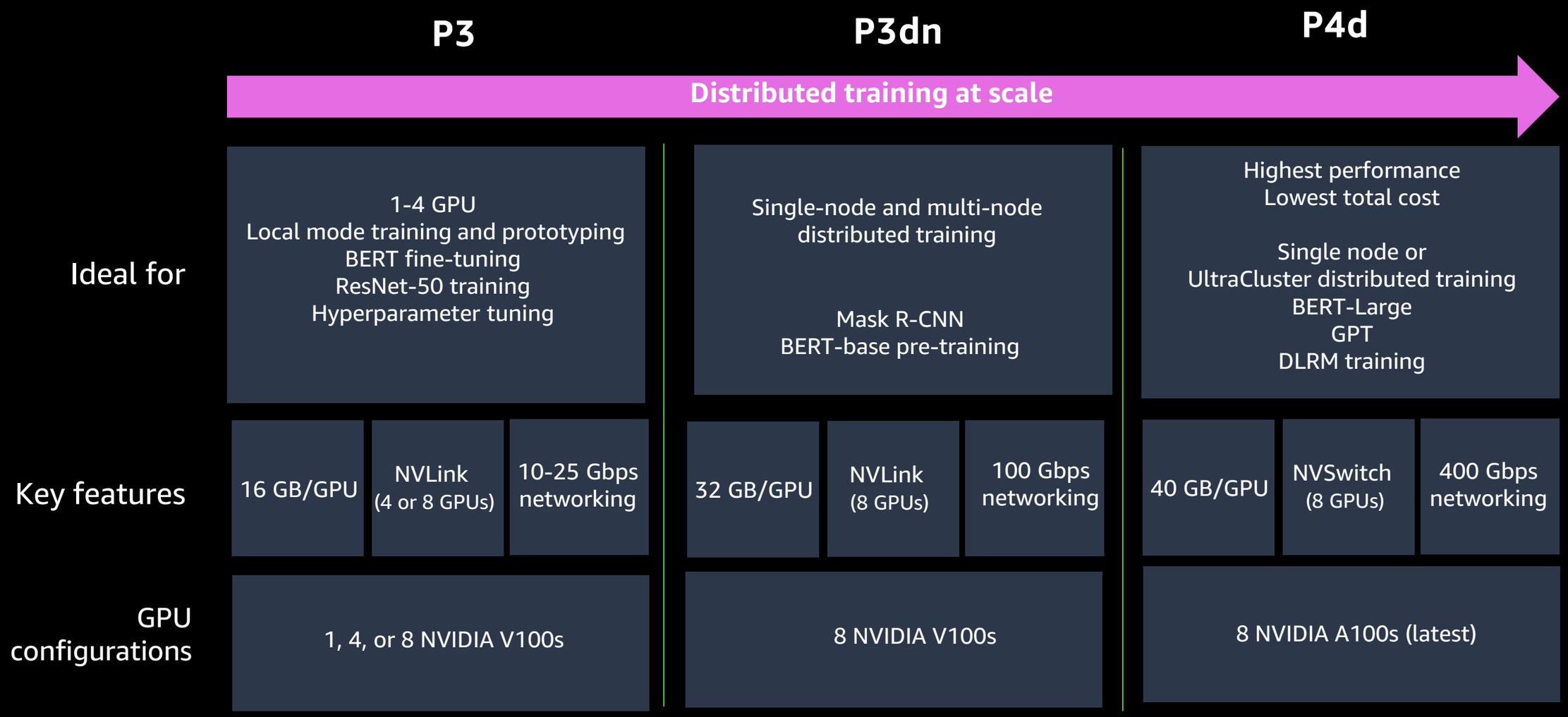


TRI-AD

"At TRI-AD, we're working to build a future where everyone has the freedom to move and explore with a focus on reducing vehicle injuries and fatalities on adaptive driving and smart city. Through the use of AWS P4d instances, we were able to reduce our training time for object recognition by 40% compared to previous generation GPU instances without any modification to existing codes."

Junya Inada
Director of Automated Driving
Toyota Research Institute – Advanced Development

Amazon EC2 P-family instances



Fastest networking in the cloud

EFA

Elastic Fabric Adapter, best for large ML and HPC workloads

Ultra-low latency for inter-instance communication

Best for tightly coupled ML and HPC workloads, such as NCCL / MPI

400 Gbps network bandwidth

< 15 microsecond network latencies

Scalable Reliable Datagram (SRD)

A RELIABLE HIGH-PERFORMANCE LOWER-LATENCY NETWORK TRANSPORT

Inspired by InfiniBand Reliable Datagram, without the drawbacks

- No limit on the number of outstanding messages per context

Out-of-order delivery – no head-of-line blocking

- Messages are independent in many cases; application / middleware can restore ordering only if and when needed

Packet spraying over multiple ECMP paths

- Fast and transparent recovery from network failures

Congestion control designed for large-scale cloud

- Maintains high throughput in the face of packet drops

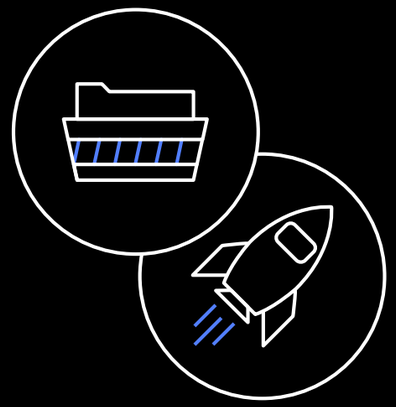
Multi-path routing

Equal-cost multi-path routing (ECMP) is a routing strategy where next-hop packet forwarding to a single destination can occur over **multiple “best paths”**

This can substantially increase bandwidth by load-balancing traffic over multiple paths

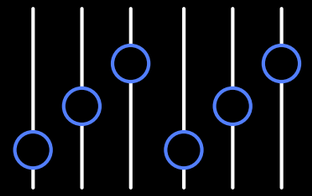
Thanks to Wikipedia and its contributors for the pithy explanation and Peter Ashwood-Smith for the snappy animated GIF explaining the concept

Amazon FSx for Lustre



High and scalable performance

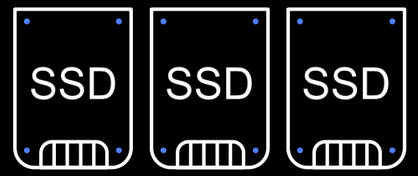
Parallel file system

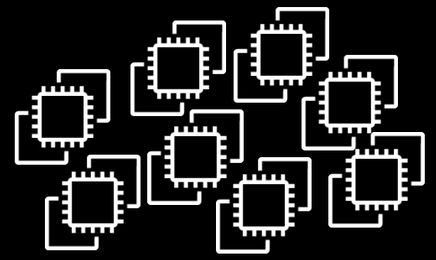




100+ GiB/s throughput
Millions of IOPS
Consistent sub-millisecond latencies

SSD-based

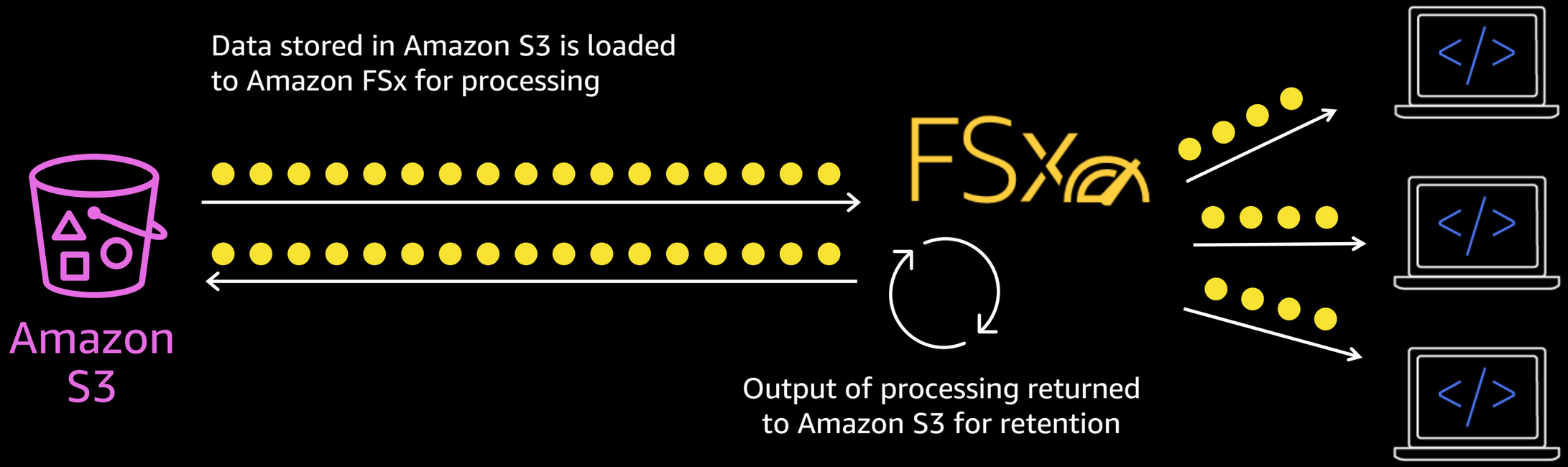




Supports concurrent access from hundreds of thousands of cores

Seamless integration with Amazon S3

Link your Amazon S3 dataset to your FSx for Lustre file system, and then …

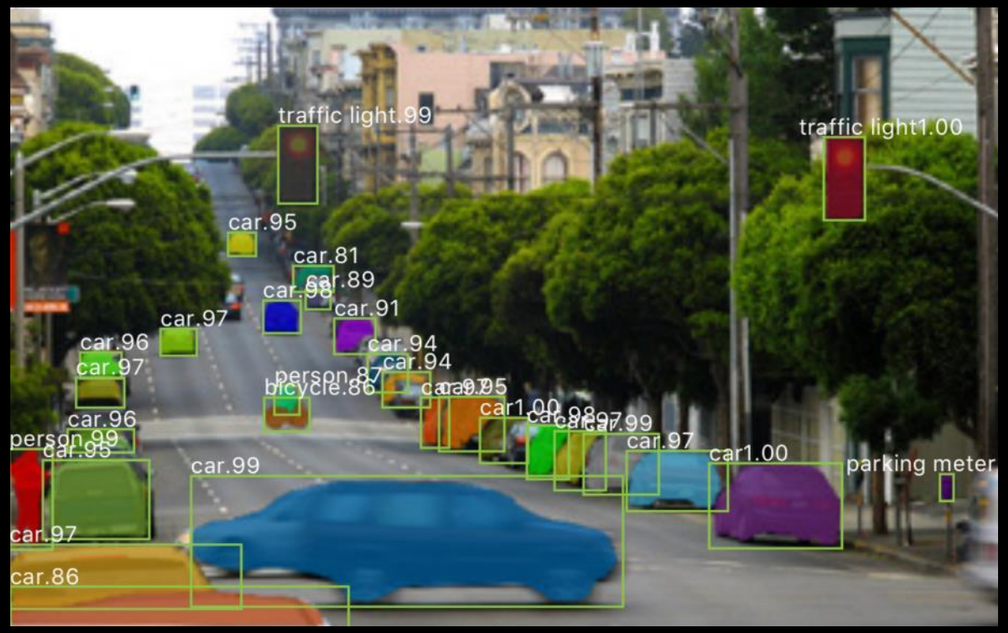


When your workload finishes, simply delete your file system

The need for training at scale

Model type	Number of parameters	Size of data	Estimated time to train
ResNet-50	23 million	300 GB	18 min – 128x V100
BERT-base	110 million	16 GB	12 days – 8x V100
BERT-Large	340 million	16 GB	3 days – 64x V100
RoBERTa	355 million	160 GB	1 day – 1024x V100
GPT-2-Large	1.5 billion	40 GB	Months – 100s of GPUs
GPT-3	175 billion	2,000 GB	44 years – 8x V100

Mask R-CNN scaling performance



re:Invent 2019

P3DN.24xlarge nodes	NVIDIA V100 GPUs	Time to train (PyTorch)
1	8	324 min
24	192	26 min

re:Invent 2020

P3DN.24xlarge nodes	NVIDIA GPUs	Time to train (PyTorch)
8	64	23 min
16	128	12 min

Using clusters of P3dn instances, each with

- 8x V100 32 GB GPUs
- 100 Gbps networking with EFA
- 300 GB/s NVLink

https://aws.amazon.com/blogs/machine-learning/amazon-web-services-achieves-fastest-training-times-for-bert-and-mask-r-cnn/

BERT-Large scaling performance at re:Invent 2019

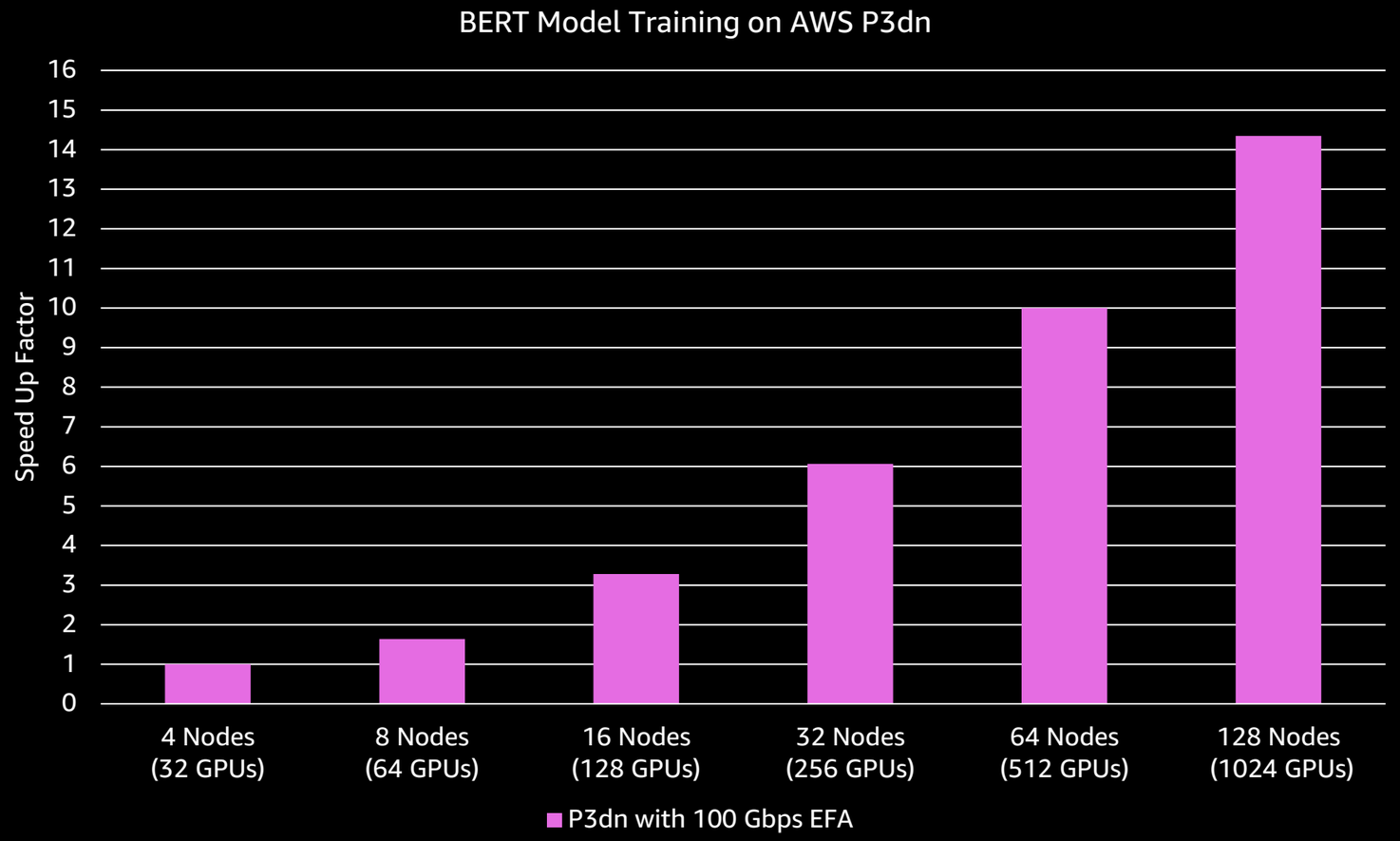
P3DN.24xlarge nodes	NVIDIA GPUs	Time to train (PyTorch)	Time to train (TensorFlow)
1	8	6.4 days	7.5 days
192	1536	69 min	
256	2048		62 min

Using clusters of P3dn instances, each with

- 8x V100 32 GB GPUs
- 100 Gbps networking with EFA
- 300 GB/s NVLink

BERT-Large

- 340 million parameters
- 16 GB dataset



https://aws.amazon.com/blogs/machine-learning/amazon-web-services-achieves-fastest-training-times-for-bert-and-mask-r-cnn/

T5 model training performance

T5 text-to-text transfer transformer

- 3 billion parameters
- 50 GB dataset

EC2 instance	# of nodes	NVIDIA GPUs	Time to train (PyTorch)
P3dn.24xlarge	256	2048 x V100	12.2 days
P4d.24xlarge	256	2048 x A100	5.9 days

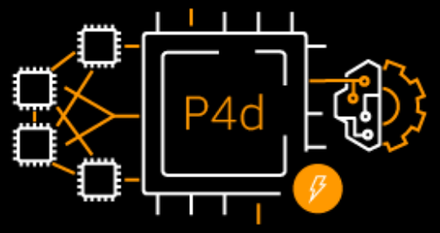
P3dn cluster, each with

- 8x V100 32 GB GPUs
- 100 Gbps networking with EFA
- 300 GB/s NVLink

P4d cluster, each with

- 8x A100 40 GB GPUs
- 400 Gbps networking with EFA and GPUDirect
- 600 GB/s NVLink

EC2 accelerated compute portfolio for ML



ML training

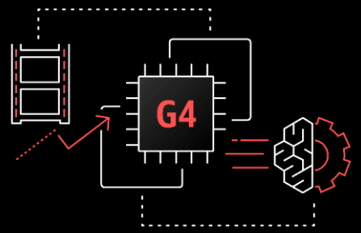
P4d: GPU compute instance

- Designed for lowest cost/highest performance single and multi-node distributed model training jobs
- 8x A100 40GB, 400 Gbps, UltraClusters



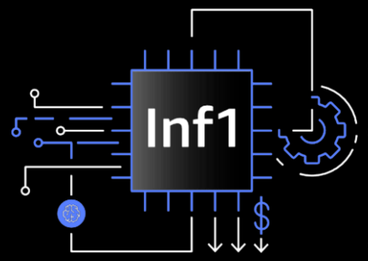
P3/P3dn: GPU compute instance

- Midrange/Single-node ML training
- 8x V100 16/32GB, 25/100 Gbps



G4dn: GPU compute instance

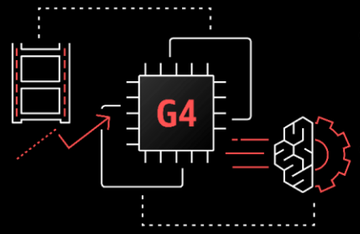
- Cost-effective small-scale training jobs
- 8x NVIDIA T4 16GB, 100 Gbps



ML inference

Inf1: Custom ML acceleration

- Lowest cost per inference in the cloud
- Up to 2,000 TOPs with AWS-designed Inferentia accelerators
- Designed for high throughput and low latency

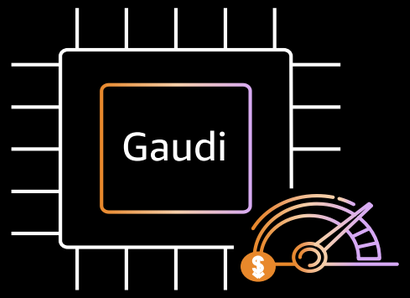


G4dn: GPU compute instance

- Up to 1030 TOPs of compute with 8x NVIDIA T4 GPUs
- Increased performance, lower latency, and reduced cost per inference compared to previous GPU-based instances

Announcing Habana Gaudi and AWS Trainium accelerators for ML training

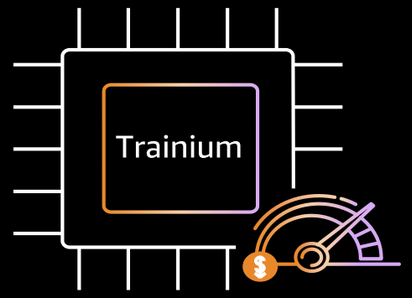
Upcoming optimized platforms for ML training



Accelerators built by Habana Labs, an Intel company, for training ML models

Up to 40% lower cost-to-train deep learning models over GPU instances

Available H1’21



Accelerators built by AWS for training ML models

Lowest cost in the cloud for training deep learning models

Available late 2021

Most comprehensive portfolio of instances for building, training, and deploying ML applications

Demo: UltraClusters

Amr Ragab
Senior Solutions Architect



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.