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

CMP207

Choosing the right accelerator for training and inference

Samir Araújo

Principal AI/ML Solutions Architect
AWS

Max Liu

Principal Specialist, Accelerated Compute
AWS

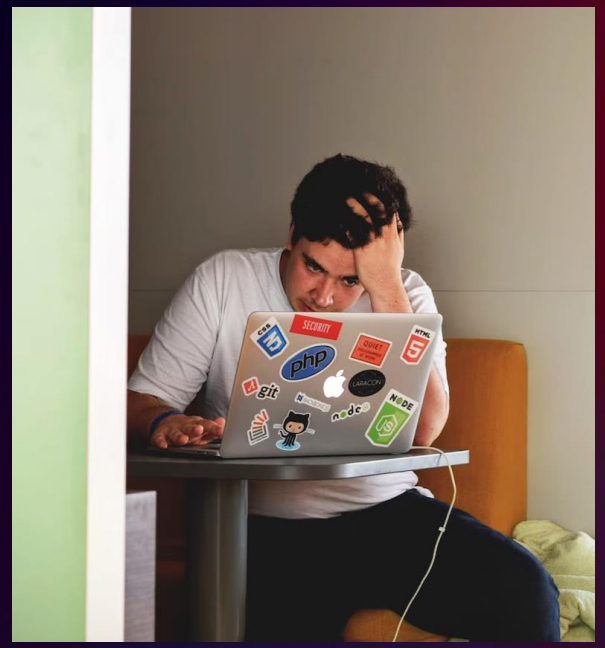
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML common scenarios

Small
Small models
Classic ML/shallow NN
Small/toy datasets



Intermediate/big
Medium/big models
DL/big trees/clusters
Intermediate/big datasets



Huge
Transformers
Models with 10 Bi+ params
Huge datasets/TBs



aws

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

Challenges

Data scientists need to build/deploy different types/sizes of ML models in less time and paying less



Best cost-performance hardware configuration to build/deploy ML models

Choosing the best hardware setup that provides best cost performance for ML model building/deploying

aws

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

Common questions

Shall I use CPU or hardware accelerator?

Do I need distributed training, or is single host training enough?

Which file system is the best option to store the dataset?

How do I optimize data transfer from the file system to the training host?

My model has 10 B+ params, and the dataset has terabytes. What should I do?

Spoiler: Amazon EC2 UltraClusters

aws

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

The AWS ML stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI SERVICES

Specialized	Business processes	Search	Code + DevOps	Industrial	Healthcare
Core	Text and documents	Chatbots	Speech		Vision

ML SERVICES

Amazon SageMaker Canvas	Amazon SageMaker Studio Lab	Amazon SageMaker Ground Truth	RStudio on Amazon SageMaker

ML FRAMEWORKS & INFRASTRUCTURE

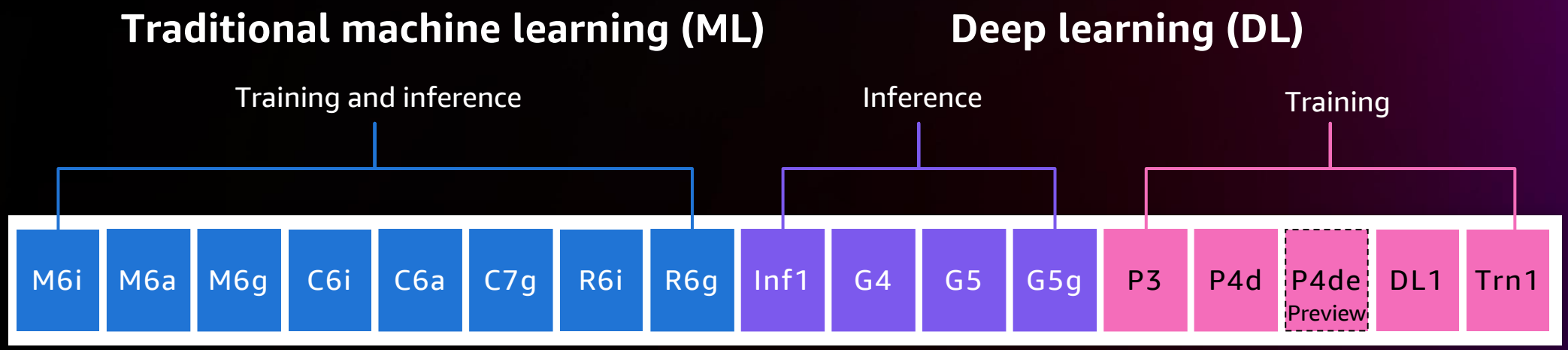
PyTorch, Apache MXNet, Hugging Face, TensorFlow	Amazon EC2	CPUs	GPUs	AWS Inferentia	AWS Trainium	Habana Gaudi	FPGA	Elastic inference

aws

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

Broadest and deepest compute infrastructure for AI/ML

CPU, GPU, AND CUSTOM ACCELERATOR EC2 INSTANCES



Xeon CPU
Habana Gaudi accelerator

EPYC CPU
Radeon GPU

Graviton CPU
Inferentia accelerator
Trainium accelerator

A100, V100, A10G, T4 GPUs

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

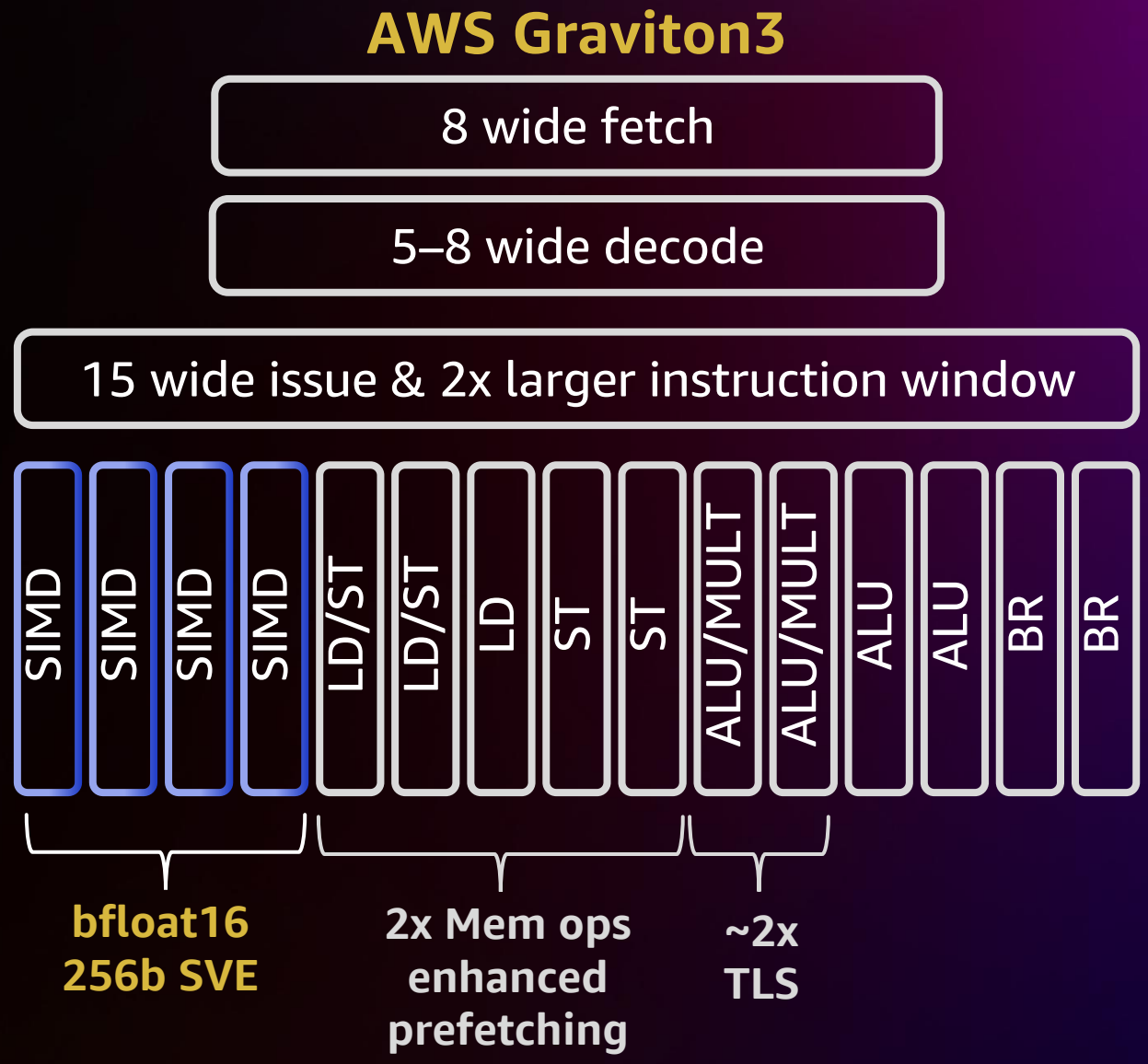
Instances and accelerators for ML inference

Inference accelerator instance type	Throughput	Latency	Cost-efficiency	Model support and programmability	Ease of use	Framework support
CPU-only C6/7 instance types	○	○	● Smaller models	●	●	●
GPU P4, G4/5 instance types	◑	◑	◑	●	◑	●
AWS Inferentia Inf1 instance type	●	●	●	◑	◑	◑

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

When to use CPU instances for ML training

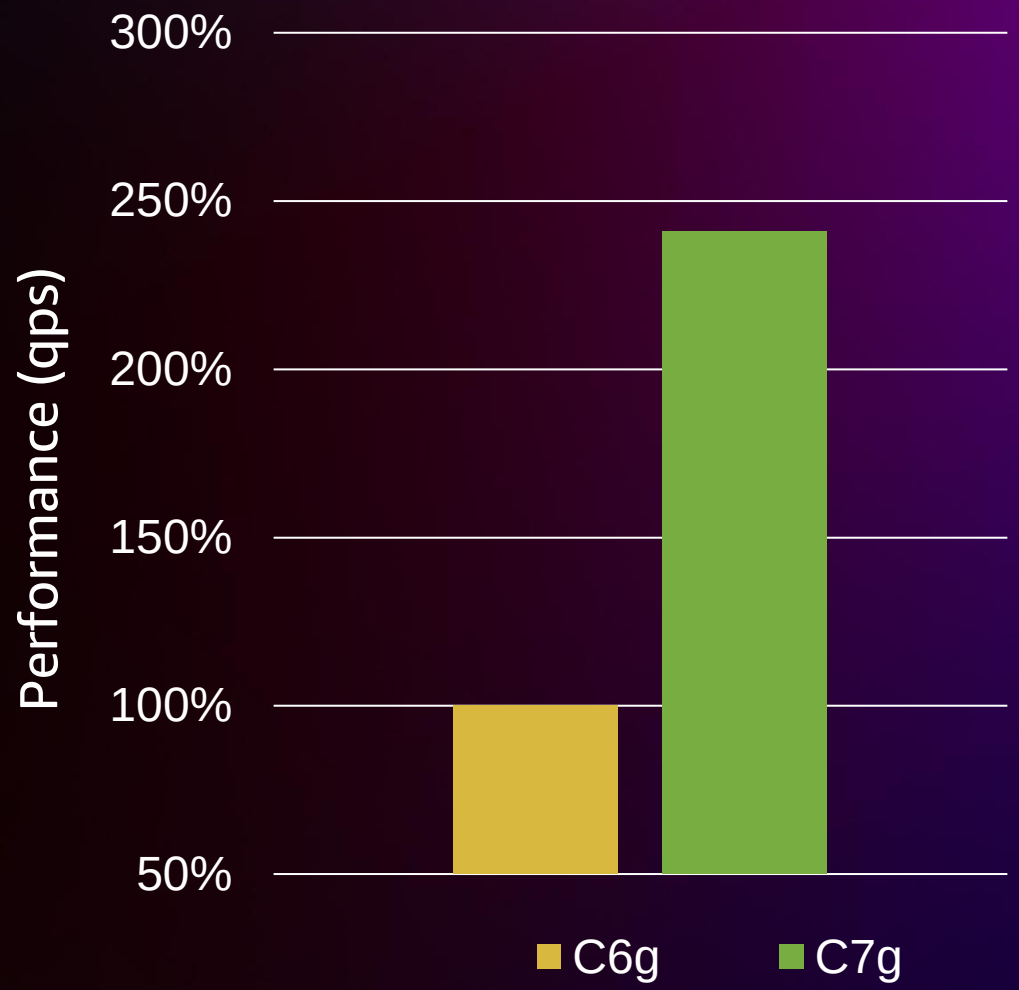
- Traditional ML models (tabular data)
  - scikit-learn
  - AutoGluon (AutoML) https://go.aws/3EkPD4k
- Tabular data pre-processing
  - Pandas
- Prototyping DL models
- Fine-tuning on a smaller dataset



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

Graviton3: Machine learning

- NLP workload using BERT model performing SQuAD v1.1 question-answering task
- Graviton3 includes support for bfloat16 and 2x the SIMD bandwidth
- C7g is nearly 2.5x higher performance than C6g



TF version: v2.7.0 (oneDNN.4, acl-21.08); arm64 container: arm tool solutions 21.11;; x86 container: x86 (sha256: 31e09cf438a4); MLPerf version: v1.0.1; 4xl instance size; TF_ENABLE_ONEDNN_OPTS=1 and DNNL_DEFAULT_FPMATH_MODE=BF16; inter_op_parallelism_threads: 1; intra_op_parallelism_threads: 16

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

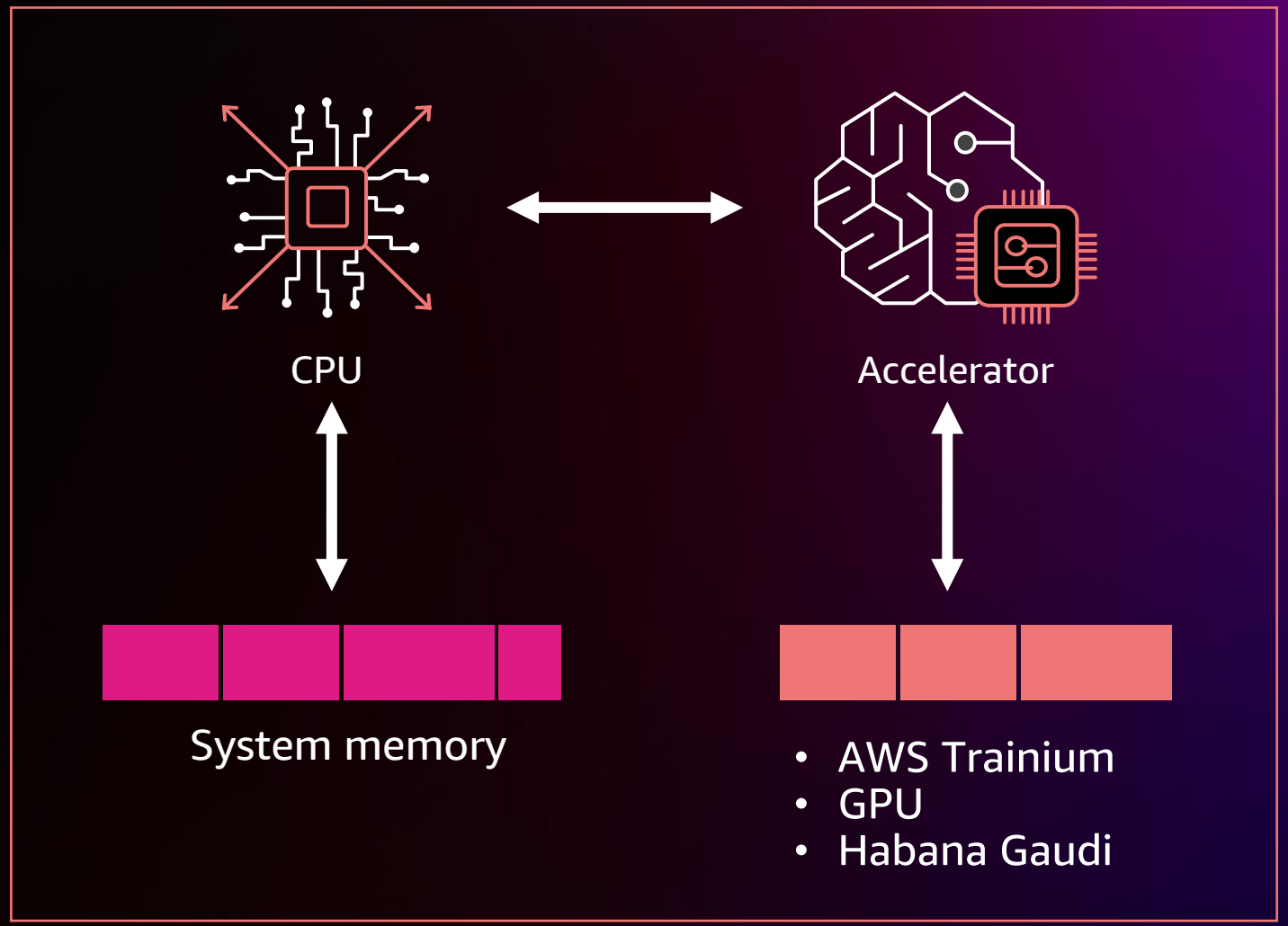
Accelerated instances for training

Accelerated instances

Trn1 — AWS Trainium

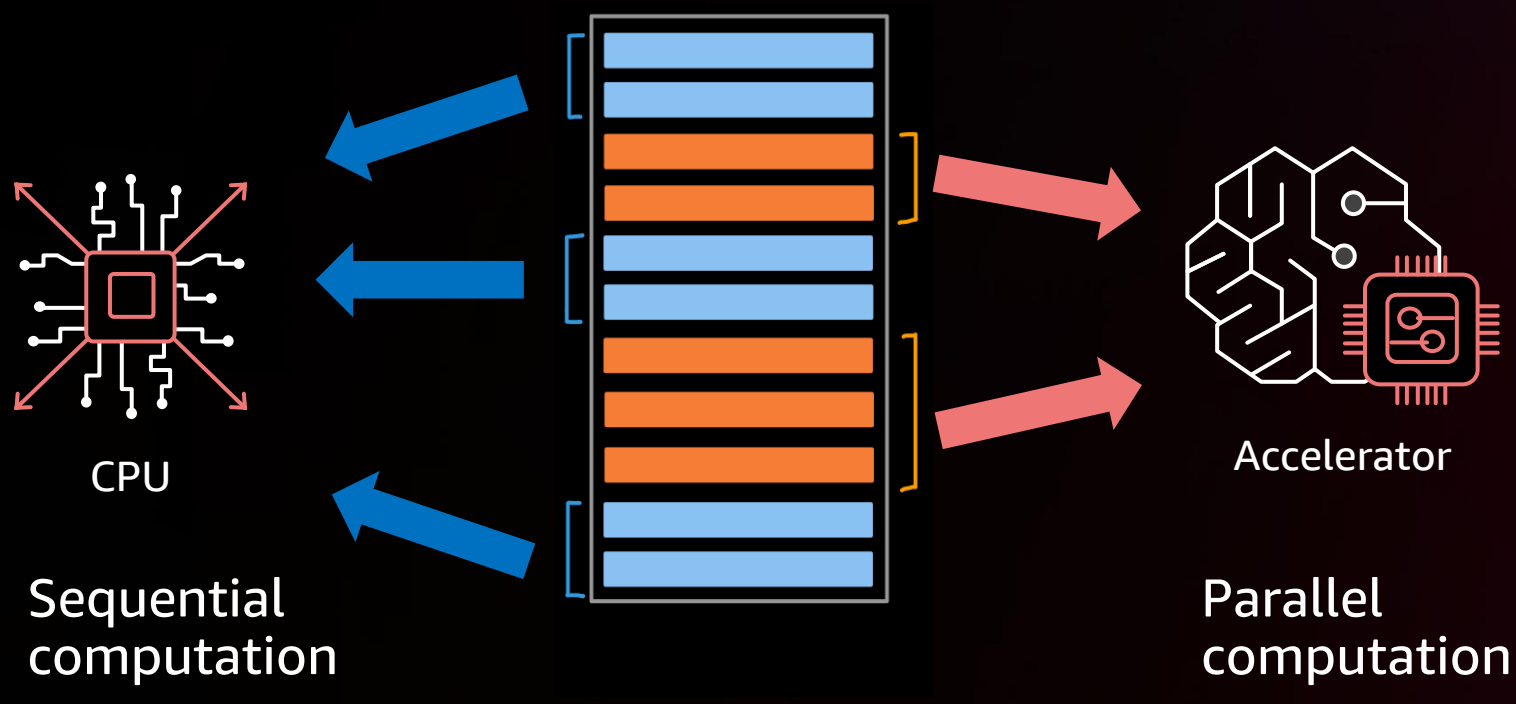
P, G — GPU

DL1 — Habana Gaudi

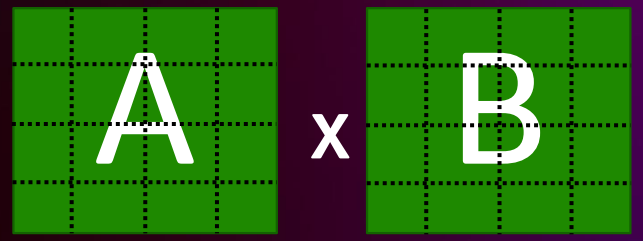


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

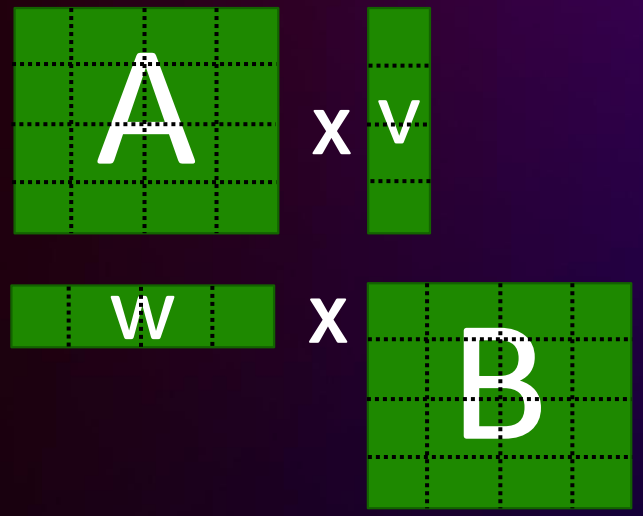
ML training on accelerated instances



Matrix-matrix



Matrix-vector



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

Amazon EC2 Trn1/Trn1n instances

THE MOST COST-EFFICIENT HIGH-PERFORMANCE TRAINING INSTANCE

Trn1(n)

BF16/FP16	TF32	FP32
3.4 PFLOPS	3.4 PFLOPS	840 TFLOPS

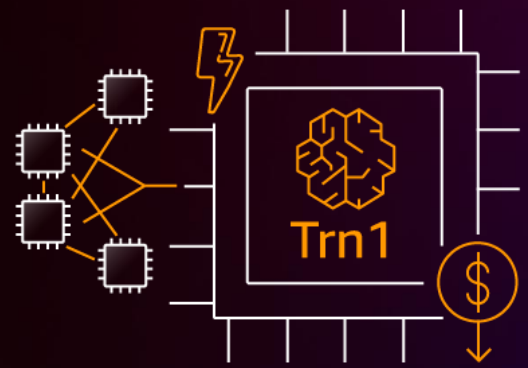
AGGREGATE ACCELERATOR MEMORY	PEAK MEMORY BANDWIDTH
512 GB	13.1 TB/sec

EFA NETWORK CONNECTIVITY	NEURON-CORE V2 NEURON-LINK V2
800/1600 Gbps	



AWS Trainium

High performance machine learning training chip, purpose-built by AWS



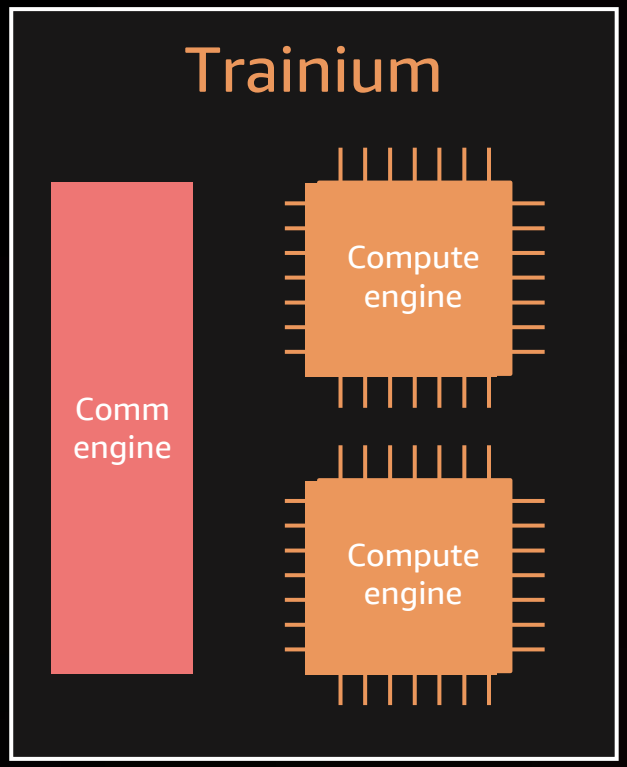
EC2 Trn1/Trn1n Instances

The most cost-efficient high-performance training instance

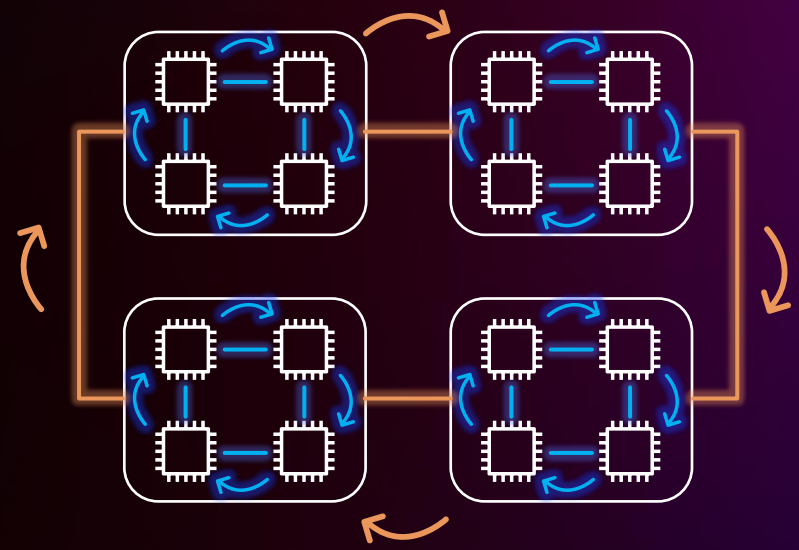
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – Built for scale-out

PARALLELIZED COMPUTATION AND COMMUNICATION



2D-RING TOPOLOGY FOR COLLECTIVE COMMUNICATION

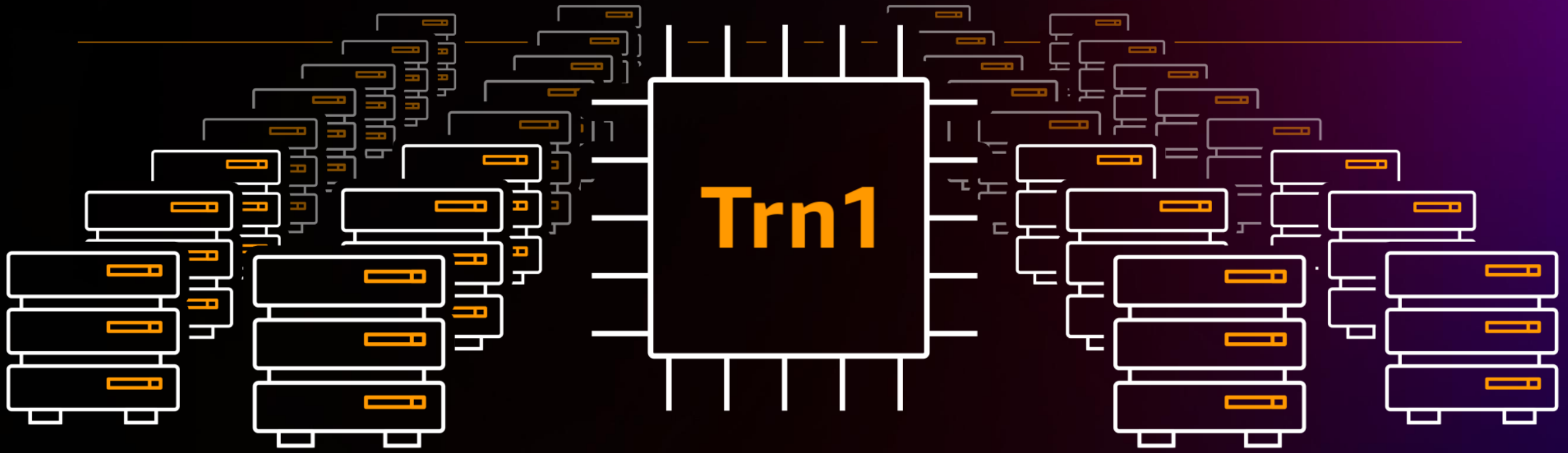


— HIGH SPEED LOCAL INTERCONNECT
— EFAv2

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

EC2 UltraCluster for ultra-large models

30K+ TRAINIUM ACCELERATORS, WITH NON-BLOCKING NETWORK CONNECTIVITY



On-demand access to a world-class supercomputer,
6 ExaFLOPS of compute

aws

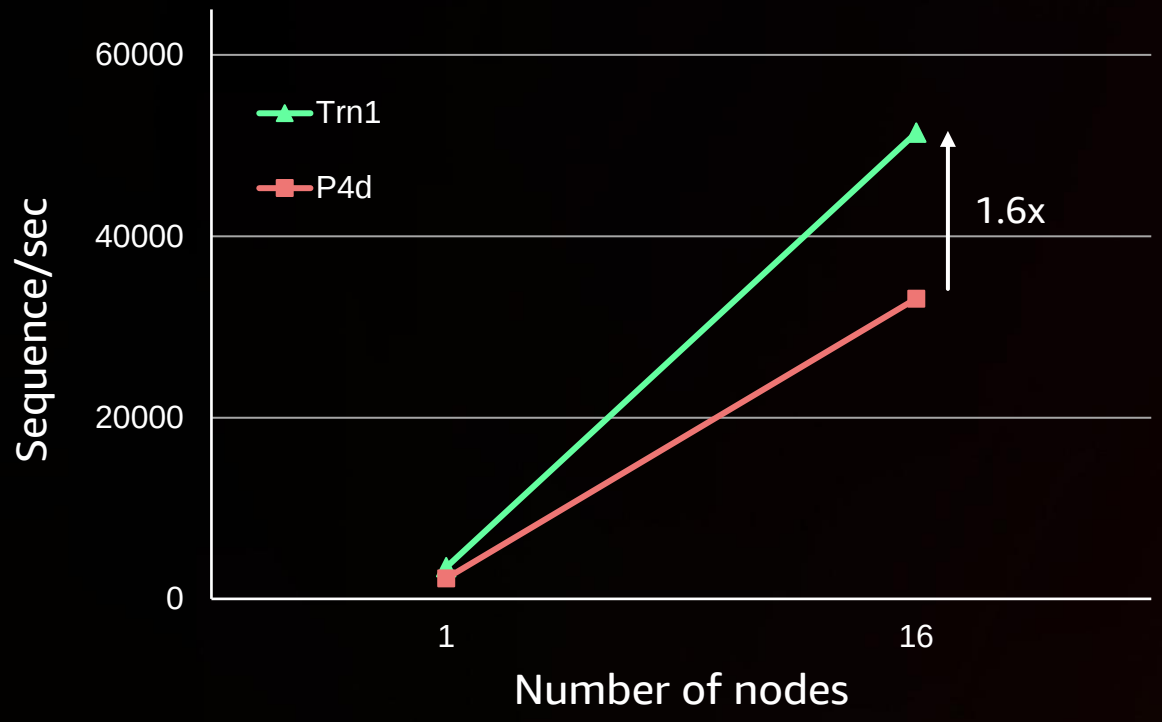
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 performance

HIGHER PERFORMANCE AT LOWER COST

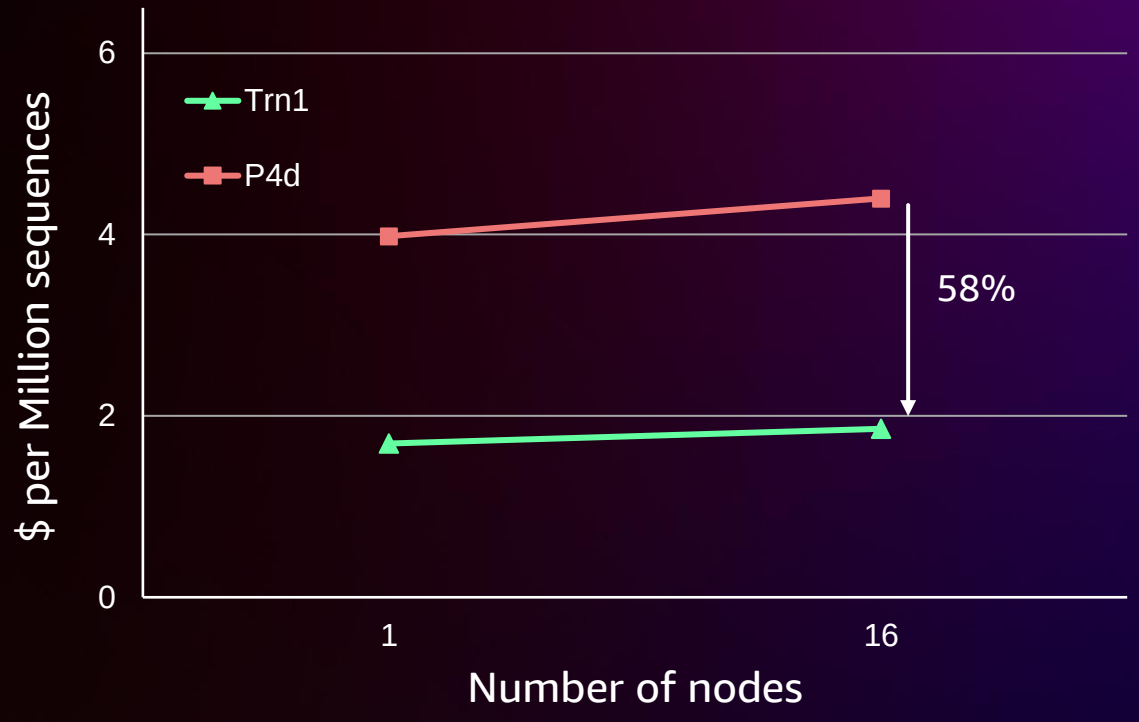
Throughput

Single Node: Trn1 **1.5x faster**
Cluster: Trn1 **1.6x faster**



Training Cost

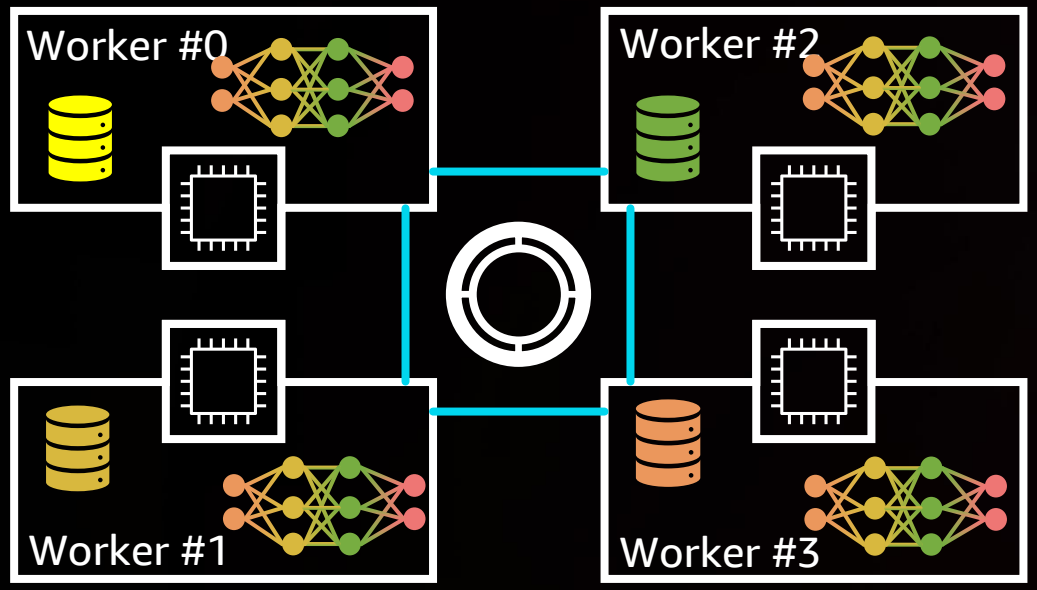
Single Node: Trn1 **57% lower cost**
Cluster: Trn1 **58% lower cost**



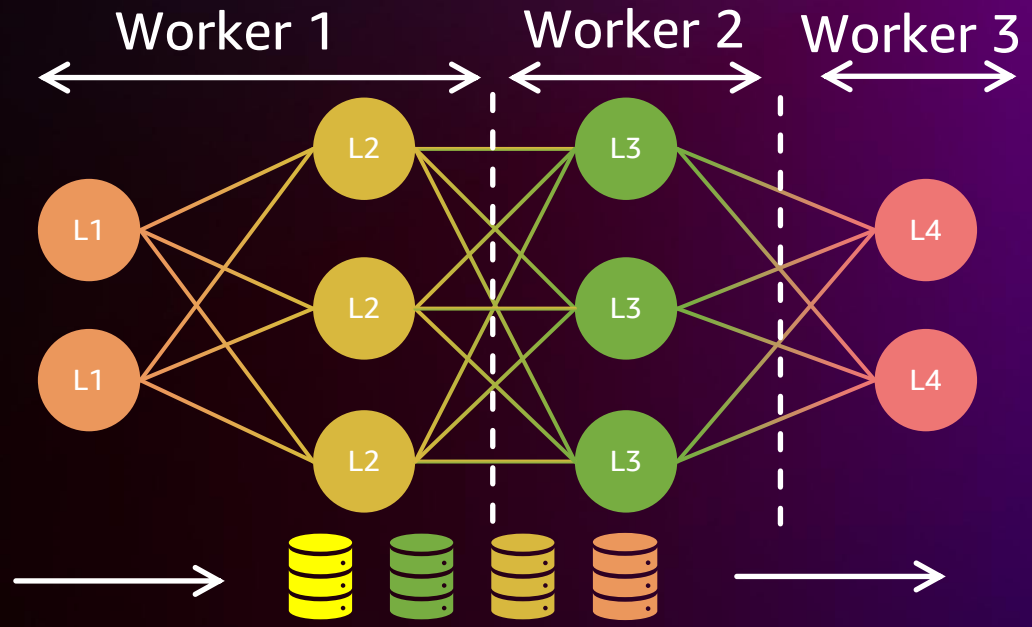
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hugging Face BERT-Large Ph1 Pretrain; 128 Seqlen
trn1 – BF16(SR), p4d – Mixed Precision (BF16/FP32)

Distributed training

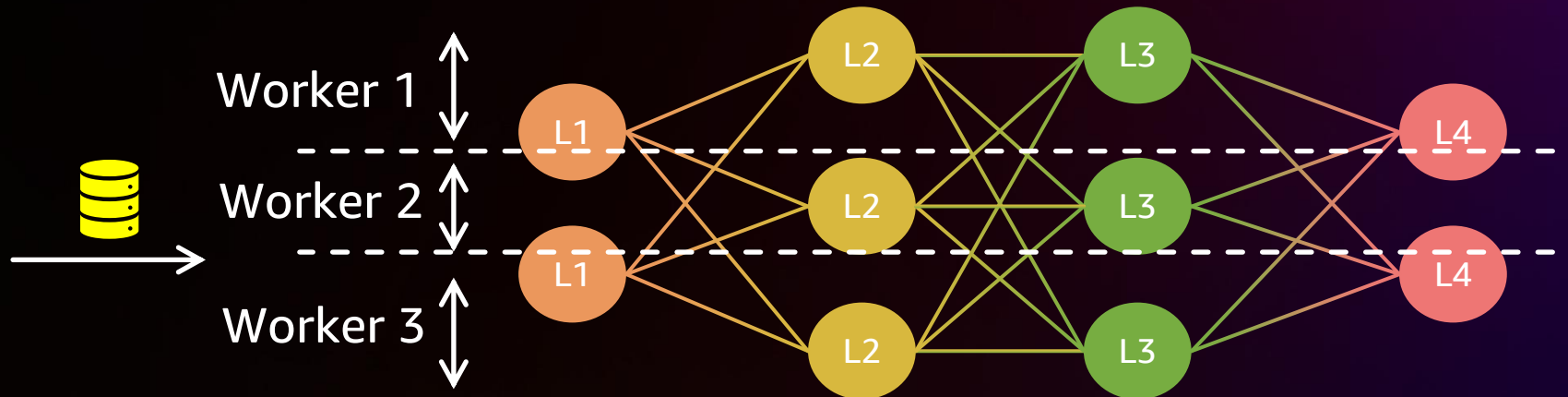


Data parallelism



Pipeline parallelism





Tensor parallelism

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

Amazon SageMaker

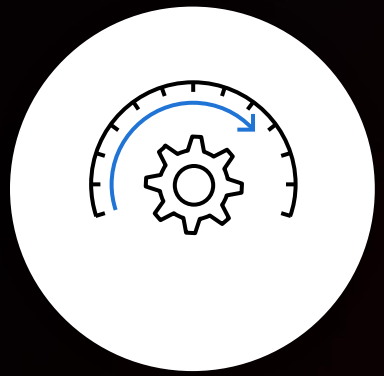
MOST COMPLETE END-TO-END ML SERVICE

PREPARE → BUILD → TRAIN & TUNE → DEPLOY & MANAGE →



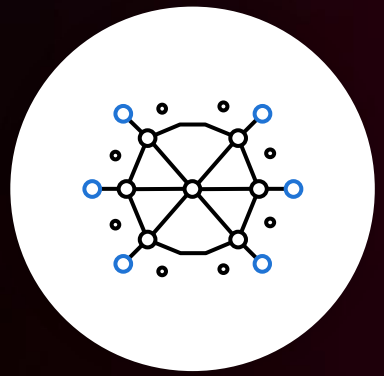
Democratize ML innovation

Empower more groups of people, including business analysts



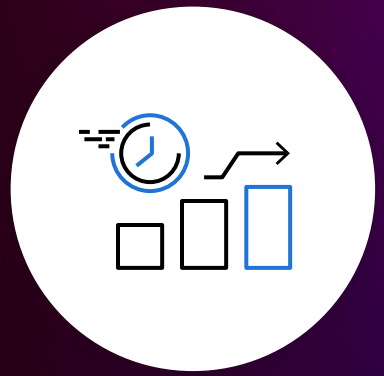
Accelerate the ML lifecycle

Reduce training time from hours to minutes



Prepare data at scale

Access, label, and process structured and unstructured data



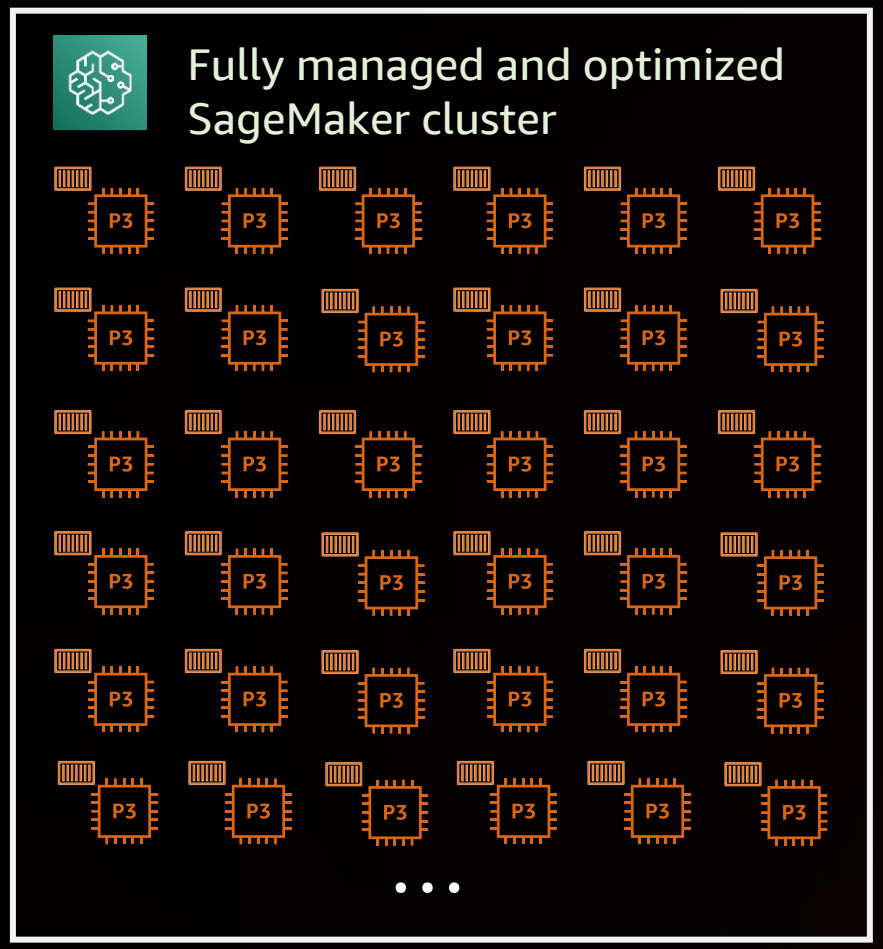
Streamline ML processes

Automate and standardize MLOps processes

aws

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

Large training datasets: What are my options?



1. Moderate and large datasets

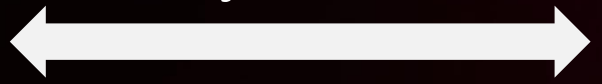




Amazon S3

- File mode: Copy entire dataset to local volume
- Fast file mode: Stream dataset from Amazon S3

2. Scalable shared file system





Amazon EFS

- No downloading or streaming
- Share file system with other services

3. High-performance file system





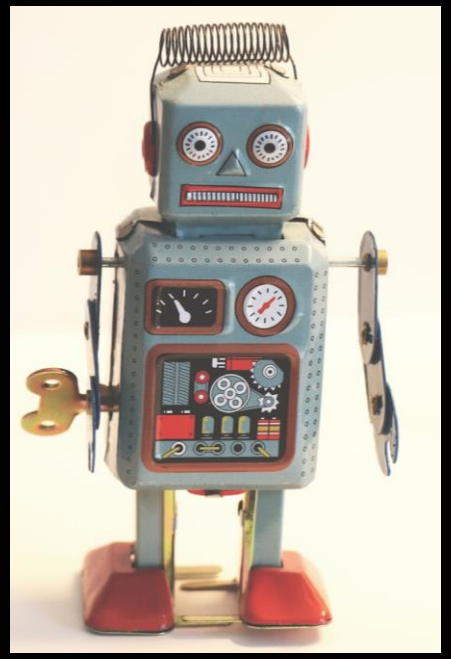
Amazon FSx for Lustre file system

- Optimized for high-performance computing
- Natively integrated with Amazon S3

aws

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

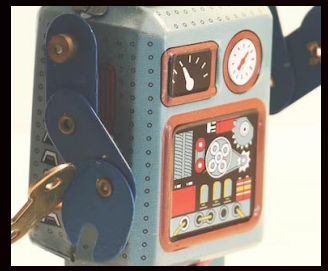
Case study: Visual search for 50K+ parts



Product with defect



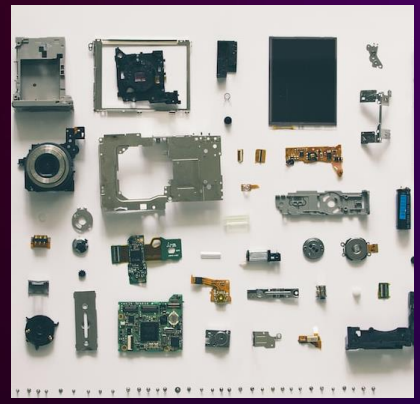
Customer



???

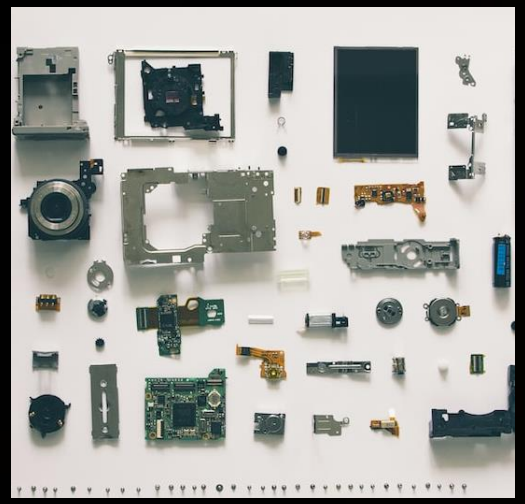


Operator



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

Case study: Visual search for 50K+ parts



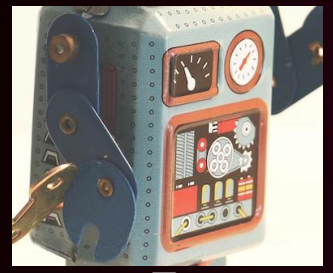
Parts images
(dataset)



Data scientist

Vision
(CNN)

Parts clusters
(K-NN)





Operator

Part ID:
123

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

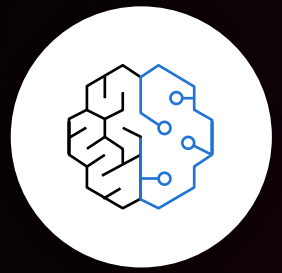
Case study: Environment setup



Data preparation

SageMaker processing job

OpenGL + Blender
# of gen images: 10,000
S3 – ShardedByS3Key
+ Pipe Mode
4 x **ml.g4dn.xlarge**



Model training

SageMaker training job

CNN: Yolov5 Large – 2 classes
Dataset: ~500GB
Image size .jpg: ~51K
2 x **ml.g4dn.xlarge**

K-NN: Scikit NearestNeighbors
n_jobs=12
Dataset: feature vectors
1 x **ml.c5.4xlarge**



Model deployment

SageMaker real-time endpoints

Yolov5: 1x **ml.inf1.xlarge**

(Payload compression: JPEG)

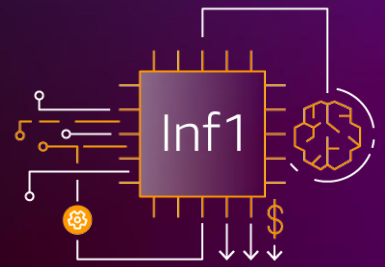
K-NN: 1x **ml.c5.2xlarge**

Time: 19.32 hours

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

Customer story

SOFTWARE AND INTERNET



Finch COMPUTING

"Given the cost of GPUs, we simply couldn't have offered our customers additional languages while keeping our product profitable. Amazon EC2 Inf1 instances changed that equation for us."

Scott Lightner
President and Founder, Finch Computing

OVERVIEW

Finch Computing develops natural language processing (NLP) technology to provide customers with the ability to uncover insights from huge volumes of text data, and it was looking to fulfill customers' requests to support additional languages

SOLUTION

Finch migrated their proprietary deep learning translation models from GPU-based instances to Amazon EC2 Inf1 instances powered by AWS Inferentia. Finch plugged the AWS Inferentia AMIs into its DevOps pipeline and updated its infrastructure-as-code templates to use AWS Inferentia to run customized containers using Amazon ECS.

IMPACT

Customer reduced its costs by more than 80 percent compared with the use of GPUs while maintaining its throughput and response times for its customers

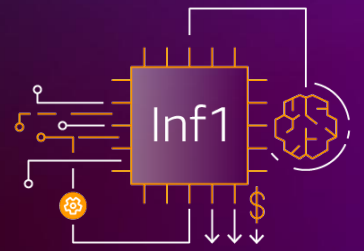
aws

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

Customer story

RETAIL

amazonrobotics

Amazon SageMaker + 

> **"Amazon SageMaker doesn't just manage the hosts we use for inferencing. It also automatically adds or removes hosts as needed to support the workload."**

Eli Gallaudet
Senior Software Manager, Amazon Robotics

OVERVIEW

Amazon Robotics used machine learning to streamline the flow of inventory in fulfillment centers. It ran into challenges scaling and managing infrastructure required to host ML models

SOLUTION

Customer modernized the ML model development process using Amazon SageMaker and migrated inference workloads to AWS Inferentia–based Inf1 instances

IMPACT

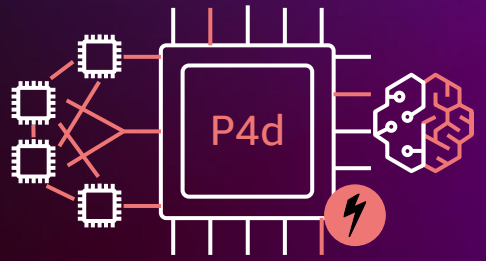
Customer saved ~70% on inferencing costs compared to self-managed solution, improved model performance by 40%, and scales to over 200 million inferences per day

aws

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

Customer story

LIFE SCIENCES



Paige®

"Using Amazon EC2 P4d instances, we increased our compute capacity while balancing costs across our on-premises and cloud environments."

Razik Yousfi
Vice President of Engineering, Paige

OVERVIEW

Paige develops complex, advanced ML applications that support healthcare professionals in delivering precision diagnoses and treatment plans. It needed to update its infrastructure to balance its growing computational requirements, and it wanted to adopt cost-effective, scalable HPC solutions in the cloud.

SOLUTION

Paige turned to Amazon Web Services and took a hybrid approach to running its ML workloads and managing its growing data footprint. For distributed training, the customer adopted Amazon EC2 P4d instances, Elastic Fabric Adapter, and FSx for Lustre.

IMPACT

Paige can run its ML jobs in parallel, speeding up its internal workflows by 72%. By scaling its compute resources in the cloud, Paige has also optimized compute costs and freed its teams to innovate.

aws

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

Learn more

Amazon EC2
accelerated computing instances



https://go.aws/3Up3BaS

Amazon SageMaker



https://aws.amazon.com/sagemaker/

aws

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

Thank you!

Samir Araújo
linkedin.com/in/samiraraujo/

Max Liu
linkedin.com/in/liujtmax



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

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