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

CMP209

AI parallelism explained: How Amazon Search scales deep-learning training

Rejith Joseph (RJ)
Principal Engineer
Amazon

James Park
Senior Solutions Architect
AWS

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

Agenda

Introduction, Trends, and Challenges

M5 and large-model training

Hardware

Performance optimization

Reproducibility, reliability, and debuggability

Model Vending and Inference Optimizations

aws

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

The AWS ML stack

THE BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI SERVICES

Vision
Amazon Rekognition

Speech
Amazon Polly
Amazon Transcribe

Chatbots
Amazon Lex

Text
Amazon Comprehend
Amazon Translate
Amazon Textract

Business tools
Amazon Personalize
Amazon Forecast
Amazon Fraud Detector
Amazon Lookout for Metrics

Contact centers
Contact Lens for Amazon Connect
Amazon Connect Voice ID

Search
Amazon Kendra

Code + DevOps
Amazon CodeGuru
Amazon DevOps Guru

Healthcare
Amazon HealthLake
Amazon Comprehend Medical
Amazon Transcribe Medical

Industrial
AWS Panorama Application SDK, Amazon Monitron, Amazon Lookout for Equipment, Amazon Lookout for Vision

AMAZON SAGEMAKER

Label data

AMAZON SAGEMAKER STUDIO IDE

Data collection prep | Store features | Detect bias and explain predictions | Visualize in notebooks | Pick algorithm | Train models faster | Tune parameters | Deploy in production | Manage & monitor

CI/CD

Manage edge devices

Frameworks & Infrastructure

PYTORCH
mxnet TensorFlow

Deep learning AMIs & containers | GPUs | Habana Gaudi Accelerators | Inferentia | Trainium | FPGA

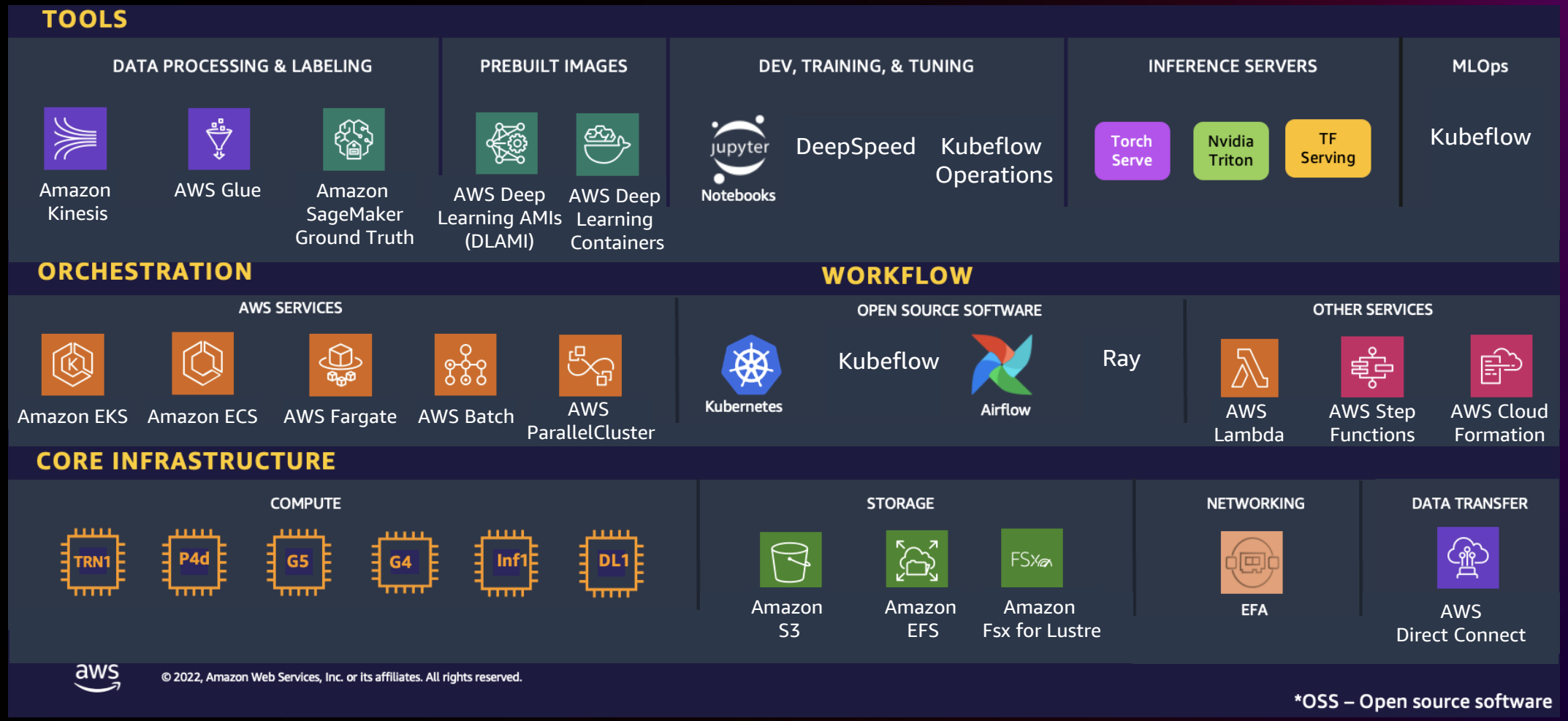
aws

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

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.

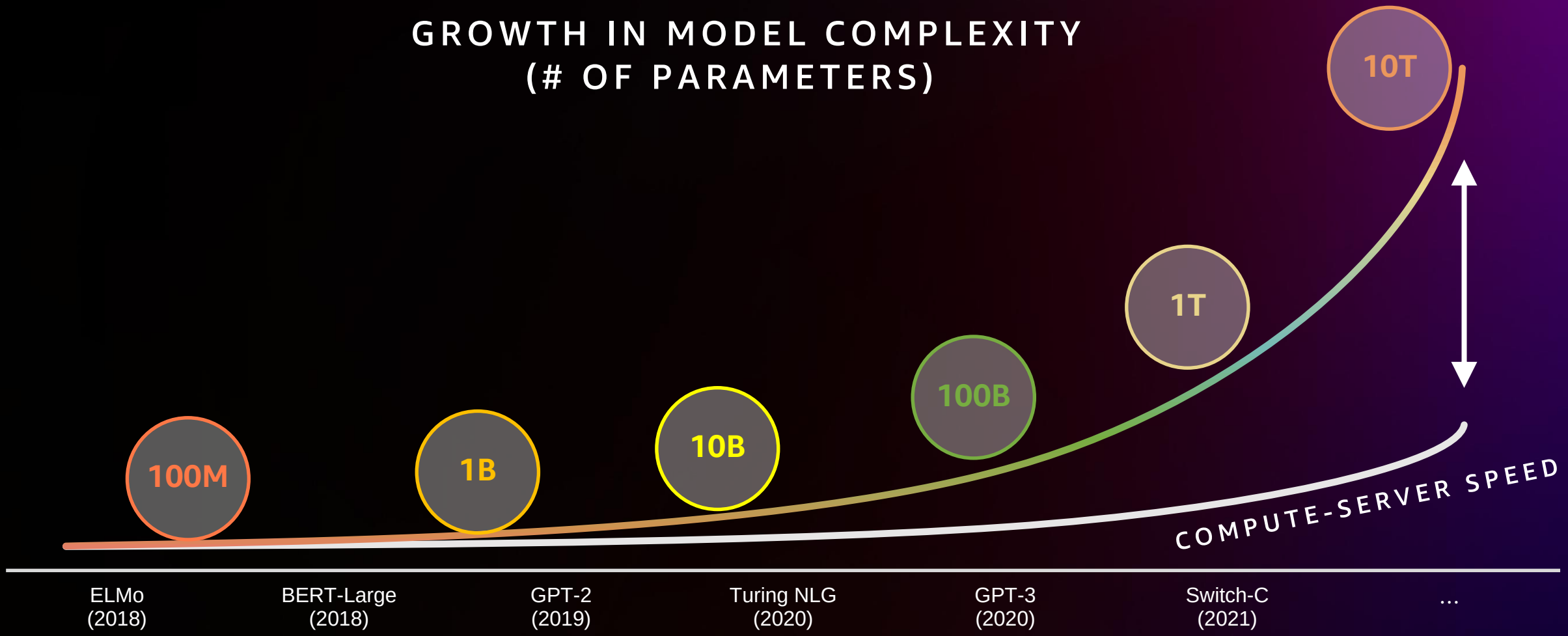
ML frameworks and infrastructure layer

FLEXIBLE OPTIONS FOR ML NEEDS



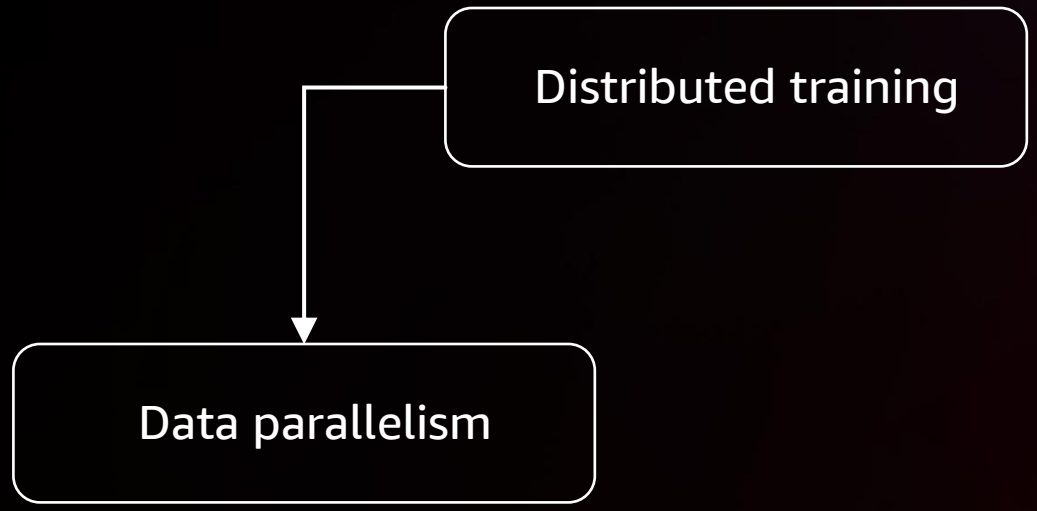
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model Size Challenges



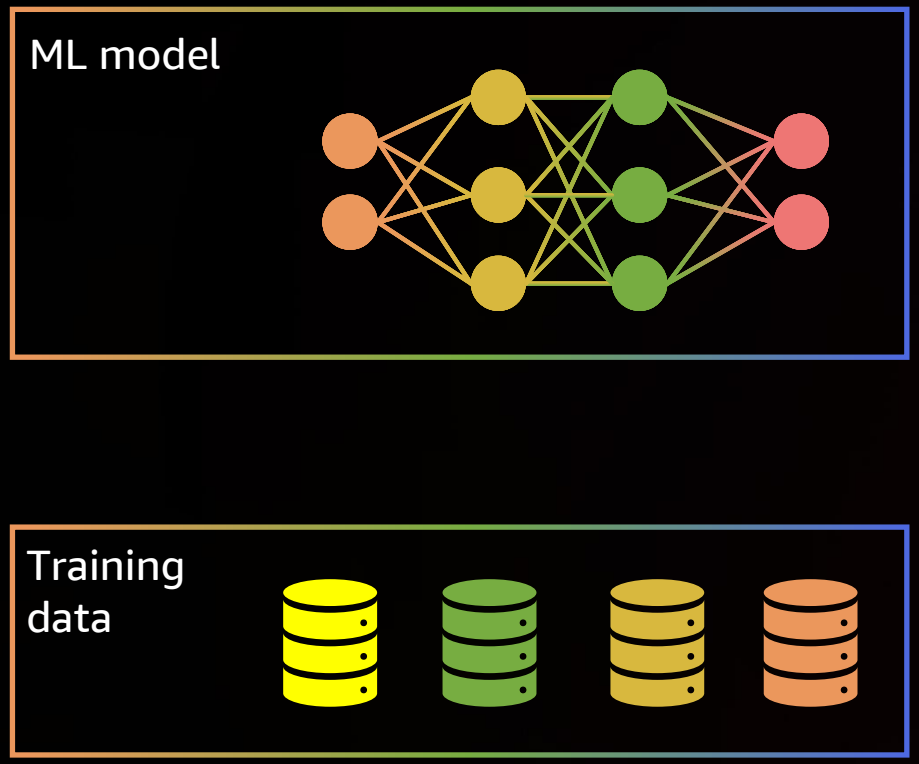
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

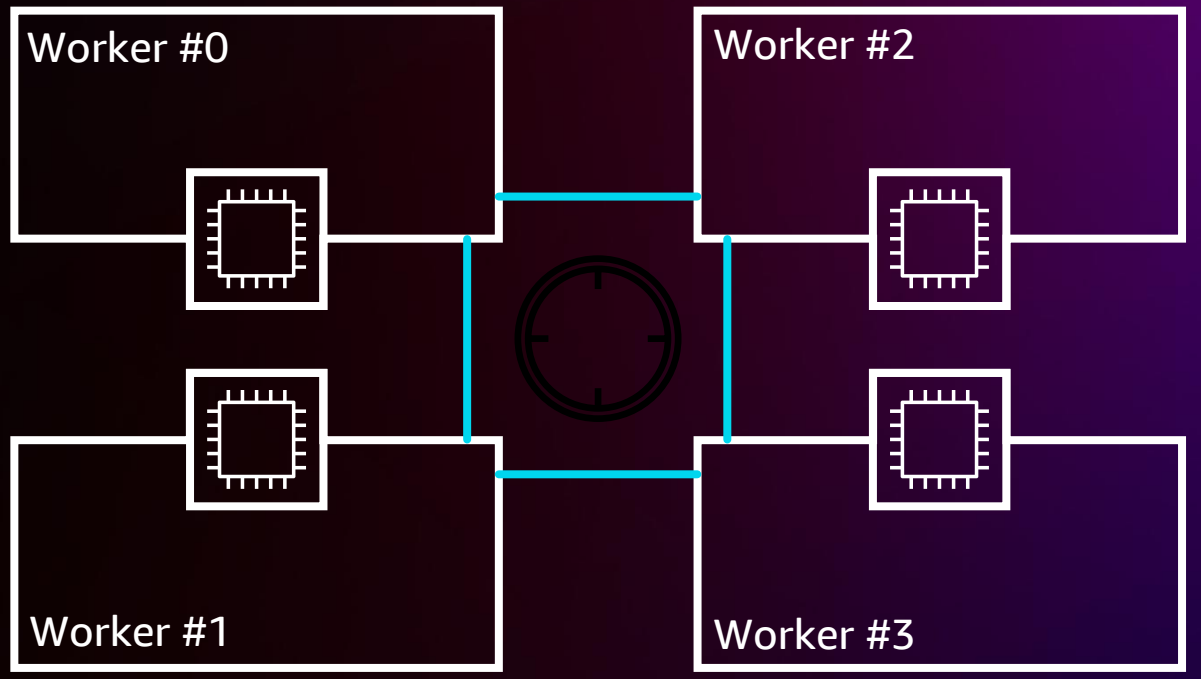
Distributed training



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

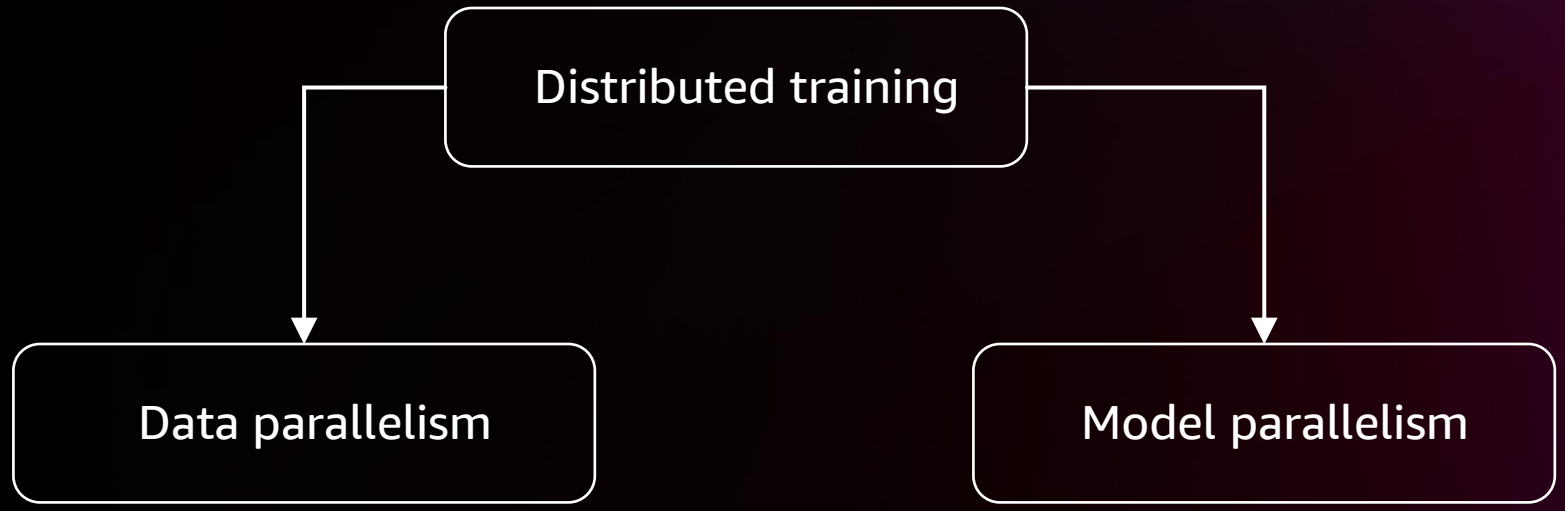
Data parallelism





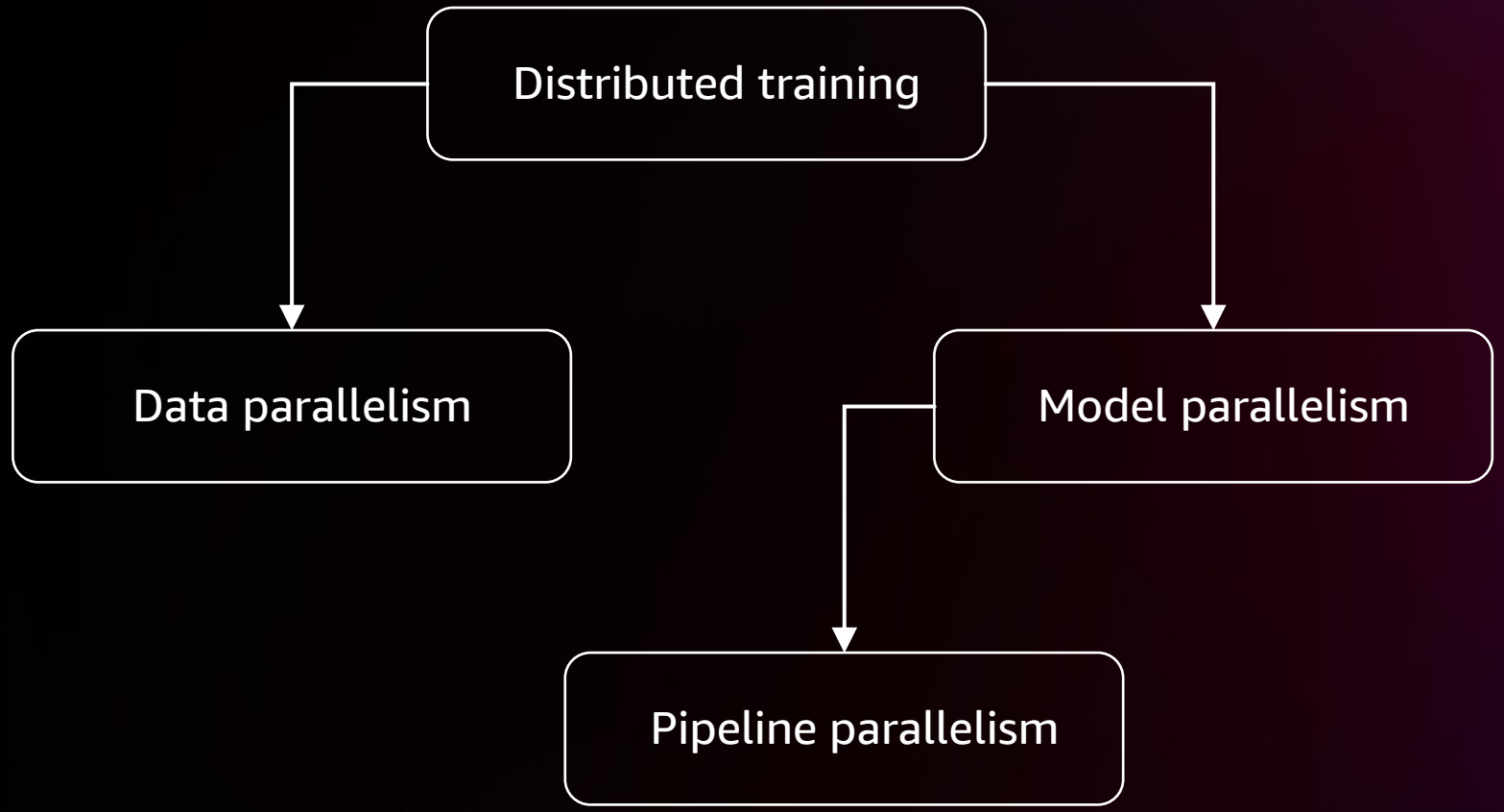
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training



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

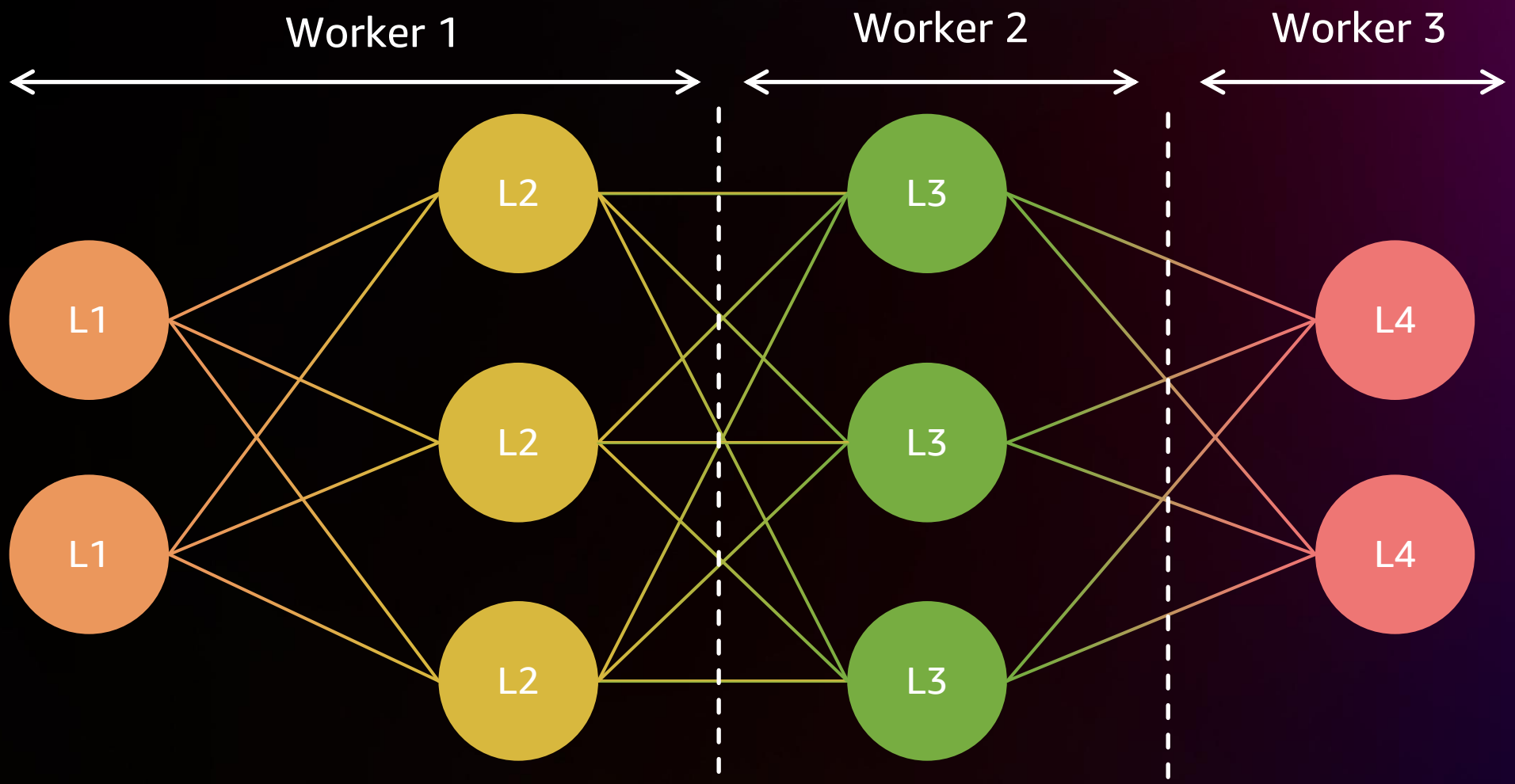
Distributed training



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

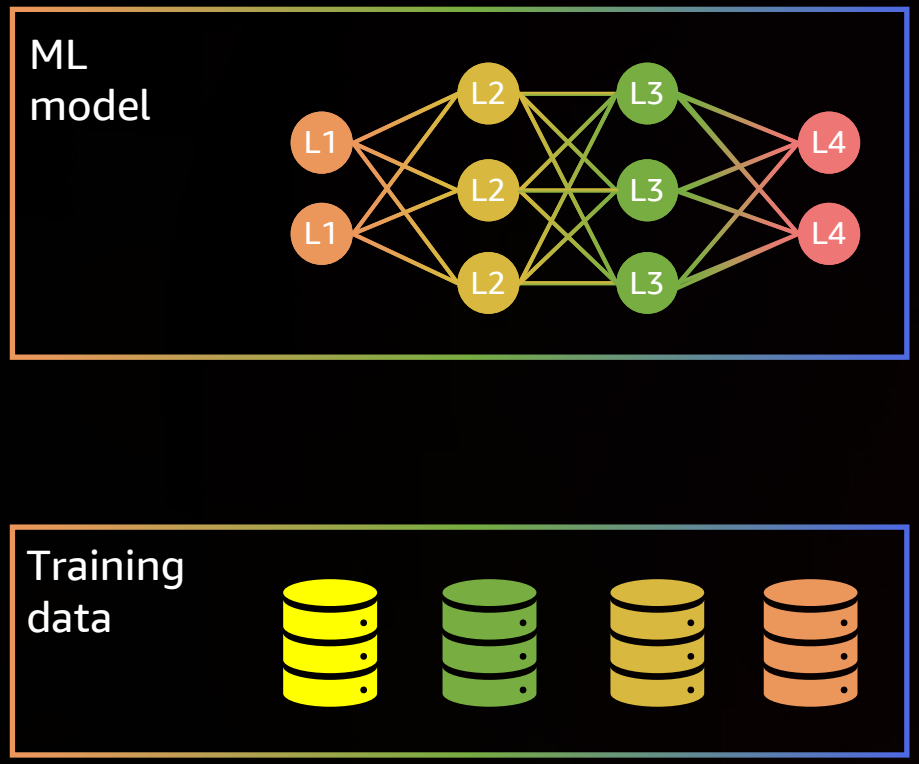
Pipeline parallelism

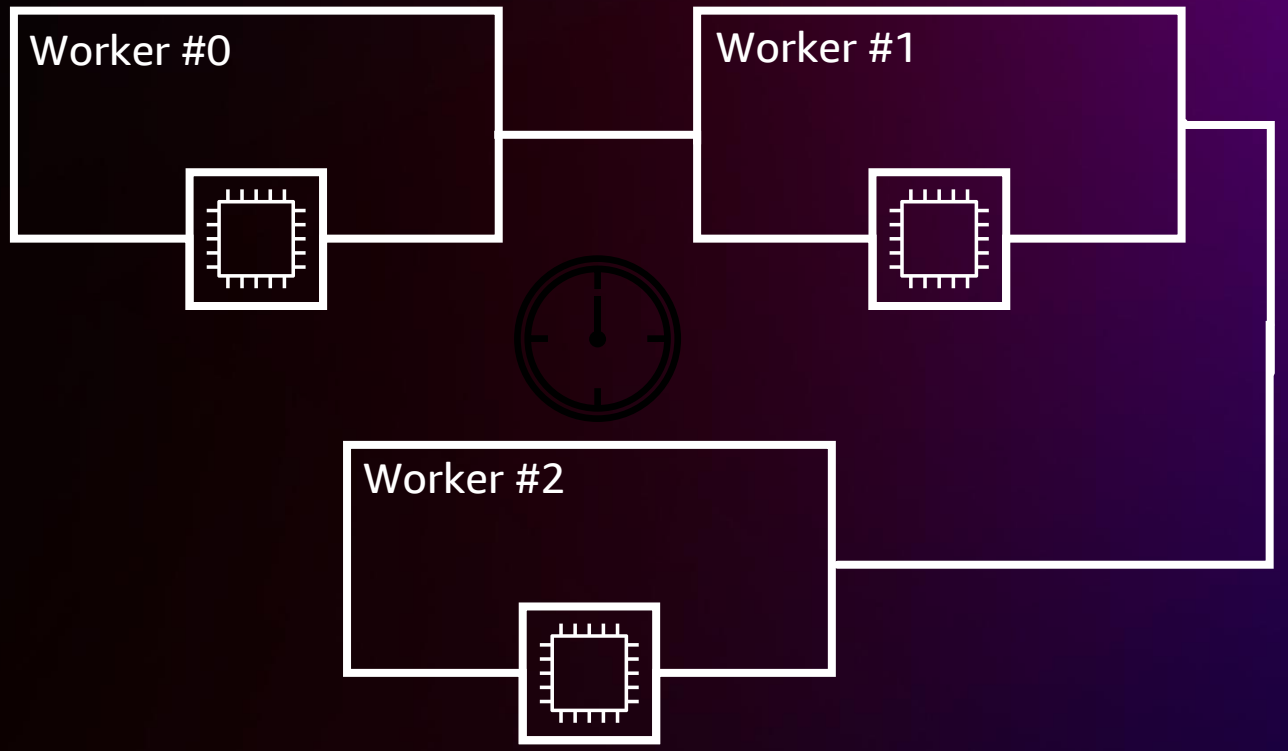
*ALSO CALLED INTER-LAYER MODEL PARALLELISM



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

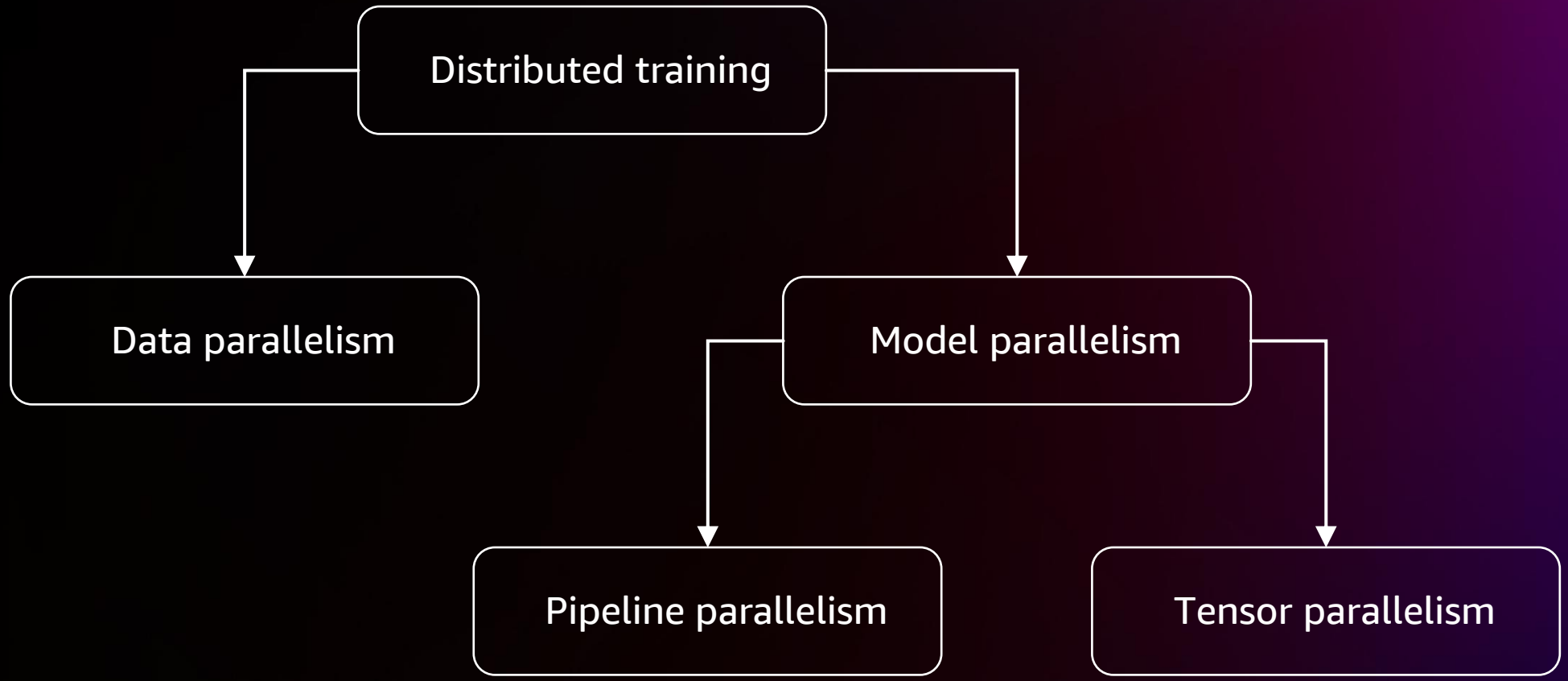
Pipeline parallelism





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

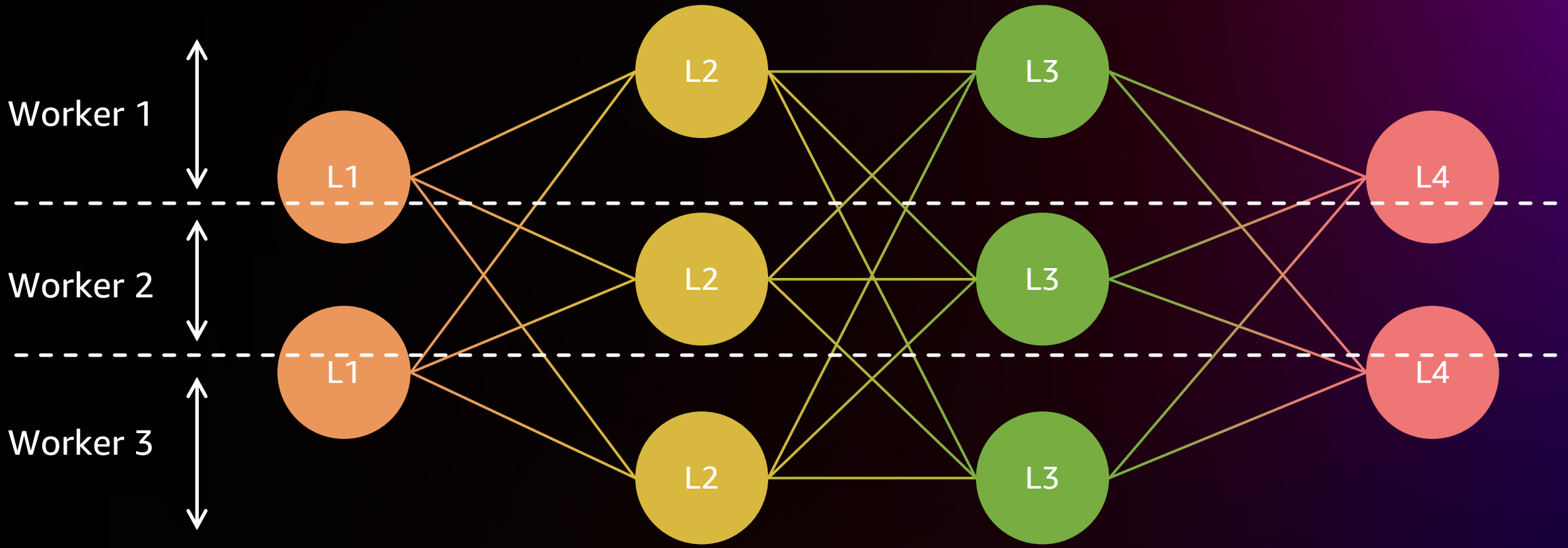
Distributed training



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

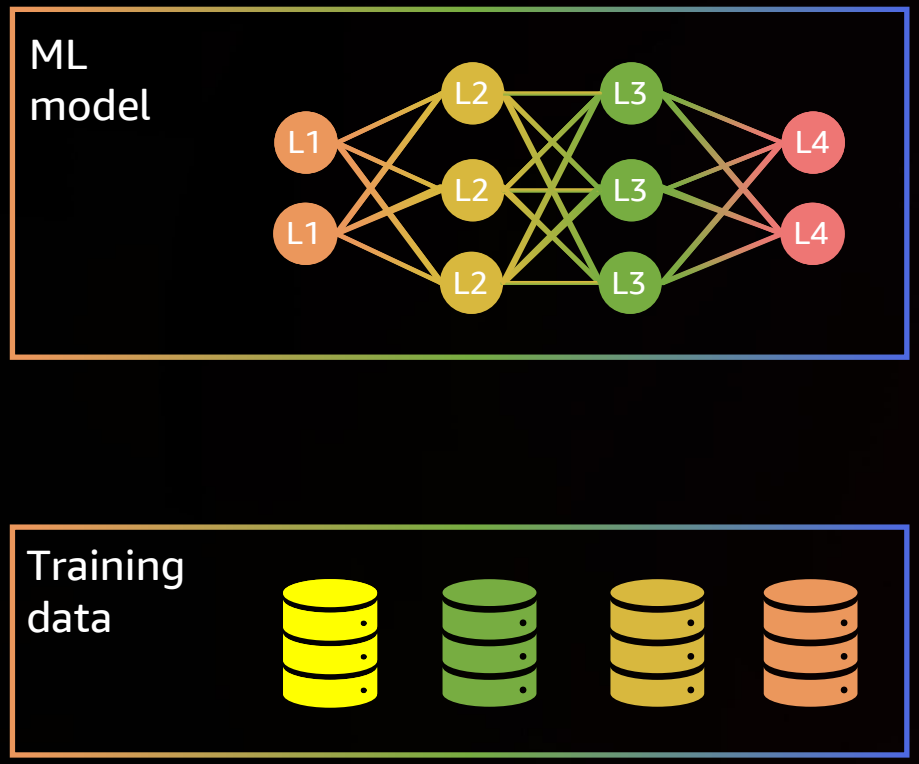
Tensor parallelism

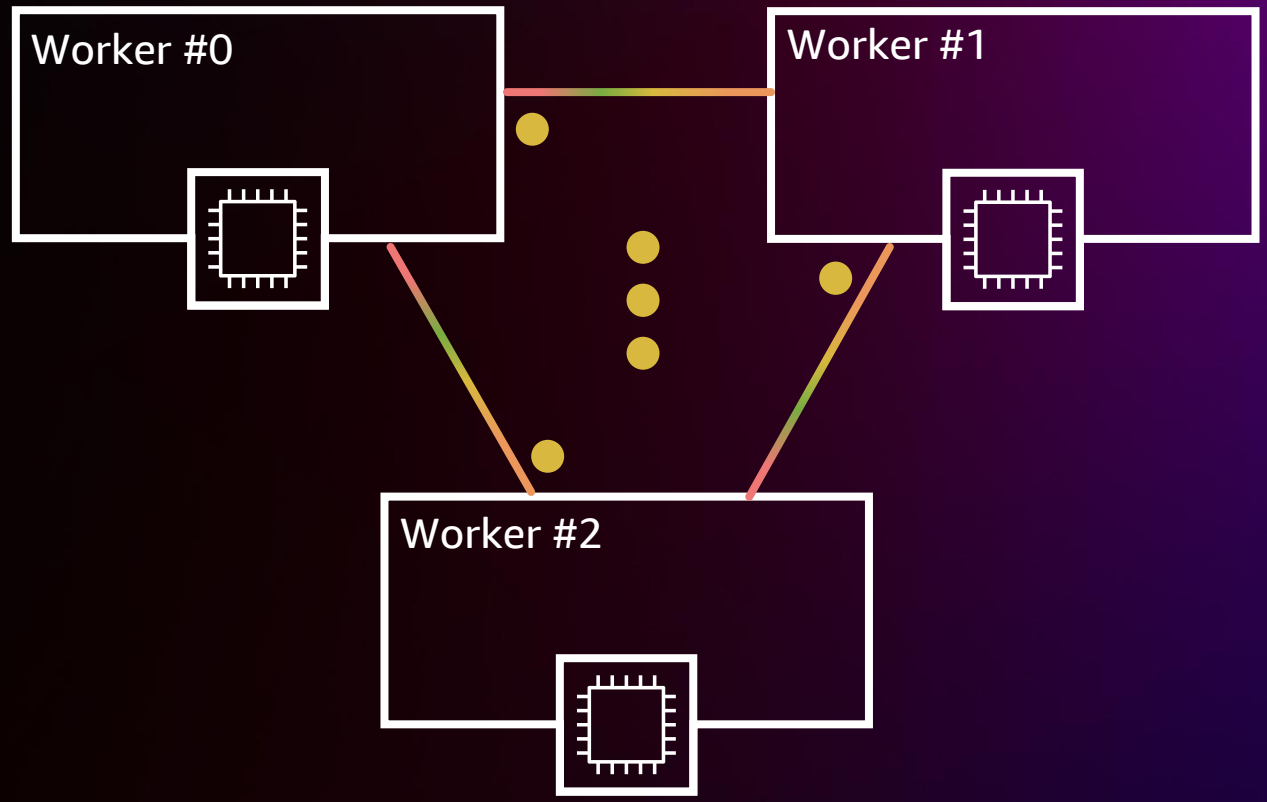
*ALSO CALLED INTRA-LAYER MODEL PARALLELISM



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

Tensor parallelism

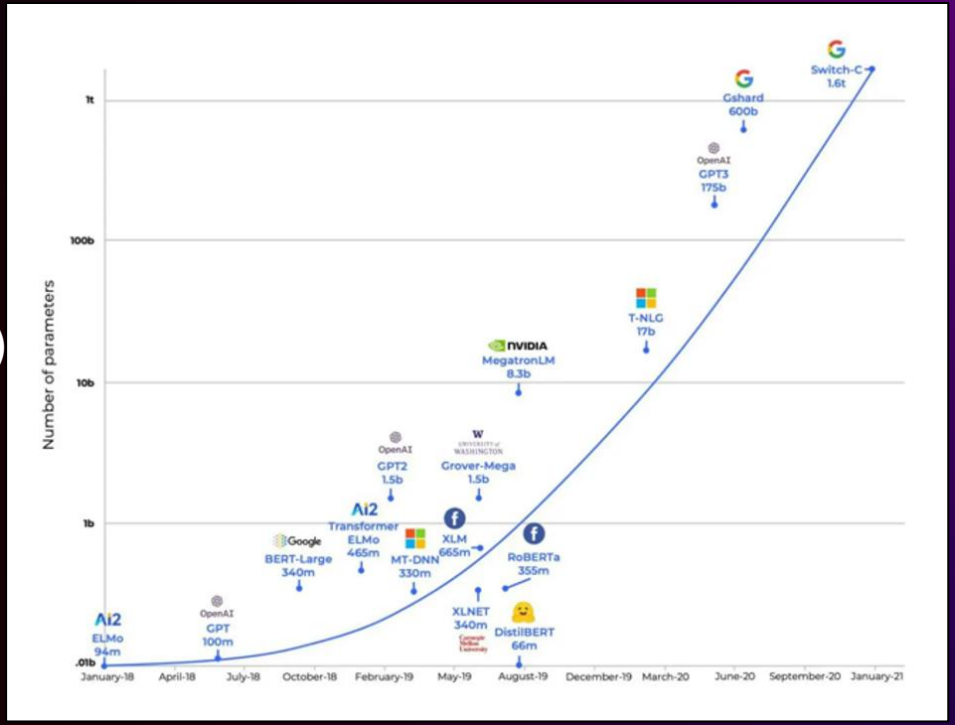




aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key trends in deep learning

- Rapid growth in size/complexity of models
  - Model/data/tensor parallelism constructions
  - Number of accelerators is increasing = inter-node/intra-node communication can become the bottleneck
  - Containers to scale workloads (AWS Batch/AWS ParallelCluster)
- Managing cost and complexity of training is challenging
  - Fault-tolerant training (checkpointing)
- Accelerator capacity is a key concern
- Proliferation of new software solutions and hardware
  - New frameworks/libraries – e.g., DeepSpeed, PyTorch FSDP



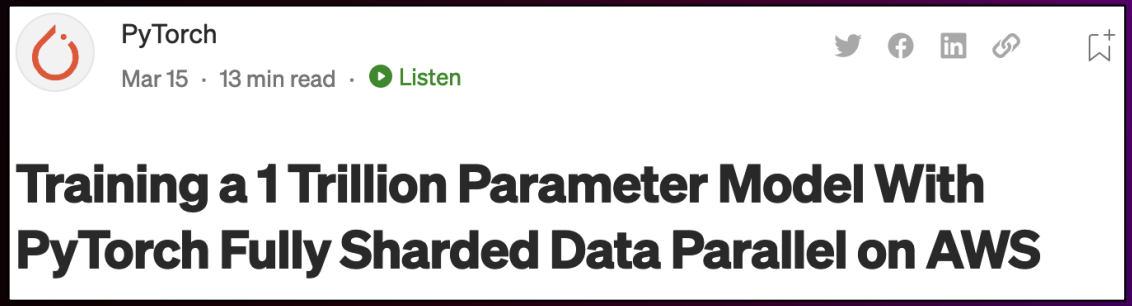
Source: https://bit.ly/3tN6UwY

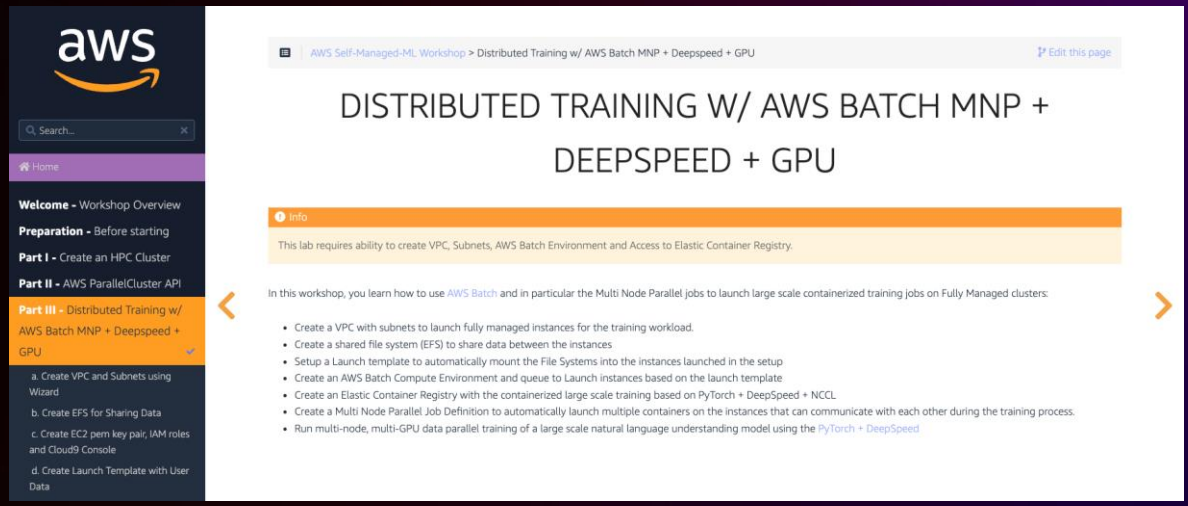
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Strategies to distributed training

REUSABLE ARCHITECTURES FOCUSING ON PRICE PERFORMANCE

- Parallelism strategies
  - Model/Data/Pipeline parallelism
- Linear scaling for high training efficiencies
  - Network bottlenecks (for example, PowerSGD)
  - Memory offloading (for example, FSDP and DeepSpeed)
- Diversification of accelerator types
  - Migration b/w accelerators (resume from checkpoints)
- Profiling
  - Node failures, resource utilization
- Data
  - Reproducibility with data
  - Fuel for large language models





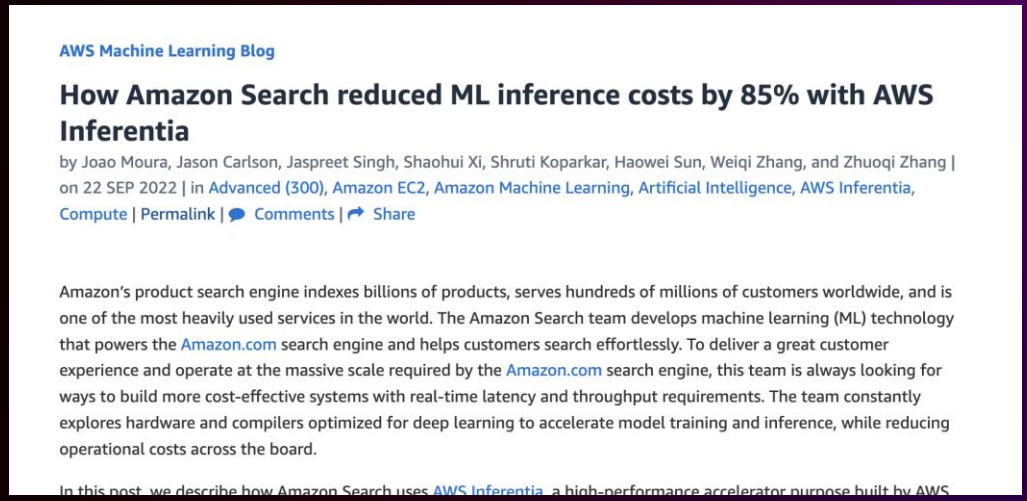
PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc.

aws

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

Challenges with ML Inference

- Decoupling training and inference
- Price-performance: latency, throughput, and cost
  - Sparse inference: medium latency/low throughput
  - Dense inference: low latency/high throughput
- Choose the right Inference option for your workload
  - Bursting vs provisioned
  - Real time vs batch/async



AWS Machine Learning Blog

How Amazon Search reduced ML inference costs by 85% with AWS Inferentia

by Joao Moura, Jason Carlson, Jaspreet Singh, Shaohui Xi, Shruti Koparkar, Haowei Sun, Weiqi Zhang, and Zhuoqi Zhang | on 22 SEP 2022 | in Advanced (300), Amazon EC2, Amazon Machine Learning, Artificial Intelligence, AWS Inferentia, Compute | Permalink | Comments | Share

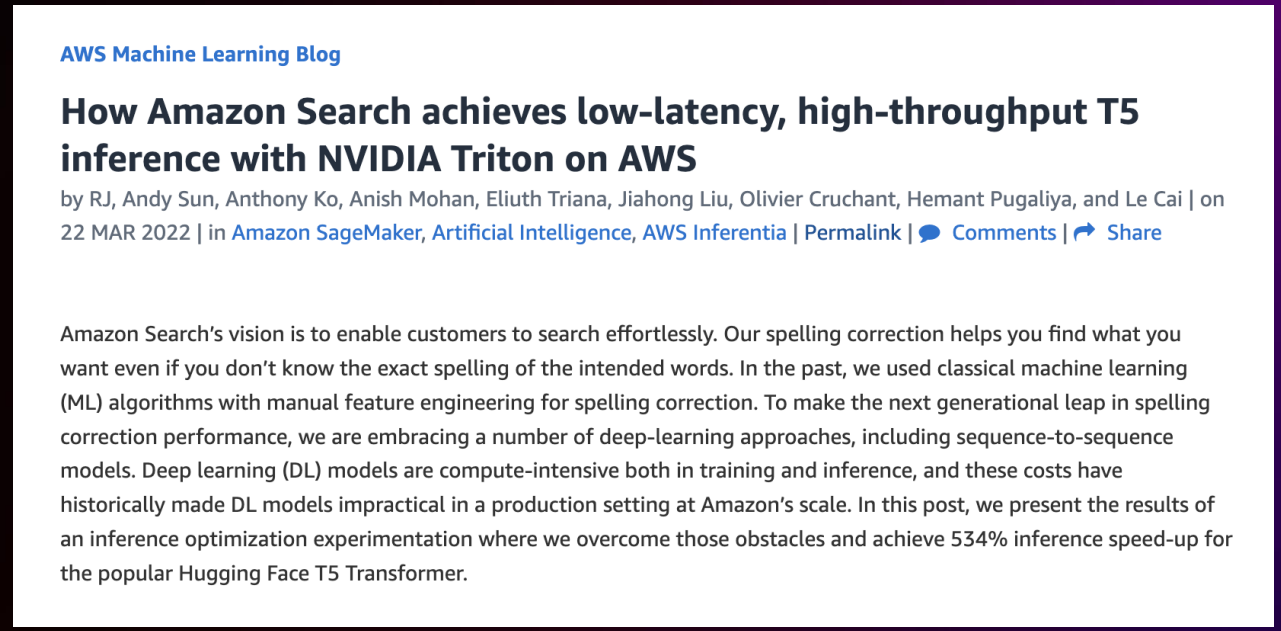
Amazon's product search engine indexes billions of products, serves hundreds of millions of customers worldwide, and is one of the most heavily used services in the world. The Amazon Search team develops machine learning (ML) technology that powers the Amazon.com search engine and helps customers search effortlessly. To deliver a great customer experience and operate at the massive scale required by the Amazon.com search engine, this team is always looking for ways to build more cost-effective systems with real-time latency and throughput requirements. The team constantly explores hardware and compilers optimized for deep learning to accelerate model training and inference, while reducing operational costs across the board.

In this post, we describe how Amazon Search uses AWS Inferentia, a high-performance accelerator purpose built by AWS

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

Triton Inference Server

- Run on SageMaker or managed services (eg. Amazon ECS)
- Multiple “backends” to support various types of models
- Support for NVIDIA GPUs and AWS Inferentia
- Optimized for throughput and latency



AWS Machine Learning Blog

How Amazon Search achieves low-latency, high-throughput T5 inference with NVIDIA Triton on AWS

by RJ, Andy Sun, Anthony Ko, Anish Mohan, Eliuth Triana, Jiahong Liu, Olivier Cruchant, Hemant Pugaliya, and Le Cai | on 22 MAR 2022 | in Amazon SageMaker, Artificial Intelligence, AWS Inferentia | Permalink | Comments | Share

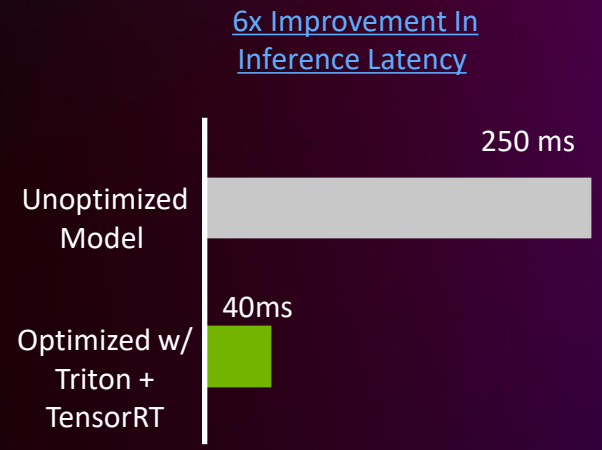
Amazon Search’s vision is to enable customers to search effortlessly. Our spelling correction helps you find what you want even if you don’t know the exact spelling of the intended words. In the past, we used classical machine learning (ML) algorithms with manual feature engineering for spelling correction. To make the next generational leap in spelling correction performance, we are embracing a number of deep-learning approaches, including sequence-to-sequence models. Deep learning (DL) models are compute-intensive both in training and inference, and these costs have historically made DL models impractical in a production setting at Amazon’s scale. In this post, we present the results of an inference optimization experimentation where we overcome those obstacles and achieve 534% inference speed-up for the popular Hugging Face T5 Transformer.

aws

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

Triton Inference Server + TensorRT

- Search services available on Amazon.com deployed worldwide with Triton and TensorRT.
- Leverage the latest TensorRT 8.2 to serve the optimized T5 models for real-time inference with Triton.
- Triton meets latency target while optimizing for throughput, meeting the real-time requirement of the services.
- Model analyzer reduced time to find optimal configuration from weeks to hours.





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

Amazon Search empowers teams across Amazon to bring deep learning to their ML applications using AWS/PyTorch

Challenges

Amazon Search wanted to create the centralized deep learning technology necessary to create large models for use by teams across Amazon

Solutions

Using various AWS services, including Amazon EC2, Amazon Search M5 achieved a technology solution for building ML models efficiently, cost-effectively, and at scale, empowering any team at Amazon to bring the power of deep learning to ML applications

Results

- Runs deep learning experiments with tens of billions of parameters
- Empowered teams across Amazon to run deep learning experiments
- Scales to thousands of training jobs per month

> *By continuing to increase our efficiency using AWS, we can unlock the possibilities of deep learning and artificial intelligence to benefit our customers.*

Rejith Joseph
Principal Engineer, Amazon Search

CUSTOMER PROFILE

INDUSTRY
Retail and Wholesale

REGION
United States

AWS Case Study

Amazon is an American multinational technology company that focuses on ecommerce, cloud computing, digital streaming, and artificial intelligence

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

M5

- Multi-modal → images, unstructured text, tables, and videos
- Multi-locale → products are sold in several countries
- Multi-lingual → several primary and secondary languages
- Multi-entity → relationships between customers, queries, sellers, and so on
- Multi-task → duplicate detection, ads, semantic matching in search, product type identification

aws

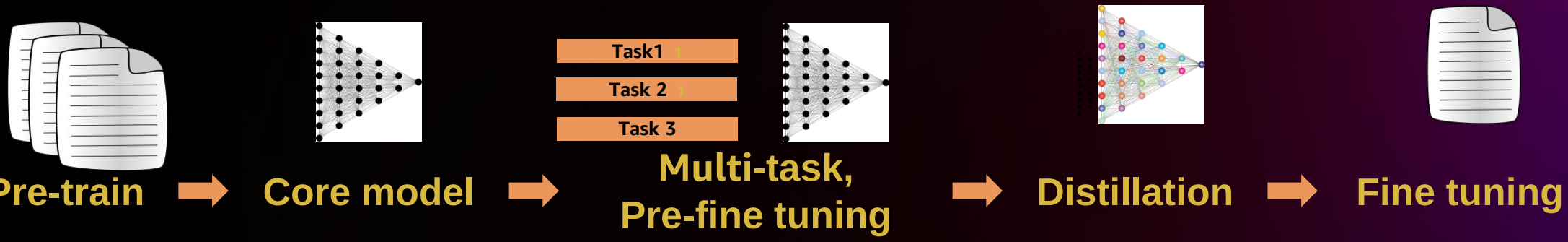
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

M5 mission and vision

- Build foundational, universal, semantic representation of important Amazon entities like products, queries, and shopping sessions
- Powered by large-scale deep learning systems
- Deploy to teams across Amazon

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

M5 life cycle



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

Large-model training frameworks

- DeepSpeed for training large models primarily due to the support of training large models with limited hardware by using ZeRO Optimization

- DeepSpeed is built on top of PyTorch distributed



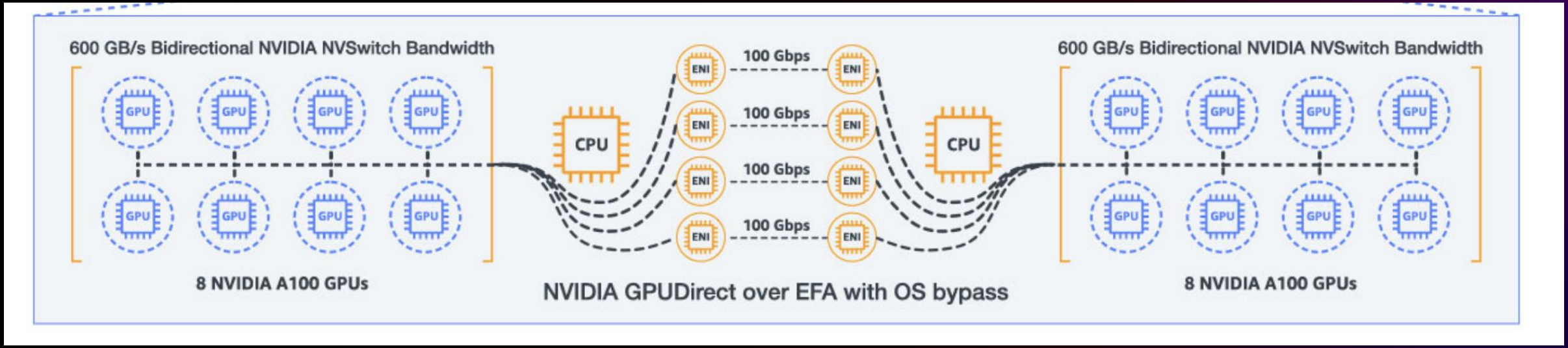
PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc.

aws

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

Hardware

- AWS GPU Hardware P3DN, P4DN, and G4DN with high throughput network
- Achieved 4.37x performance for P4DN cluster over P3DN cluster using AWS-optimized NCCL and EFA



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

Framework optimizations

Scaling to trillion-parameter model training on AWS

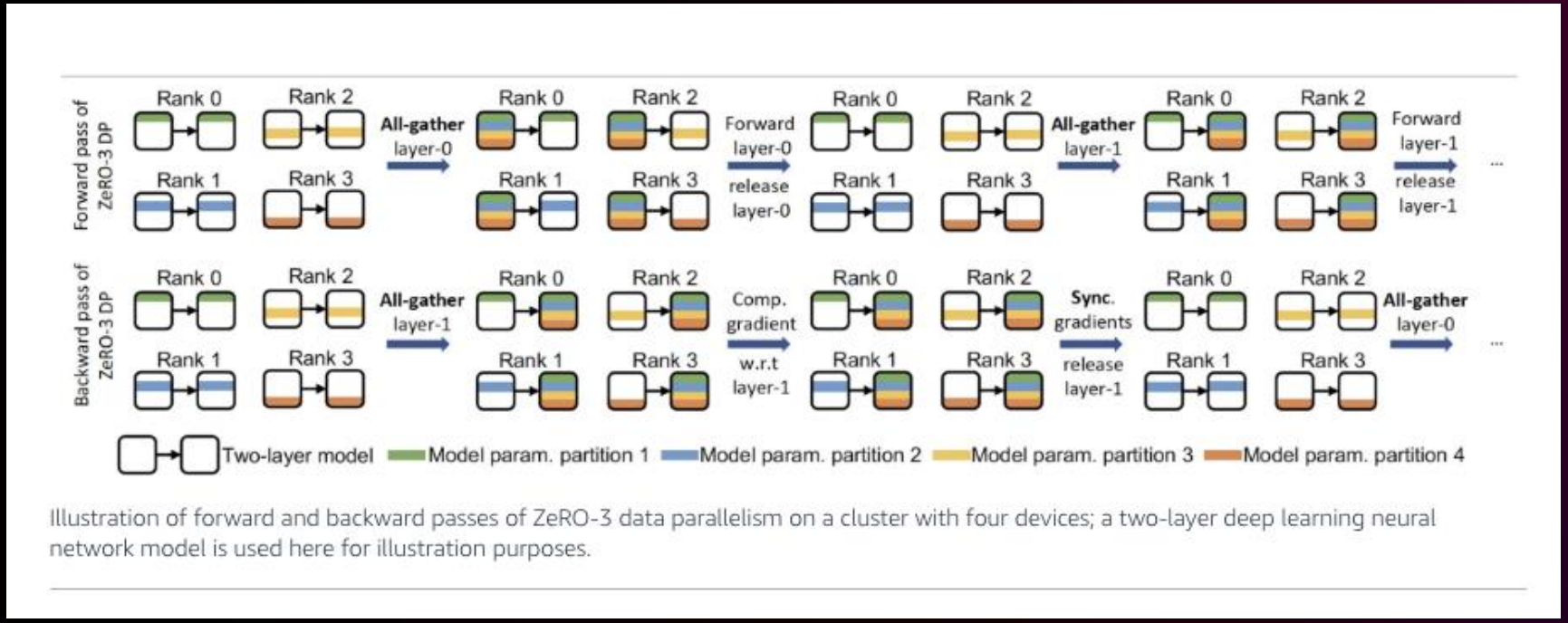


Illustration of forward and backward passes of ZeRO-3 data parallelism on a cluster with four devices; a two-layer deep learning neural network model is used here for illustration purposes.

https://www.amazon.science/blog/scaling-to-trillion-parameter-model-training-on-aws

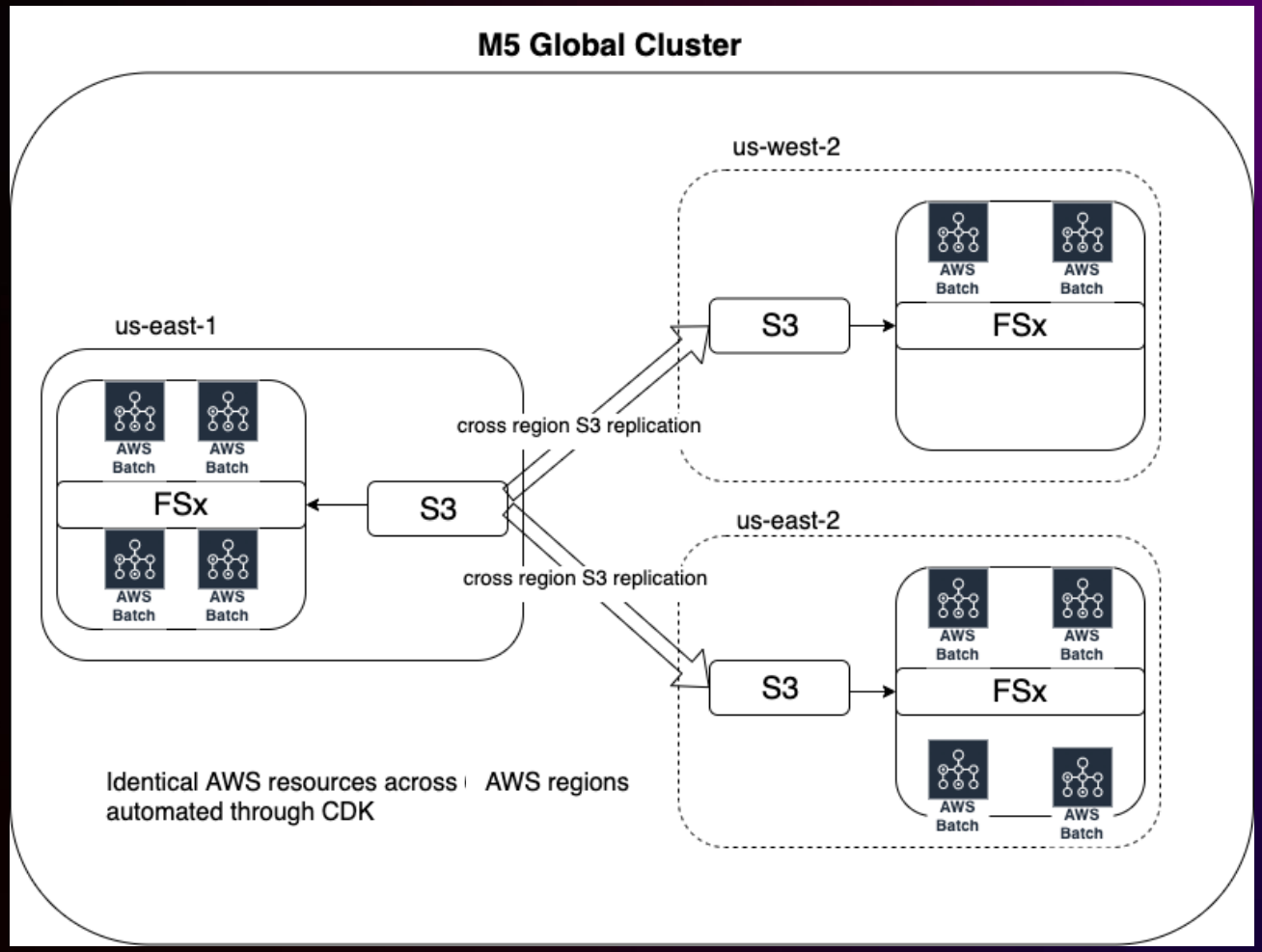
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Training workflow

GPU instances are scarce and hard to get provisioned in one Region

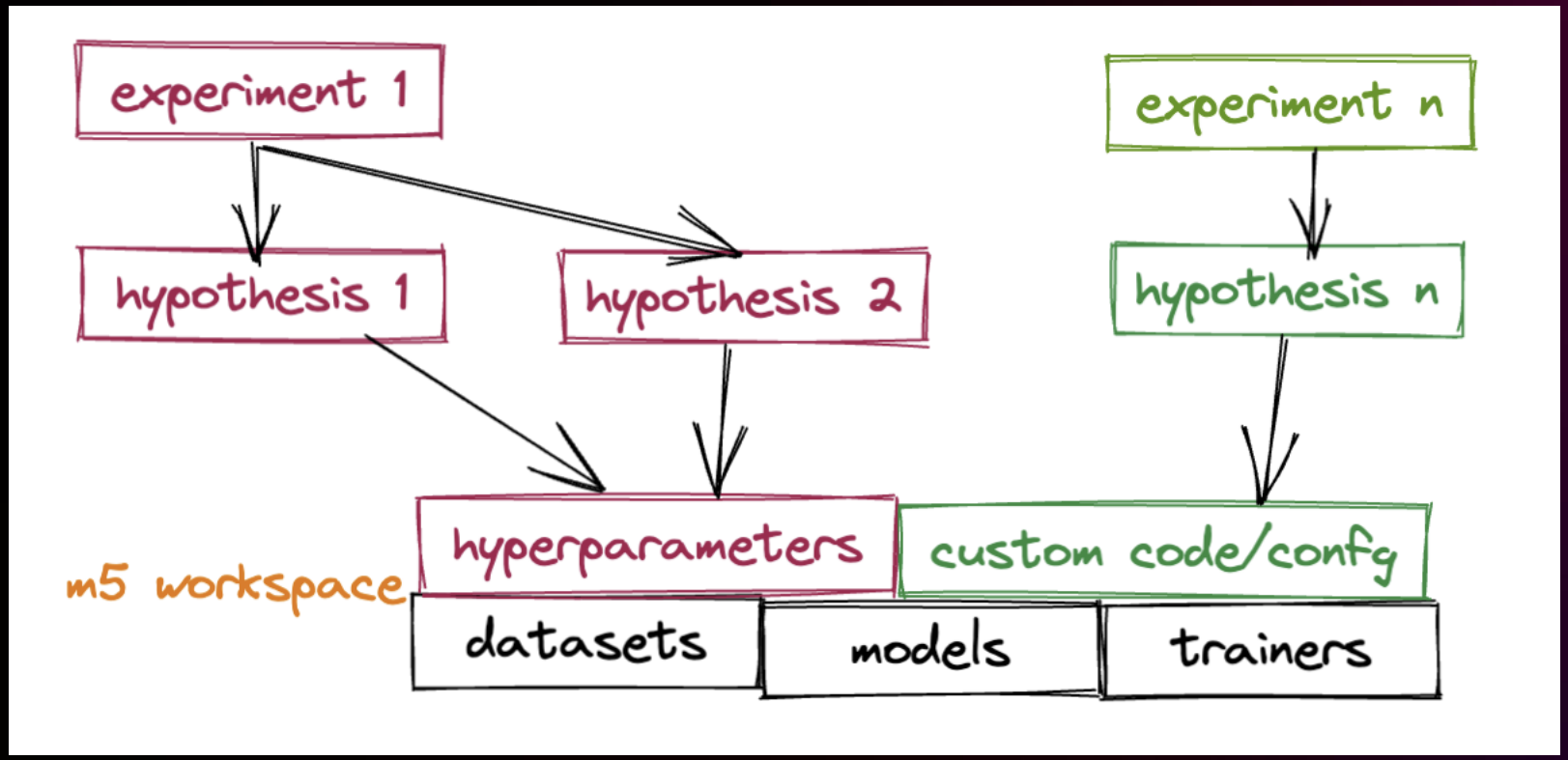
Developers get access to Amazon EC2 across multiple AWS Regions

AWS Batch for queueing



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

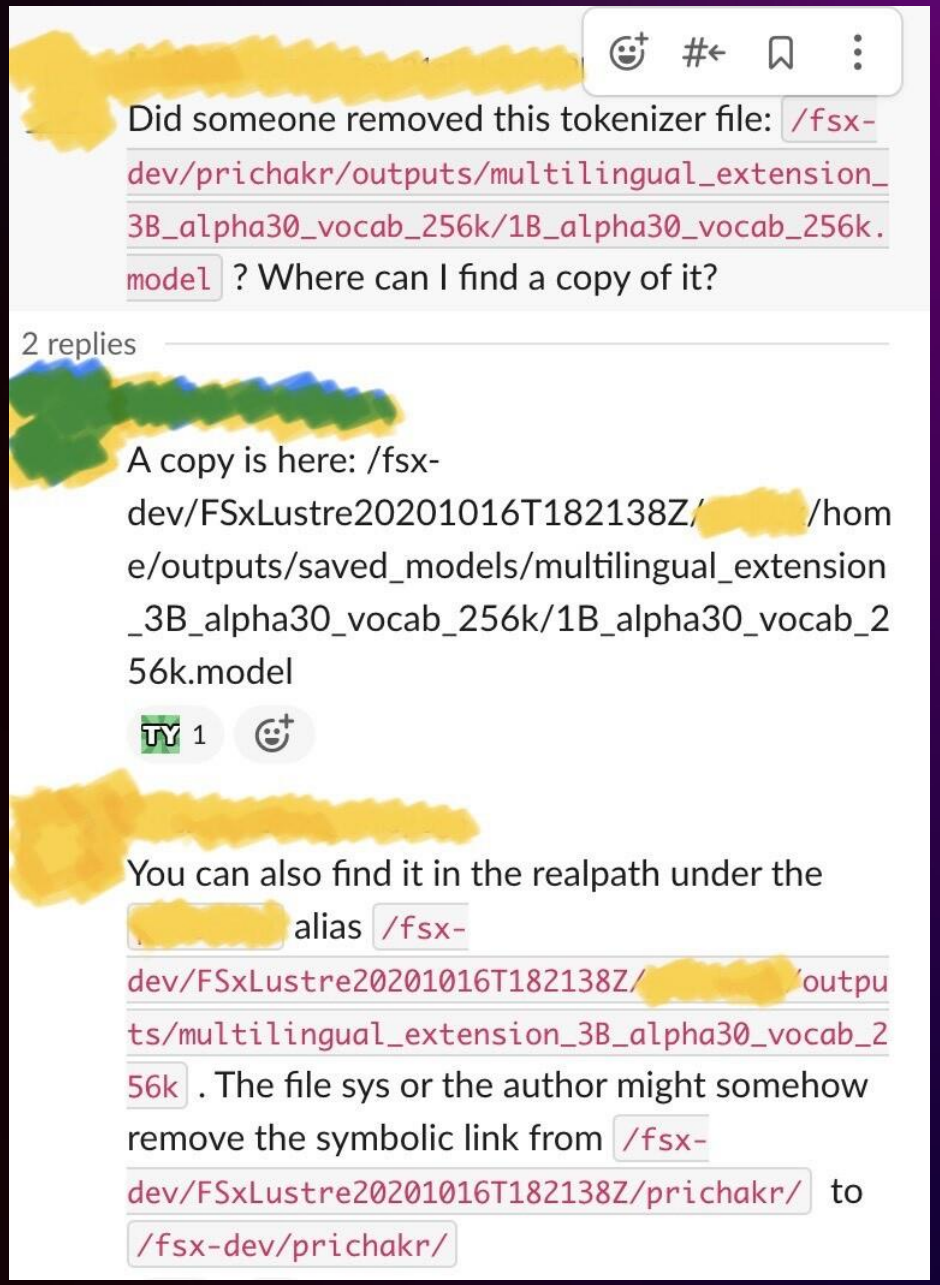
M5 Experiment



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

Reproducibility

- Reproducibility – How do things get lost?
  - Tracking on docs
  - Artifacts in per-region silos
  - No versioning on storage
  - Losing track of data pre-processing
- Why is this important?
  - Partner trust on M5 results
  - Incremental improvements to models



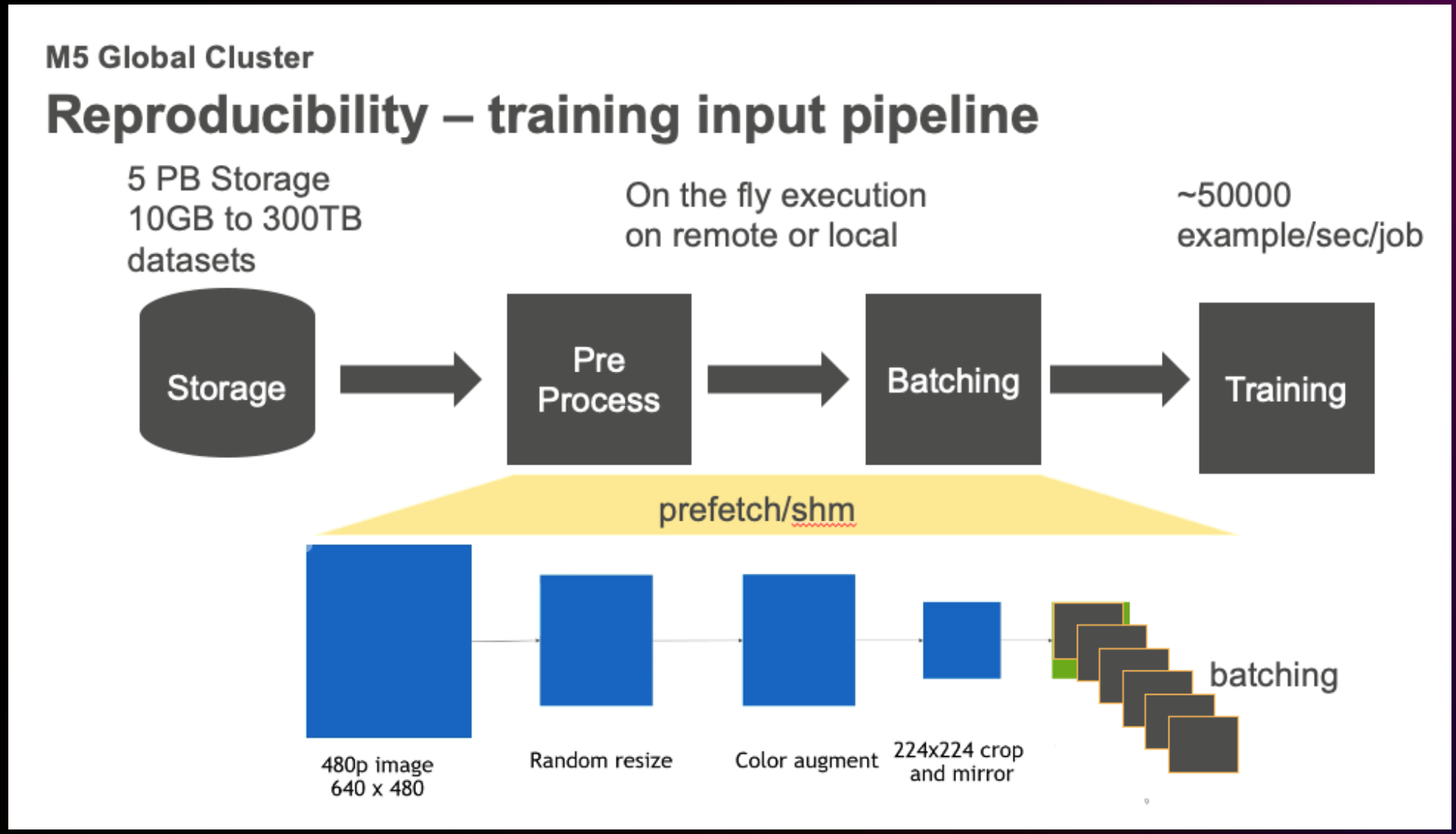
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reproducibility (solution)

- Global access to data and artifacts
- On-the-fly data processing on CPU cores
- Experiment tracking – MLflow
- Workspace snapshotting

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

Reproducibility storage



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

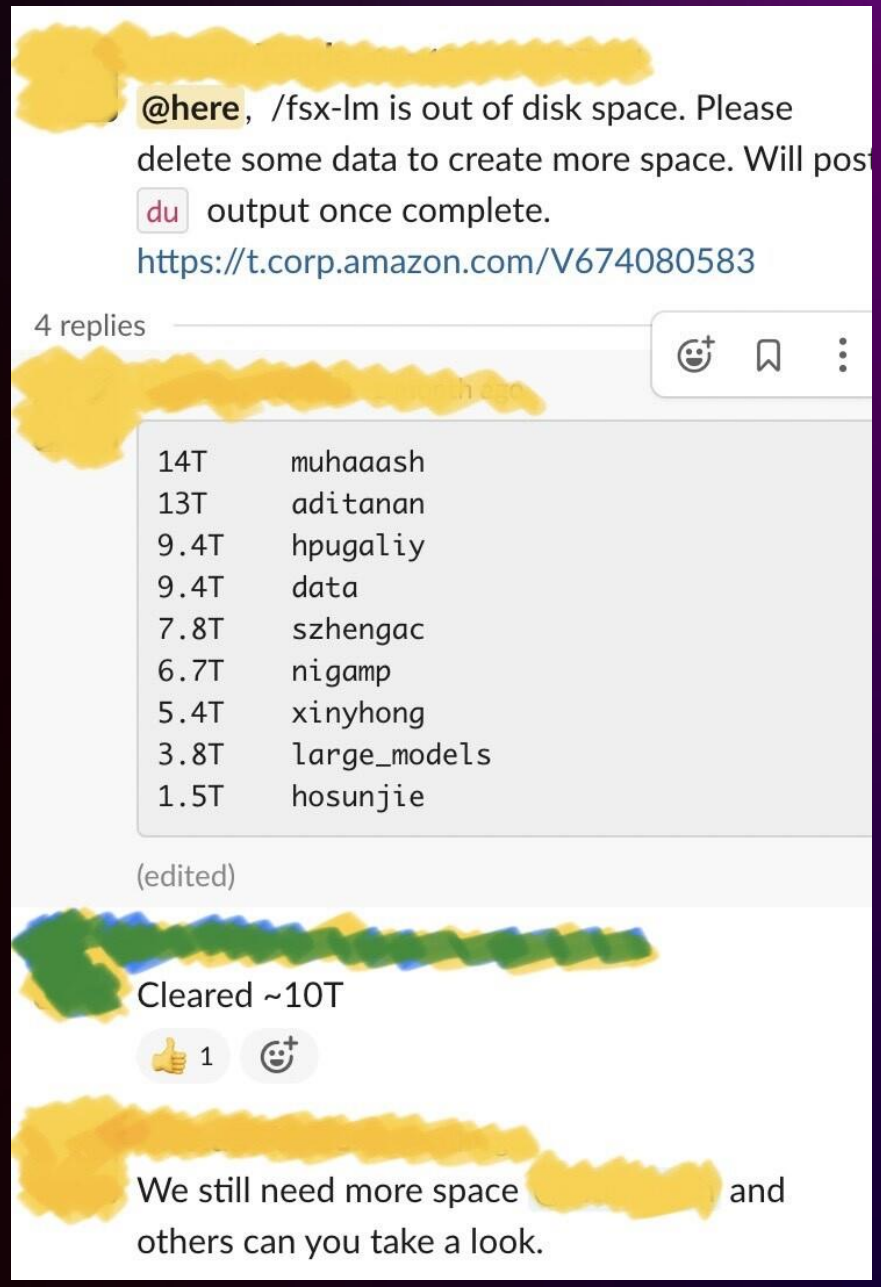
Reliability

- Long-running clusters and jobs
- Stuck-ness across the stack
  - Bad nodes and stuck GPUs

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

Reliability

- Reliability – How do jobs get stuck?
  - Long-running clusters and jobs
  - Stuck-ness across the stack
    - Noisy neighbors on shared resources



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

Reliability solution

- Stuck jobs detectors (job heartbeats)
- Canary jobs (detect perf regressions)
- Per-user isolation (storage/compute)
- Frequent checkpoints (recover)

aws

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

Debuggability

- Experiment Tracker through MLFlow
- Automatic error classification with Slack Notification
- CI/CD frequently – pre-flight CR checks
- Performance debugging

aws

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

Model vending

- Amazon SageMaker model vending
- PyTorch model serialization

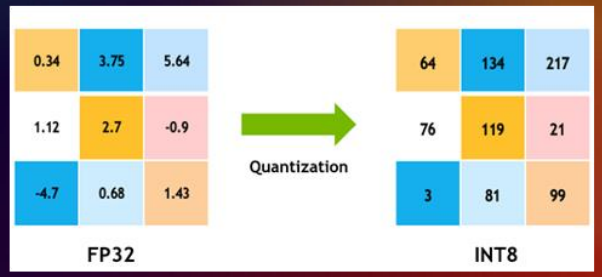
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

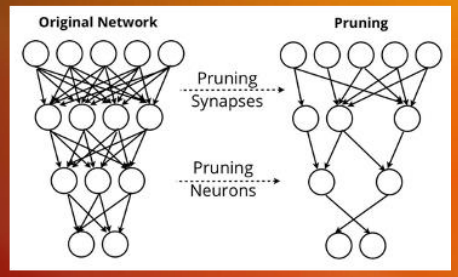
M5 inference optimization

Algorithm and data structure

SDK and model format

Hardware

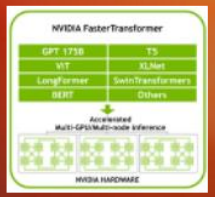




AWS Neuron

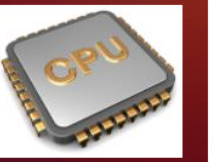
TensorRT

Onyx



AWS Inferentia





NVIDIA
TRITON INFERENCE SERVER

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

Results

- Multiple 100B Bert encoder Models trained and converged
- 10k+ training experiments/jobs per month
- Models trained on AWS Trainium, P4DN, P3DN and DL1
- Less than 10 ms latency for 1B encoder models

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

Collaborations

- AWS AI services
- AWS Batch
- Nvidia
- Meta PyTorch
- Amazon FSx

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

Takeaways

- As your team grows, reduce cost and friction to experiments
  - M5 did this by focusing on reproducibility, reliability, and debuggability
- Training large models to convergence remains a challenge
  - M5 picked two-click debugging and checkpoints, and did a lot of experiments
- Profile, observe, and build performance expertise – plus engage with AWS experts

aws

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

Thank you!

Rejith Joseph (RJ)
rgeorgej@amazon.com

James Park
jamesprk@amazon.com



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

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