AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

CMP334

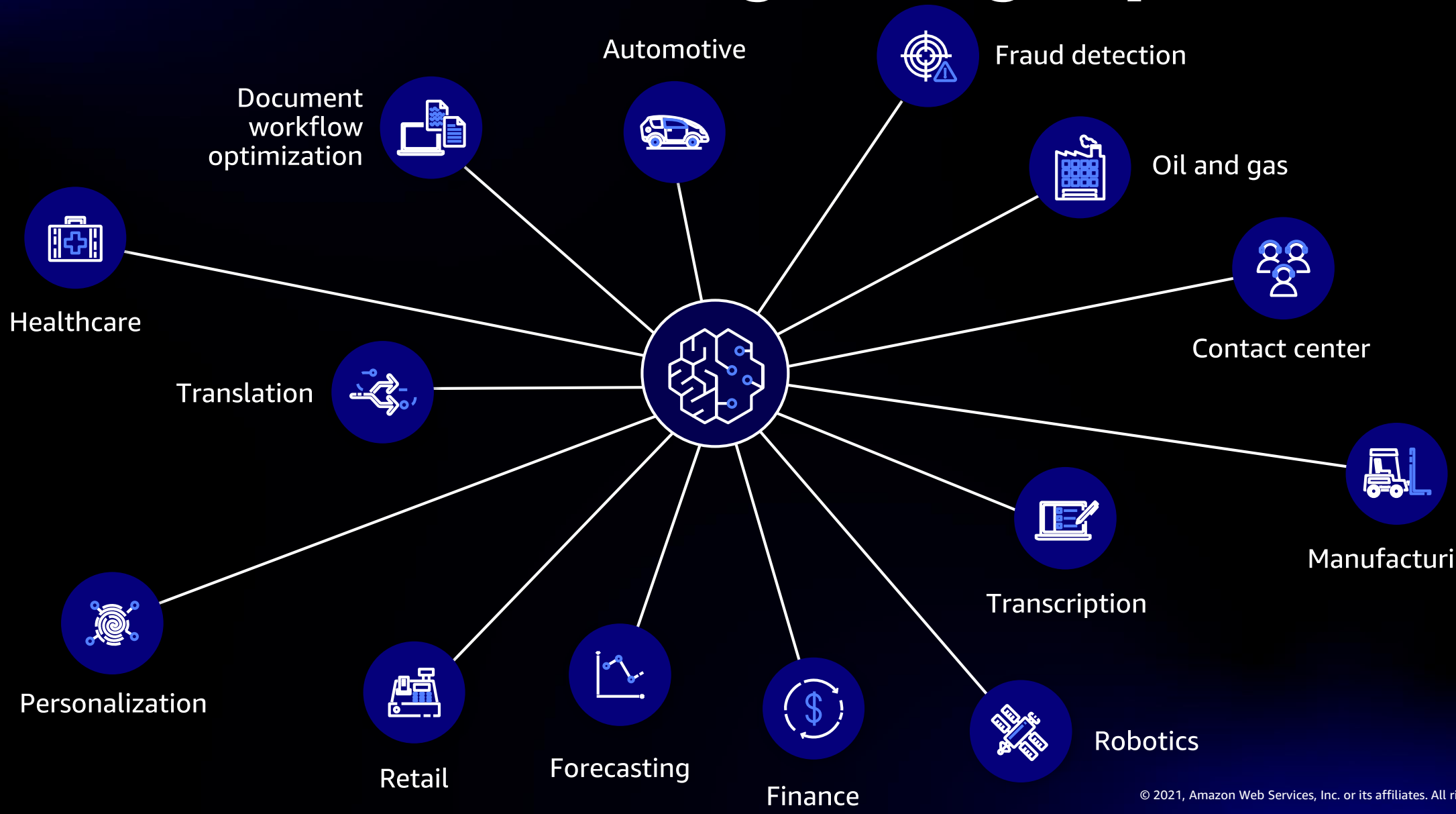
Introducing AWS Trainium-based Amazon EC2 Trn1 instances

Chetan Kapoor
Director, EC2 Core Product Management
AWS

Ron Diamant
Sr. Principal ML Engineer
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use cases for AI/ML are growing exponentially



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Companies of every size and industry are developing and deploying ML

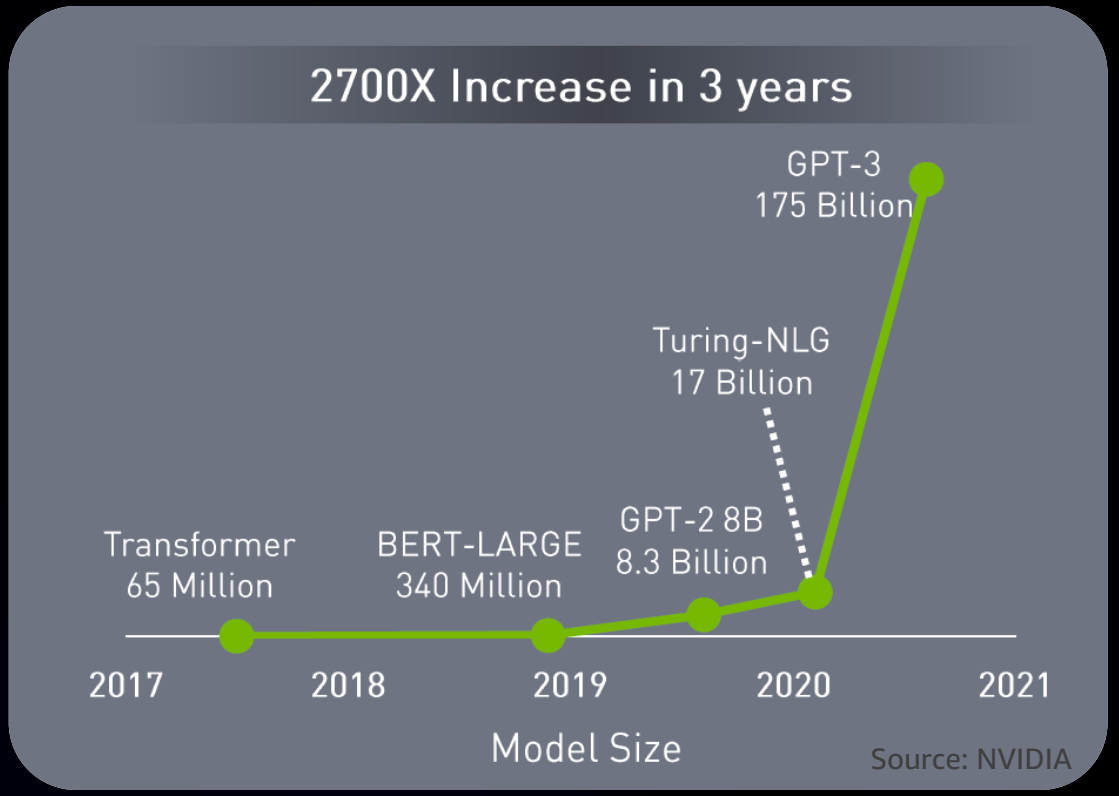
TENS OF THOUSANDS CUSTOMERS HAVE CHOSEN AWS FOR THEIR ML/DL WORKLOADS



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key trends in AI/ML infrastructure

1. Models are becoming more complex, with end users moving from classical ML to deep learning
2. Time and cost to train these models have exploded from days to weeks
3. Similar challenges with running these models in production, serving customer traffic
4. Data scientists and ML engineers want to continue using software tools and hardware platforms that best fit their use case and experience



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

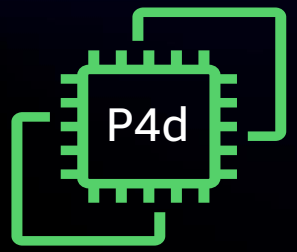
Innovation across stack for ML/DL acceleration

Managed services	Amazon SageMaker for end–end ML/AI services	EKS, ECS support	Native ML framework integration
Data center	EC2 UltraClusters with 10K+ Accelerators	Petabit Scale non-blocking network	MPI style network interface
Server	AWS Nitro with AWS storage and network adapters	Up to 16 interconnected accelerators per server	Up to 800 Gbps network, 80 Gbps EBS
Chip	annapurnalabs an amazon company — CPUs and DL accelerators	NVIDIA — GPUs	AMD — GPUs	intel — Custom CPUs and DL accelerators

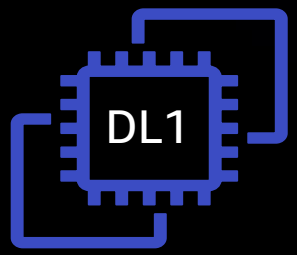
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Industry leading DL acceleration choices

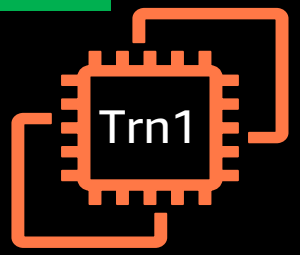


8 x NVIDIA A100,
400 Gbps, UltraClusters multi-node, high-performance training

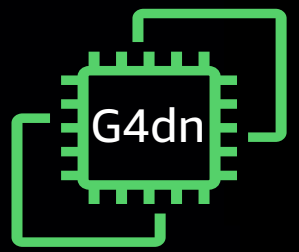


8 x Intel/Habana Gaudi,
400 Gbps, cost-efficient training

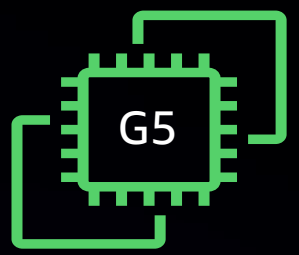
NEW



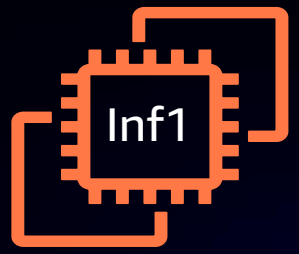
16 x AWS Trainium,
800 Gbps, UltraClusters, high performance, best price/performance



8 x NVIDIA T4,
100 Gbps, inference and cost-efficient small scale training



8 x NVIDIA A10G,
100 Gbps, cost-efficient small scale training, and high-performance inference



16 x AWS Inferentia,
100 Gbps, lowest cost DL inference in cloud

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Trn1 instances

THE MOST COST-EFFICIENT DL INSTANCE IN THE CLOUD

Trn1

MATH ENGINE FREQUENCY
3 GHz

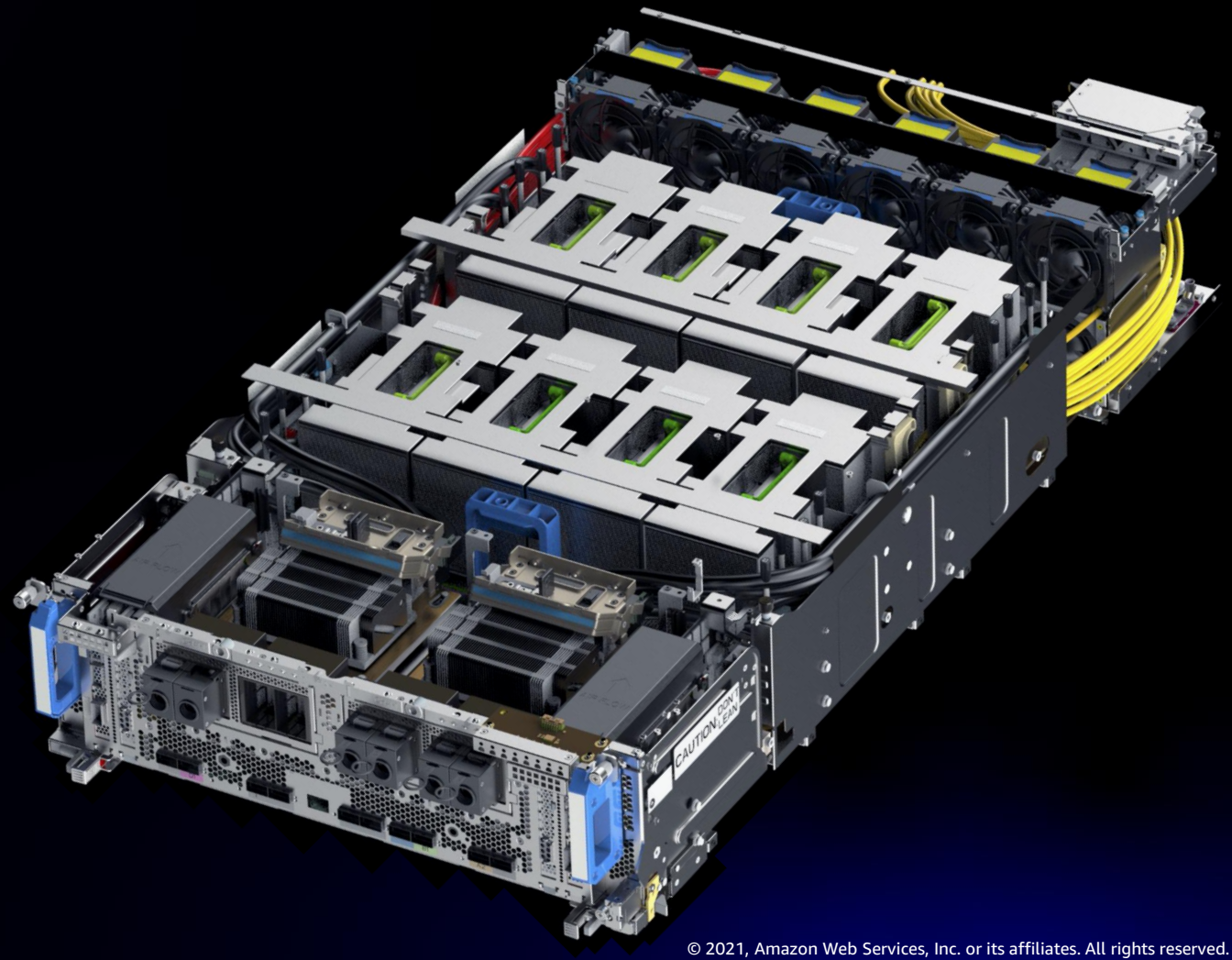
BF16/FP16	TF32	FP32
3.4 PFLOPS	**3.4 PFLOPS**	**840 TFLOPS**

AGGREGATE ACCELERATOR MEMORY
512 GB

PEAK MEMORY BANDWIDTH
13.1 TB/sec

NEURONLINK BANDWIDTH BETWEEN CHIPS
768 GB/sec

NETWORK CONNECTIVITY
800 Gbps EFA

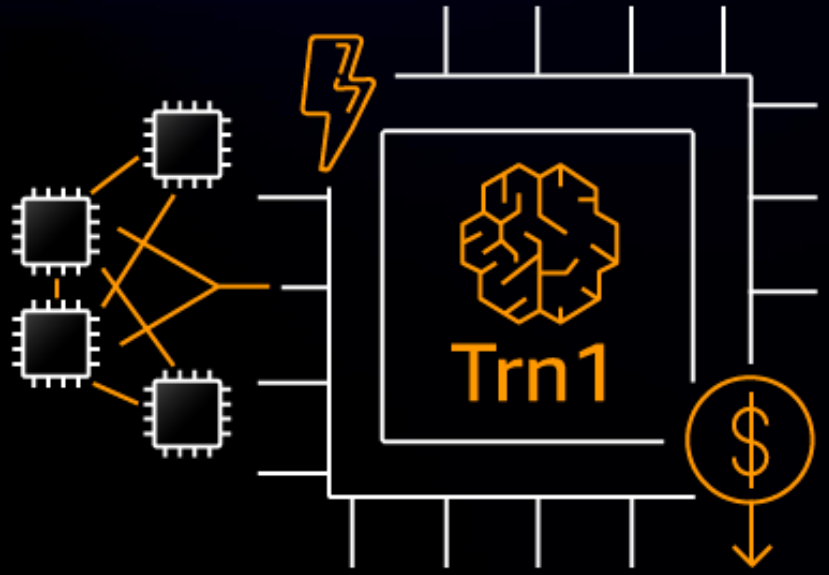


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Trn1 instances

In preview **now** – general audience in H1'22

THE MOST COST-EFFICIENT DL TRAINING IN THE CLOUD



- 60% higher accelerator memory (vs. P4d)
- 2x network BW (vs. P4d)
- Native support for PyTorch and TensorFlow
- Train on Trn1 and deploy anywhere

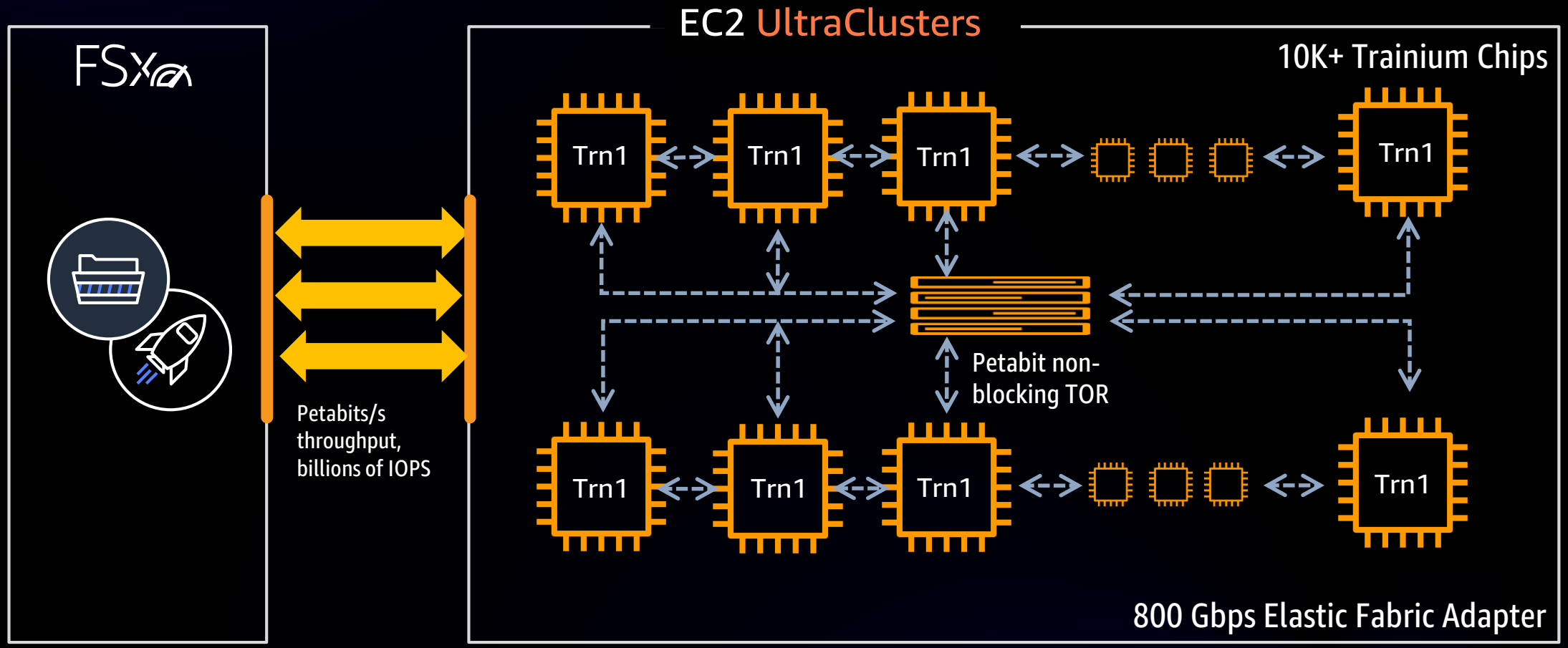
Instance size*	vCPUs	Trainium chips	Accelerator memory	NeuronLink	Instance memory	Instance networking	Local instance storage
Trn1.2xlarge	8	1	32 GB	N/A	32 GB	Up to 10Gbps	500 GB NVMe
Trn1.32xlarge	128	16	512 GB	**768 GB/sec**	512 GB	**800 Gbps**	8 TB NVMe

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

UltraCluster scale out for ultra-large models

10K+ TRAINIUM ACCELERATORS PER EC2 ULTRACLUSTERS



On-demand access to a world-class supercomputer 2.1 exaflops of compute

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS portfolio integration

Frameworks & workflow services

Pytorch

TensorFlow

Amazon Sagemaker

AWS Deep Learning AMIs

Amazon EKS

Amazon ECS

AWS Deep Learning Containers

Networking & storage

Elastic Fabric Adapter

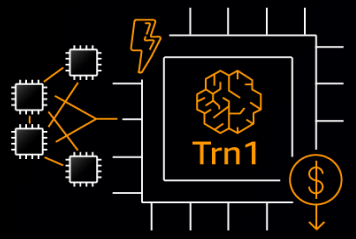
Amazon S3

Amazon EBS

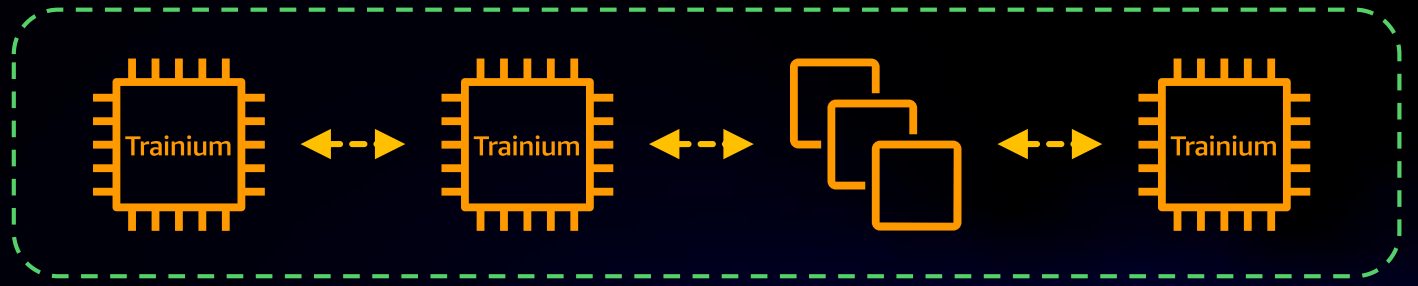
Amazon FSx for Lustre

Amazon EFS

Compute acceleration



Amazon EC2 Trn1

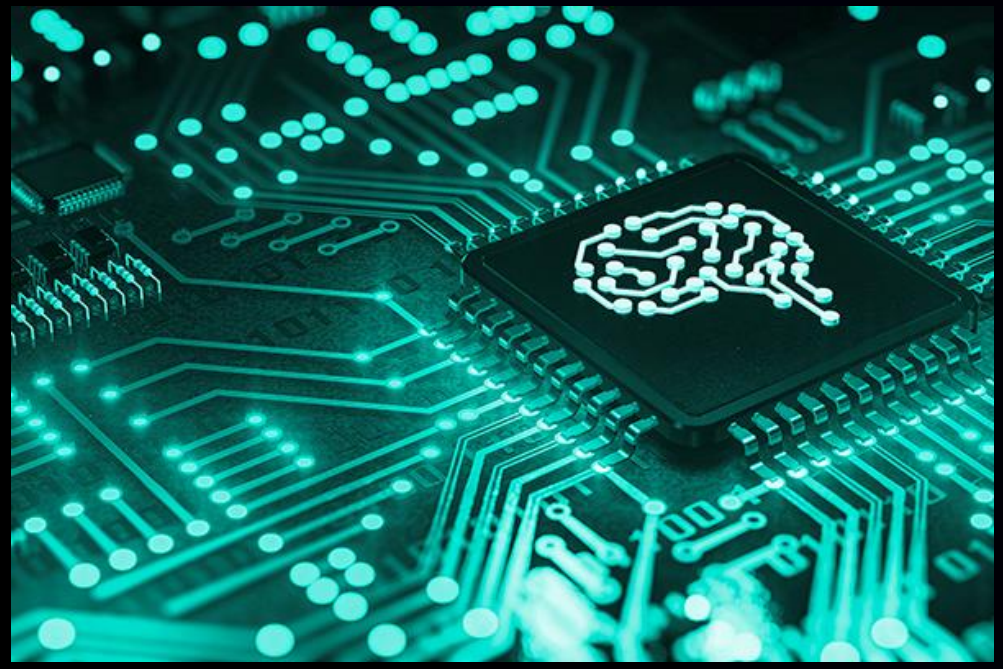


EC2 Trn1 UltraCluster

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers



ANTHROP\C

“A major key to our success is access to modern infrastructure that allows us to spin up very large fleets of high-performance deep learning accelerators. We are looking forward to using Amazon EC2 Trn1 instances powered by AWS Trainium, as their unprecedented ability to scale to tens of thousands of nodes and higher network bandwidth will enable us to iterate faster while keeping our costs under control.”

Tom Brown, Co-Founder at Anthropic

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customers



“Sprinklr's natural language processing and computer vision ML models analyze different data formats sourced from publicly available social media posts, blog posts, video content, and other content available on public domains across more than 30 channels. Based on our value from using AWS Inferentia, we are eager to try AWS Trainium to improve time to train and lower training costs for our models. We look forward to developing our models on these high-performance and low-cost training instances.”

Vasant Srinivasan, Senior Vice President of Product Engineering at Sprinklr

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

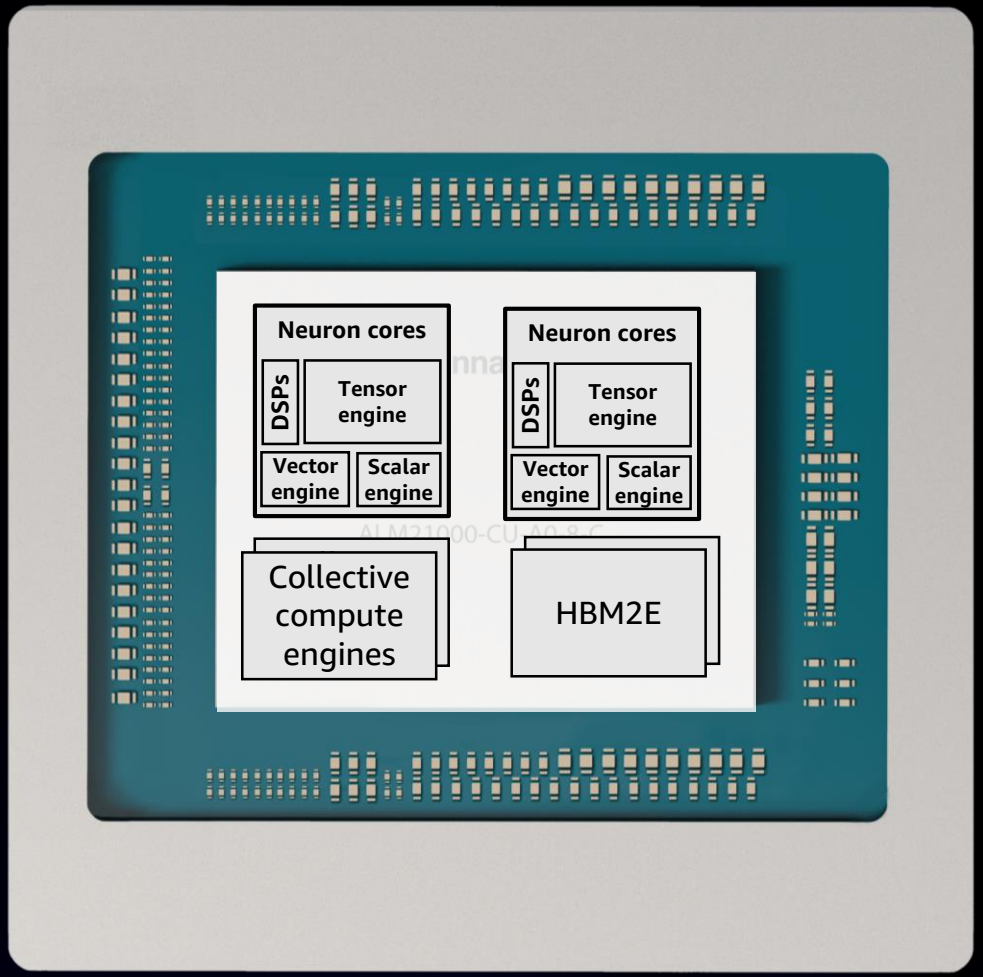
Trainium architecture

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Trainium

- 2 neuron cores
- Dedicated collective compute engines
- Tensor, scalar, and vector engines
- Embedded general purpose DSPs

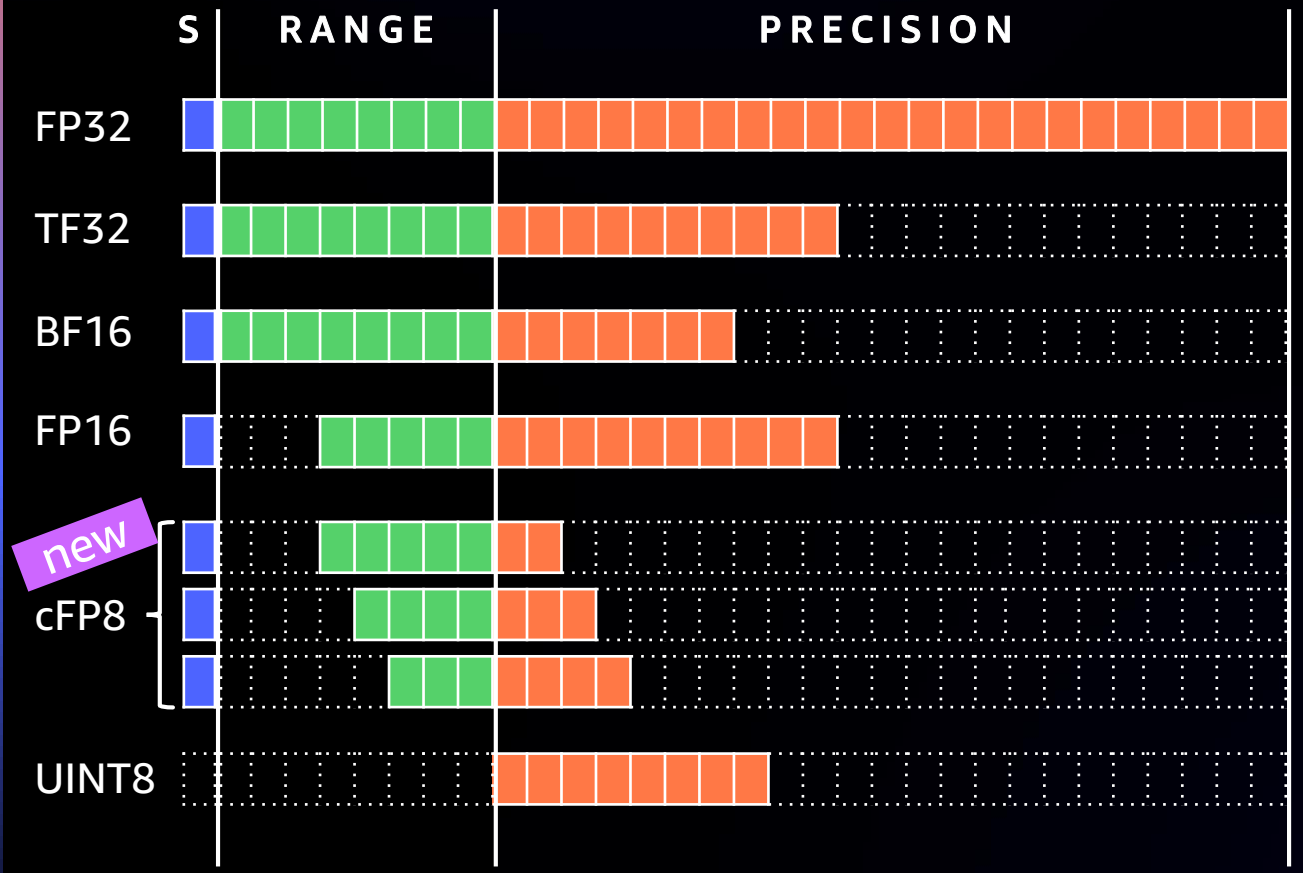
  Support for custom operators



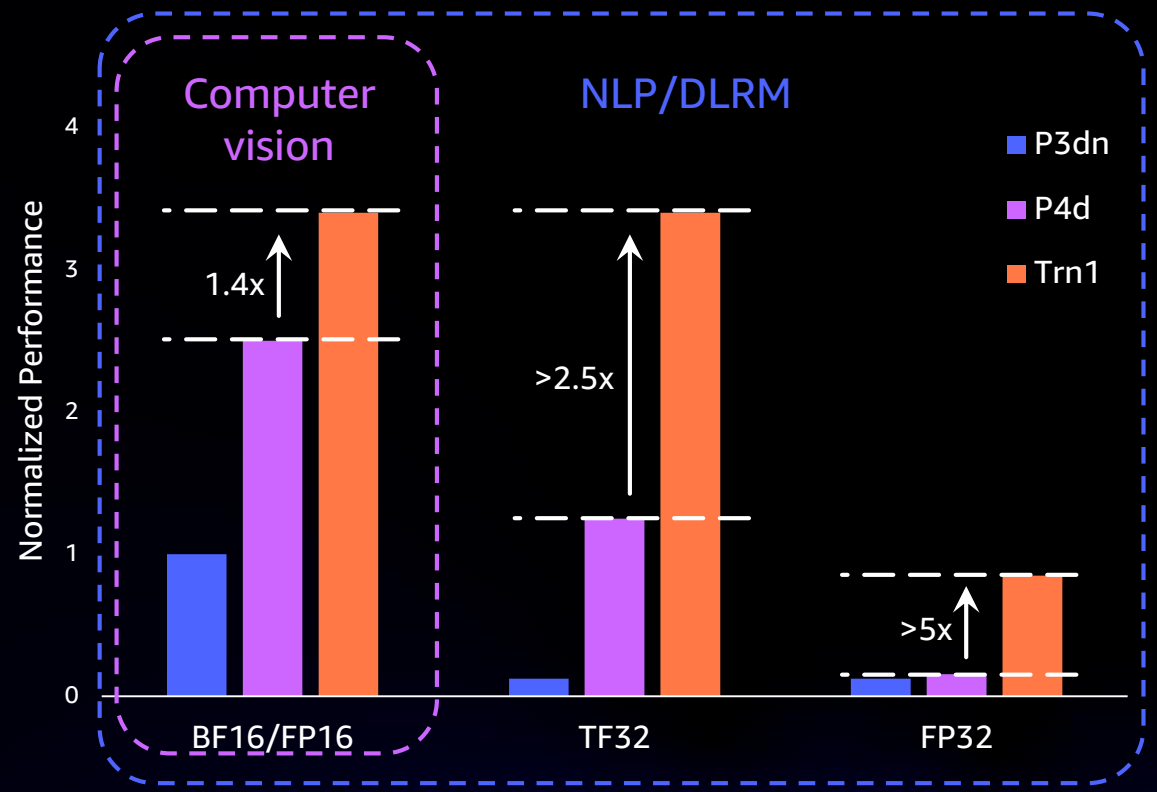
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – fastest across data types

SUPPORTED DATA TYPES



TFLOPS PER DATA TYPE



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

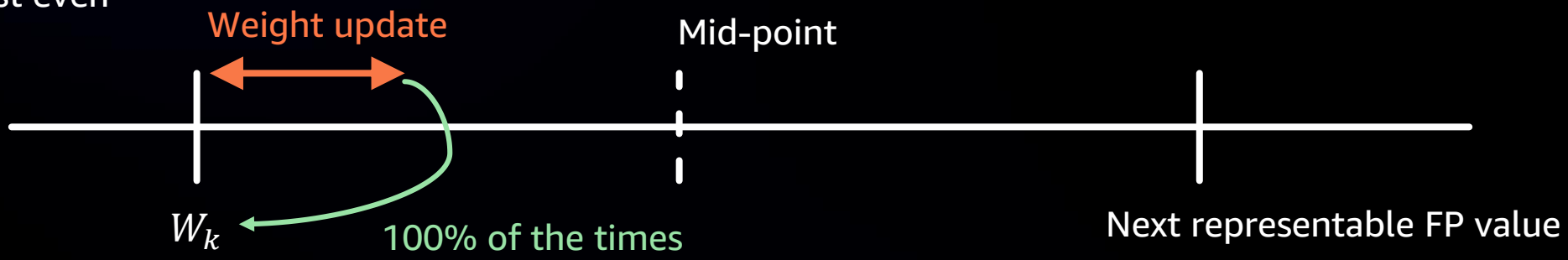
Stochastic rounding

- Programmable rounding modes in hardware:

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Stochastic rounding

- Programmable rounding modes in hardware:
  - Round nearest even



aws

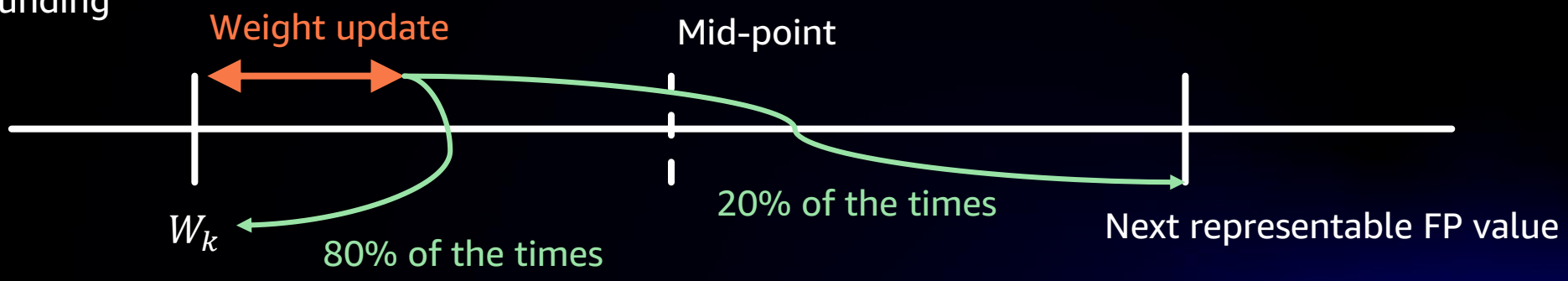
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Stochastic rounding

- Programmable rounding modes in hardware:
  - Round nearest even



  - Stochastic rounding

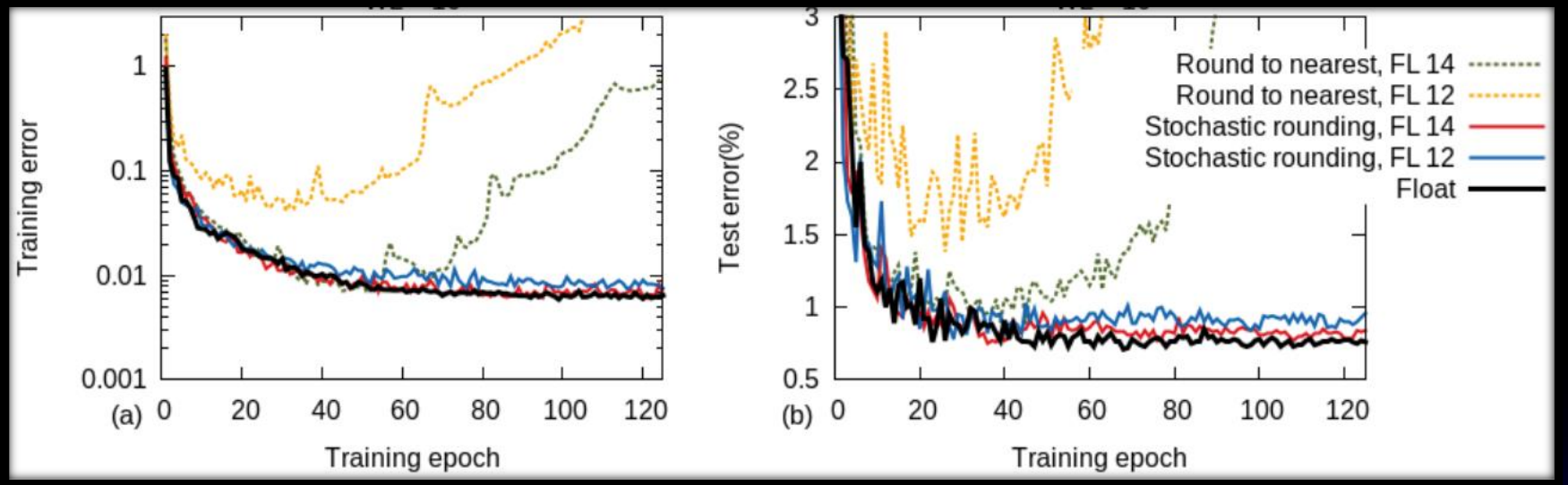


aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Stochastic rounding

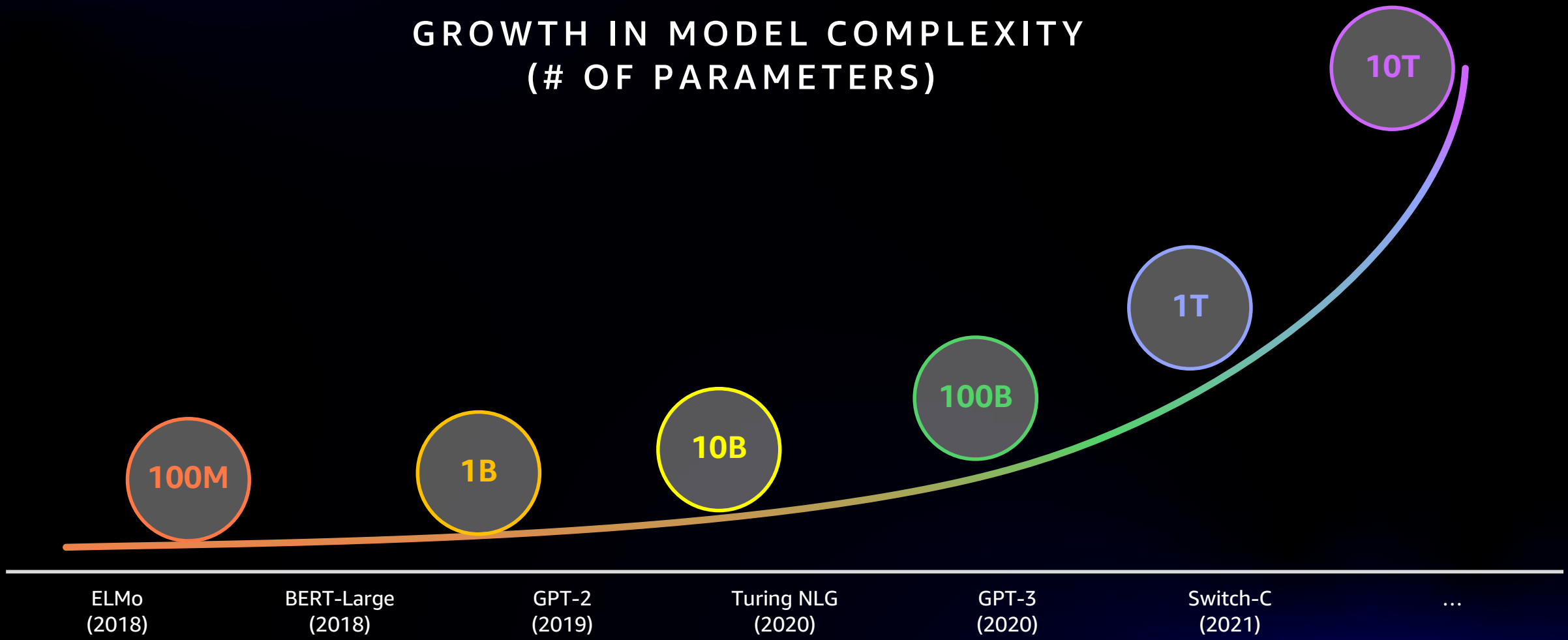
- Programmable rounding modes in hardware:
  - Round nearest even
  - Stochastic rounding
- Higher accuracy with stochastic rounding!



https://arxiv.org/pdf/1502.02551.pdf

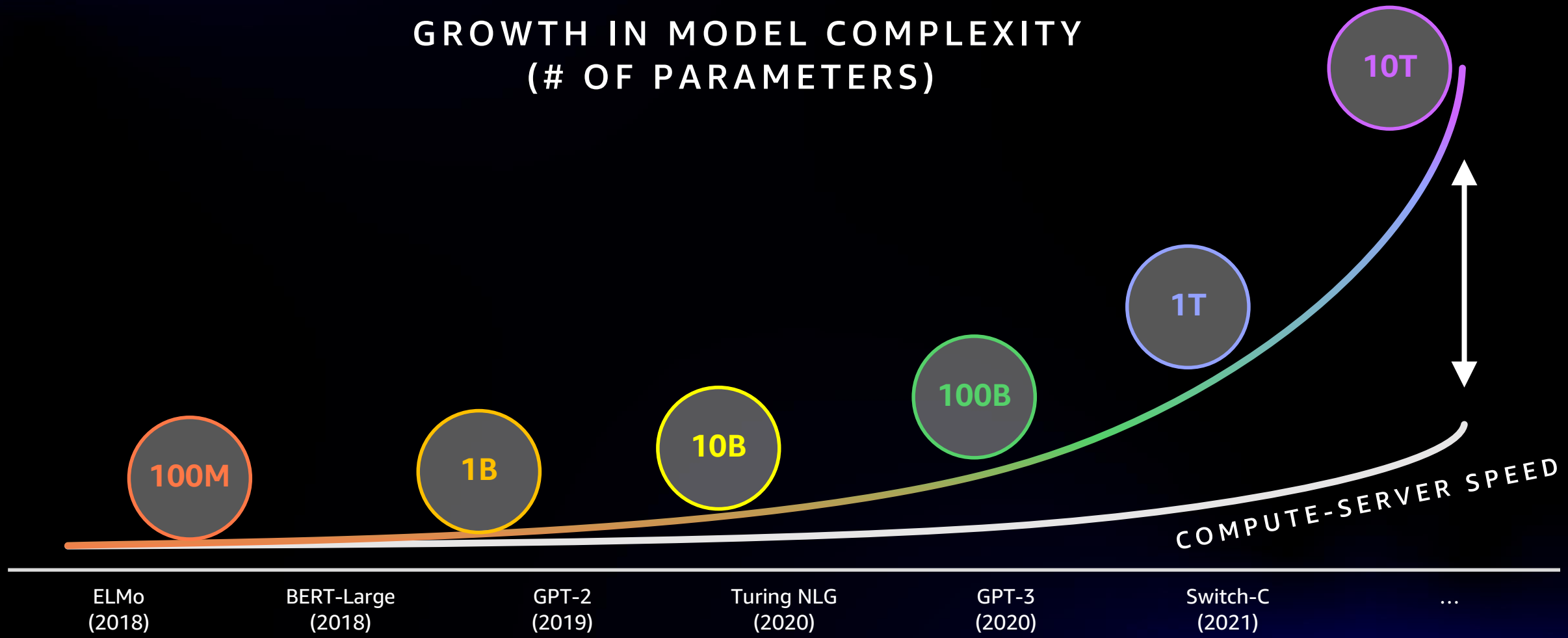
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The need for scale-out training



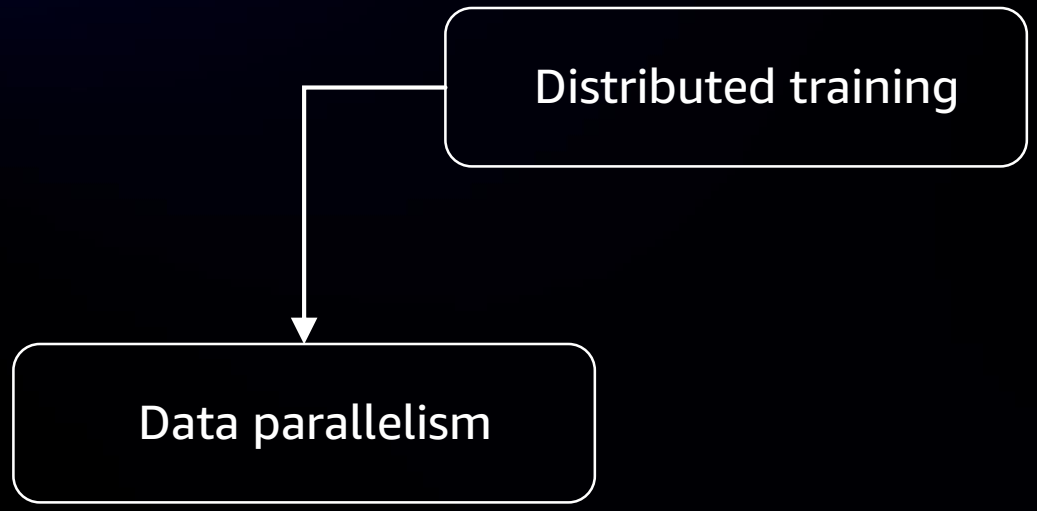
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The need for scale-out training



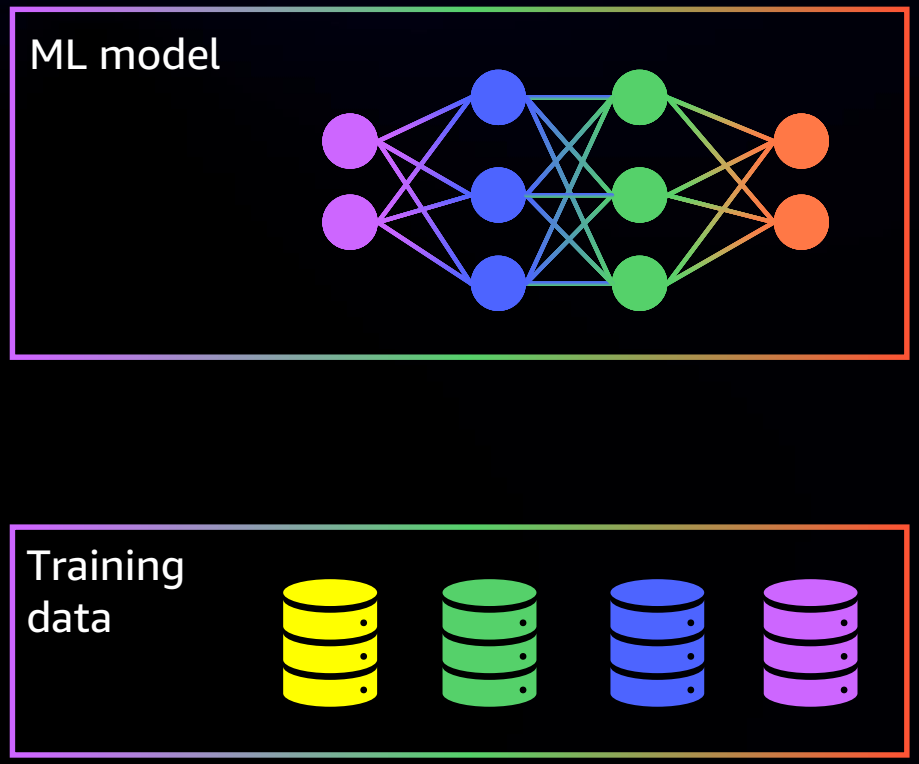
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

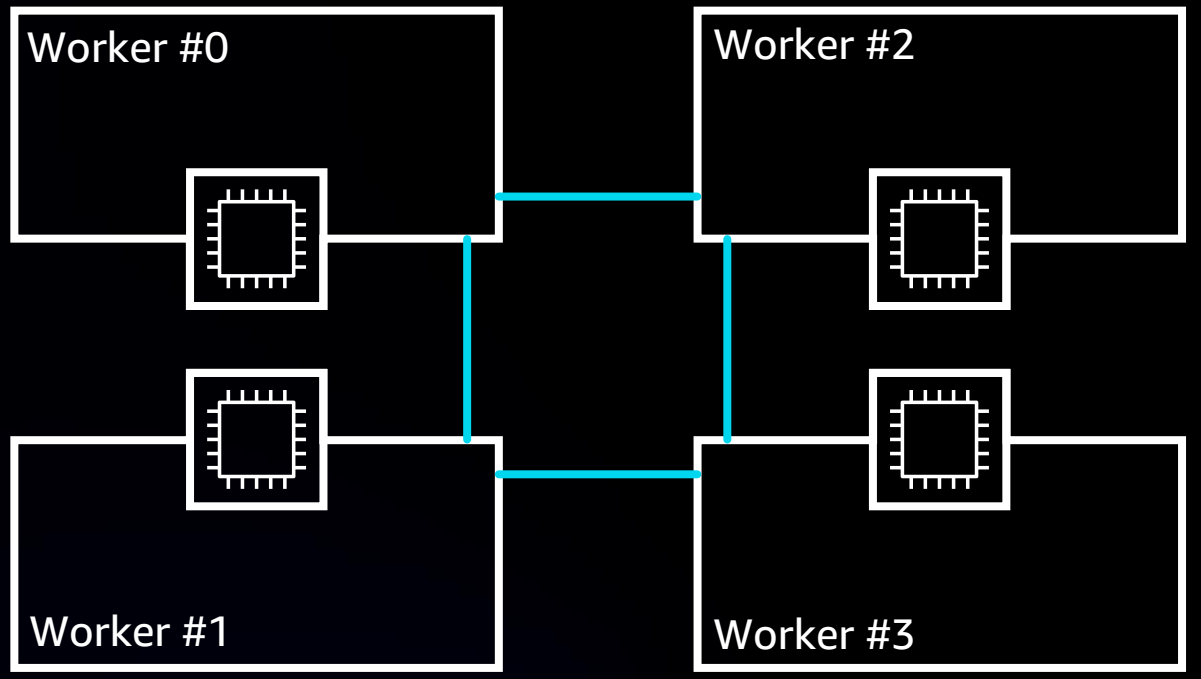
Distributed training



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data parallelism





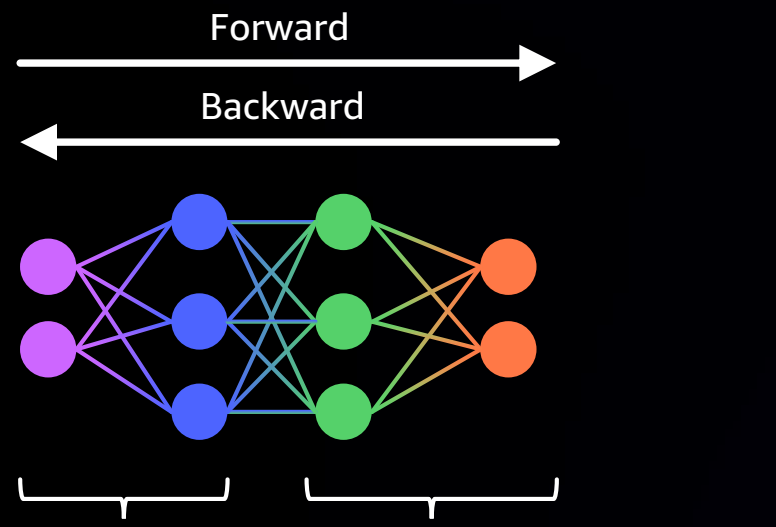
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data parallelism

Benefits:

- Simple, scalable
- Simple computation/communication overlap:

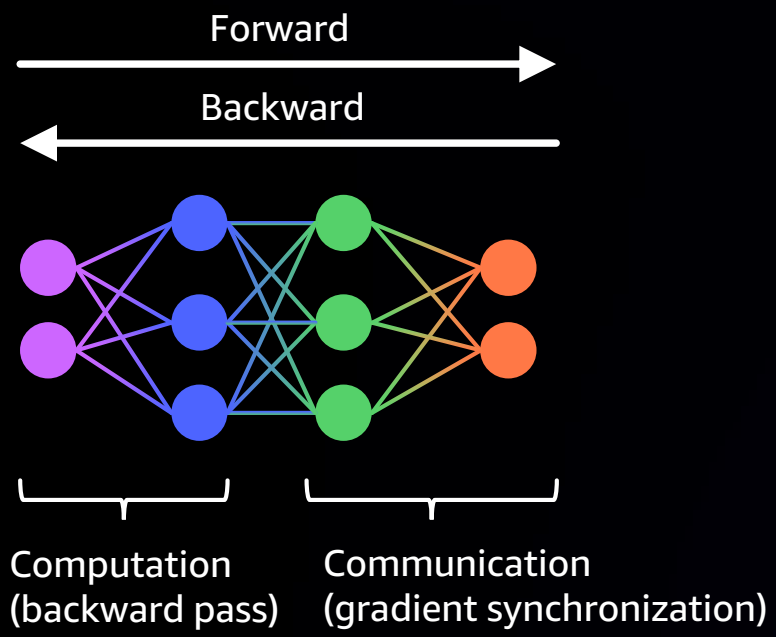


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data parallelism

Benefits:

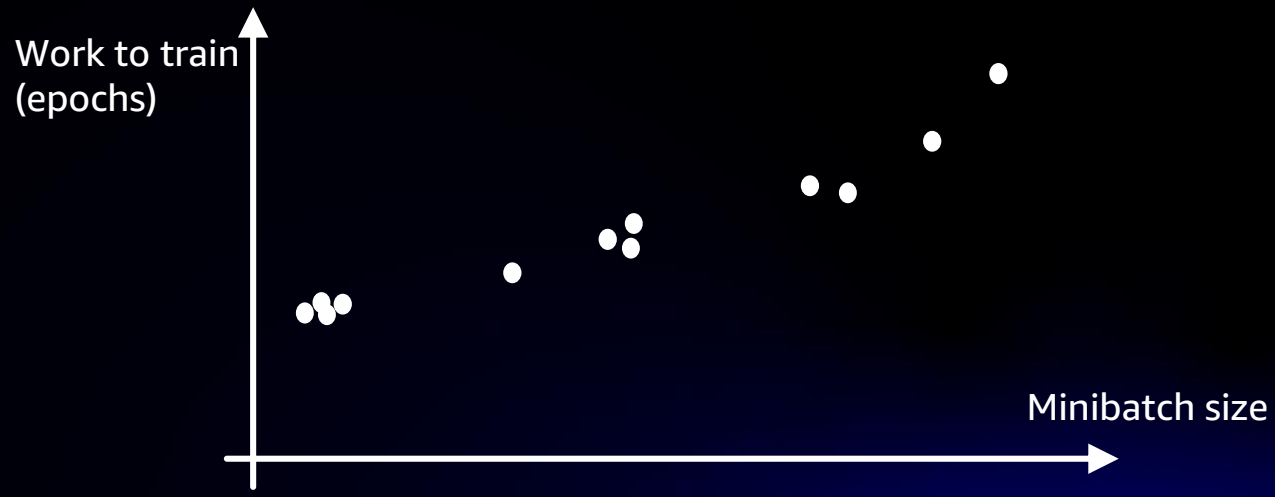
- Simple, scalable
- Simple computation/communication overlap:



Challenges:

- What if the model doesn't fit?
  → Large memory
- Weak vs. strong scaling:

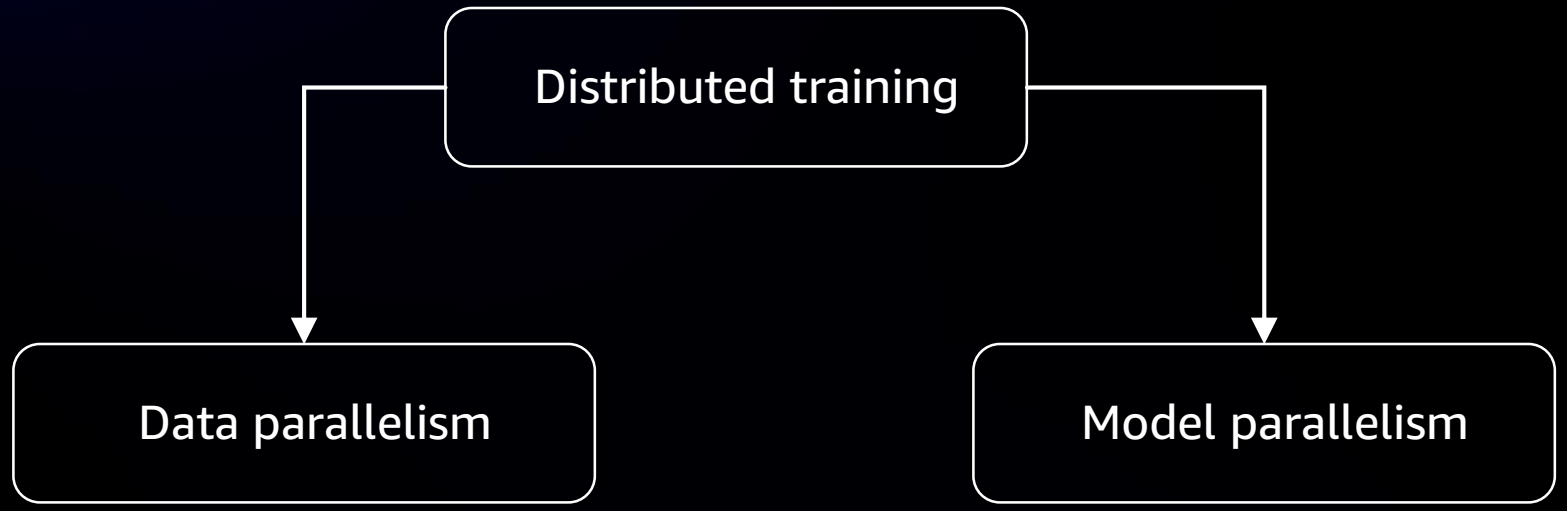
  Work to train increases with minibatch size
  → Strong scaling



aws

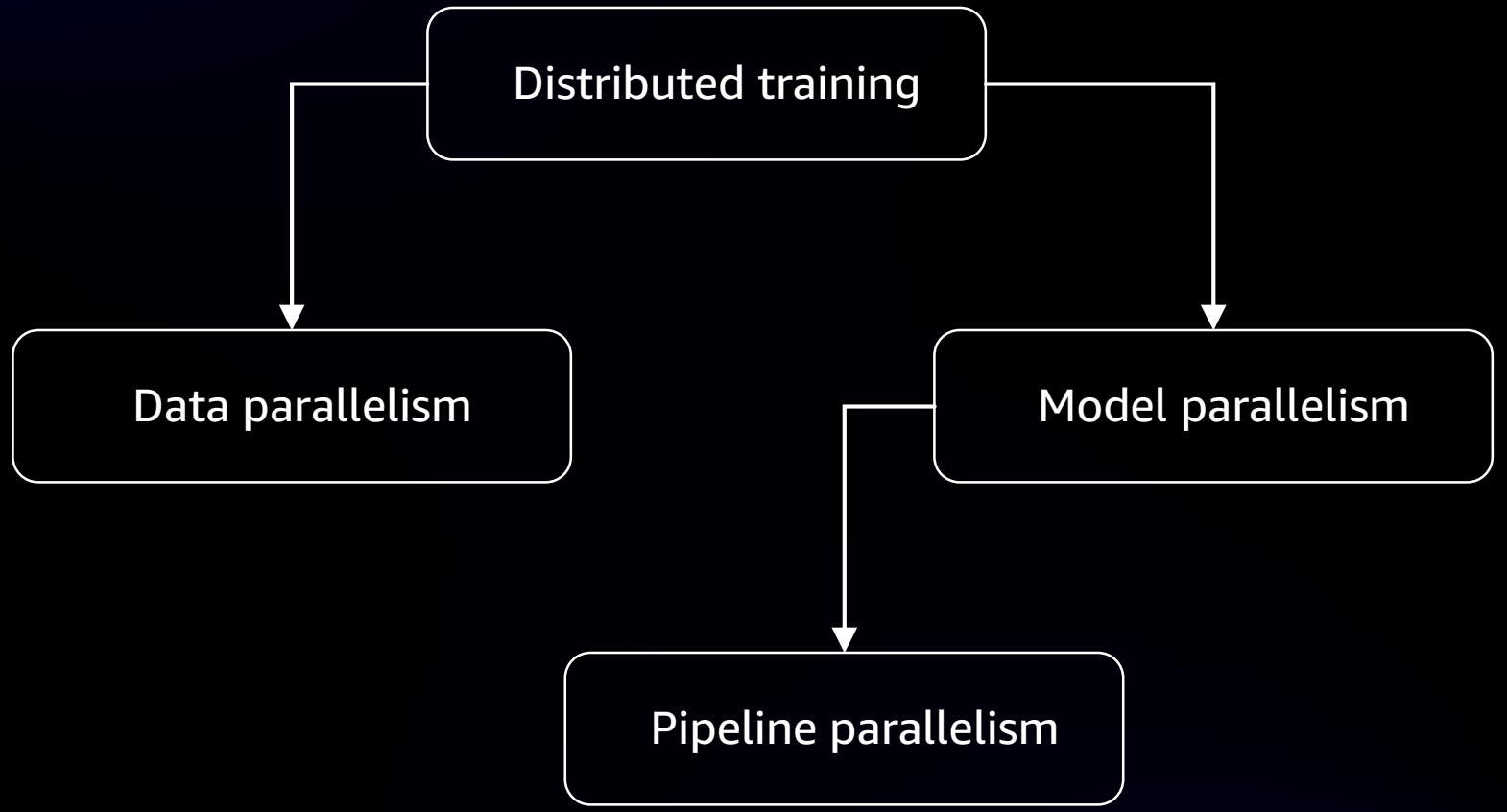
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

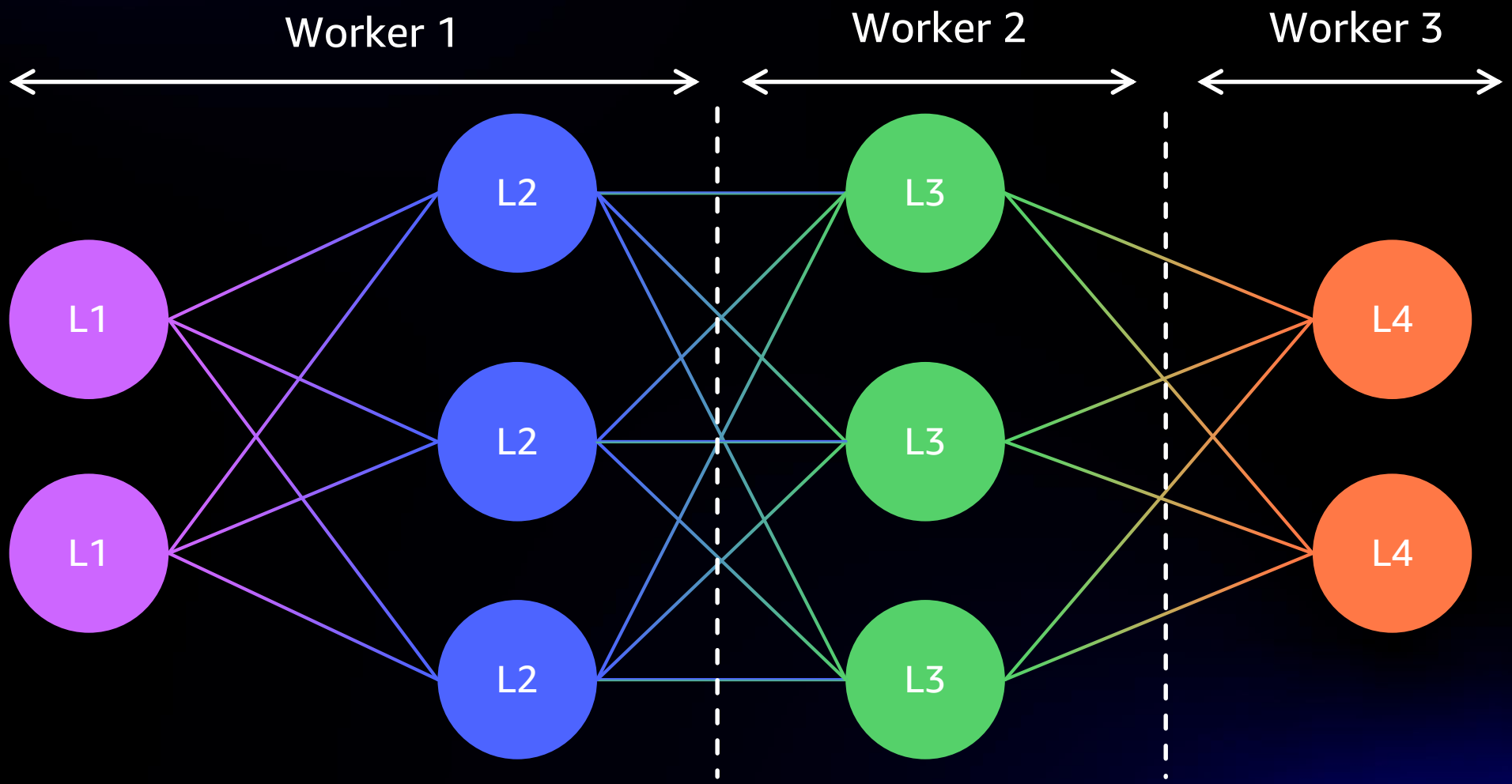
Distributed training



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

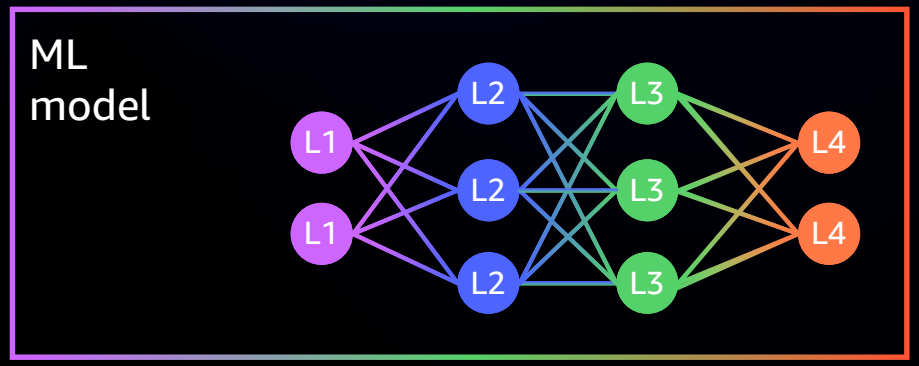
Pipeline parallelism

*ALSO CALLED INTER-LAYER MODEL PARALLELISM

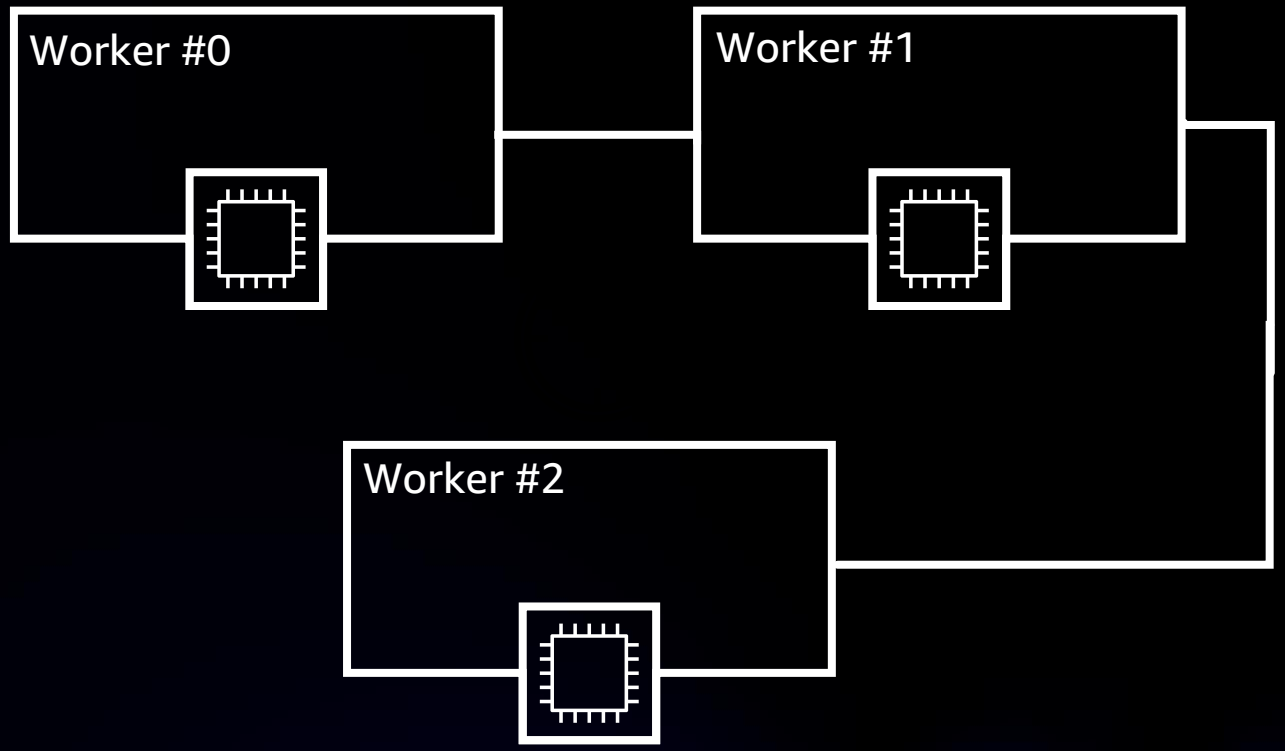


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pipeline parallelism







© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pipeline parallelism

Benefits:

- Fits larger models
- P2P communication

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

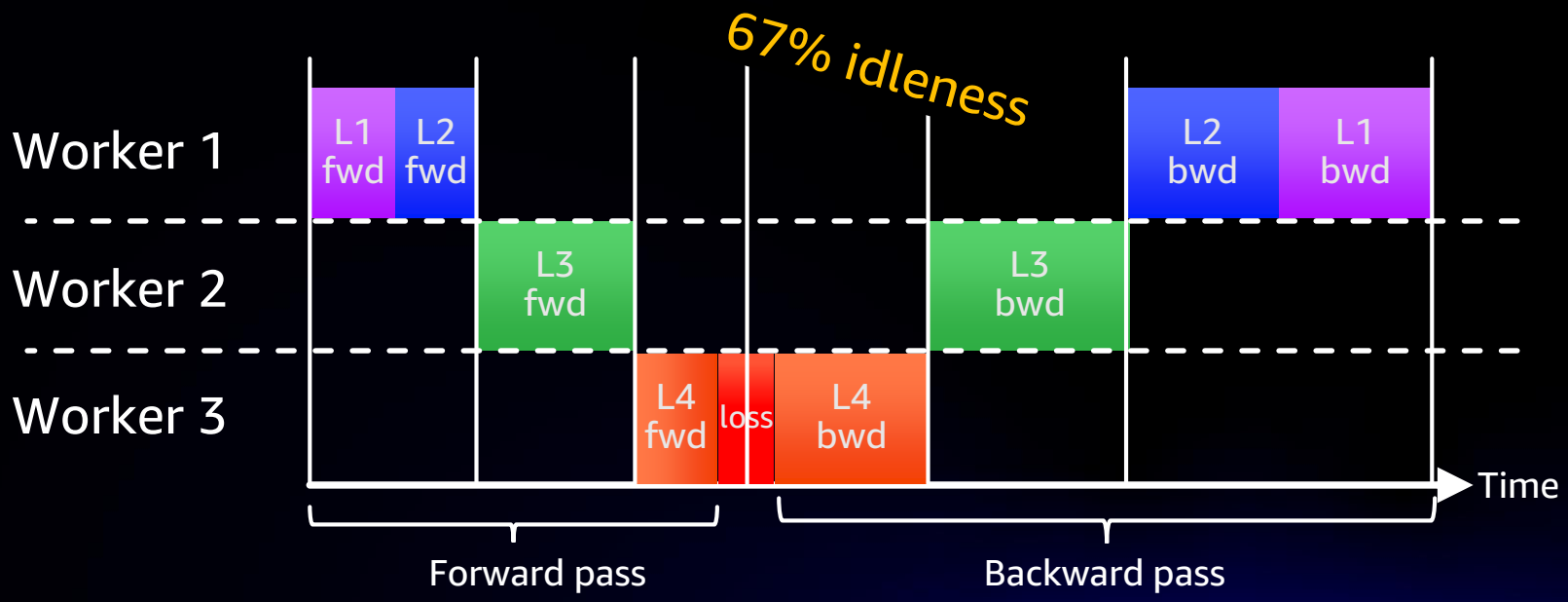
Pipeline parallelism

Benefits:

- Fits larger models
- P2P communication

Challenges:

- What if a layer doesn't fit?
  → Large memory
- Pipeline idleness reduces scaling efficiency



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

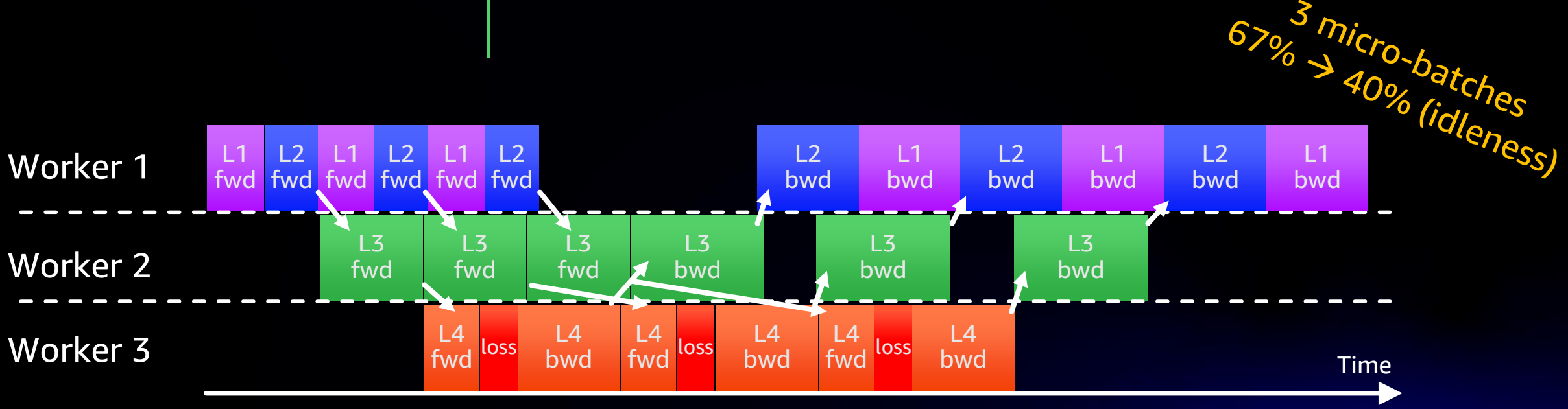
Pipeline parallelism

Benefits:

- Fits larger models
- P2P communication

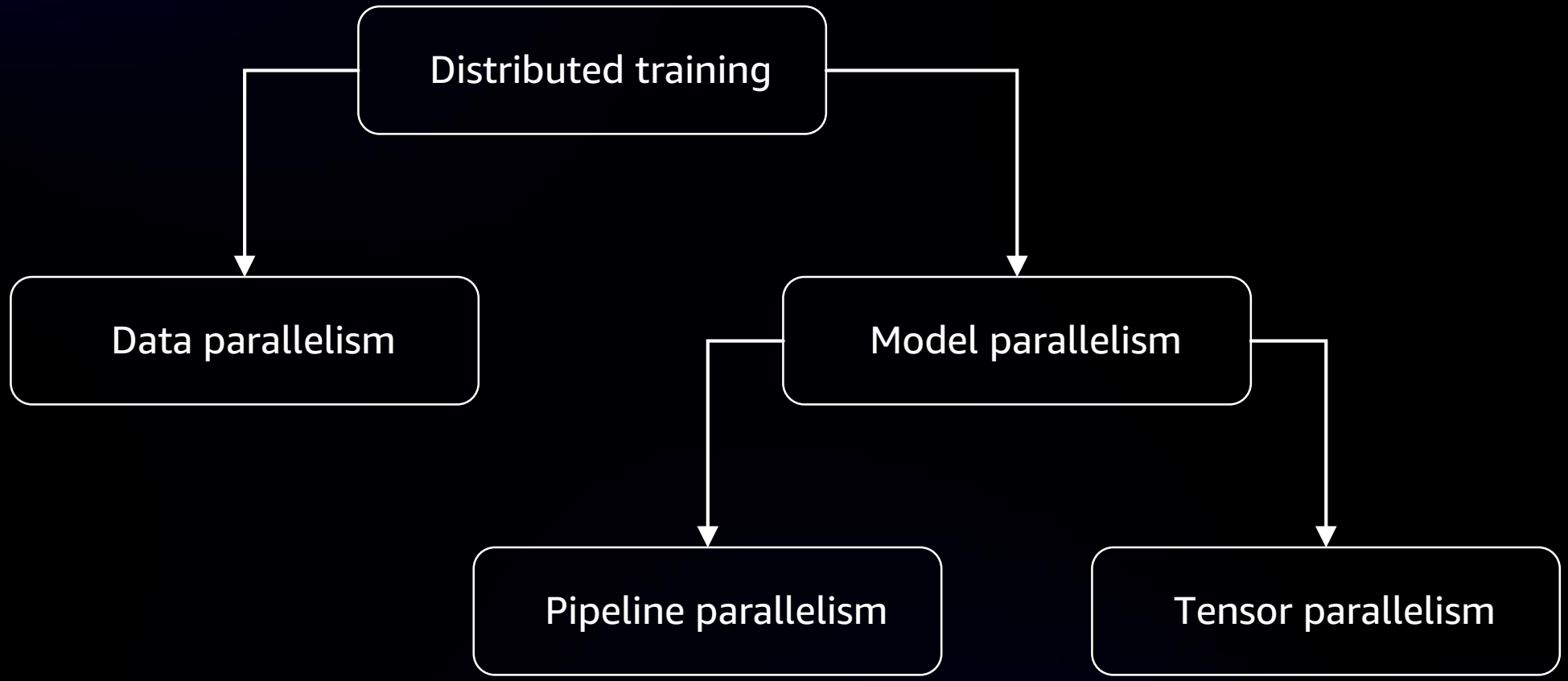
Challenges:

- What if a layer doesn't fit?
  → Large memory
- Pipeline idleness reduces scaling efficiency
  → Overlapped comp/comm



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

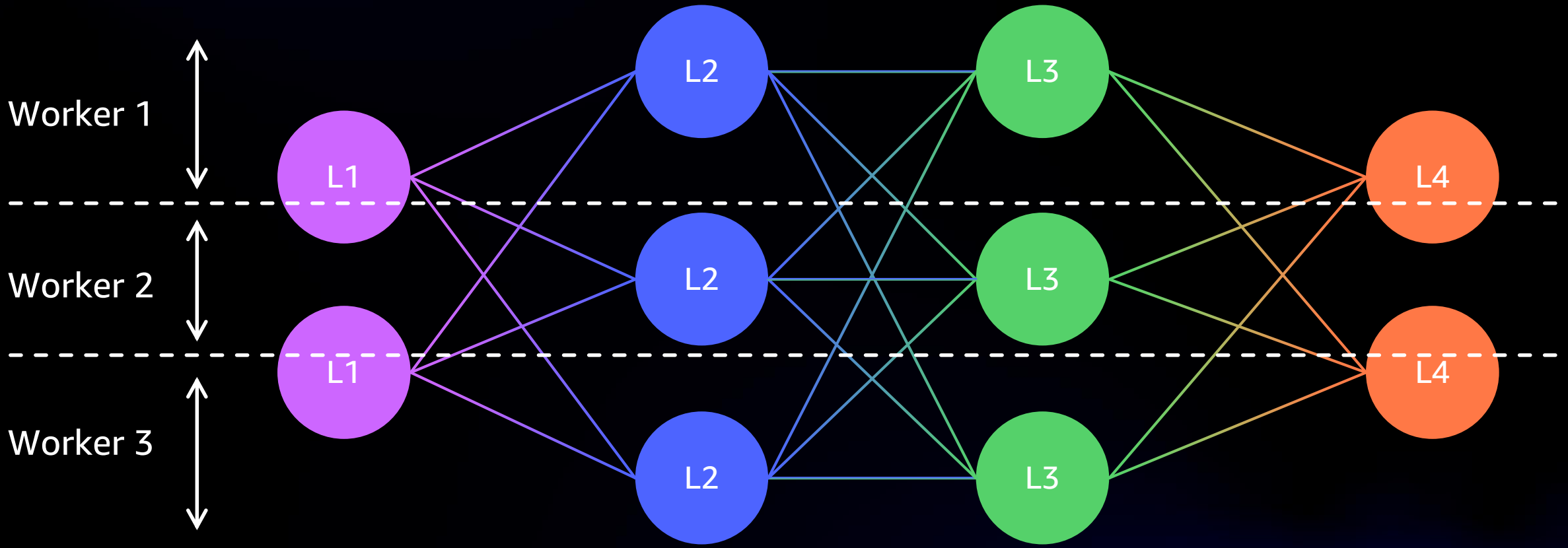
Distributed training



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

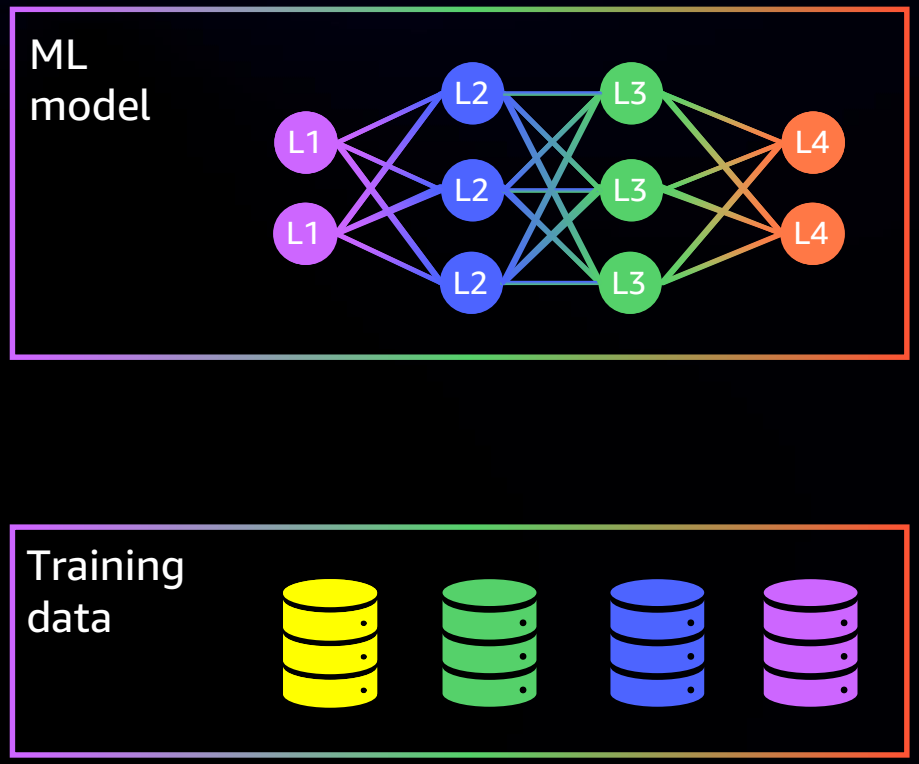
Tensor parallelism

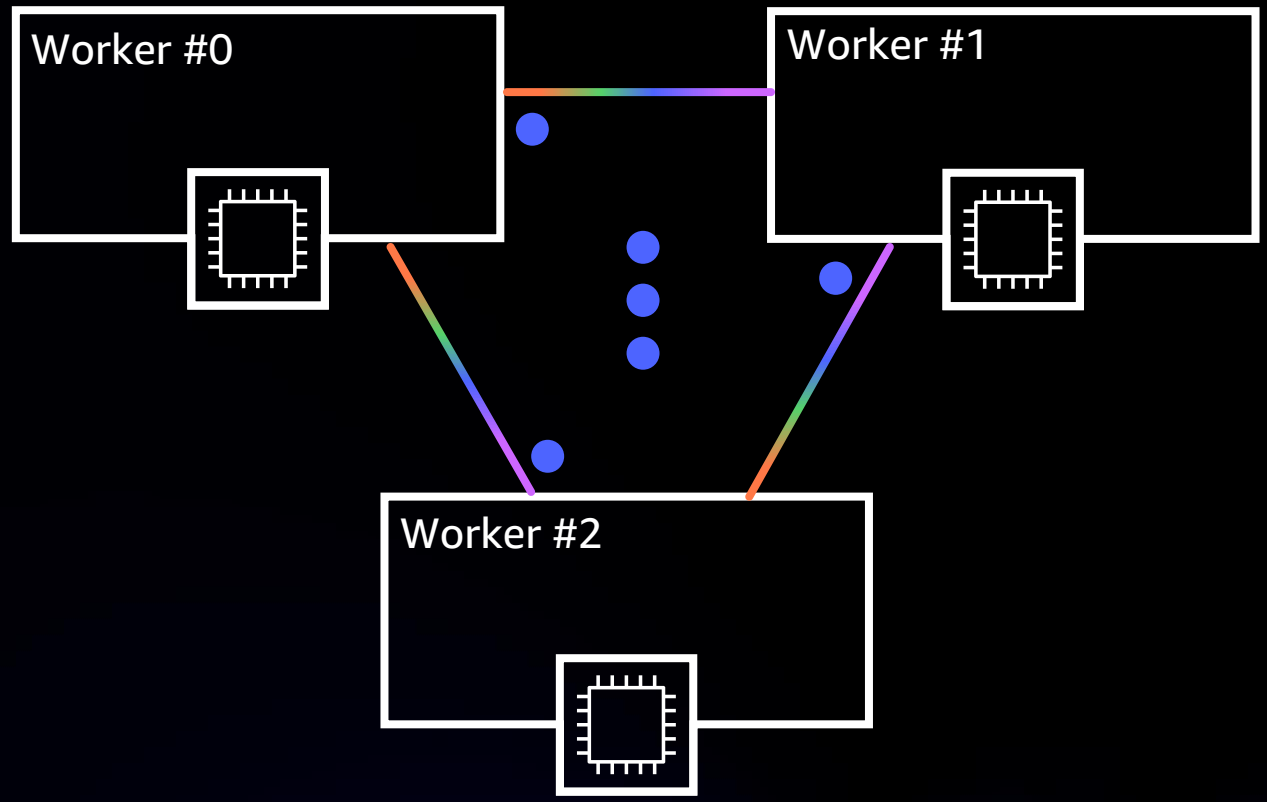
*ALSO CALLED INTRA-LAYER MODEL PARALLELISM



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tensor parallelism





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tensor parallelism

Benefits:

- Fits large models/layers
- Reduces minibatch size

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tensor parallelism

Benefits:

- Fits large models/layers
- Reduces minibatch size

Challenges:

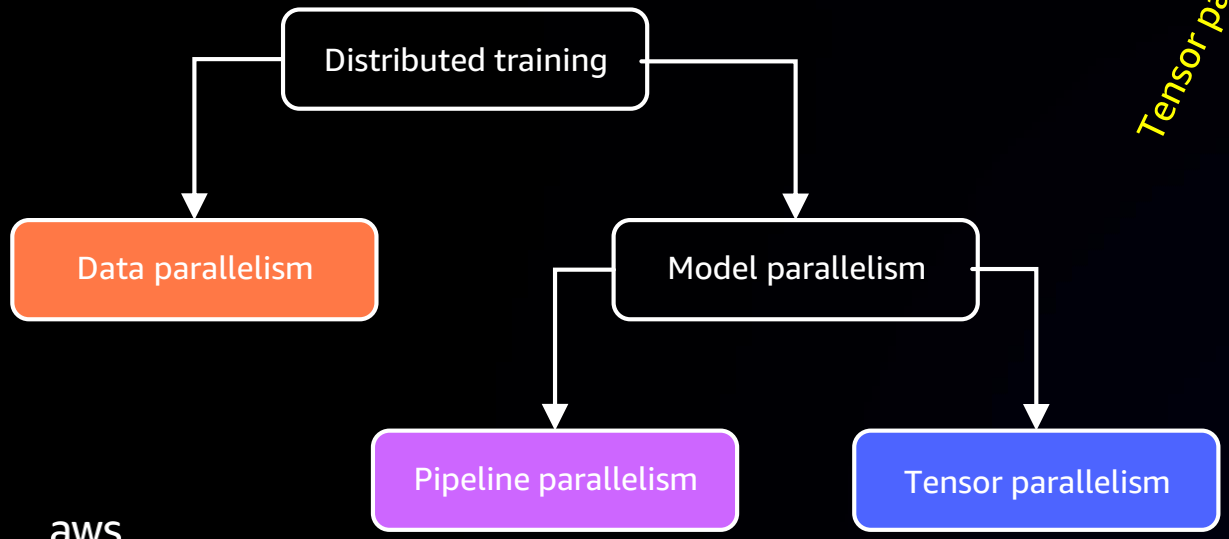
- All-to-all communication
  → High-bandwidth, low-latency interconnect
- Difficult to overlap computation and communication
  → Overlapped comp/comm

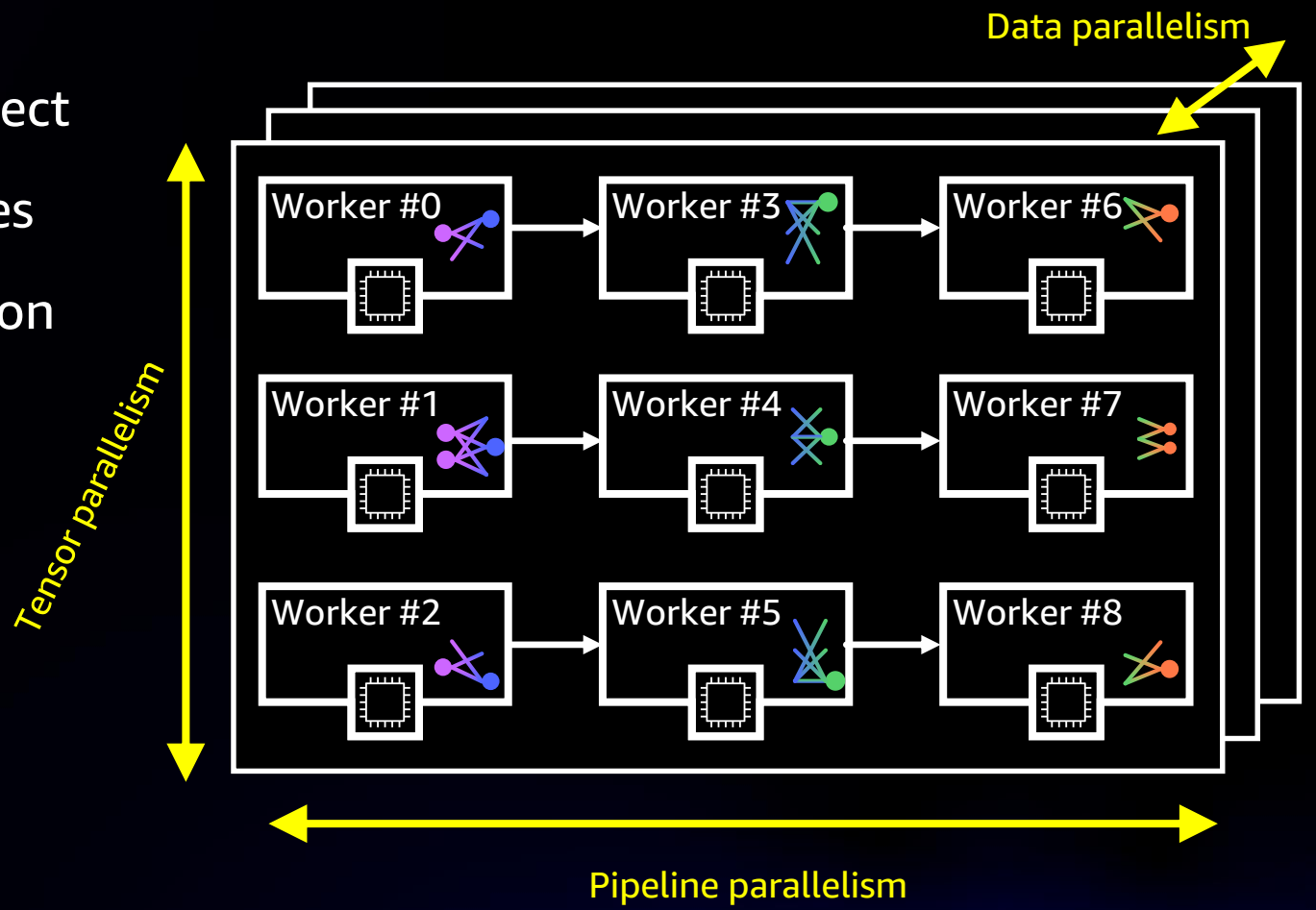
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recap – scale out requirements

- Large in-server accelerator memory
- High-bandwidth and low-latency interconnect
- Strong scaling: efficient for small batch sizes
- Parallelized computation and communication

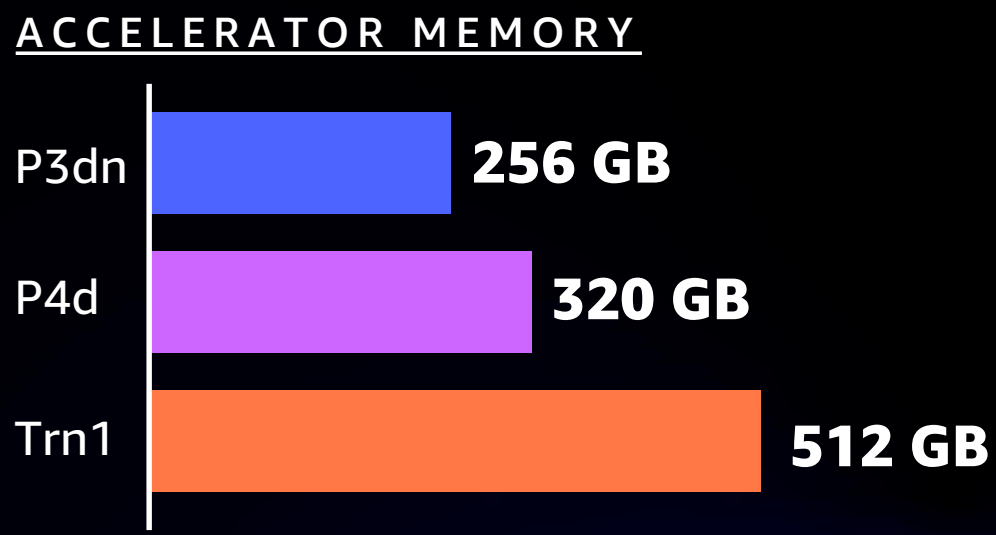




© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – built for scale-out

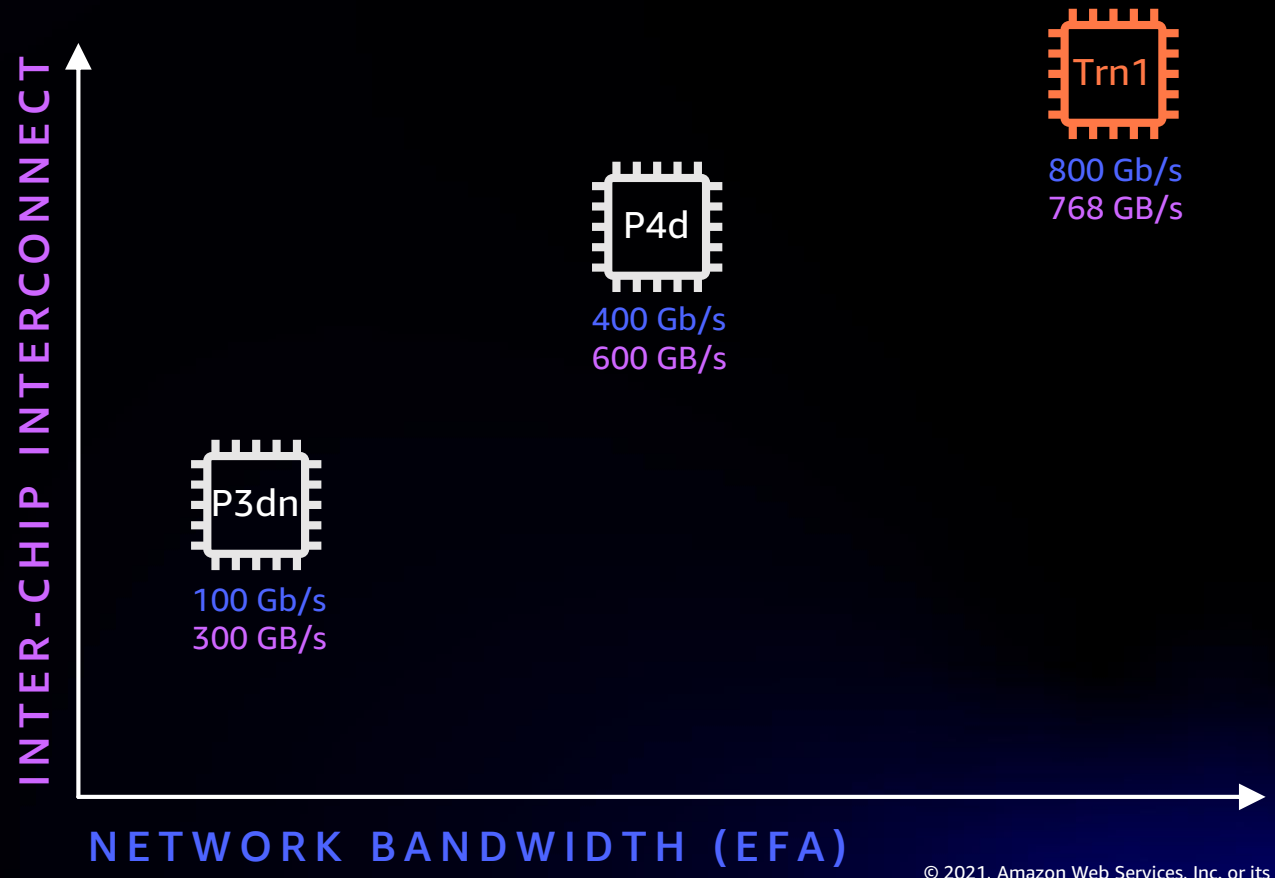
- Large in-server accelerator memory



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – built for scale-out

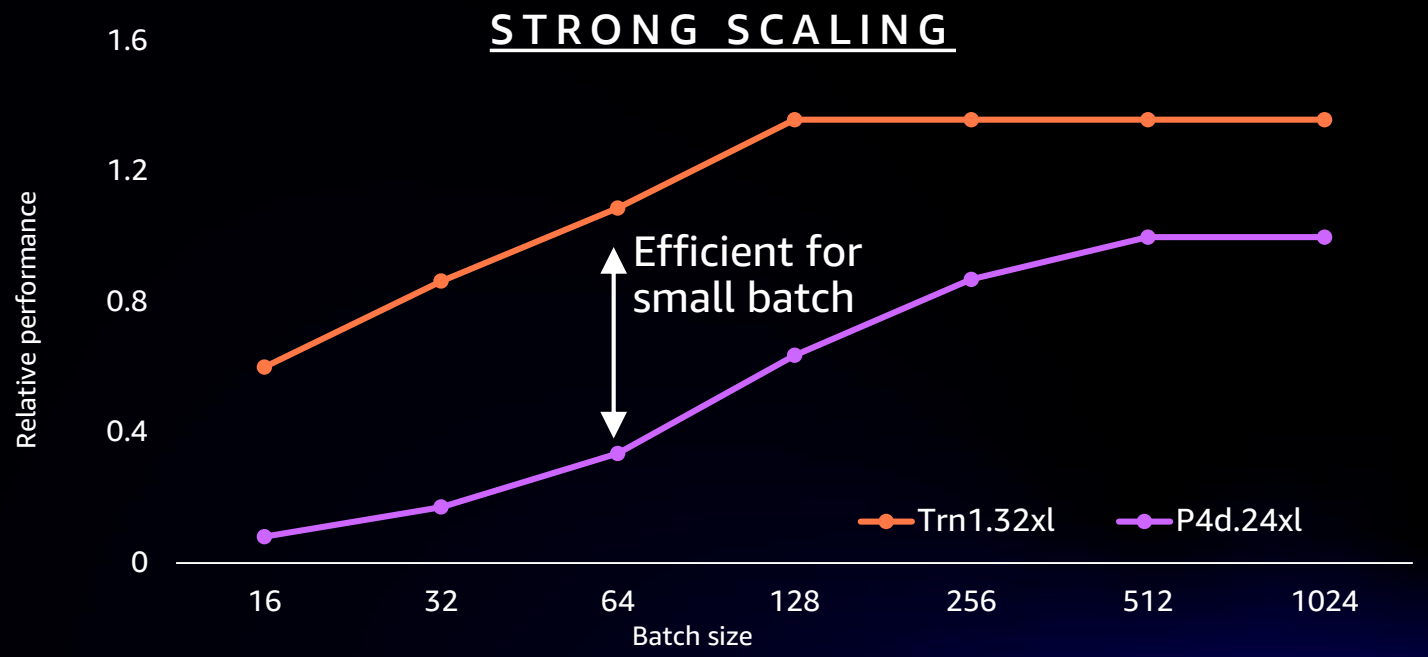
- Large in-server accelerator memory
- High-bandwidth and low-latency interconnect



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – built for scale-out

- Large in-server accelerator memory
- High-bandwidth and low-latency interconnect
- Strong scaling: efficient for small batch sizes



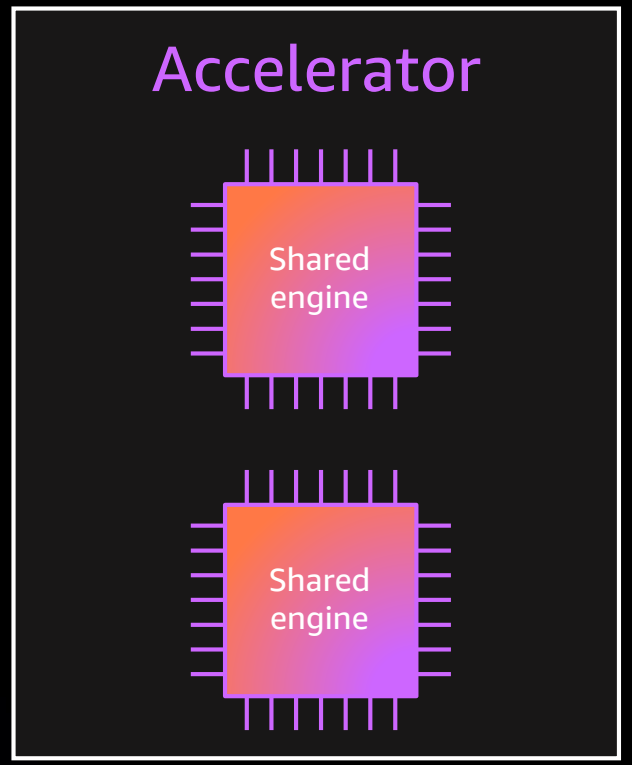
*BERT-Large relative performance

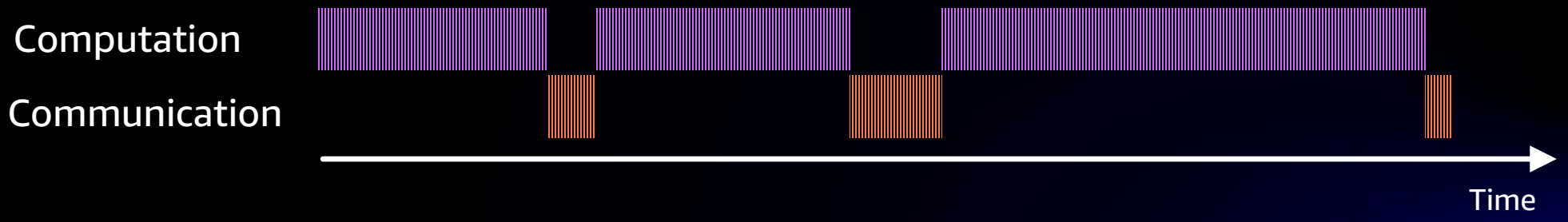
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – built for scale-out

- Large in-server accelerator memory
- High-bandwidth and low-latency interconnect
- Strong scaling: efficient for small batch sizes
- Parallelized computation and communication

  Separate engines for core-computation and communication primitives



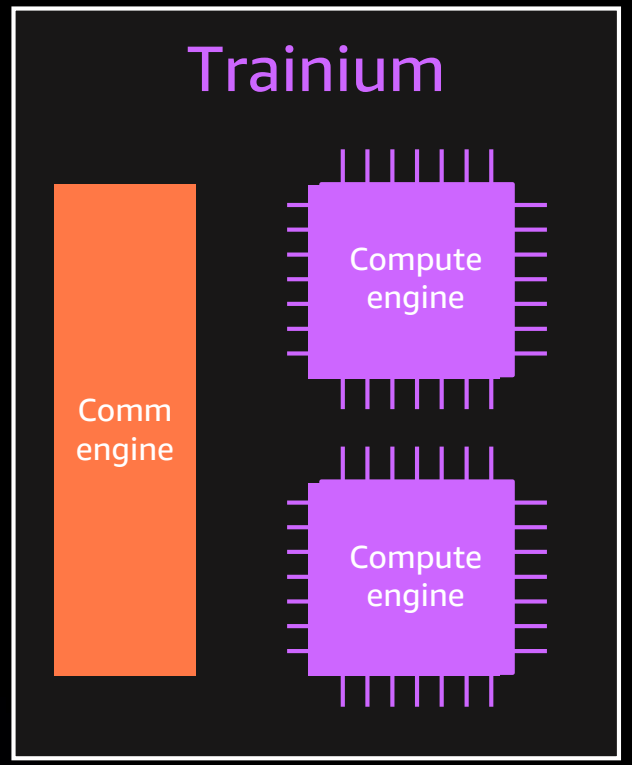


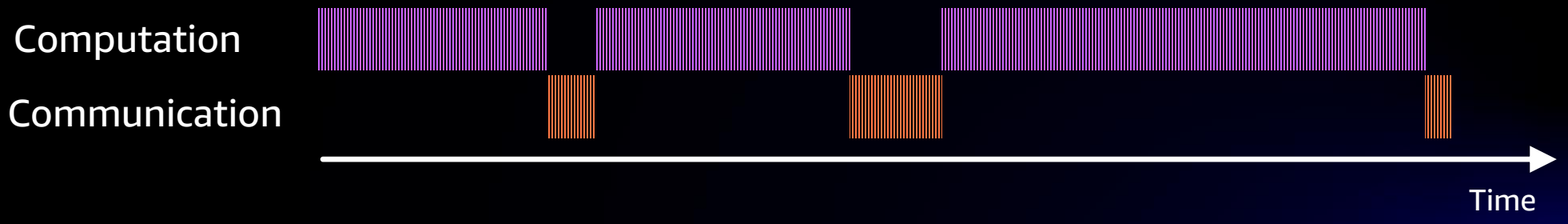
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trn1 – built for scale-out

- Large in-server accelerator memory
- High-bandwidth and low-latency interconnect
- Strong scaling: efficient for small batch sizes
- Parallelized computation and communication

  Separate engines for core-computation and communication primitives





aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Trn1 instances

THE MOST COST-EFFICIENT DL INSTANCE IN THE CLOUD

Trn1

MATH ENGINE FREQUENCY
3 GHz

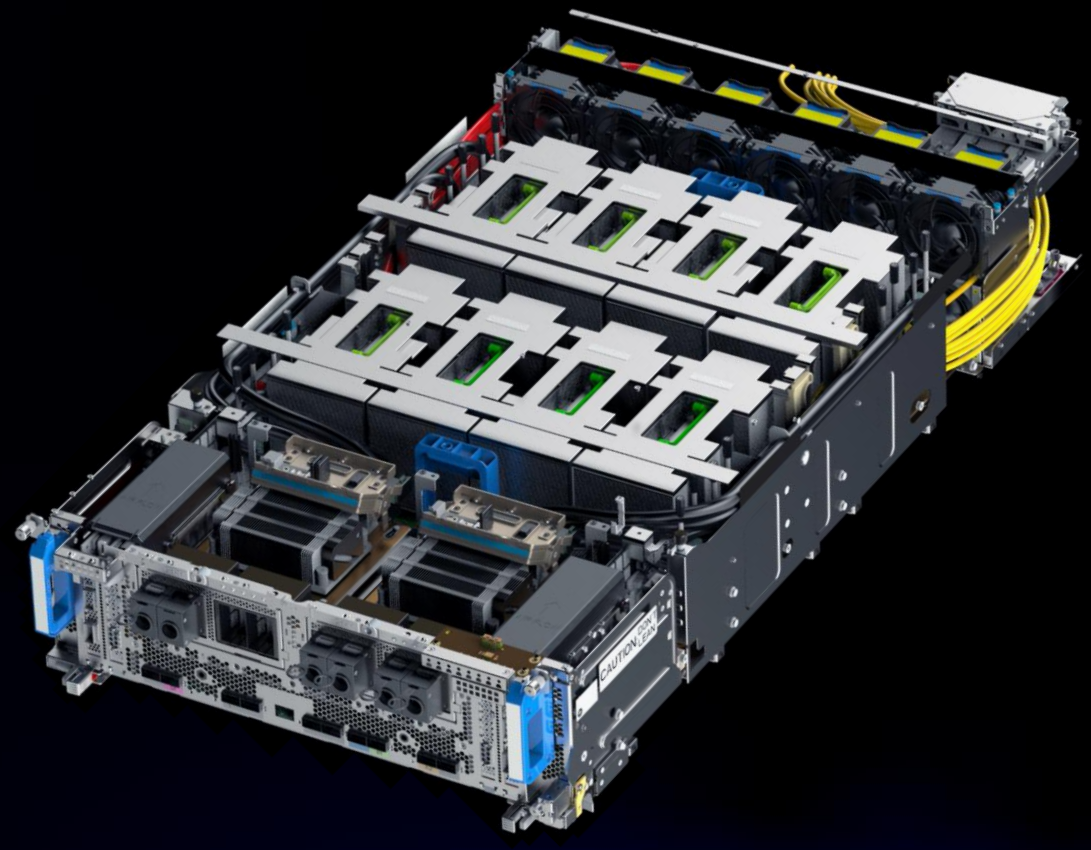
BF16/FP16	TF32	FP32
3.4 PFLOPS	**3.4 PFLOPS**	**840 TFLOPS**

ACCELERATOR MEMORY
512 GB

PEAK MEMORY BANDWIDTH
13.1 TB/sec

NEURONLINK BANDWIDTH BETWEEN CHIPS
768 GB/sec

NETWORK CONNECTIVITY
800 Gbps EFA

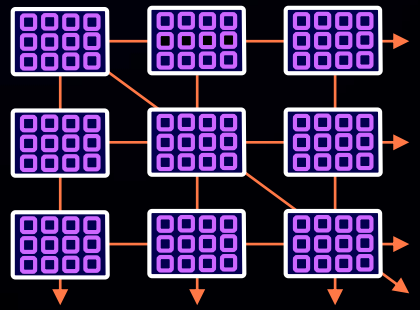


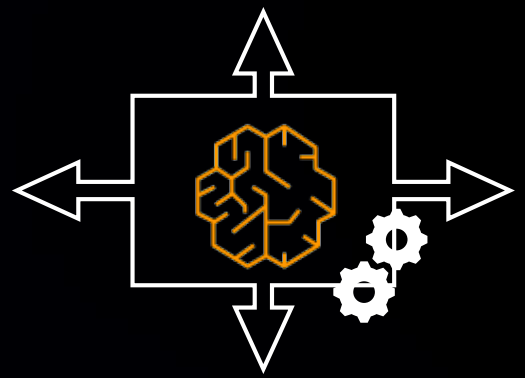
aws

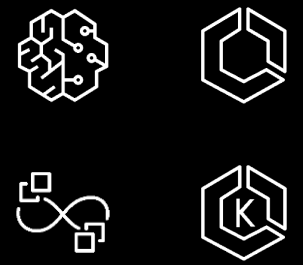
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Neuron highlights









Framework integration

Full framework integration, JIT, Eager mode, collective compute

Distributed training

Scale up to 10K+ devices, integration with distributed training libraries, and EFA

Flexible and extendable

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

Fully integrated with AWS

SageMaker, EKS, ECS, ParallelCluster, AMIs

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Neuron highlights

Neuron

Framewo
integrati

ated
s

Full framew
integration,
Eager mod
collective con

, ECS,
AMIs

```python
import os
...
import torch
import torch_xla
import torch_xla.core.xla_model as xm
...
from transformers import BertForPreTraining

model = BertForPreTraining.from_pretrained('bert-large-uncased')

def train_loop_fn(model, optimizer, train_loader, device, epoch, global_step, training_ustep, running_loss):
    max_grad_norm = 1.0
    for i, data in enumerate(train_loader):
        training_ustep += 1
        input_ids, segment_ids, input_mask, masked_lm_labels, next_sentence_labels = data
        outputs = model(input_ids=input_ids,
                        attention_mask=input_mask,
                        token_type_ids=segment_ids,
                        labels=masked_lm_labels,
                        next_sentence_label=next_sentence_labels)
        loss = outputs.loss / flags.grad_accum_usteps
        loss.backward()
        running_loss += loss.detach()

        if (training_ustep + 1) % flags.grad_accum_usteps == 0:
            xm.mark_step()
            running_loss_cpu = running_loss.detach().cpu().item()
            running_loss.zero_()
            torch.nn.utils.clip_grad_norm_(model.parameters(), max_grad_norm)
            xm.optimizer_step(optimizer)
            optimizer.zero_grad()
            scheduler.step()
            global_step += 1
            if global_step >= flags.steps_this_run:
                break

    return global_step, training_ustep, running_loss
```

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Wrap-up

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Trn1

POWERED BY AWS TRAINIUM

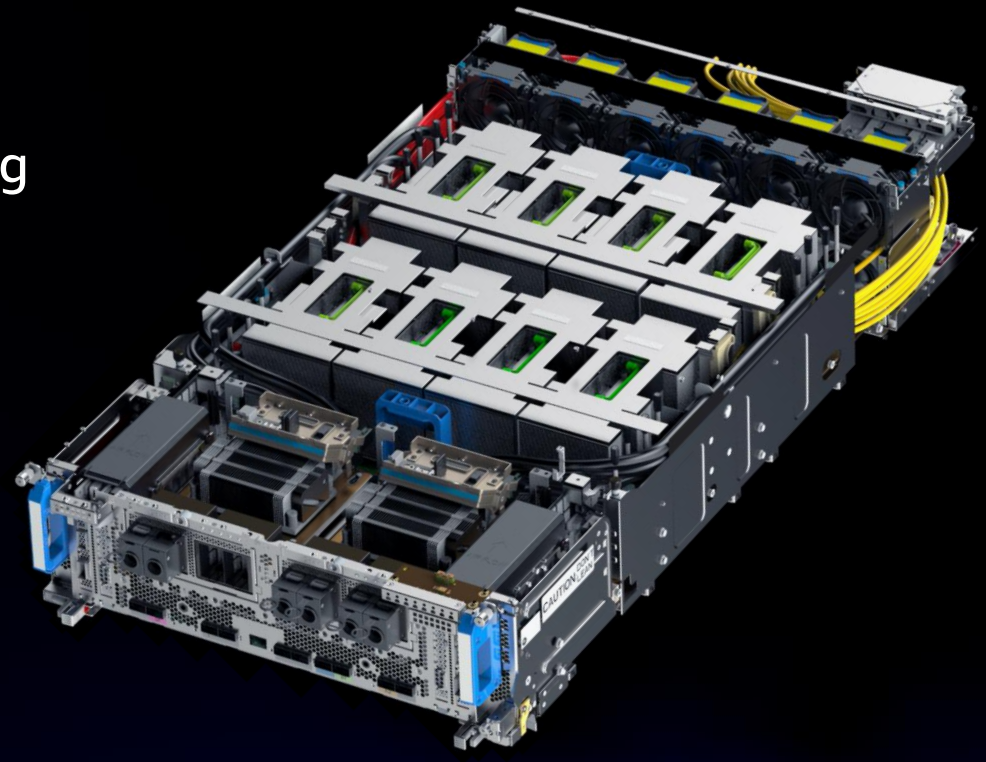
Purposely built for the **most cost-efficient and fast DL training in the cloud** for a broad spectrum of applications

AWS is **innovating across the chips, servers, and data center** layers to provide end users with access to cutting edge hardware on-demand

Max developer efficiency with Neuron SDK providing full integration into PyTorch and TensorFlow

Seamless integration with AWS services like SageMaker, Amazon ECS, ParallelCluster and more

In preview **now –** general availability in H1’22



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Chetan Kapoor

Ron Diamant

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.