生成系 AI の学習向け Amazon EC2 の選択肢

Keita Watanabe

Senior Solutions Architect, AIML Frameworks

渡辺啓太 （Keita Watanabe）

Senior Solutions Architect, AI/ML Frameworks

略歴

- 現職では Solutions Architect としてGenAI/LLM の分散学習やモデルのデプロイメントなどを支援
- 日本最大級の E コマース サイトを手掛ける企業にて ML Researcherとして 商品画像検索サービスの研究開発に従事
- 自動運転を手掛けるスタートアップ企業にてML Researcher として自動運転車両の意思決定システムの研究開発に従事

好きなAWS Service

- AWS ParallelCluster
- Amazon EKS
- Amazon EC2

本セッションについて

目的

生成系AI の学習に使用できる、各種アクセラレータ搭載インスタンスについて紹介する

対象者

- NVIDA GPU や AWS Custom Silicon などの機械学習向けアクセラレータに興味をお持ちの方
- AWS 上で大規模生成モデルの事前学習を行うことに興味をお持ちの方

話さないこと

- AWS ML Service について
- 各 ML モデルの詳細

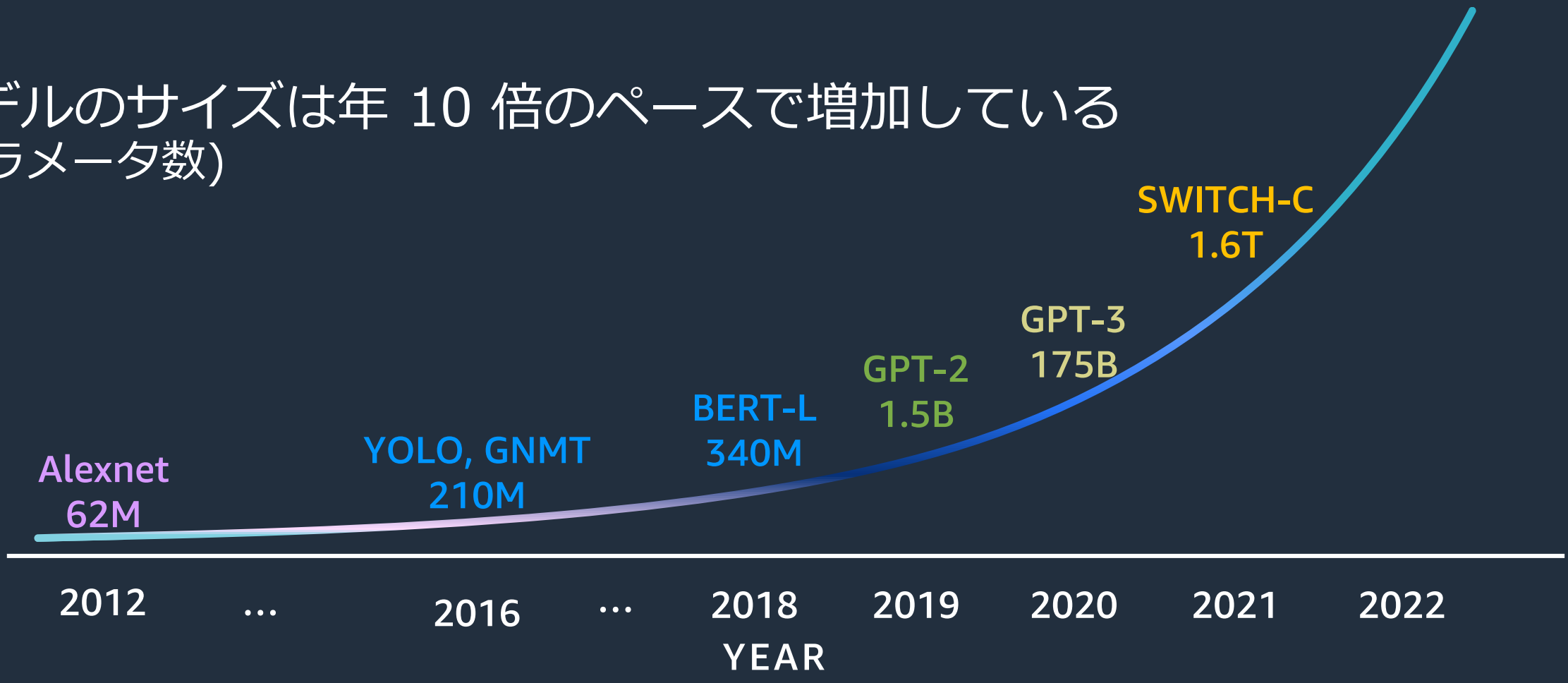
Generative AI(生成系AI)とは?

- 会話、ストーリー、画像、動画、音楽など、新しいコンテンツやアイデアを創造
- 一般に基盤モデル（FM:ファウンデーションモデル）と呼ばれる
  膨大なデータに基づいて事前にトレーニングされた大規模モデルを搭載

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

基盤モデルの性能を支える機械学習モデルの大規模化

モデルのサイズは年 10 倍のペースで増加している
(パラメータ数)



大規模モデルの学習には大規模な計算資源が必要

Q. GPT-3 (175 B) に一ヶ月で300 B tokens 学習させたいとき、必要となる NVIDIA A100（80GB） は何枚か？
※ Sequence Length は2048, Batch size は1536 とする

大規模モデルの学習には大規模な計算資源が必要

Q. GPT-3 (175 B) に一ヶ月で300 B tokens 学習させたいとき、必要となる NVIDIA A100（80GB） は何枚か？
※ Sequence Length は2048, Batch size は1536 とする
A. 1024 枚用いて34 日間のトレーニングが必要[1]

[1] Narayanan, Deepak, et al. "Efficient large-scale language model training on gpu clusters using megatron-lm." *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021.

大規模モデルの学習には大規模な計算資源が必要

Q. GPT-3 (175 B) に一ヶ月で300 B tokens 学習させたいとき、必要となる NVIDIA A100 (80GB) は何枚か？
※ Sequence Length は2048, Batch size は1536 とする
A. 1024枚用いて34 日間のトレーニングが必要[1]

高速なアクセラレータを搭載したインスタンスを多数用いて
トレーニングをスケールさせる必要がある

[1] Narayanan, Deepak, et al. "Efficient large-scale language model training on gpu clusters using megatron-lm." *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021.

大規模学習を支える技術要素

Compute

- 大容量デバイス
  メモリ
- 高速アクセラレータ

Network

- 広帯域
  インターコネクト

大規模学習を支える技術要素

Compute

- 大容量デバイス
  メモリ
- 高速アクセラレータ

Network

- 広帯域
  インターコネクト

大規模モデル学習に適した幅広い選択肢

NVIDIA

H100/A100/V100

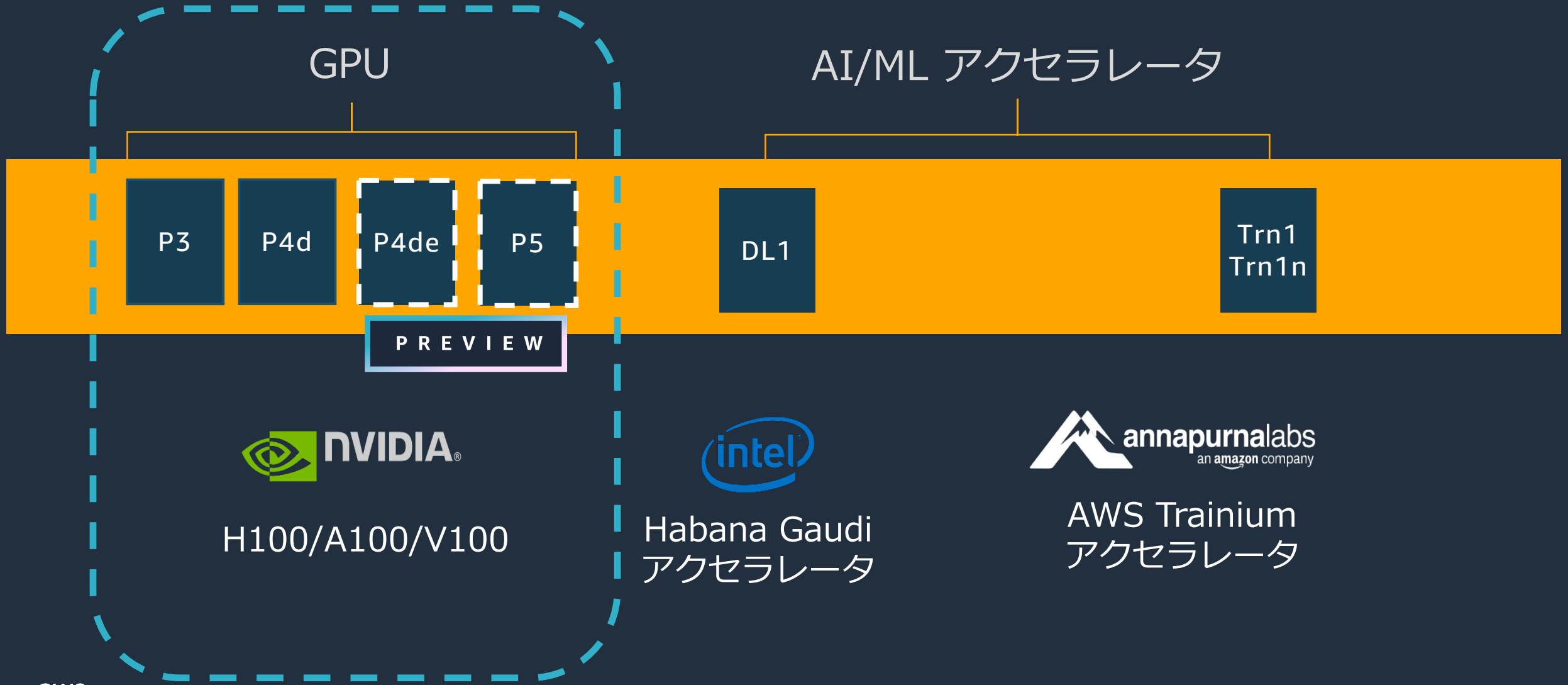
intel

Habana Gaudi
アクセラレータ

annapurnalabs
an amazon company

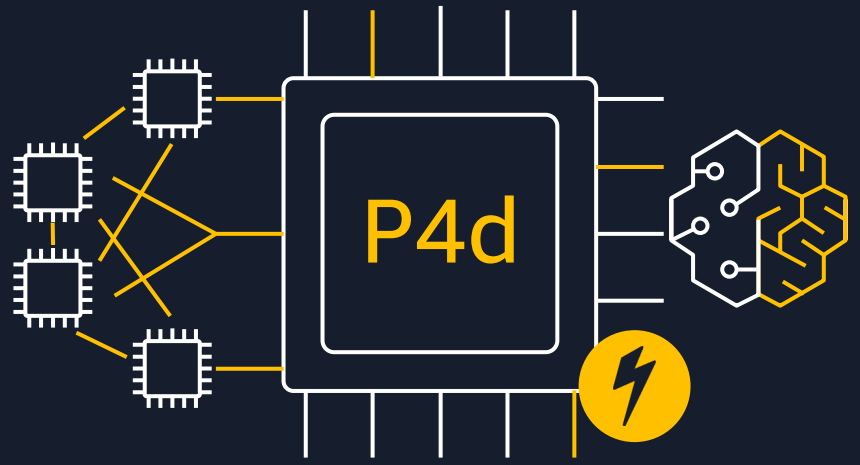
AWS Trainium
アクセラレータ

大規模モデル学習に適した幅広い選択肢



Amazon EC2 P4d/P4de インスタンス

＊ N. Virginia



- 8枚の NVIDIA A100 GPU アクセラレータ搭載
- 各40/80GB、合計320/640 GB の高帯域幅（HBM2）デバイスメモリ
- 8TB のローカル NVMe SSD
- 600GB/s のアクセラレータ間通信

Instance Size	GPU	アクセラレータメモリ	vCPU	メモリ	アクセラレータ P2P BW	オンデマンド価格* (USD/時間)
P4d.24xlarge	8	320 GB	96	1152 GiB	600 GB/s	32.77
P4de.24xlarge	8	640 GB	96	1152 GiB	600 GB/s	40.96

PREVIEW

https://aws.amazon.com/ec2/instance-types/p4/

AI21 Labs では、Amazon EC2 と PyTorch を使用して、1,780 億のパラメータの基礎となるモデルを構築しました。

CHALLENGE

AI21 Labs は、1,780 億個のパラメーターを持つ自己回帰型言語モデルである Jurassic-1 Jumbo をトレーニングしながら、強力なコンピューティング機能とネットワーク機能を実現し、効率を最大化したいと考えていました。

SOLUTION

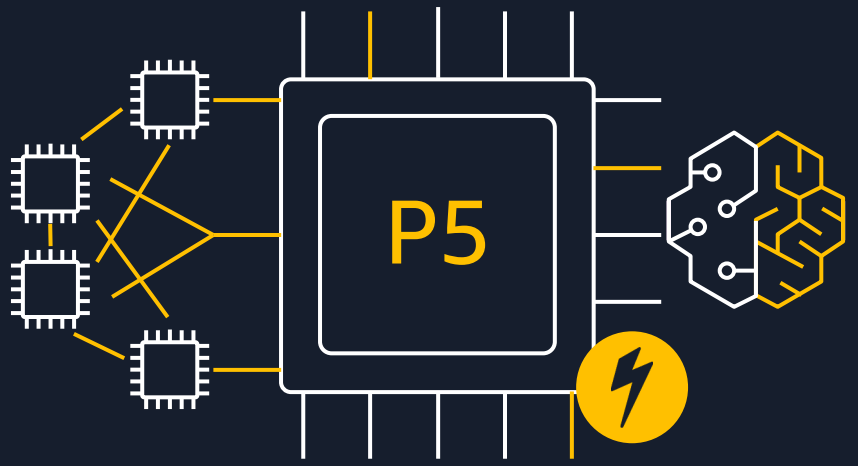
同社は Amazon EC2 P4d インスタンスを使用しました。これにより、モデルトレーニングを数百の GPU に分散させることで、自然言語処理をサービスとして提供するために必要なパフォーマンスとメモリを獲得できました。

OUTCOME

- ✓ 1,780億のパラメータと256,000項目の語彙を含む言語モデルを開発しました
- ✓ 効率的かつコスト効率よく数百個の GPU に拡張可能
- ✓ 大規模なモデル開発に関する知識の確立

Amazon EC2 P5 インスタンス

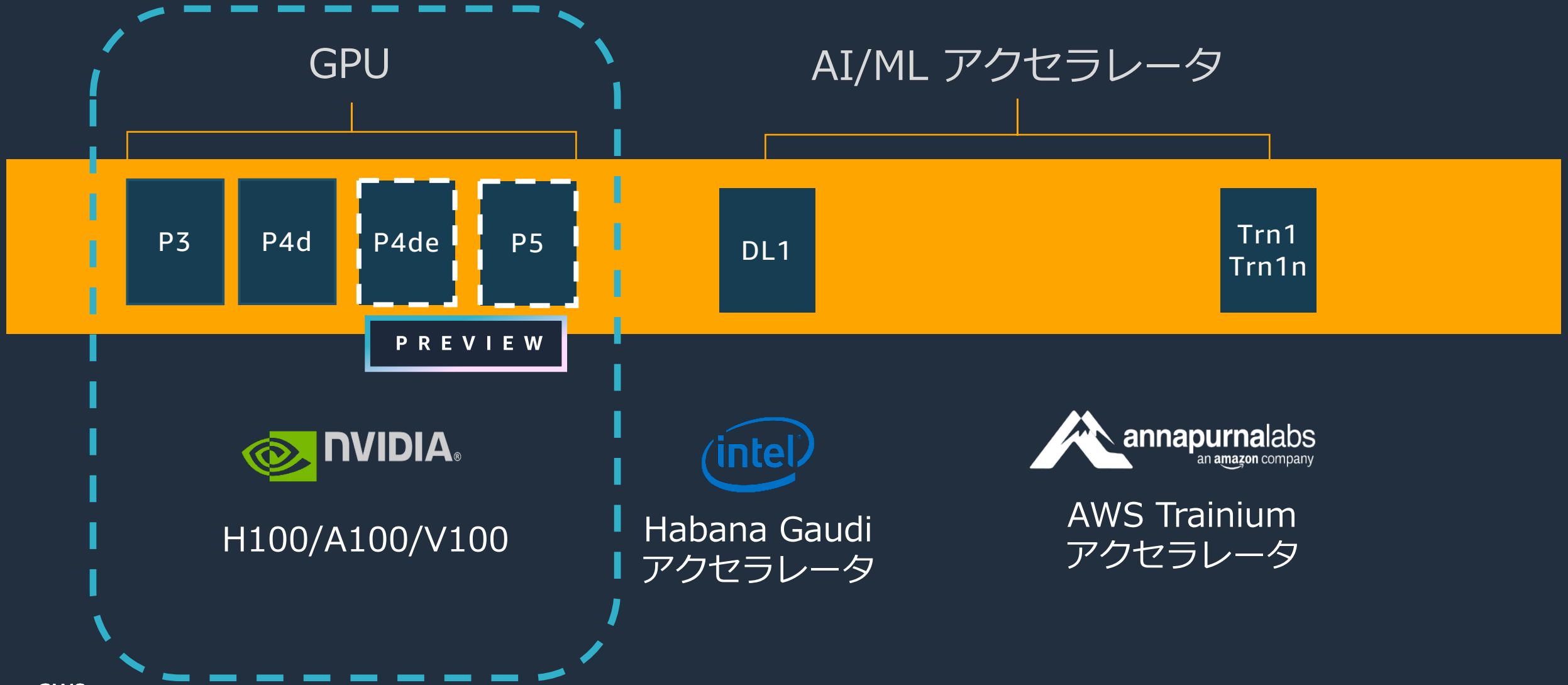
PREVIEW



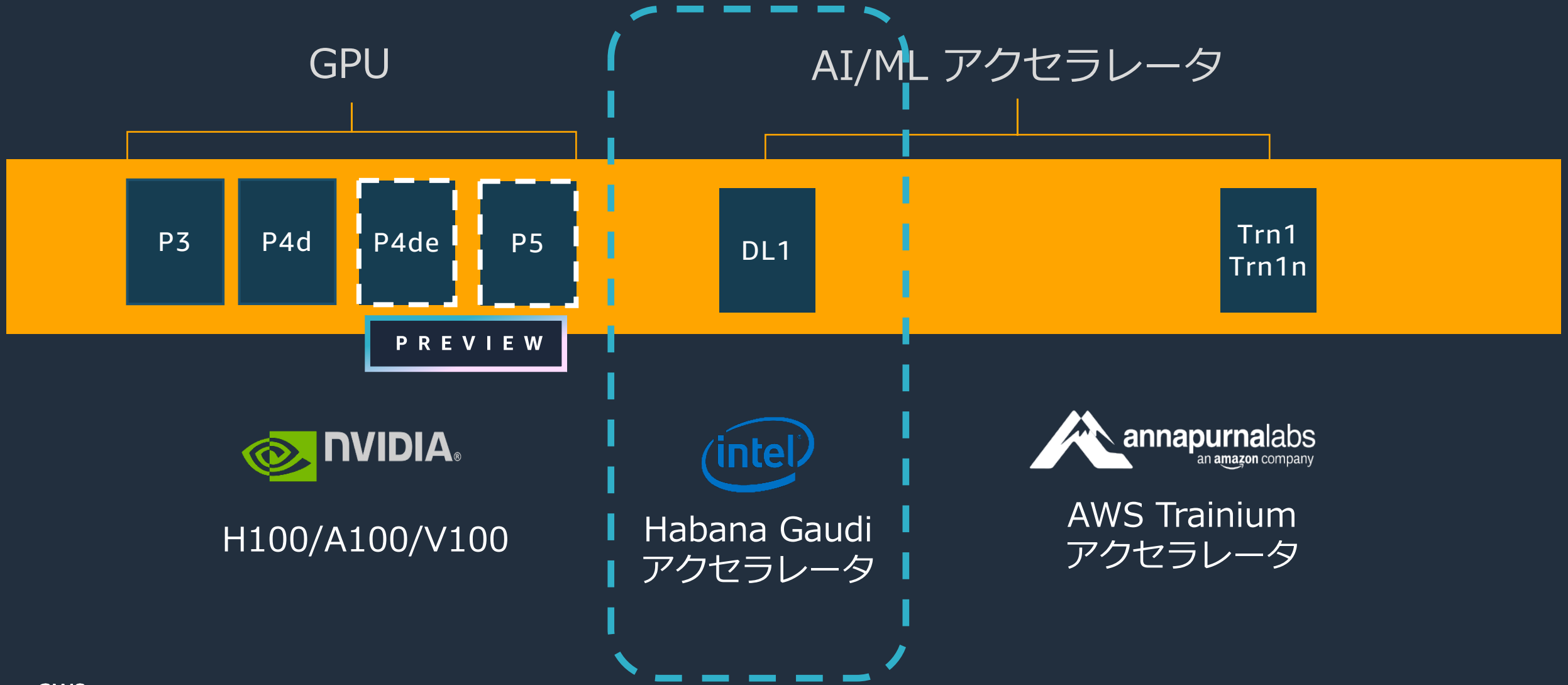
- 最大8枚の NVIDIA H100 GPU
- 各80GB、合計640 GB のデバイスメモリ
- 前世代のGPU インスタンスと比較して最大40% 低コスト
- P4d インスタンスと比較して最大6倍、P4de インタンスと比較して最大4倍高速

https://press.aboutamazon.com/2023/3/aws-and-nvidia-collaborate-on-next-generation-infrastructure-for-training-large-machine-learning-models-and-building-generative-ai-applications

大規模モデル学習に適した幅広い選択肢



大規模モデル学習に適した幅広い選択肢



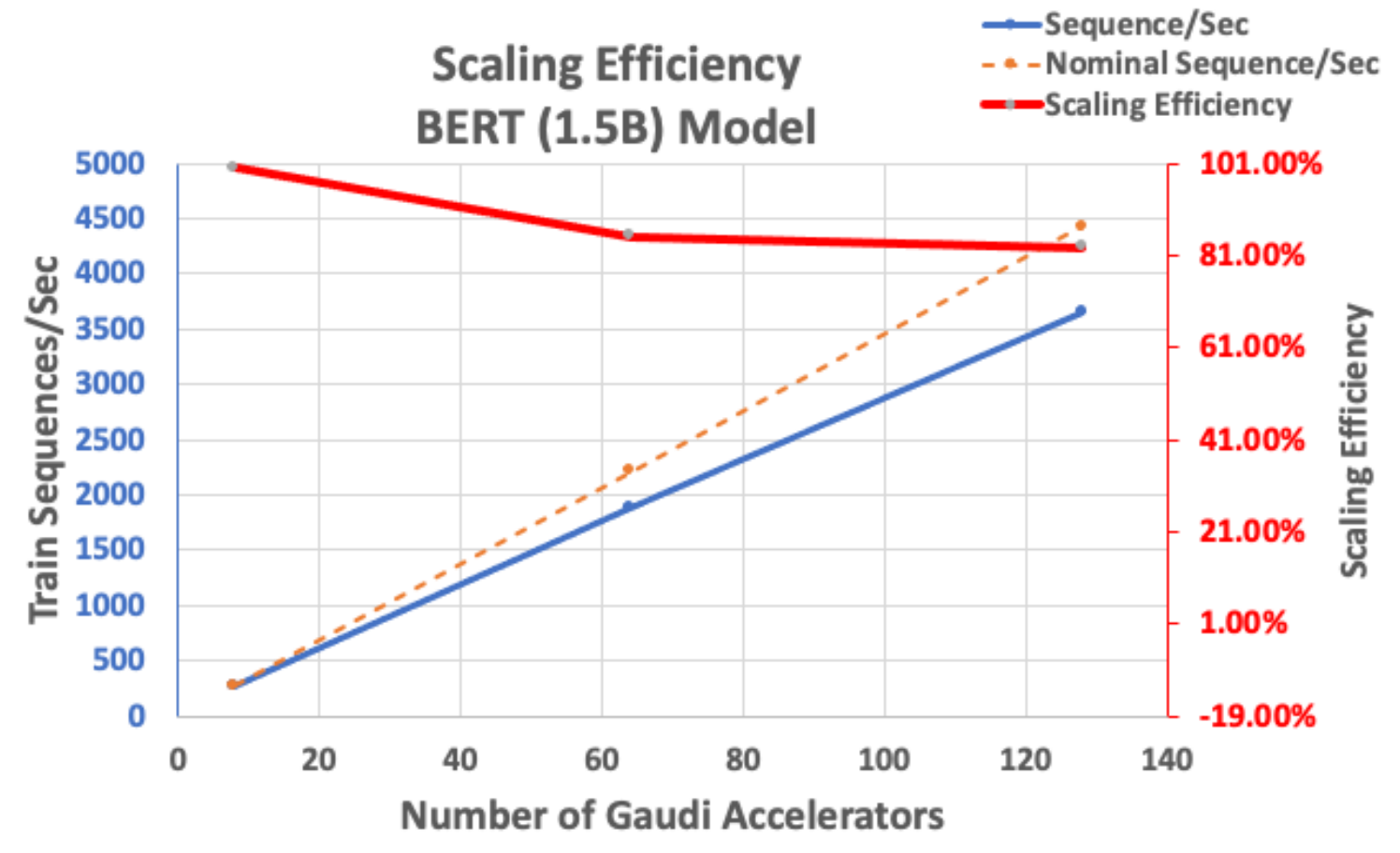
Amazon EC2 DL1インスタンス

- 8枚の Habana Gaudi アクセラレータ搭載
- 各32GB 、合計256GB の高帯域幅（HDM2）デバイスメモリ
- 100 Gbps のアクセラレータ間通信
- 4TB のローカルNVMe SSD
- PyTorch およびTensorFlow と統合されたSDK

Instance Size	Gaudi	アクセラレータメモリ	vCPU	メモリ	アクセラレータP2P BW	オンデマンド価格* (USD/時間)
DL1.24xlarge	8	256 GB	96	768 GB	100 Gbps	13.11

* N. Virginia

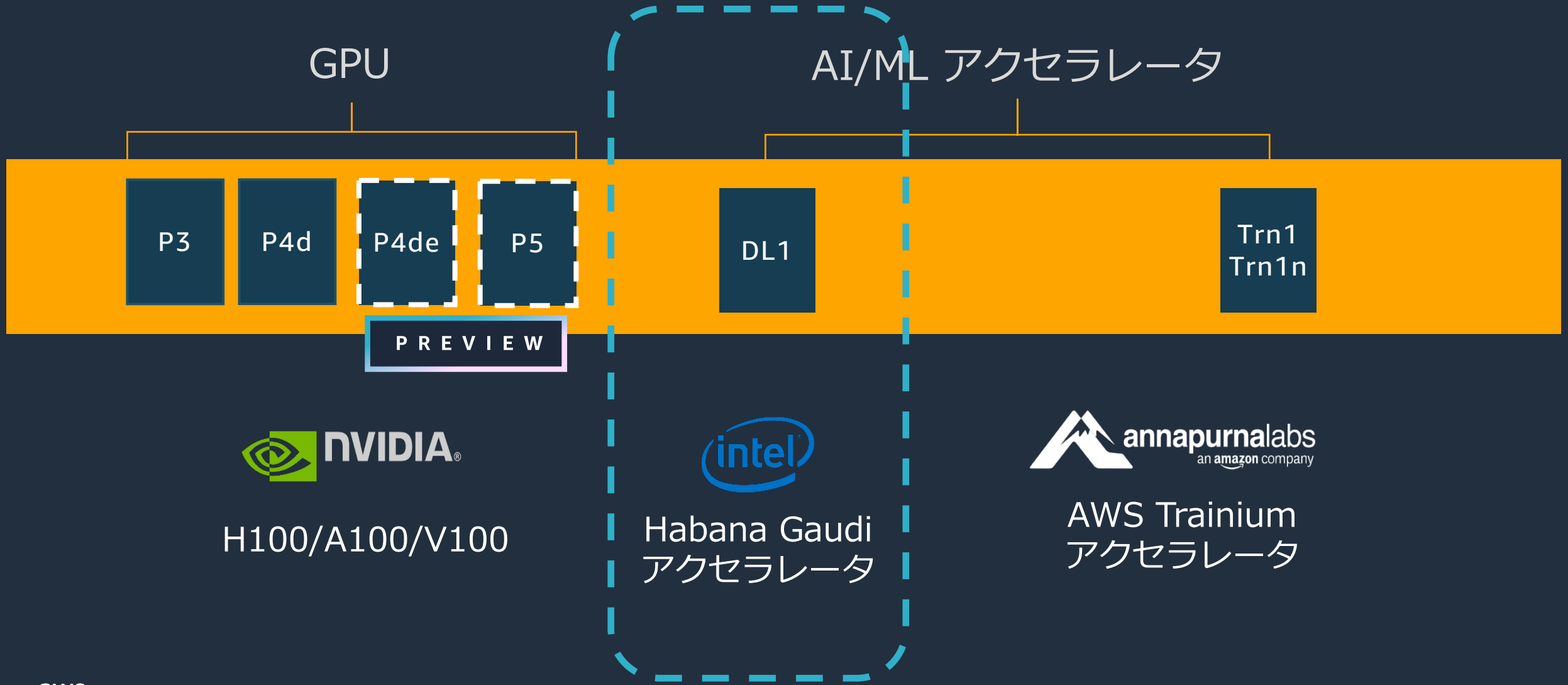
DL1 上での PyTorch DeepSpeed を用いた分散学習



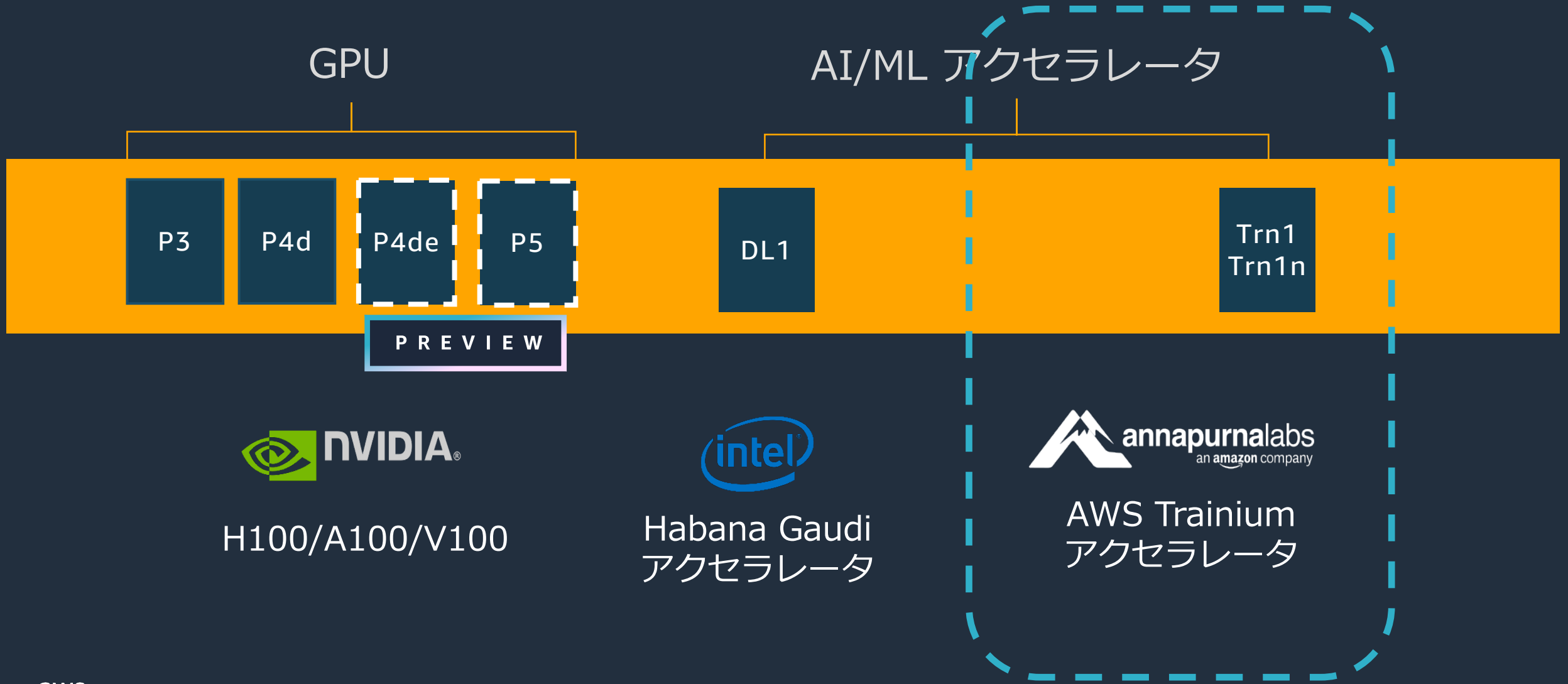
- Habana SynapseAI v1.6
- 128枚の Gaudi アクセラレータを用いて学習に 16 時間 85 % のスケーリング効率を達成

https://aws.amazon.com/blogs/machine-learning/accelerate-pytorch-with-deepspeed-to-train-large-language-models-with-intel-habana-gaudi-based-dl1-ec2-instances/

大規模モデル学習に適した幅広い選択肢

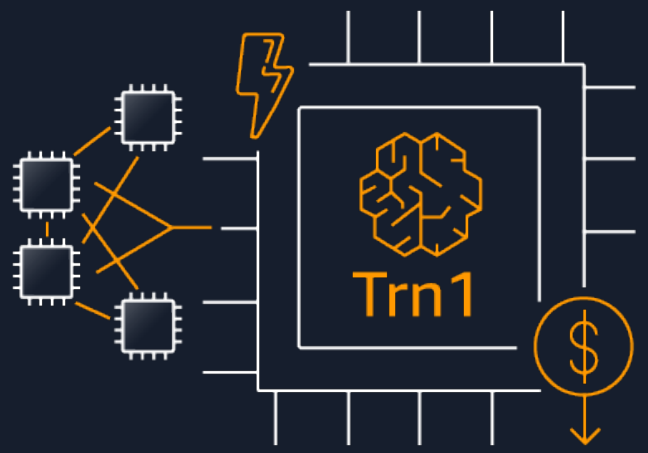


大規模モデル学習に適した幅広い選択肢



Amazon EC2 Trn1/Trn1n インスタンス

＊ N. Virginia



- 最大16個の AWS Trainium アクセラレータを搭載 **512GB** の高帯域幅（HBM2）デバイスメモリ
- 最大8TB のローカル NVMe SSDを搭載
- 768 GB/s のアクセラレータ間通信
- PyTorch、TensorFlow と統合されたSDK

Instance Size	Trainium	アクセラレータメモリ	vCPU	メモリ	アクセラレータ P2P BW	オンデマンド価格 (USD/時間)
Trn1.2xlarge	1	32 GB	8	32 GB	-	1.34
Trn1.32xlarge	16	512 GB	128	512 GB	768 Gbps	21.5
Trn1n.32xlarge	16	512 GB	128	512 GB	768 Gbps	24.78

https://aws.amazon.com/jp/ec2/instance-types/trn1/

様々なお客様が AWS Trainium を活用開始

Magic

Helixon
华深智药

PyTorch

amazon

sprinklr

Money Forward

CACTUS

Trn1 の使用例（一部抜粋）

Point 1:
モデルやデータを
Trainium上に配置
（GPUと同様の手続き）

```
device = "xla"
model = MLP().to(device)
... # 以下学習のループ内で
  train_x = train_x.to(device)
  train_label = train_label.to(device)
  output = model(train_x)
  loss = loss_fn(output, train_label)
  loss.backward()
  optimizer.step()
  xm.mark_step()
```


https://github.com/aws-neuron/aws-neuron-samples/blob/master/torch-neuronx/training/mnist_mlp/train.py

Trn1 の使用例（一部抜粋）

```
device = "xla"
model = MLP().to(device)
... # 以下学習のループ内で
  train_x = train_x.to(device)
  train_label = train_label.to(device)
  output = model(train_x)
  loss = loss_fn(output, train_label)
  loss.backward()
  optimizer.step()
  xm.mark_step()
```

Point 2:
コンパイルと
トレーニングステップ
の実行

https://github.com/aws-neuron/aws-neuron-samples/blob/master/torch-neuronx/training/mnist_mlp/train.py

国内のお客様での Trn1 活用事例

「私たちはユーモアを取り入れて、その場で面白い答えを出す、革新的でインタラクティブな AI チャットボットサービス「大喜利 AI」を提供するため、大規模言語モデルを採用しています。

テンソル並列、データ並列を活用して、Trn1.32xlarge インスタンスで GPT ベースの日本語モデルを事前トレーニングしました。トレーニングは 28 日以内に完了し、以前の GPU ベースのインフラストラクチャよりも **33% のコスト削減**を実現しました。モデルが急速に複雑化し続けているため、大規模なモデルのトレーニングをスピードアップするために、Trn1 の **2 倍のネットワーク帯域幅を持つ Trn1n インスタンス**を楽しみにしています。」

株式会社わたしは 最高技術責任者（CTO）小橋 洋平 様

大規模学習を支える技術要素

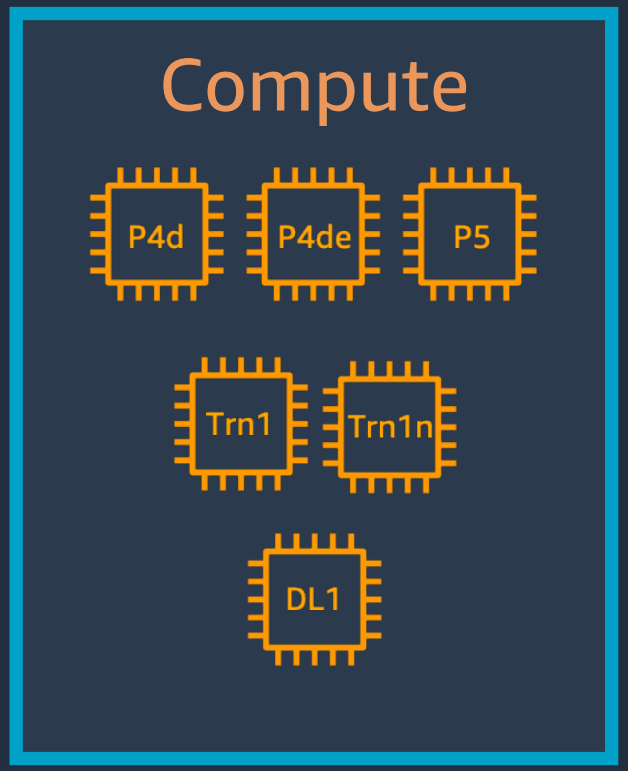
Compute

- 大容量デバイス
  メモリ
- 高速アクセラレータ

Network

- 広帯域
  インターコネクト

大規模学習を支える技術要素

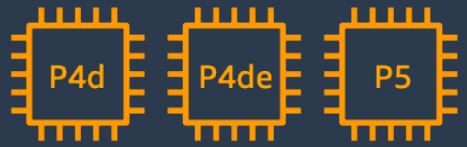


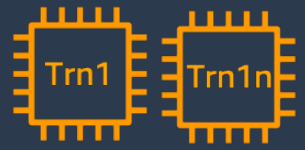
Network

- 広帯域
インターコネクト

大規模学習を支える技術要素

Compute

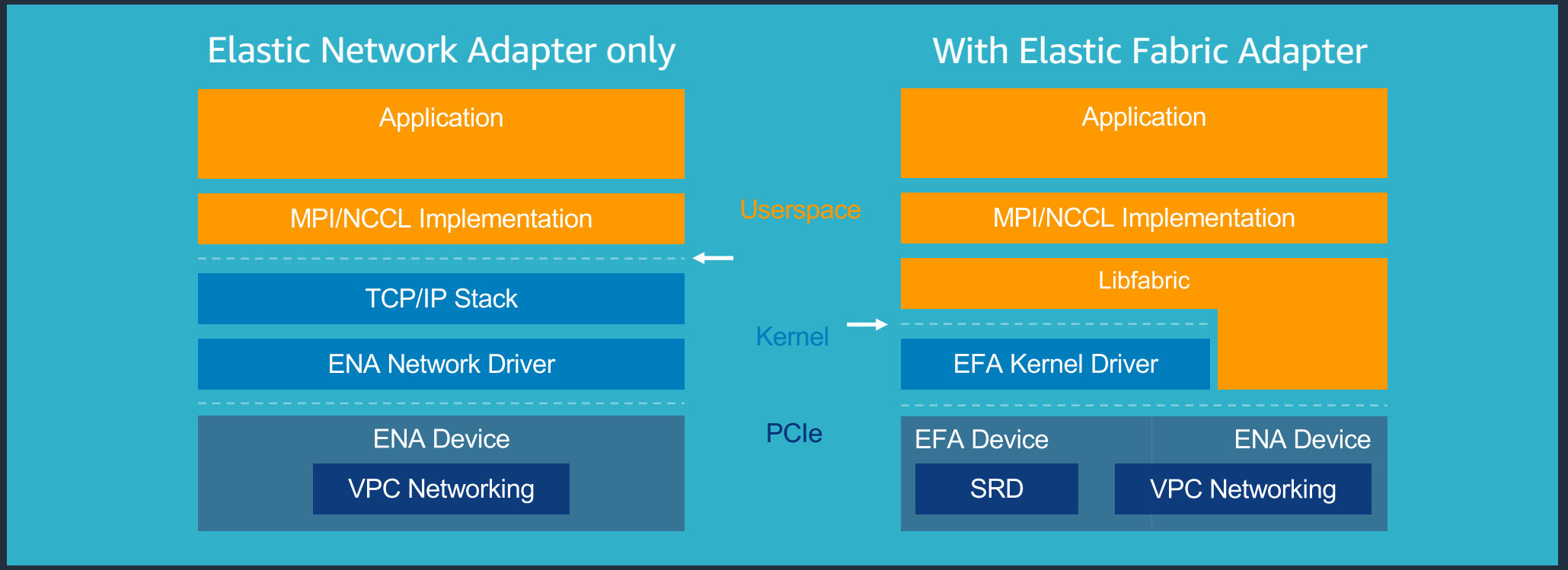




Network

- 広帯域
  インターコネクト

Elastic Fabric Adapter (EFA)



1. Libfabric API によるOSカーネルをバイパスした通信
2. Out of order での転送： Head of blocking 問題を回避
3. マルチパスルーティングによる安定した低レイテンシーの実現

アクセラレータ別 EFA 対応状況まとめ

Instance	アクセラレータ	アクセラレータメモリ	EFA version	ネットワーク帯域
DL1.24xlarge	Gaudi	256 GB	V1	400 Gbps
Trn1.32xlarge	Trainium	512 GB	V2	800 Gbps
Trn1n.32xlarge	Trainium	512 GB	V2	1600 Gbps
P4d.24xlarge	GPU	320 GB	V1	400 Gbps
P4de.24xlarge*	GPU	320 GB	V1	400 Gbps
P5*	GPU	640 GB	V2	3200 Gbps

* Preview

アクセラレータ別 EFA 対応状況まとめ

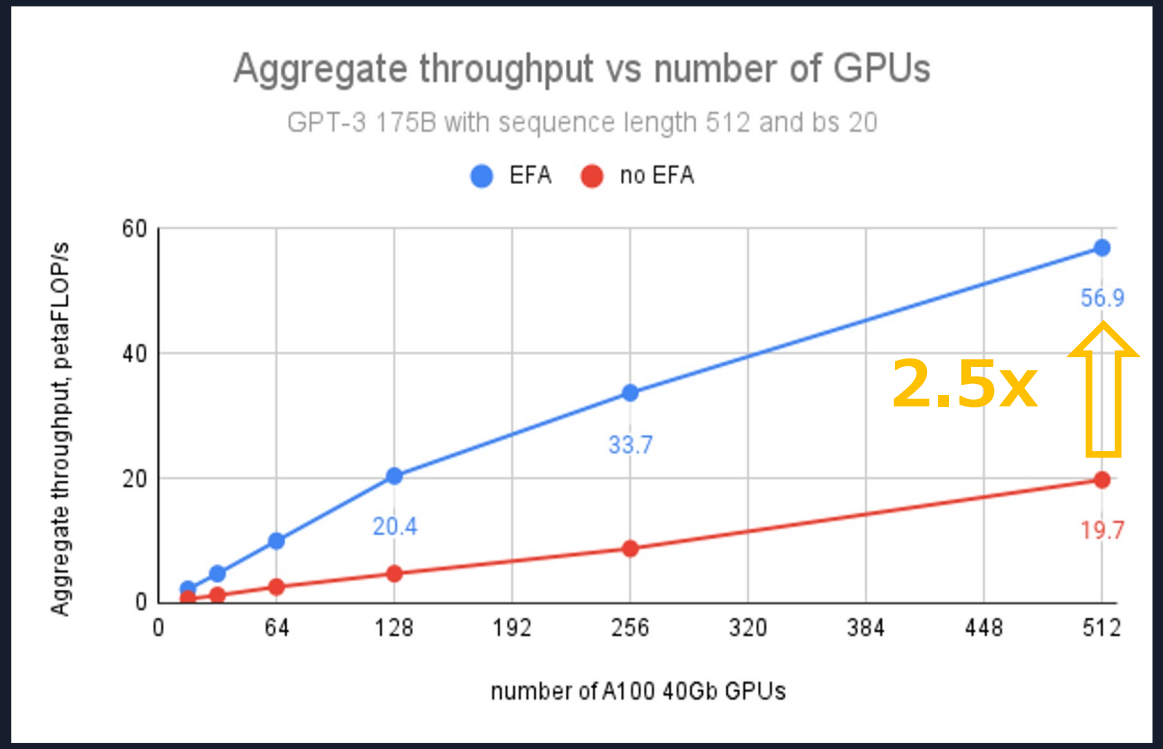
Instance	アクセラレータ	アクセラレータメモリ	EFA version	ネットワーク帯域
DL1.24xlarge	Gaudi	256 GB	V1	400 Gbps
Trn1.32xlarge	Trainium	512 GB	V2	800 Gbps
Trn1n.32xlarge	Trainium	512 GB	V2	1600 Gbps
P4d.24xlarge	GPU	320 G		Gbps
P4de.24xlarge*	GPU	320 G		Gbps
P5*	GPU	640 GB	V2	3200 Gbps

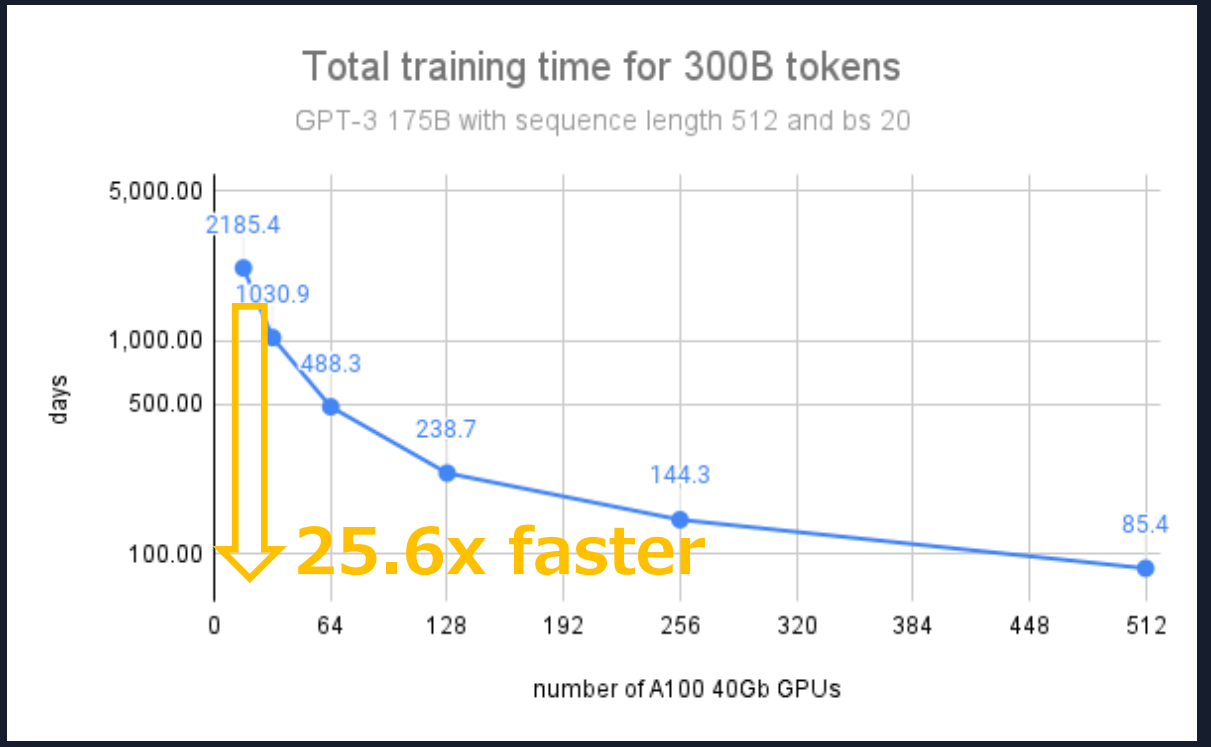
大規模モデルの学習に特にオススメ

* Preview

PyTorch FSDP を用いた分散学習のパフォーマンス検証

GPT-3 175B parameters model





512 GPUs = 64 instances

https://medium.com/pytorch/training-a-1-trillion-parameter-model-with-pytorch-fully-sharded-data-parallel-on-aws-3ac13aa96cff

記事公開日: 2022/05/16

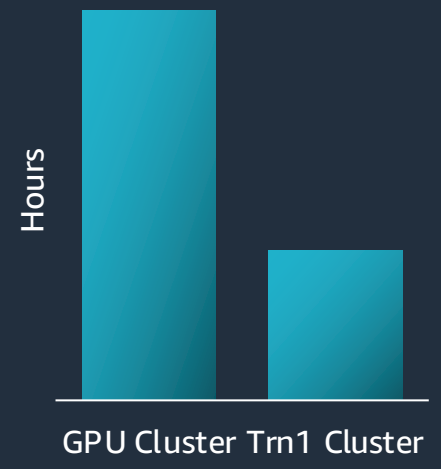
AWS Trainium: 高性能、低電力、低コストを両立

Training BERT Large with AWS Trainium

2.3x

高速なトレーニング

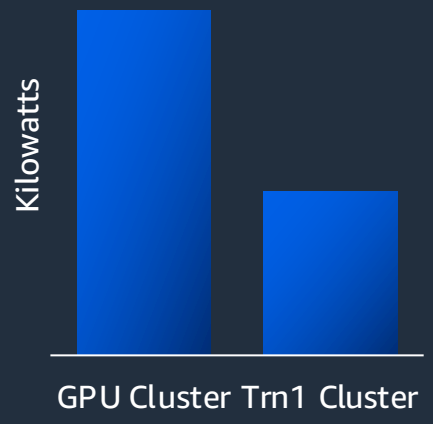
トレーニング時間



47%

低電力

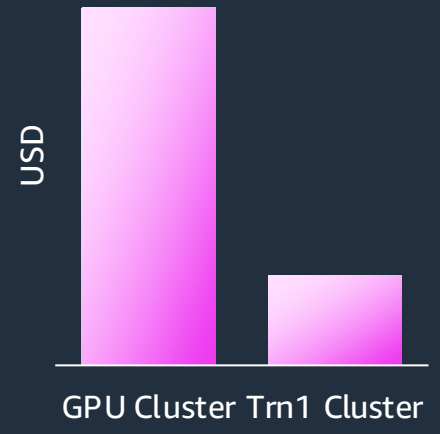
電力



72%

低コスト

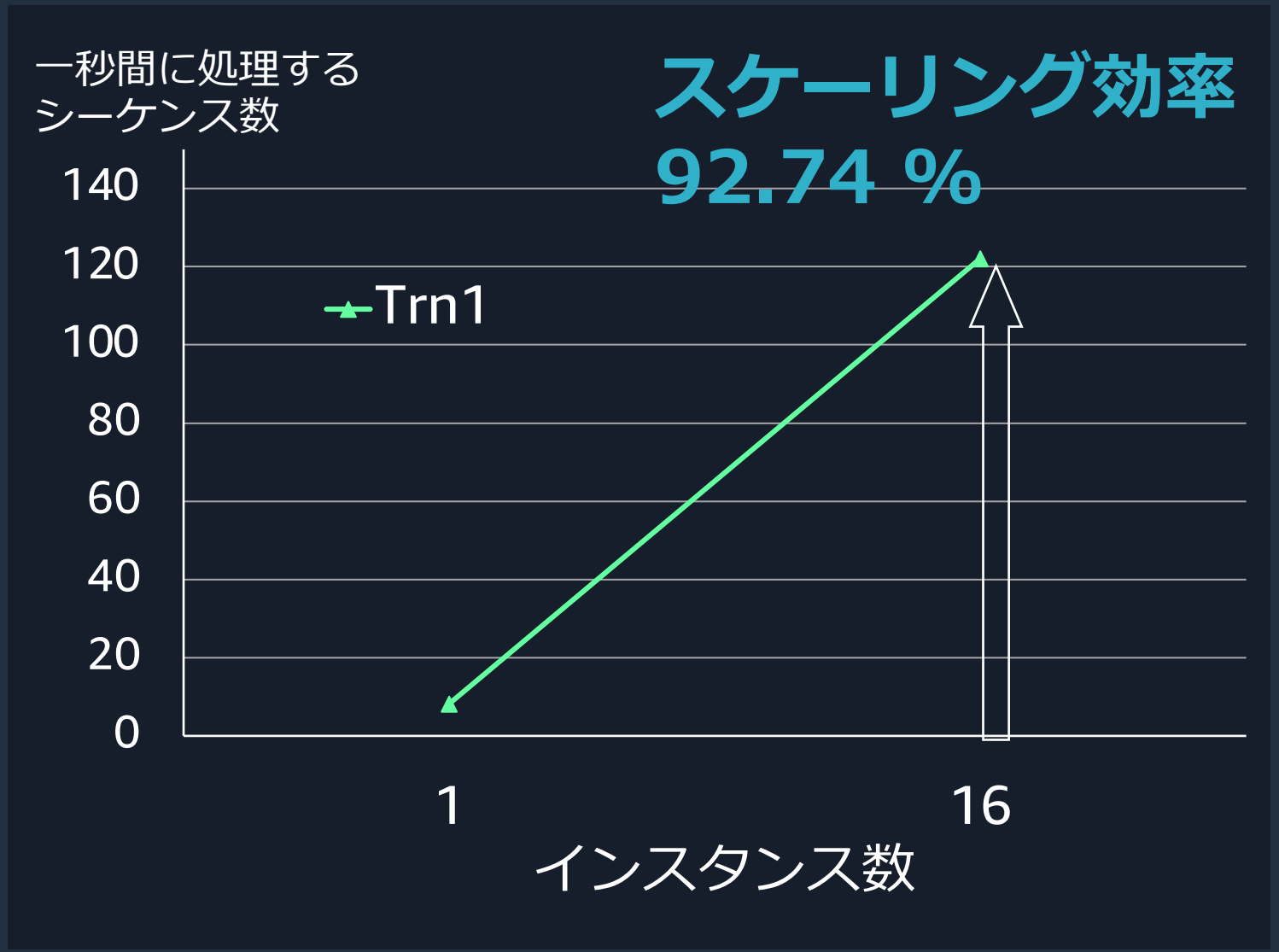
トレーニングコスト



Details: Hugging Face Bert-Large, FP32, On-Demand EC2 pricing

第２世代 **EFA** によるスケーリングの例 **GPT-3**

- Megatron-LM を用いたTrn1上での分散学習例
- 1インスタンス
  - Global minibatch 64
- 16 インスタンス
  - Global minibatch 1024



https://awsdocs-neuron.readthedocs-hosted.com/en/latest/frameworks/torch/torch-neuronx/tutorials/training/megatron_lm_gpt.html#megatron-lm-pretraining-tutorial

https://awsdocs-neuron.readthedocs-hosted.com/en/latest/general/benchmarks/trn1/trn1-performance.html#trn1-performance

大規模学習を支える技術要素

Compute

- 大容量デバイス
  メモリ
- 高速アクセラレータ

Network

- 広帯域
  インターコネクト

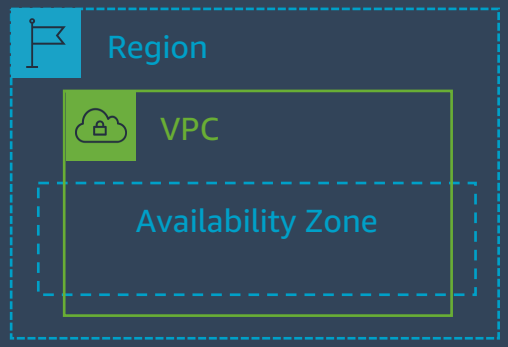
大規模学習を支える技術要素

Compute

- 大容量デバイスメモリ
- 高速アクセラレータ

Network

Elastic Fabric Adapter



まとめ：大規模学習を支える技術要素