



生成AIを加速するNVIDIAの取組み

エンタープライズ事業本部 事業本部長 井崎 武士

NVIDIA —A COMPUTING PLATFORM COMPANY

- > 1993 年創業
- > 創業者及び CEO ジェンスン ファン
- > 従業員 27,000 人
- > 2023 会計年度売上高 270億ドル
- 時価総額 1兆ドル



NVIDIA のプラットフォーム戦略

4つのコアコンピタンス

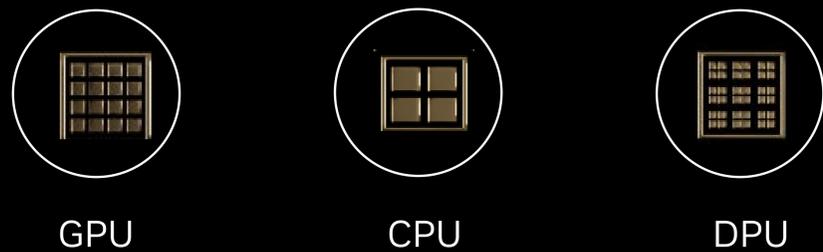
NVIDIA アプリケーション フレームワーク



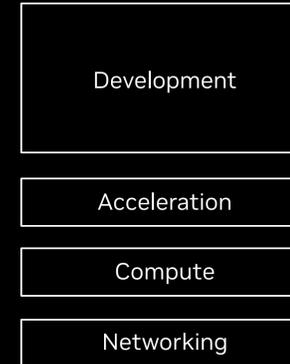
NVIDIA プラットフォーム ソフトウェア ライブラリ



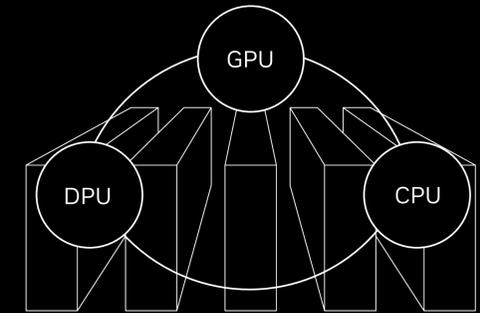
NVIDIA アクセラレーテッドコンピューティング インフラ



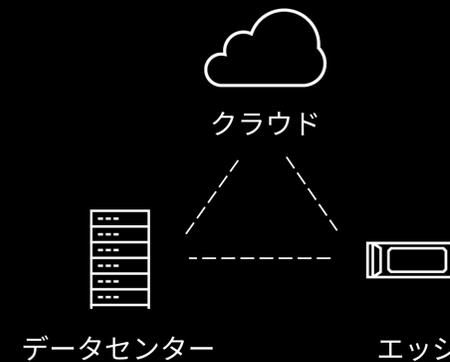
フルスタック



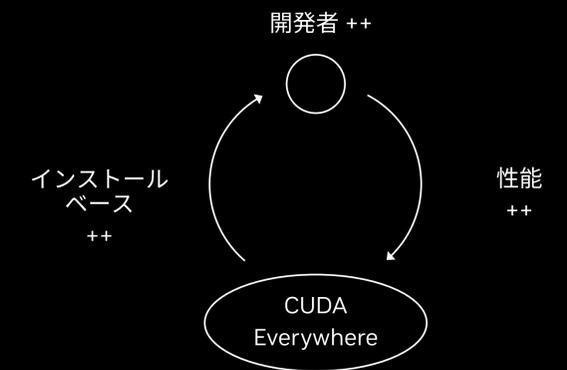
データセンタースケール コンピューティング



エンド・ツー・エンド



単一アーキテクチャ



開発者数: 400万人 / 利用企業: 4万社 / アプリケーション数: 3000超 / スタートアップ: 13,000社



生成AIの衝撃

iPhone moment

Stable Diffusion



Water front city landscape, sunshining, photo realistic, summer, 4k, rule of thirds, extreme detail, detailed drawing, trending artstation, hd, realistic lighting, by Alphonse Mucha, Greg Rutkowski, sharp focus, backlit

ChatGPT

プロの経済アナリストの立場で教えてください。AIの活用による経済効果を分野ごとにリストアップしてください

AIの活用による経済効果は、さまざまな分野で期待されています。以下に、いくつかの分野とそれに伴う経済効果をリストアップします。

1. 製造業:
 - 自動化とロボット化による生産性向上
 - プロセスの最適化による品質向上
 - 機械学習による予測メンテナンスによる機械の故障
 - データ分析による在庫最適化と生産計画の最適化
2. 金融業:
 - 高速なデータ処理とパターン認識によるリスク管理
 - モデルベースの予測による株式市場や為替市場の
 - ロボアドバイザーによる個人投資家の助言やポート
 - 不正検知によるセキュリティの向上
3. 医療業界:
 - パターン認識と画像解析による病気の早期検出
 - データマイニングによる臨床試験の最適化と新薬開
 - ヘルスケアデータの分析による予防医療の向上とコ
 - ロボット手術による手術の精度向上と手術時間の
4. 小売業:

クルマの車種を分類するViTを使ったPyTorchのスク립ト書いて

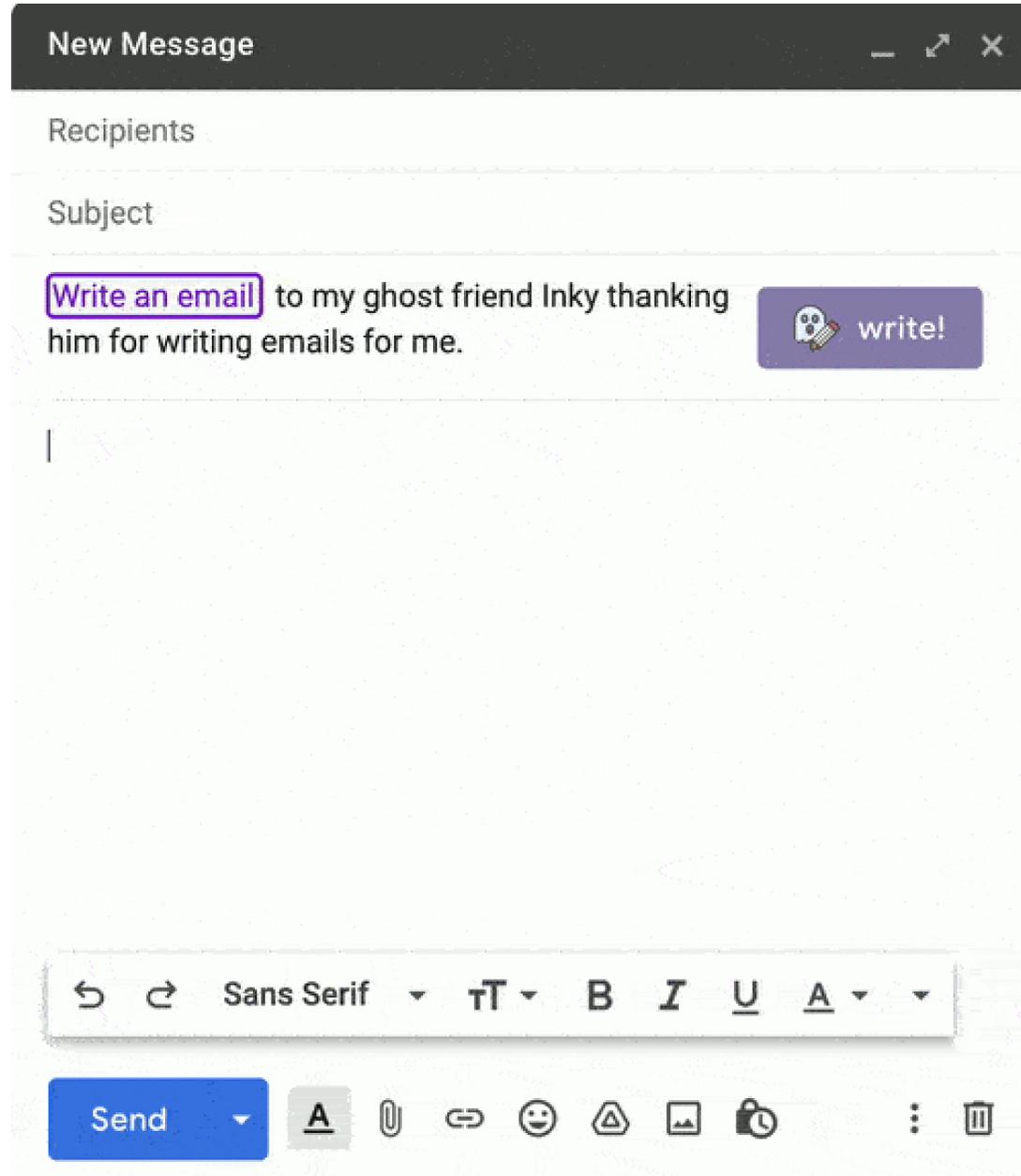
以下は、PyTorchを使用してViTを使ったクルマの車種分類スク립トの例です。このスク립トは、CIFAR-10データセットで訓練されたもので、ViTのパラメータは事前に学習されたものを使用しています。

```
python
import torch
import torch.nn as nn
import torch.optim as optim
import torchvision.datasets as datasets
import torchvision.transforms as transforms
from torch.utils.data import DataLoader
from einops.layers.torch import Rearrange

# Define the ViT model architecture
class ViT(nn.Module):
    def __init__(self, image_size=32, patch_size=4, num_classes=10, dim=768, depth=12, head_dim=64):
        super(ViT, self).__init__()

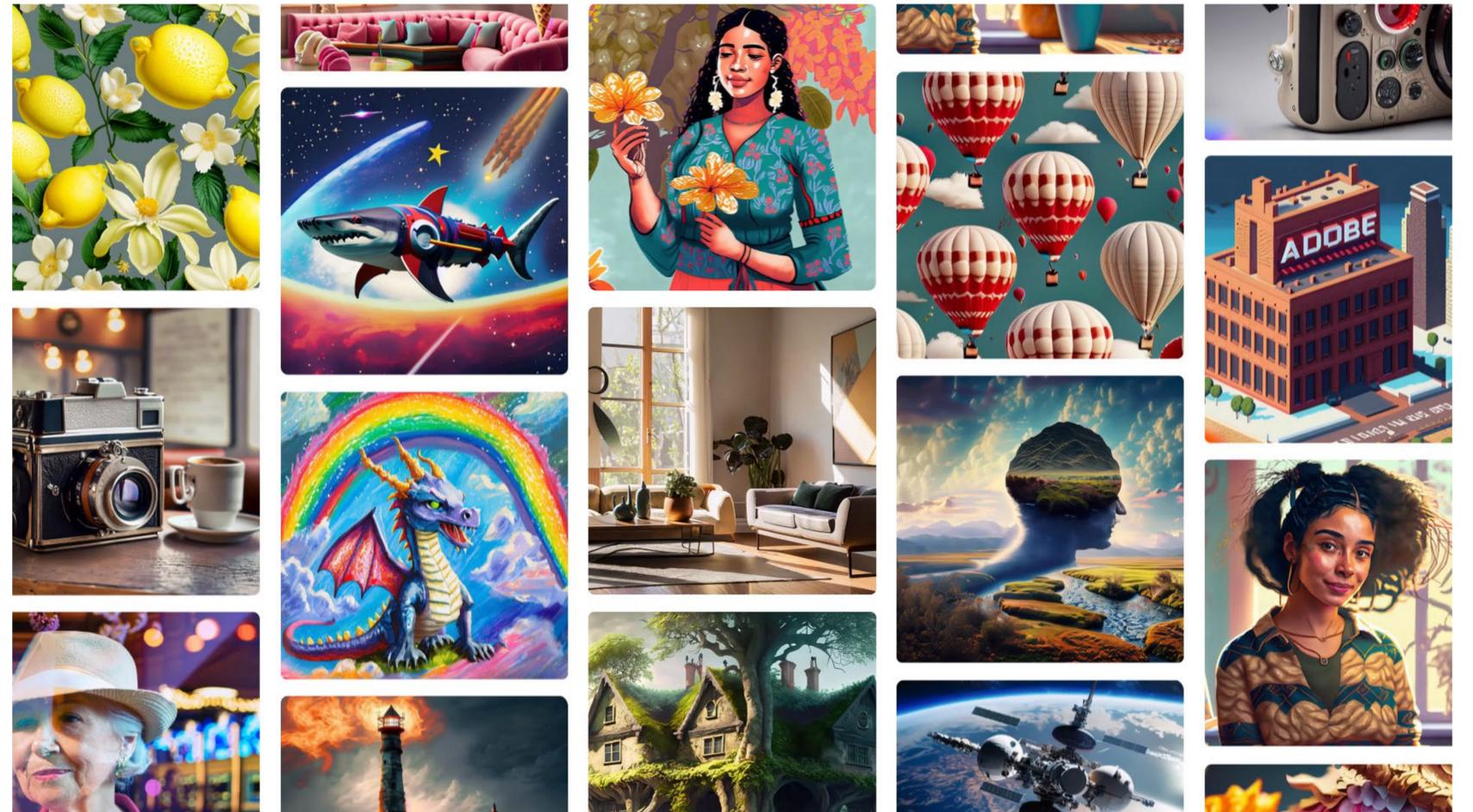
        # Calculate the number of patches
        self.num_patches = (image_size // patch_size) ** 2
        self.patch_dim = 3 * patch_size ** 2
```

生成AI 事例



Ghostwrite

ChatGPTを活用した、e-mail 作成アシスタント

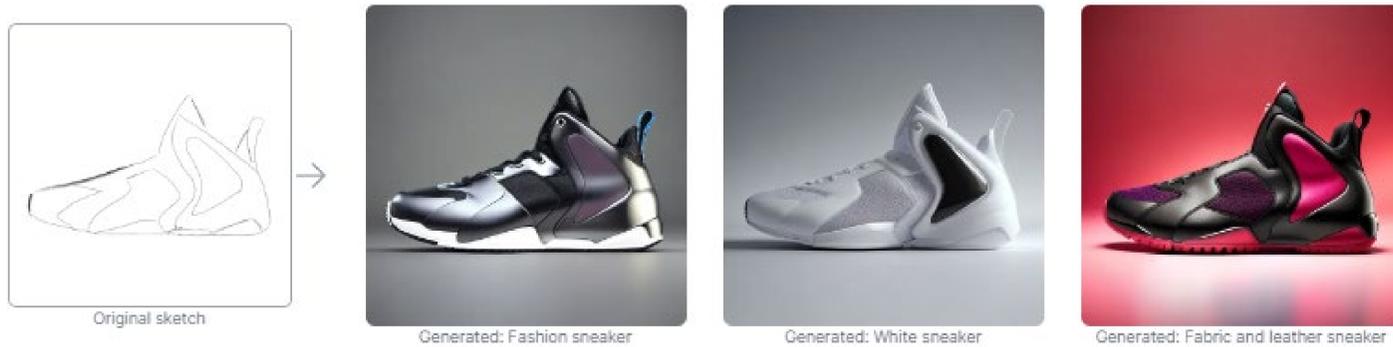


Adobe Firefly

ベータリリース

Adobe.comより

生成AI 事例



newarc.ai

Describe your music

A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a Samurai battle.

Condition on a melody (optional) File or Mic

file mic

File

Generate

A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a Samurai battle.

Music gen

生成AIをどのように使えばいいのか？

利用者数

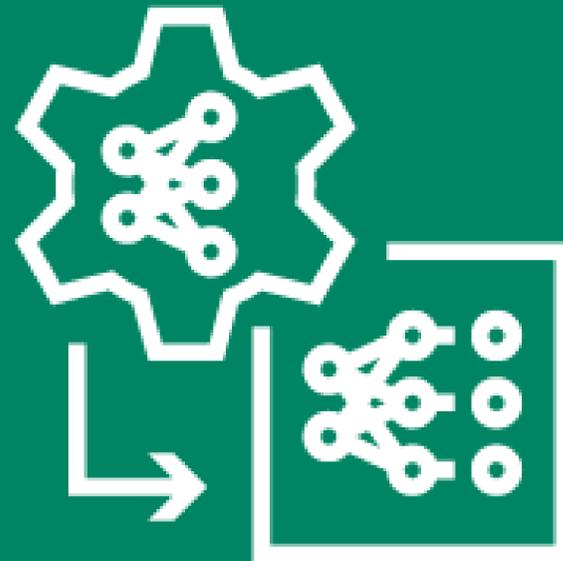
最小限のカスタム化

生成AIサービスを利用する - ChatGPT, Google Bard, Amazon Bedrockなどの既存サービス
サービス利用に応じたコンサンプションモデル
早期の市場投入が可能



中間のカスタム化

事前学習モデルのファインチューニング
インフラやリソースに数億の投資が必要
数週間から数ヶ月の開発期間



広範囲なカスタム化

独自の基盤モデル構築もしくは広範囲なファインチューニング
インフラやリソースに数十億の投資が必要
半年以上の開発期間



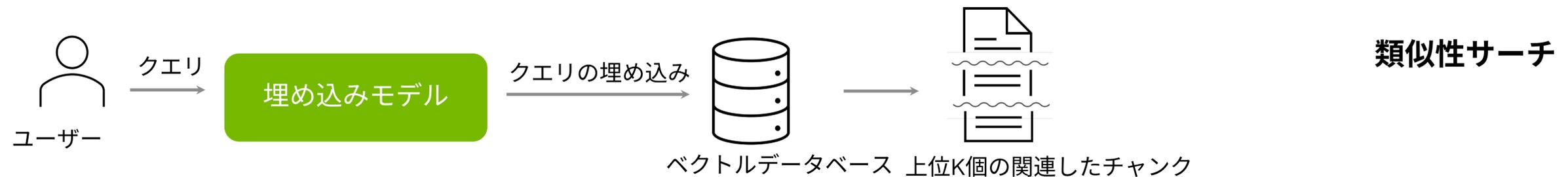
生成AIのカスタム化

RAG (Retrieval Augmented Generation)

- ステップ 1: 知識ベースを埋め込む



- ステップ 2: ユーザーのクエリに従ってベクトルデータベースから関連した文書を抜き出す



- ステップ 3: 抜き出したチャンクからのコンテキストを落ちて回答を生成するためにLLMを使用



エンタープライズに向けた生成AIアプリケーションの構築

エンド・ツー・エンド、構築、カスタマイズ、デプロイのためのクラウドネイティブなフレームワーク

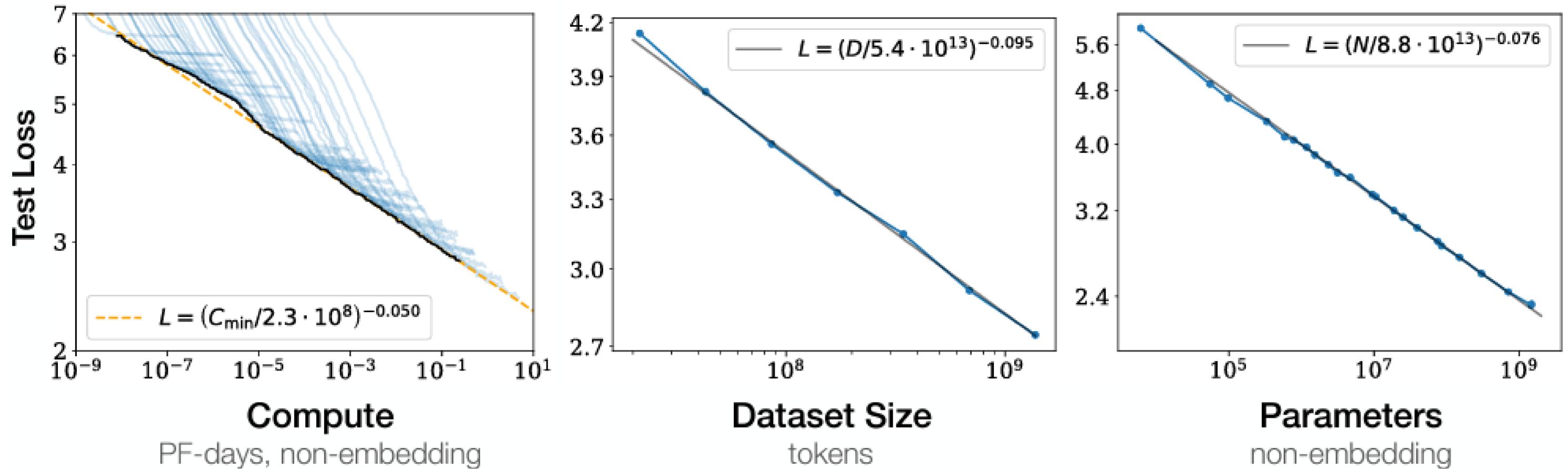


データ準備

学習とカスタマイズ

デプロイ

自然言語モデルのスケール則



NVIDIA H100

世界の AI インフラを支える新たなエンジン

最高の AI/HPC 性能

4PF FP8 (6X) | 2PF FP16 (3X) | 1PF TF32 (3X) | 60TF FP64 (3X)
3TB/s (1.5X), 80GB HBM3 memory

TRANSFORMER モデルへの最適化

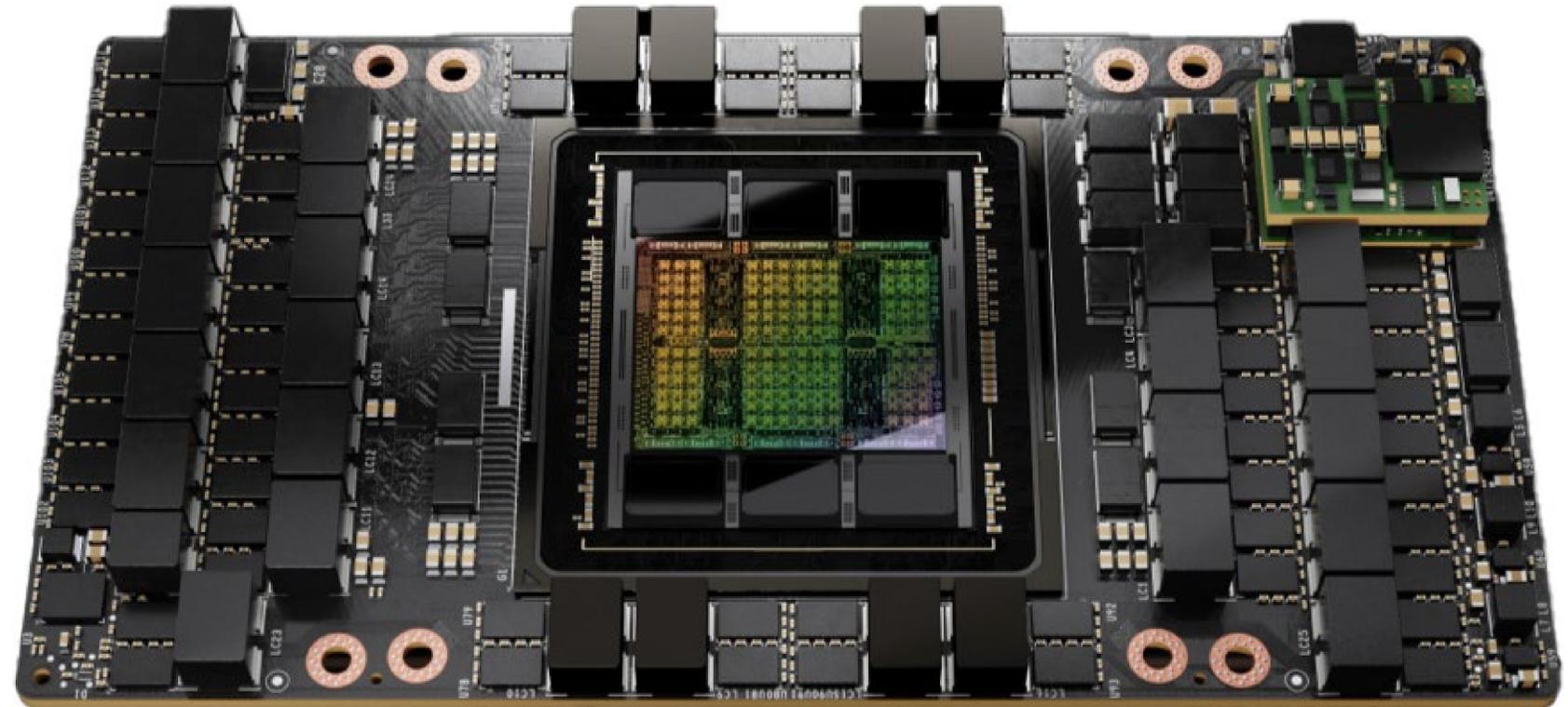
6X faster on largest transformer models

高い稼働率とセキュリティ

7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

史上最速でスケーラブルなインターコネクト

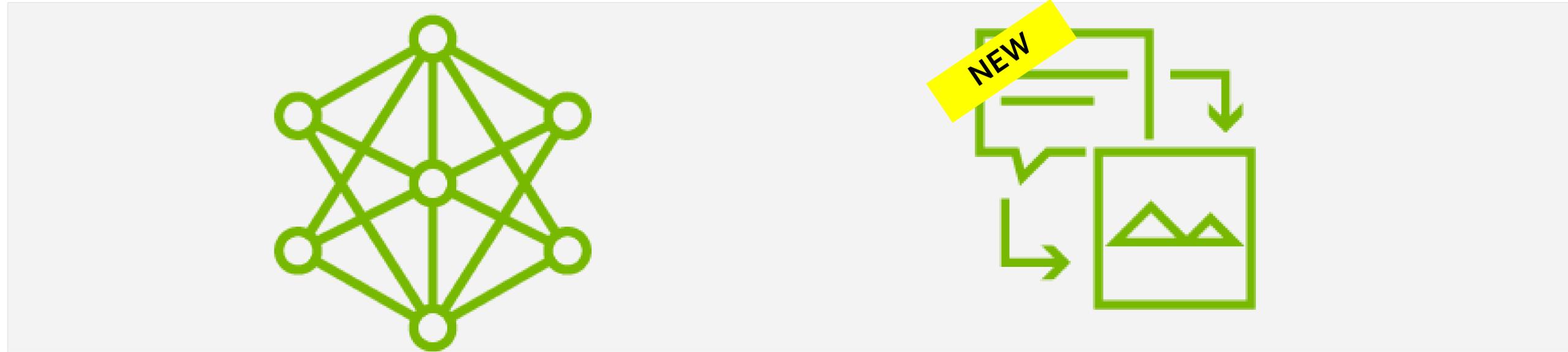
900 GB/s GPU-2-GPU connectivity (1.5X) | 128GB/s PCI Gen5



Custom TSMC 4N Process | 4.9 TB/s Total External B/W

業界標準の生成AI学習ベンチマーク

MLPerf Training v3.1



GPT-3 175B
大規模言語モデル

Stable Diffusion
Text-to-Image



DLRMv2
Recommendation



BERT-Large
NLP



RetinaNet
Object Detection,
Lightweight



Mask R-CNN
Object Detection,
Heavyweight



3D U-Net
Biomedical Image
Segmentation



RNN-T
Speech Recognition



Cat
ResNet-50 v1.5
Image Classification

新たな6つの性能記録

No.1の記録をさらに更新

GPT-3 175B (1B トークン)

3.9 分

2.8倍高速化

Stable Diffusion

2.5 分

新たなワークロード

DLRM-dcnv2

1 分

1.6倍高速化

BERT-Large

7.2 秒

1.1倍高速化

RetinaNet

55.2 秒

1.8倍高速化

3D U-Net

46 秒

1.07倍高速化



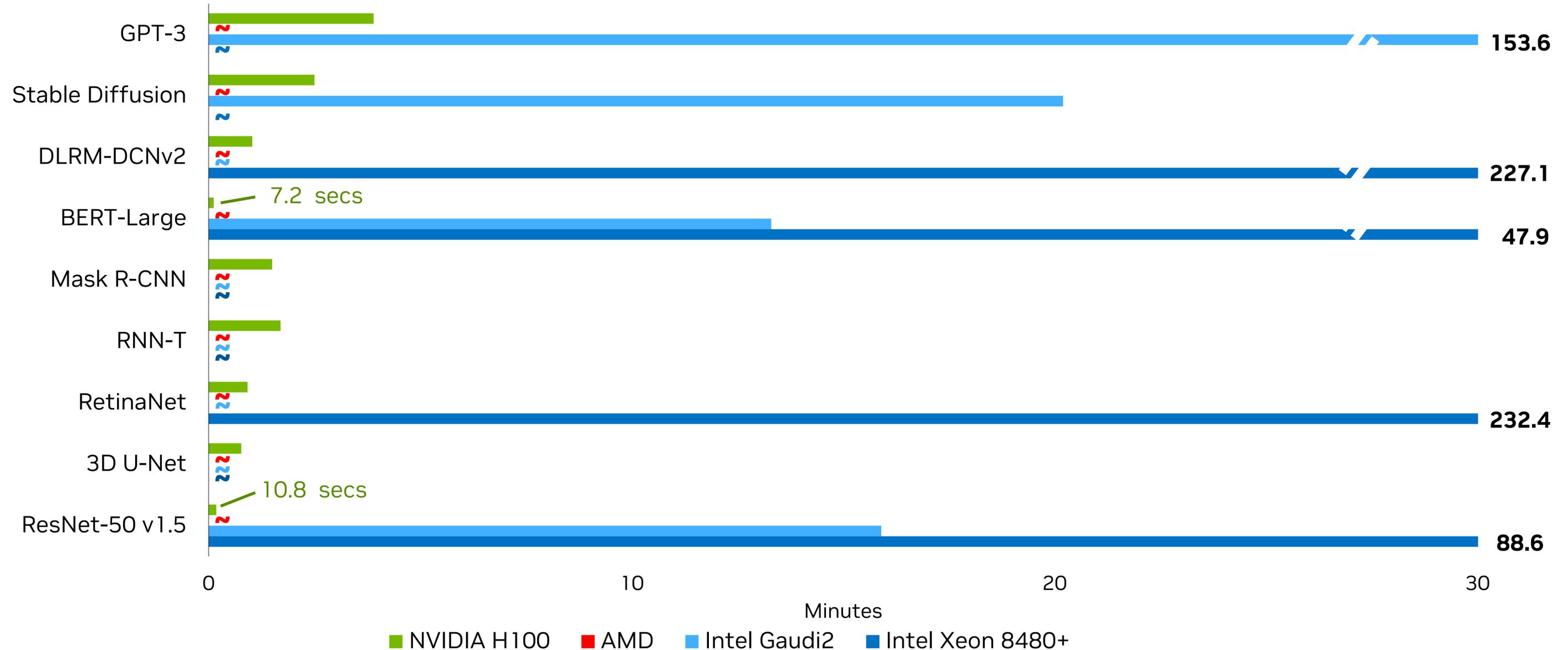
H100 がすべてのワークロードで最高性能を記録

過去最高の性能と高い汎用性

提出された最大規模の学習時間

~ = 提出無し

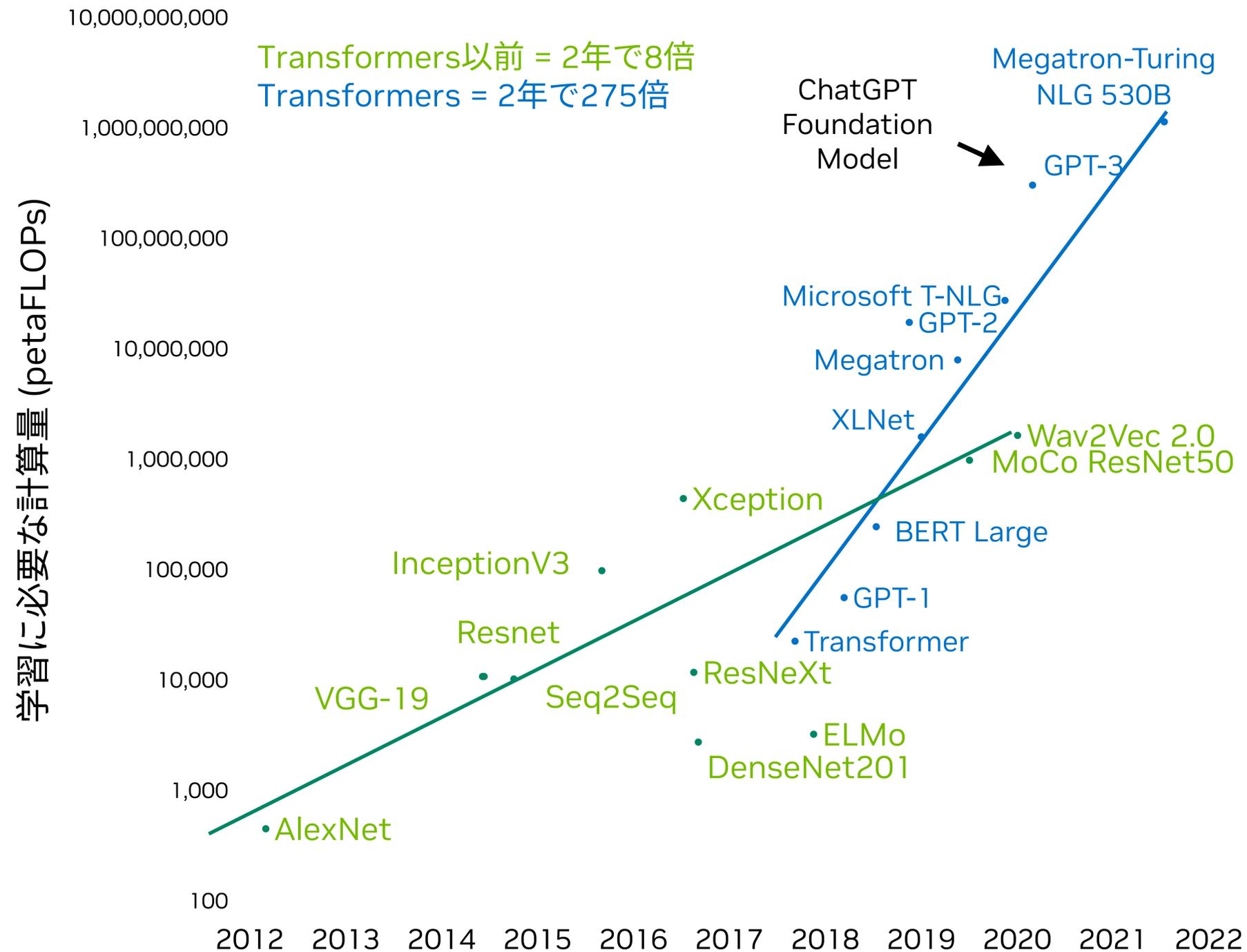
低い方が高い性能を示す



指数関数的に増大するLLMモデルサイズ

LLMが必要な計算能力は指数関数的に伸びている

AIの学習に必要な計算量



代表的なLLMと学習

モデル	発表時期	モデルサイズ(B)	事前学習 データサイズ	ハードウェア	学習時間
OPT (Meta)	May-2022	175	180B Token	992 80G A100	-
GLM (清華大学)	Oct-2022	130	400B Token	768 40G A100	60日
BLOOM (BigScience)	Nov-2022	176	366B Token	384 80G A100	105日
LLaMA (Meta)	Feb-2023	65	1.4T Token	2048 80G A100	21日
LLaMA2(Meta)	Jul-2023	70	2T Token	2000 80G A100	
MT-NLG (MS/NVIDIA)	Jan-2022	530	270B Token	4480 80G A100	-

新発表： NVIDIA HGX H200

世界をリードするAIコンピューティングプラットフォーム

メモリ
141GB
HBM3e

メモリ帯域
4.8 TB/s
HBM3e

GPT-3 175B 推論
1.6X
H100対性能

Llama 2 70B 推論
1.9X
H100対性能

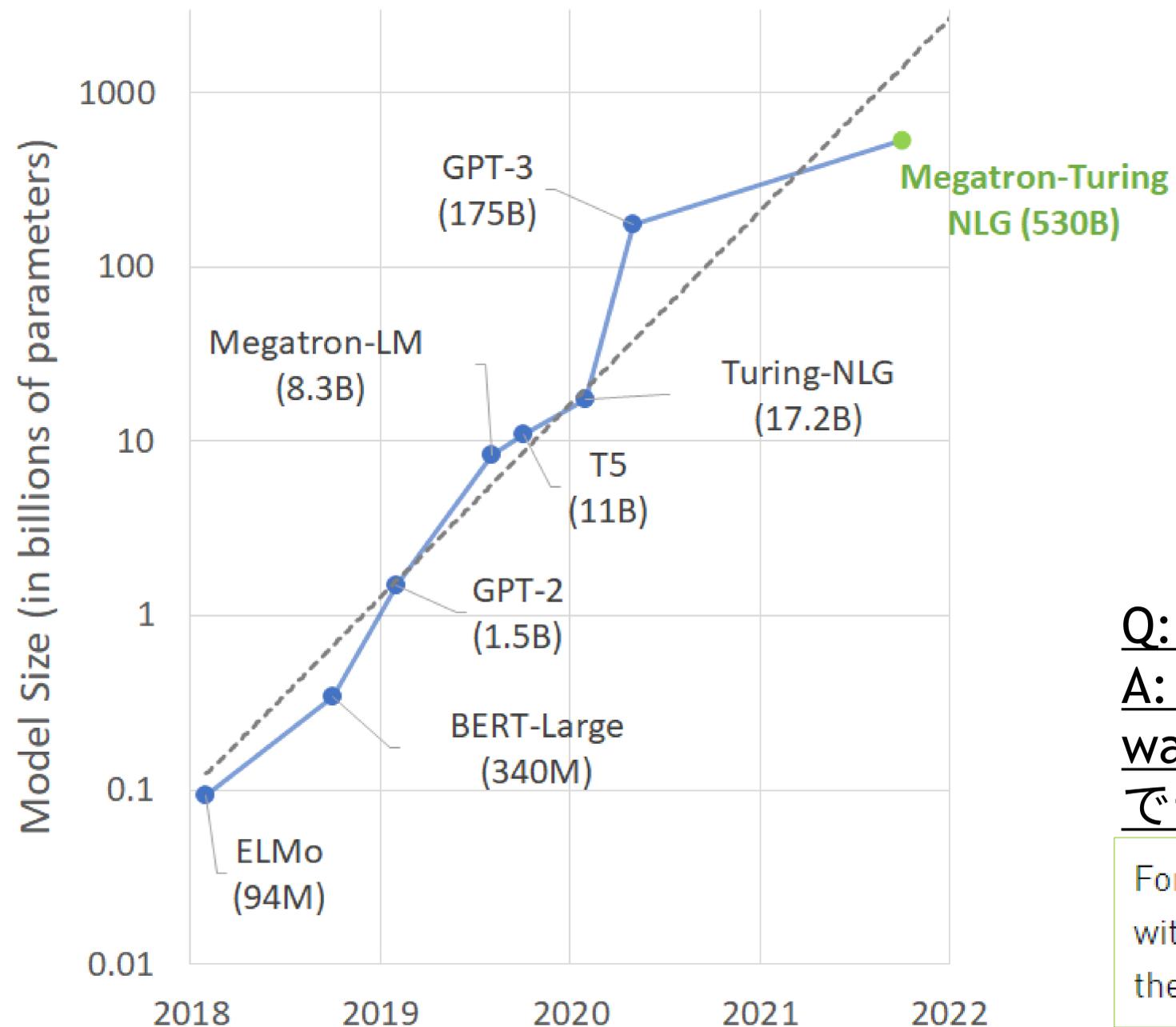


1.1 TB HBM3e | 32 PF FP8
1.8X More Memory Capacity | 1.4X More HBM Bandwidth

Projected performance, subject to change
GPT-3 175B: ISL 80, OSL 200 | Throughput | HGX H100 8-GPUs BS 64 | HGX H200 8-GPUs BS 128
Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 1x GPU BS 8 | H200 1x GPU BS 32
HPC MILC- dataset NERSC Apex Medium | HGX H200 4-GPU | dual Sapphire Rapids 8480

モデルがGPUに載らない？

分散学習の必要性



MT-NLGは530Bパラメータ

- 単純計算で 2,120 GB (in FP32)
- モデルをすべてメモリにロードするだけで、8xA100 (80GB) サーバが、3 台強必要
- ワーキングメモリも当然必要

Q: 実際どうやって扱っているの？

A: 280xA100 (=35 nodes) で 8-way tensor parallel & 35-way pipeline parallel のモデル並列 & 4480GPUのクラスターでデータ並列

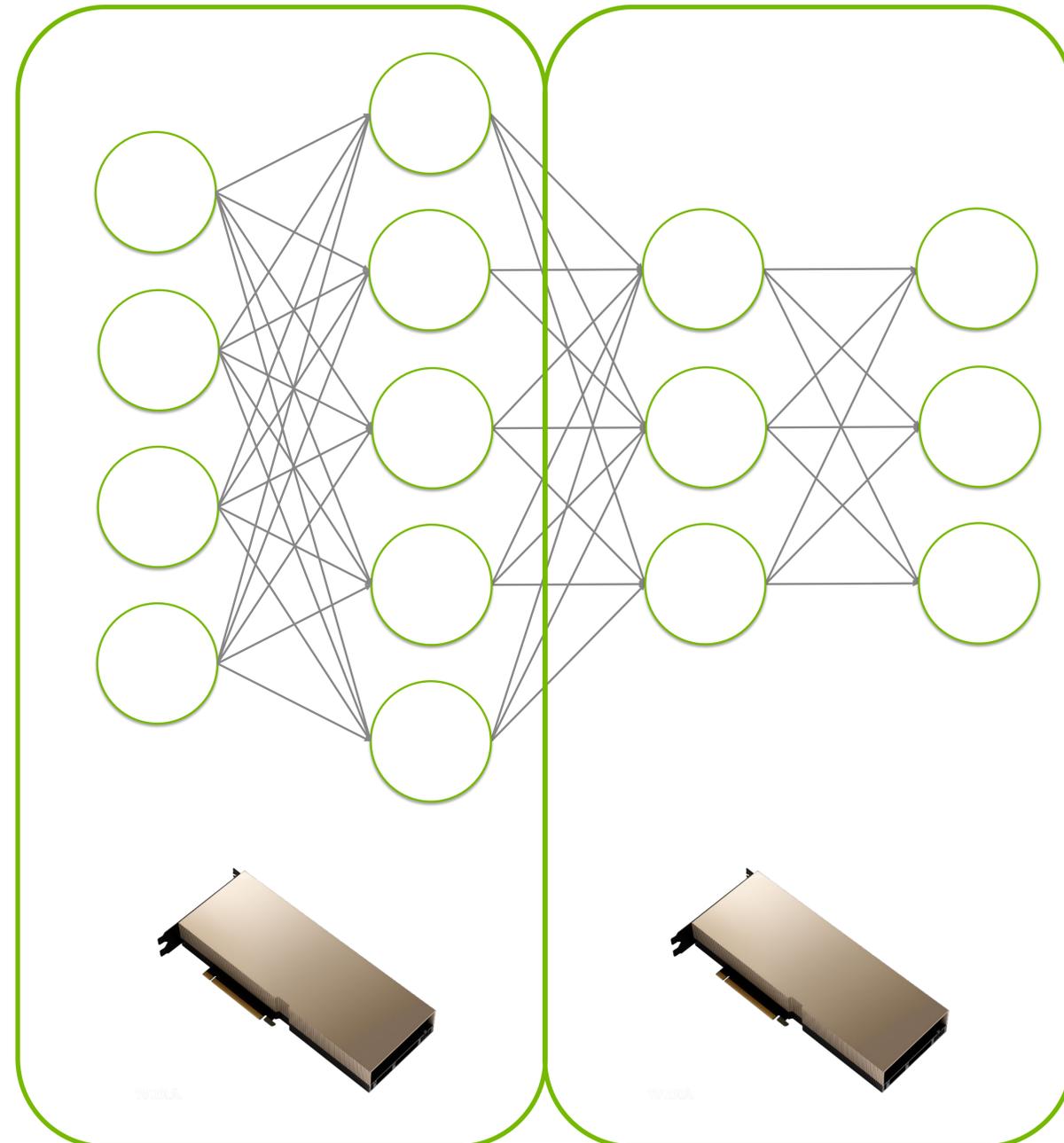
For example, for the 530 billion model, each model replica spans 280 NVIDIA A100 GPUs, with 8-way tensor-slicing within a node and 35-way pipeline parallelism across nodes. We then use data parallelism from DeepSpeed to scale out further to thousands of GPUs.

<https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/>

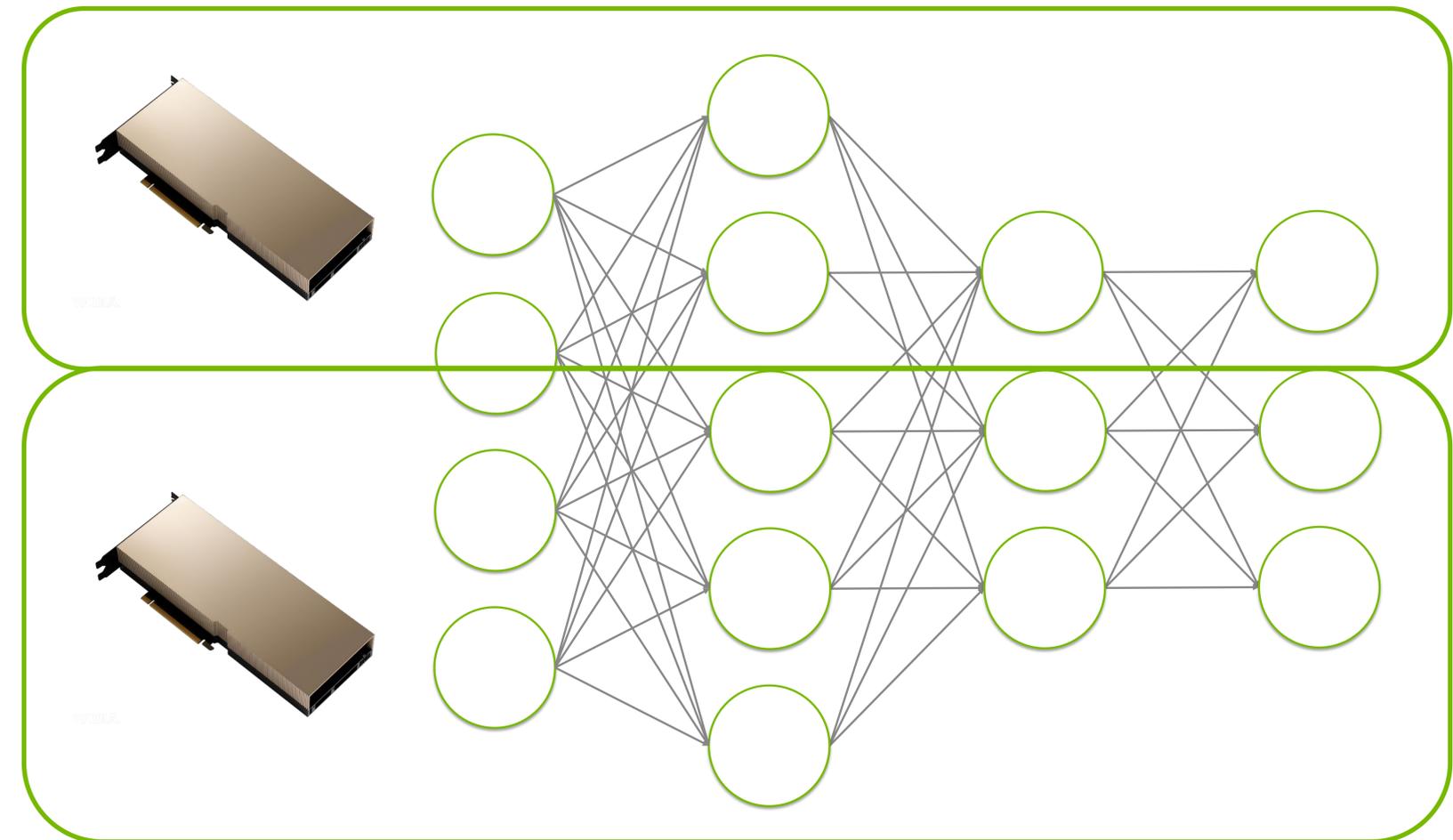
モデル並列の考え方

大規模なモデルを扱えるサイズに分割する

レイヤーごとに別デバイスへ割り当て
(pipeline parallel とも)



各デバイスでレイヤーの一部を担当
(tensor parallel とも)



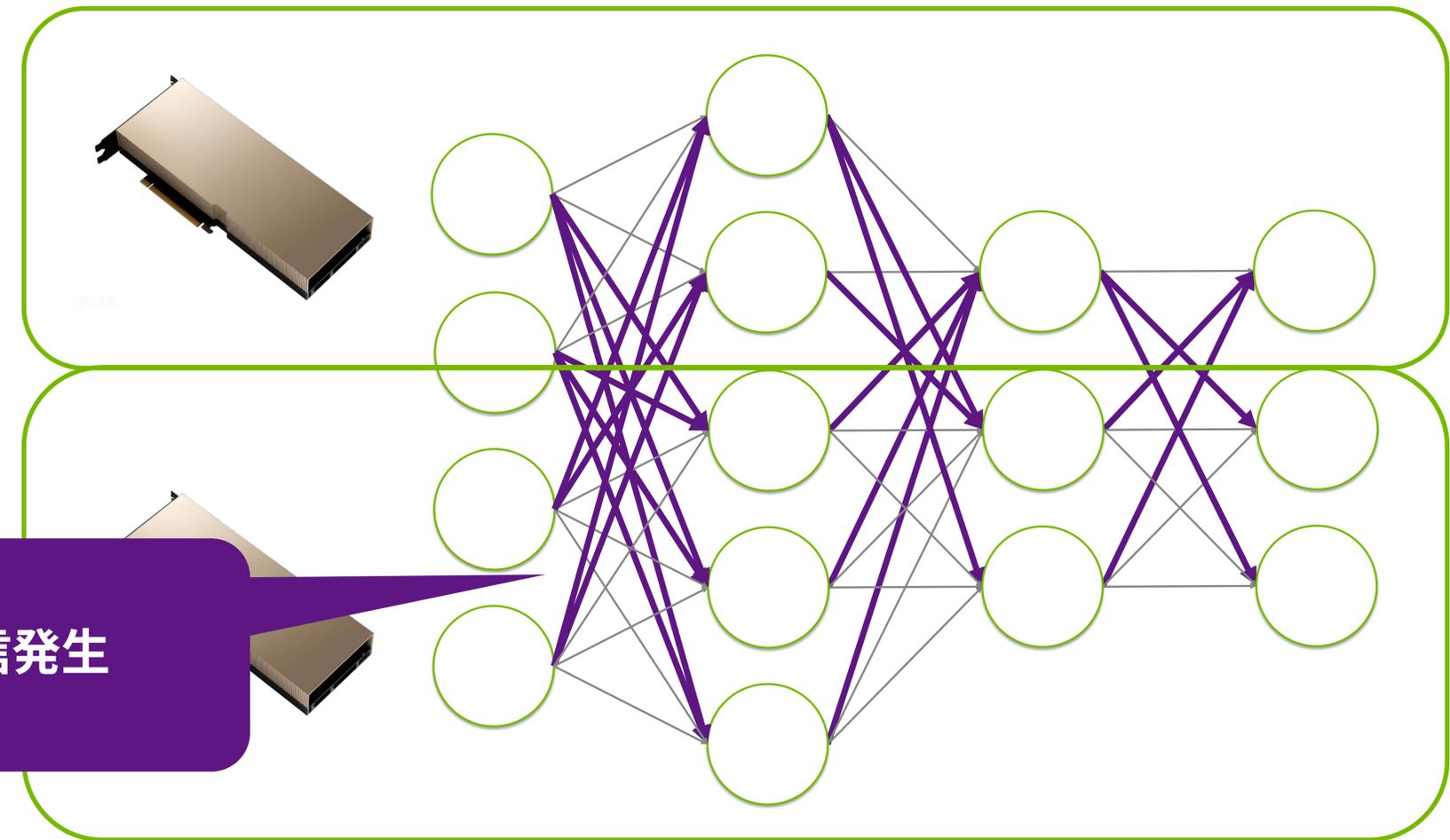
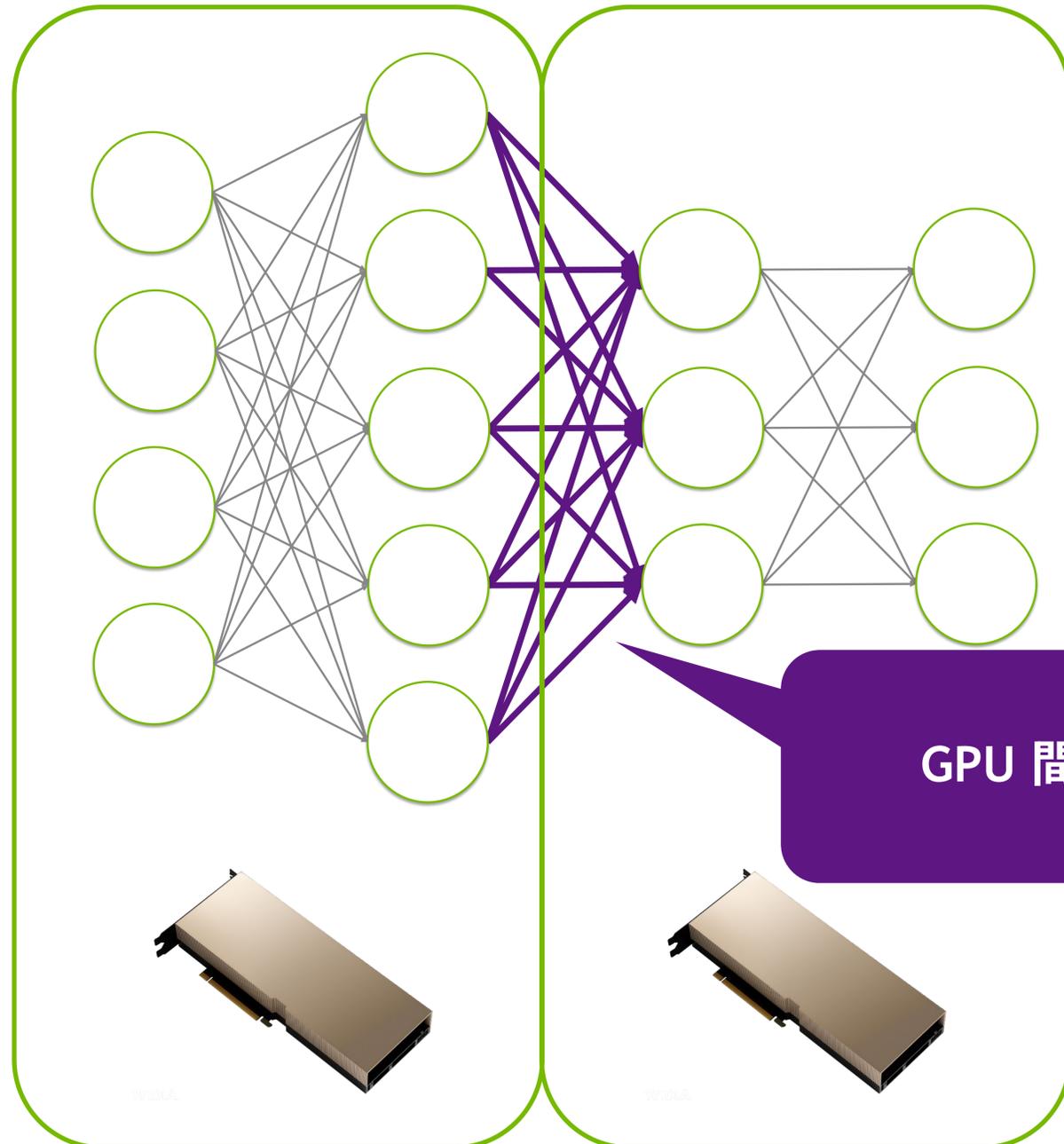
その他、畳み込みのチャンネルを
別デバイスに割り当てるとも.....

モデル並列の考え方

大規模なモデルを扱えるサイズに分割する

レイヤーごとに別デバイスへ割り当て
(pipeline parallel とも)

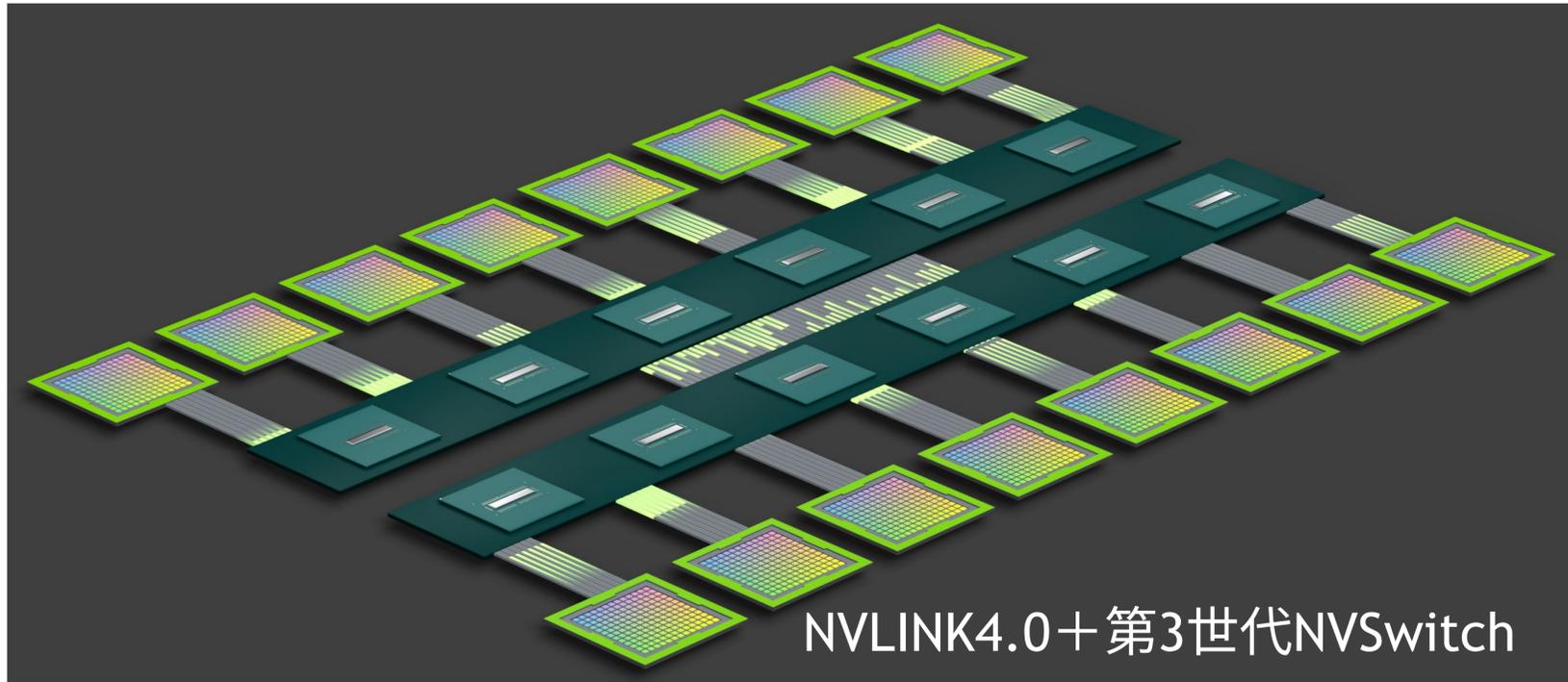
各デバイスでレイヤーの一部を担当
(tensor parallel とも)



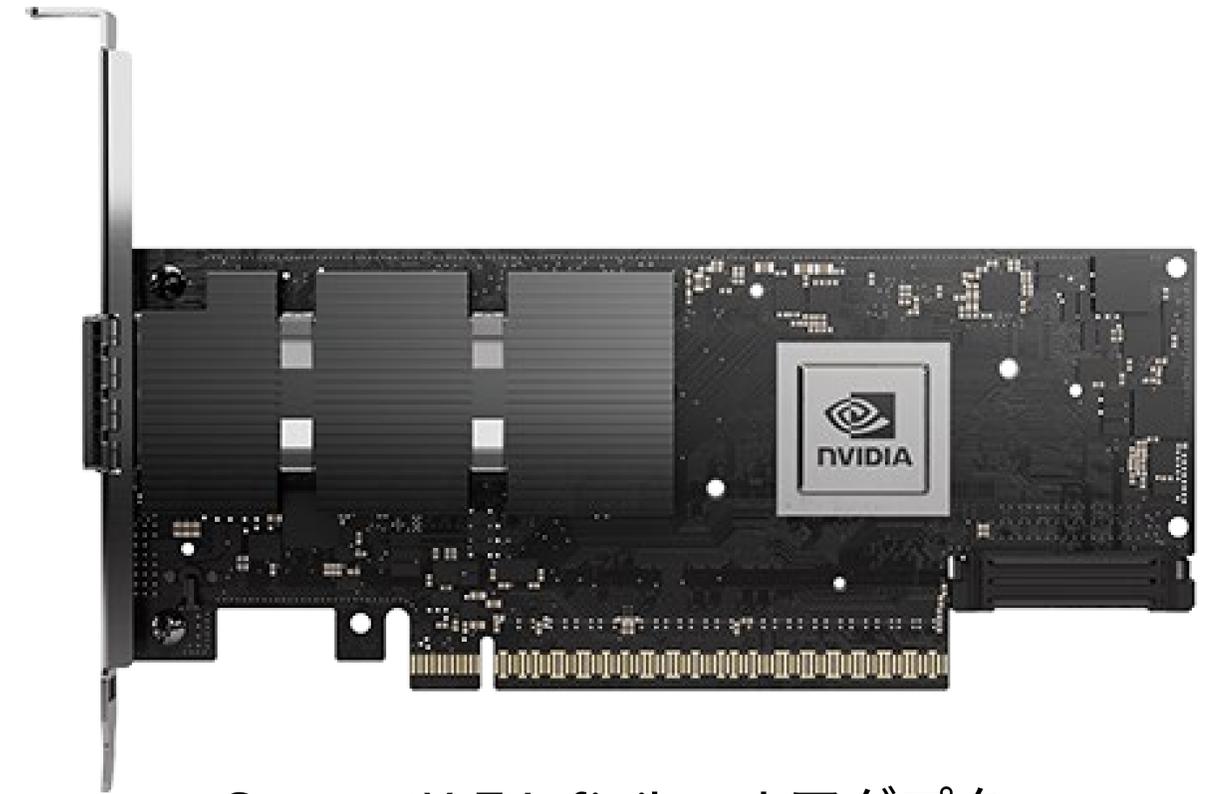
その他、畳み込みのチャンネルを
別デバイスに割り当ててるなどなど.....

GPU-GPU インターコネクト/ノード間通信

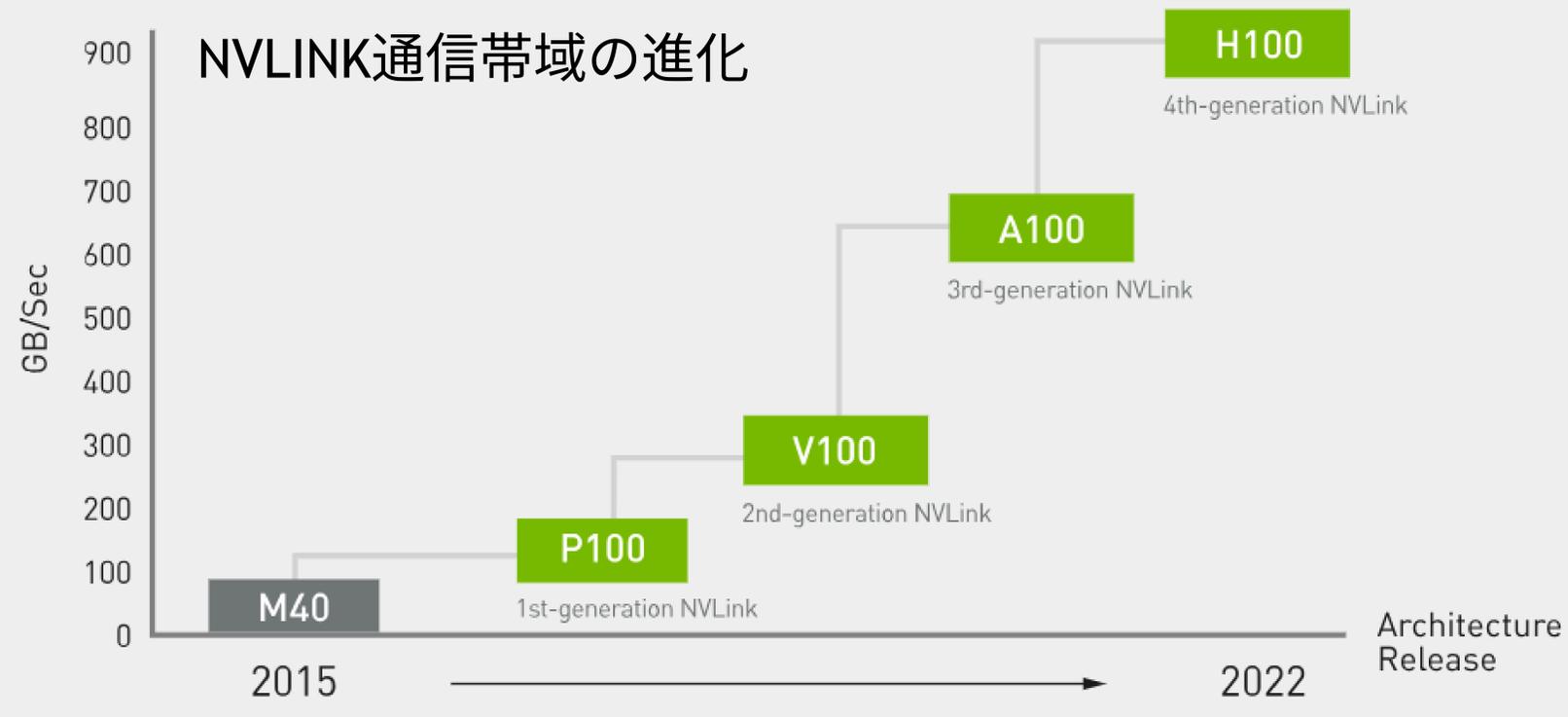
ノード内



ノード間



ConnectX-7 InfiniBand アダプター



NDR 400Gb/s InfiniBand
PCIe Gen5 最大x32 レーン
GPUDirect® RDMA
GPUDirect Storage
In-Network Computing

エンタープライズに向けた生成AIアプリケーションの構築

NVIDIA Nemoによる生成AIモデルの構築、カスタムズ、デプロイ

