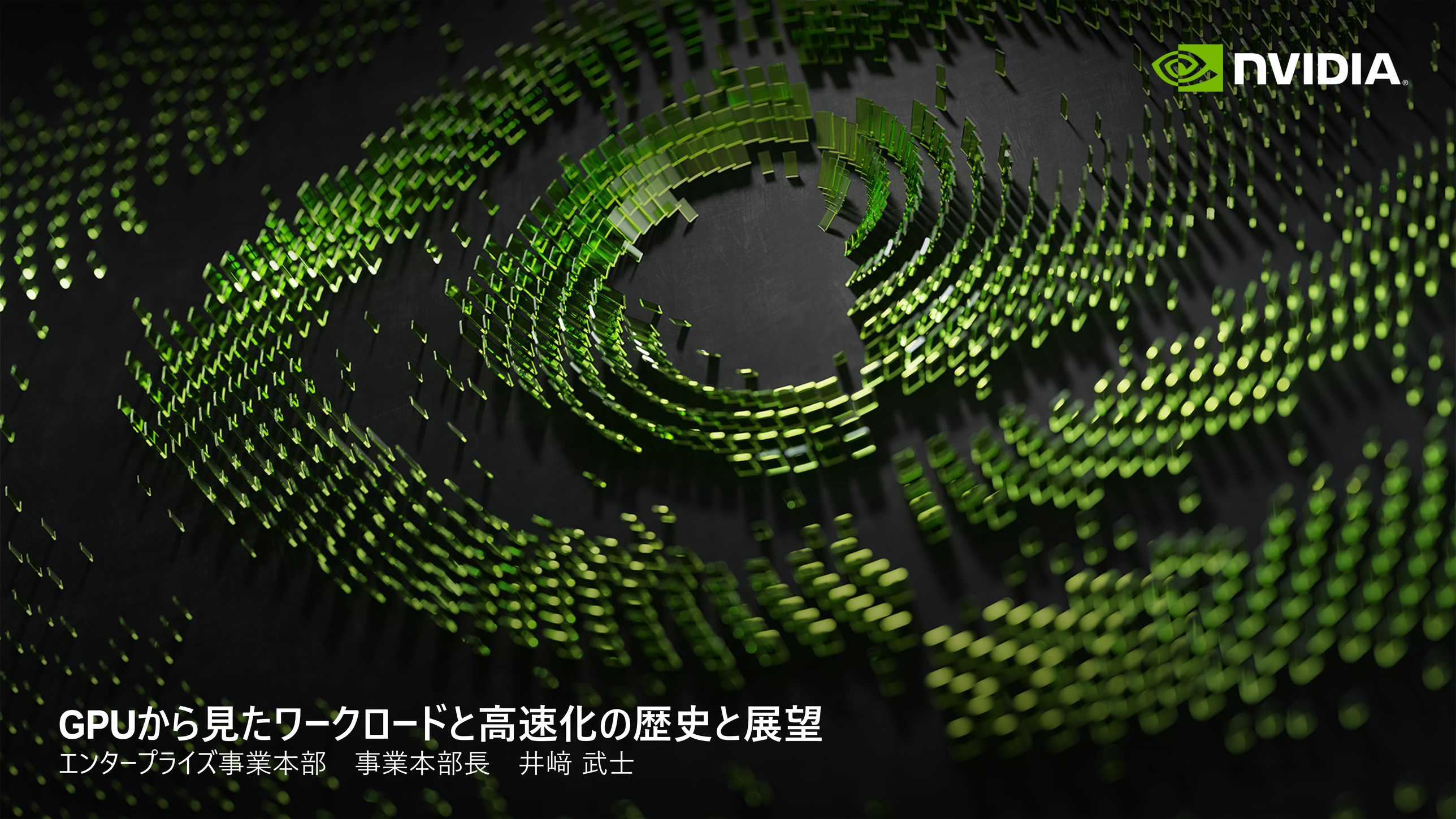
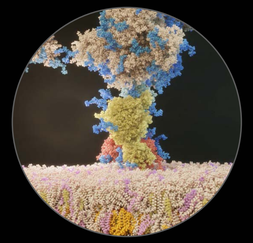


GPUから見たワークロードと高速化の歴史と展望

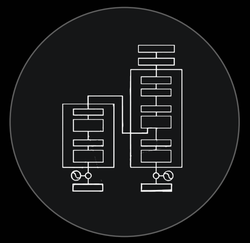
エンタープライズ事業本部　事業本部長　井﨑 武士

アプリケーションの多様化と支えるインフラの柔軟性

大規模シミュレーション



AI学習



リアルタイム処理

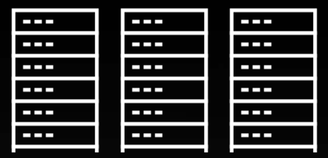
デジタル・ツイン

クラウドストレージ

オンライン購買

オンライン予約

ソーシャル
プラットフォーム



HPC + AI スーパーコンピューター

クラウド

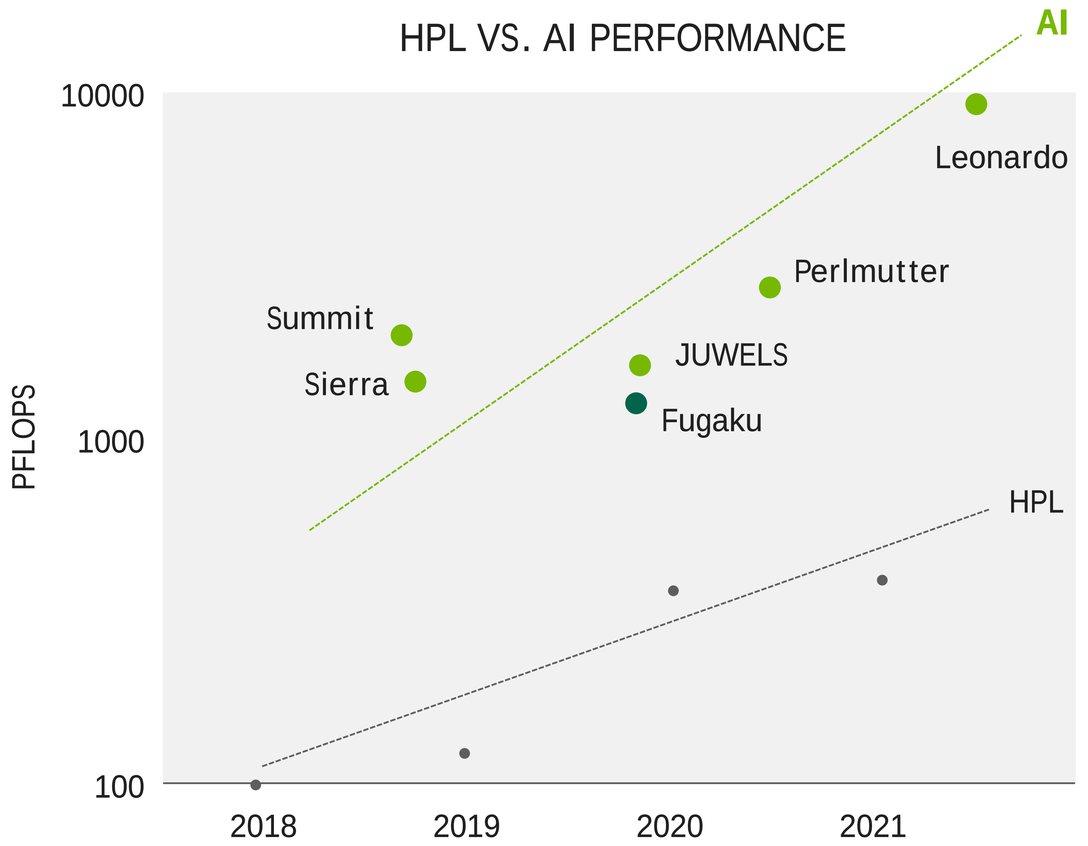
エッジ

ネットワーク

高性能 IO

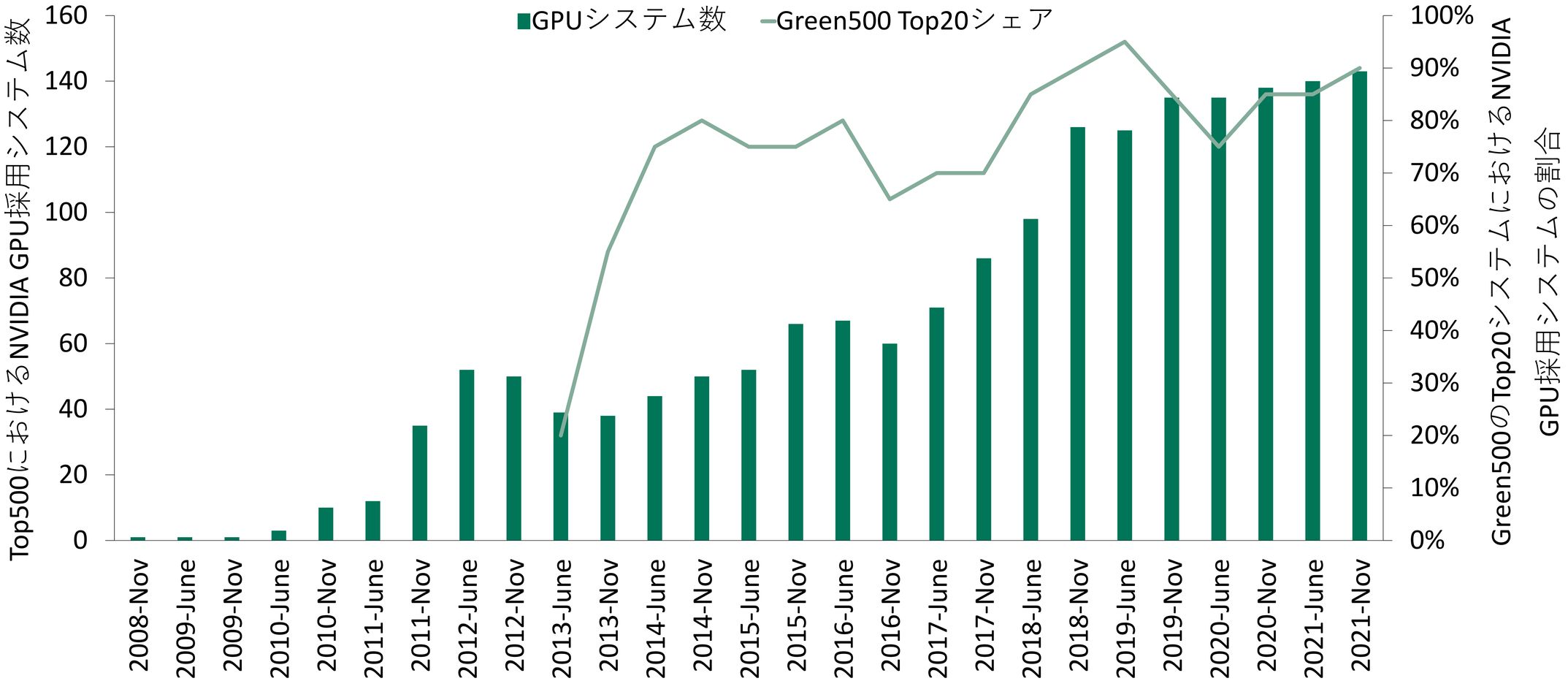
エクサスケールの時代
AI スーパーコンピューティング

AI が現代のHPCの新たな成長ドライバー

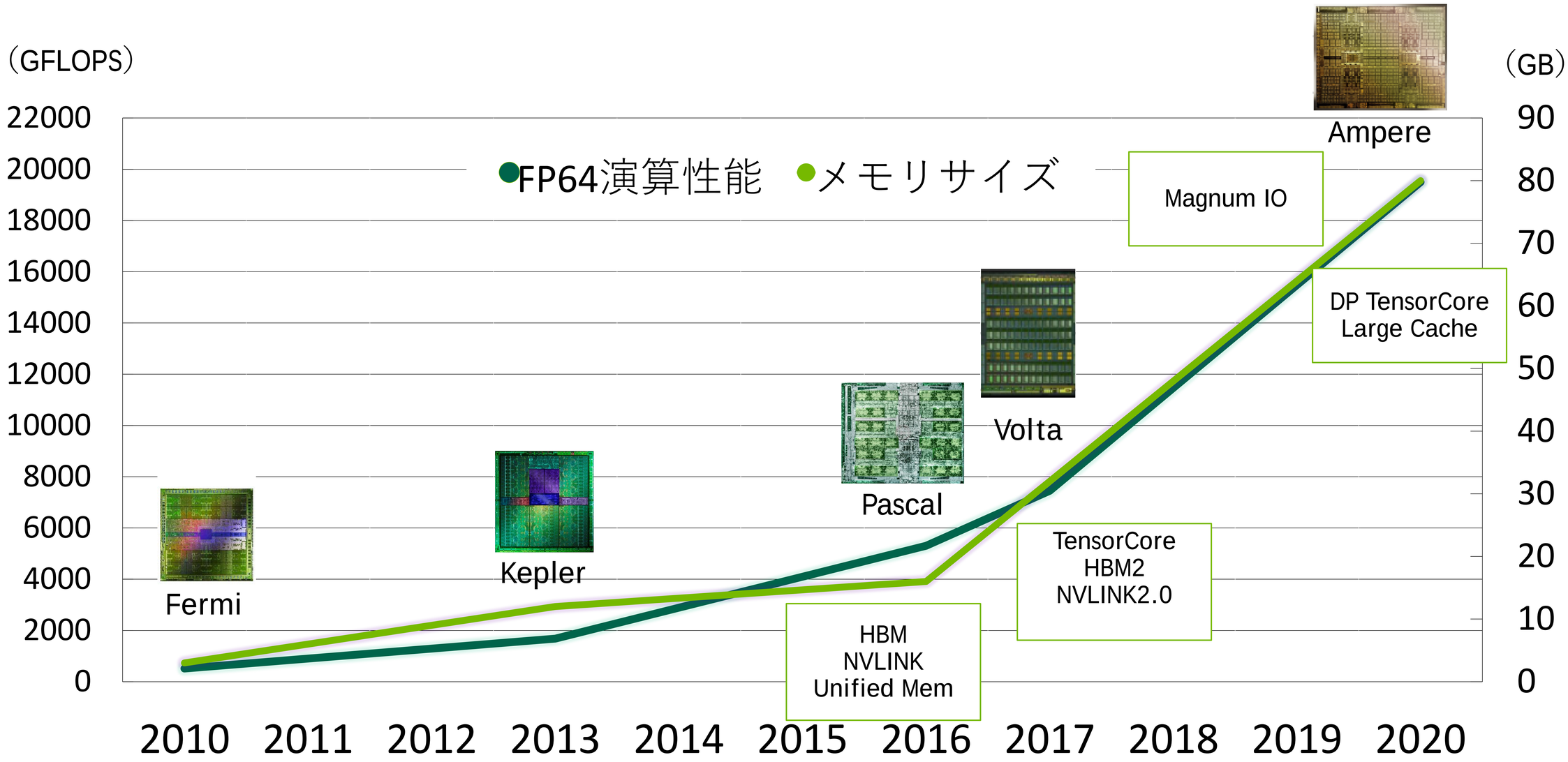


HPL: Based on #1 system in June Top500
AI: Peak system FP16 FLOPS

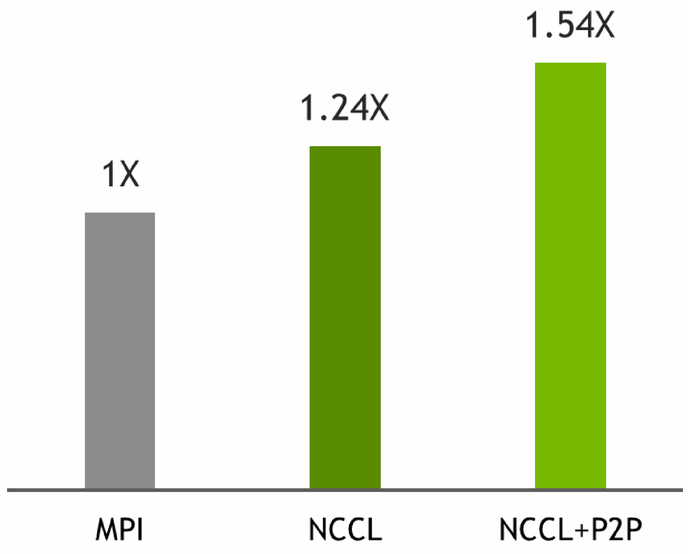
TOP500・GREEN500におけるNVIDIA GPU採用システム数の推移



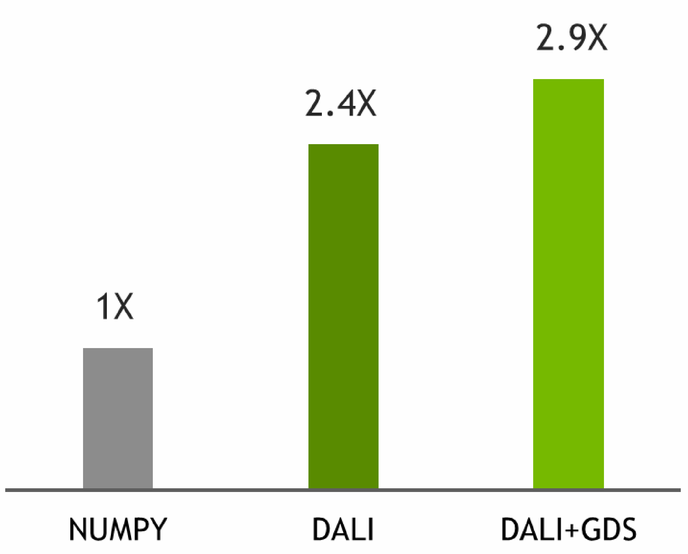
GPUハードウェア性能の推移



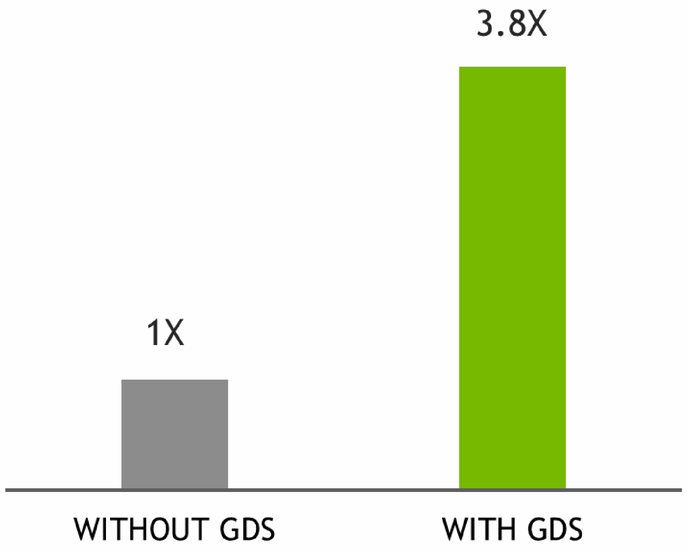
MAGNUM IO によるベネフィット



NCCLE P2Pを用いてリンモノレイヤーの
シミュレーション性能を向上



気象シミュレーションにおける異常気象現象の
セグメンテーション



DGX A100における帯域最大の
リモートファイルリード

NVIDIA データセンタープラットフォーム

ビジネス
アプリケーション

カスタマー
エンゲージメント | 患者診断 | 不正検知 | 品質保証 | 産業の自動化 | プレシジョン
マーケティング | 分子
シミュレーション | ++

NGC

ソフトウェア
ハブ

HELM

*認証済み
コンテナ*

学習済みモデル

SDK

アプリケーション
フレームワーク

スマートシティ *Metropolis* | 対話 AI *Riva* | 自動運転 *Drive* | 推薦システム *Merlin* | ヘルスケア *Clara* | サイバーセキュリティ *Morpheus* | ++ …

開発者
ツールキット

機械学習 & データ分析: RAPIDS, dmlc XGBoost

AI 学習 & 推論: TensorFlow, *TensorRT*, PYTORCH, mxnet

ハイパフォーマンス
コンピューティング: *NVIDIA HPC SDK*

レンダリング &
可視化: *IndeX*, *OptiX*, *CloudXR*, *MDL*

高速化
ライブラリ

コンピューティング: *CUDA-X*

ネットワーク, ストレージ & セキュリティ: *DOCA*, *MAGNUM IO*

オペレーション

VIRTUAL GPU SW

TRITON
INFERENCE
SERVER

FLEET
COMMAND

NVIDIA GPU
Operator

Red Hat

vmware

NVIDIA 認証

検証済みの
ソリューション

サーバー &
クラウド

DGX
HGX

高密度

EGX

メインストリーム & エッジ

クラウド インスタンス

監視

DCGM

Grafana

Prometheus

UFM

ハードウェア
テクノロジー

GPU

NVSwitch

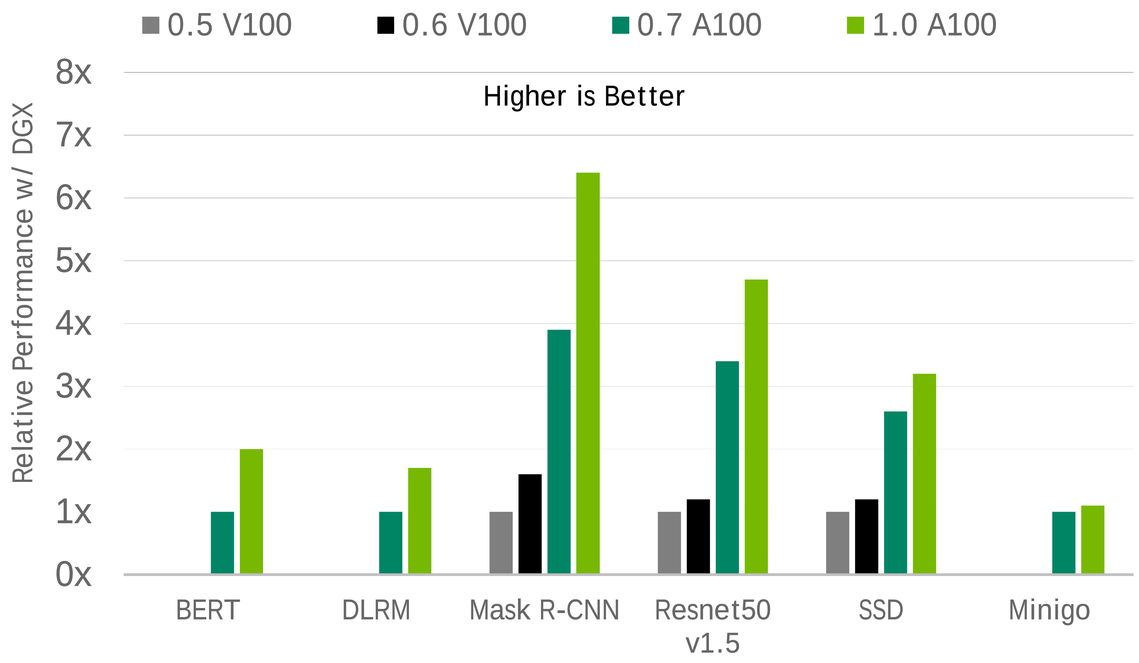
BF DPU

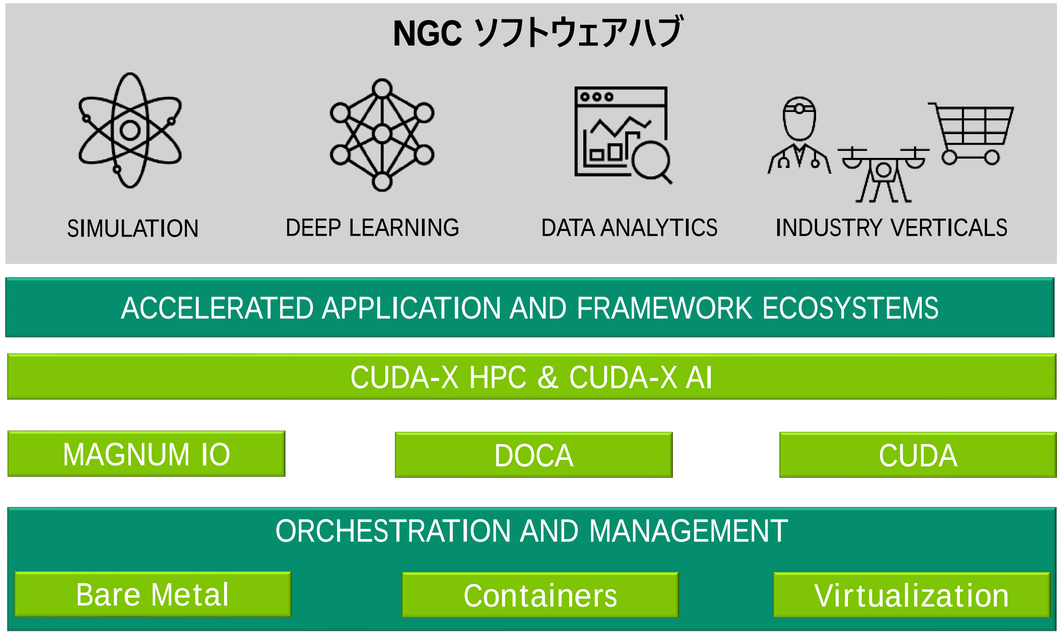
SMART NIC

NVIDIA Switch

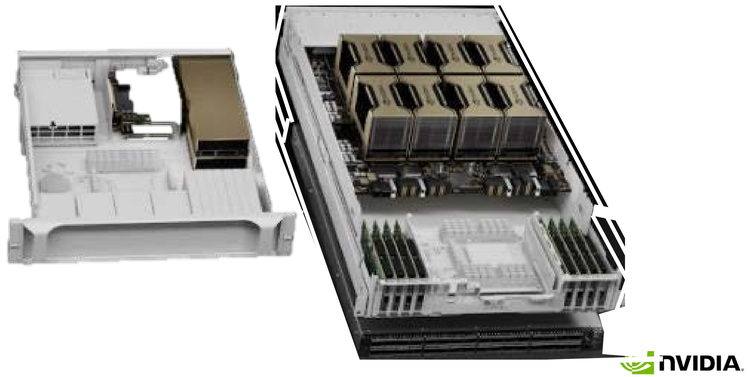
2.5年でGPU当たり最大6.5倍の性能向上

NVIDIA AI はフルスタックの改善で継続的に向上





NVIDIA CERTIFIED



Results normalized for throughput due to higher accuracy requirements on latest round of MLPerf 1.0
MLPerf ID 0.5/0.6/0.7/1.0 comparison:
BERT: 0.7-19/1.0-1060 | DLRM: 0.7-17/1.0-1058 | Mask R-CNN: 0.5-13/0.6-9/0.7-19/1.0-1060 | ResNet50v1.5: 0.5-12/0.6-8/0.7-18/1.0-1059 | SSD: 0.5-13/0.6-9/0.7-19/1.0-1059 | MiniGo: 0.7-20/1.0-1059
MLPerf name and logo are trademarks. See www.mlperf.org for more information.

NVIDIA プラットフォームの多様なプログラミング手法

CPU, GPU, Network

標準言語を使った高速化

ISO C++, ISO Fortran

```
std::transform(par, x, x+n, y, y,
    [=](float x, float y){ return y +
a*x; }
);
```

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

```
import cunumeric as np
…
def saxpy(a, x, y):
    y[:] += a*x
```

インクリメンタル ポータブル 最適化

OpenACC, OpenMP

```
#pragma acc data copy(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}

#pragma omp target data map(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}
```

プラットフォーム特化

CUDA

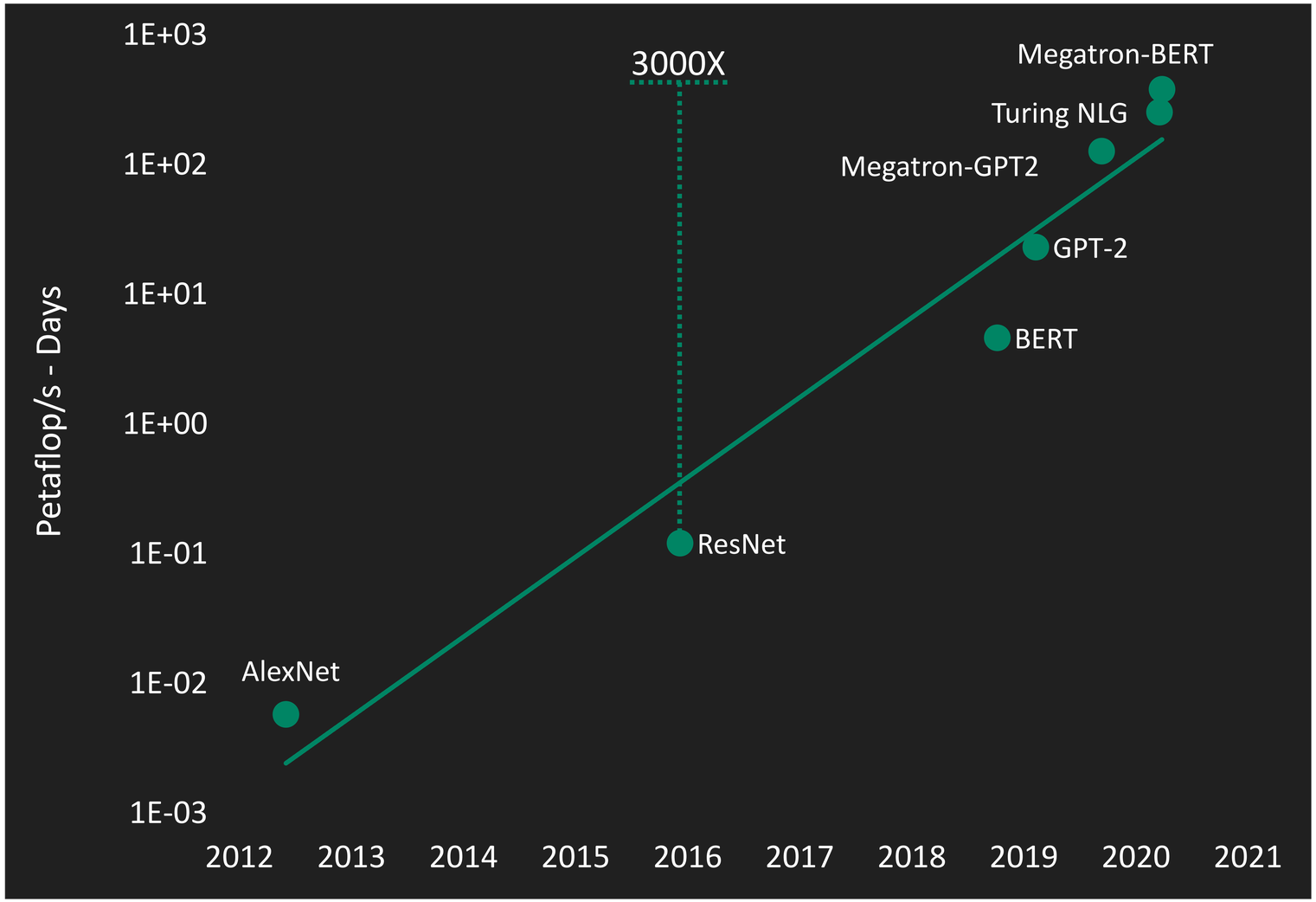
```
__global__
void saxpy(int n, float a,
           float *x, float *y) {
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] += a*x[i];
}

int main(void) {
  ...
  cudaMemcpy(d_x, x, ...);
  cudaMemcpy(d_y, y, ...);

  saxpy<<<(N+255)/256,256>>>(...);

  cudaMemcpy(y, d_y, ...);
```

高速化ライブラリ

Core	Math	Communication	Data Analytics	AI	Quantum

学習に要求される計算性能の飛躍的拡大

指数関数的な計算要求

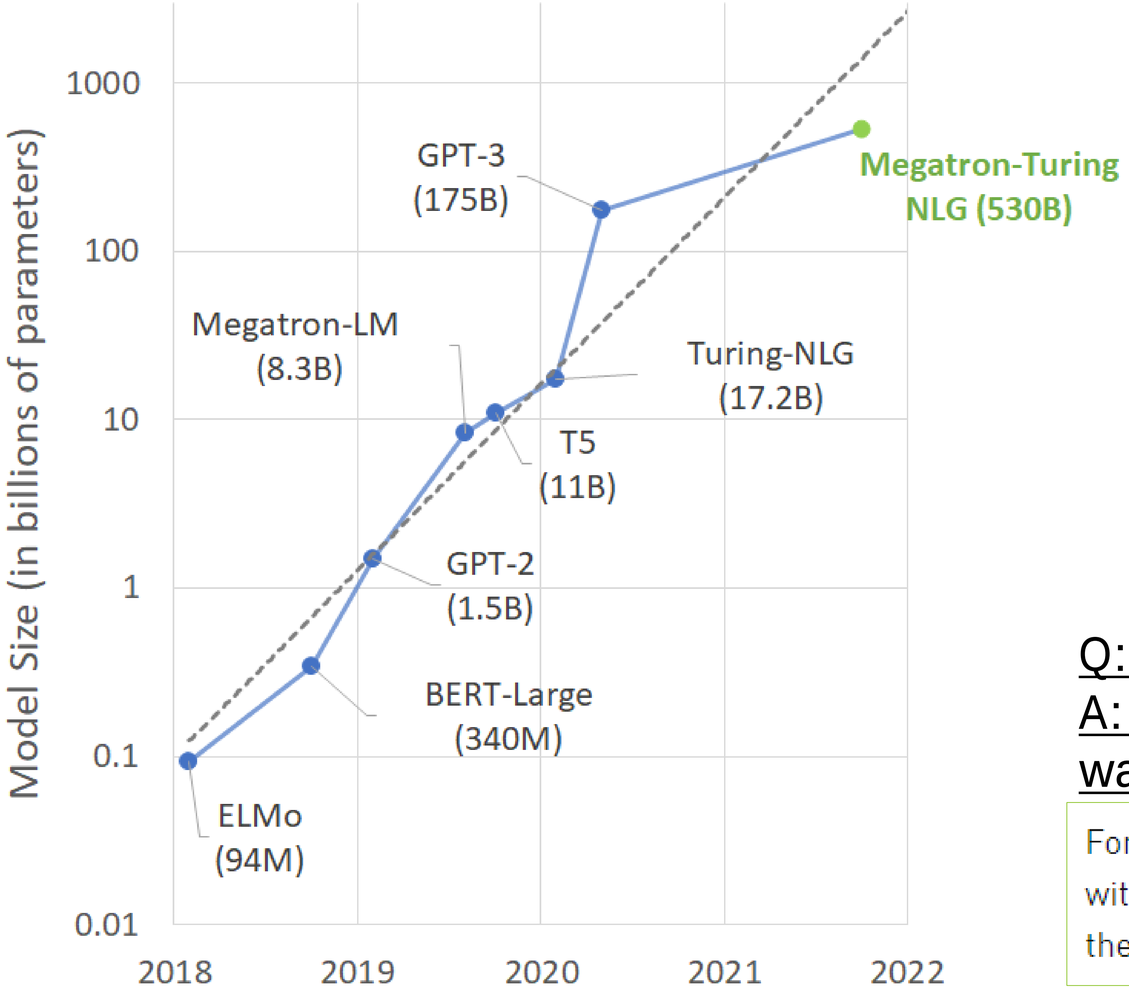


Voltaの時代から大規模モデルの学習に必要な計算性能は3000倍になっている

Source: OpenAI, NVIDIA

深層分散学習の状況

モデルは引き続き巨大に



- 2021/10/11時点の最新モデルは 530B パラメータにまで拡大
  - 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

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＆クラスタ全体でデータ並列

> 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.

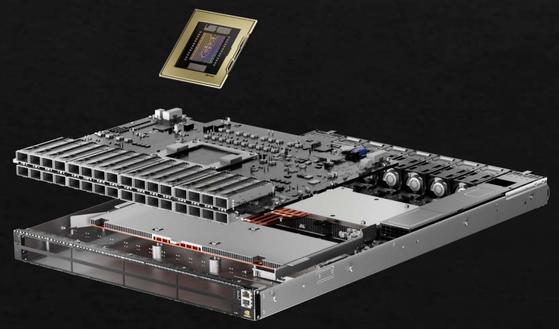
https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

NVIDIA

NVIDIA QUANTUM-2

400G NDR InfiniBand

QUANTUM-2 SWITCH



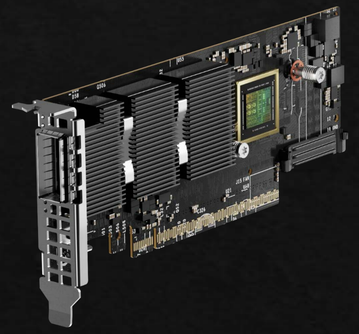
400 Gbps 64 ポート / 200 Gbps 128 ポート

32倍超の AI 高速化エンジン

3倍のスイッチングスループット

サンプル出荷中

CONNECTX-7 INFINIBAND

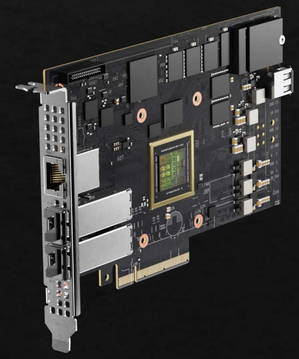


4倍のネットワーク内コンピューティング性能

2倍の GPUDirect スループット

2022年1月サンプル予定

BLUEFIELD-3 INFINIBAND



16 Arm 64 ビットコア | 400 Gbps 暗号化アクセラレータ

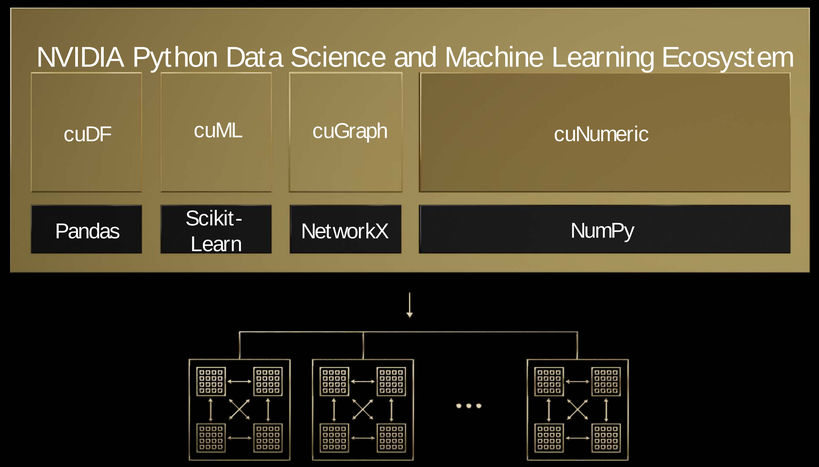
4倍のネットワーク内コンピューティング性能

2倍のGPUDirect スループット

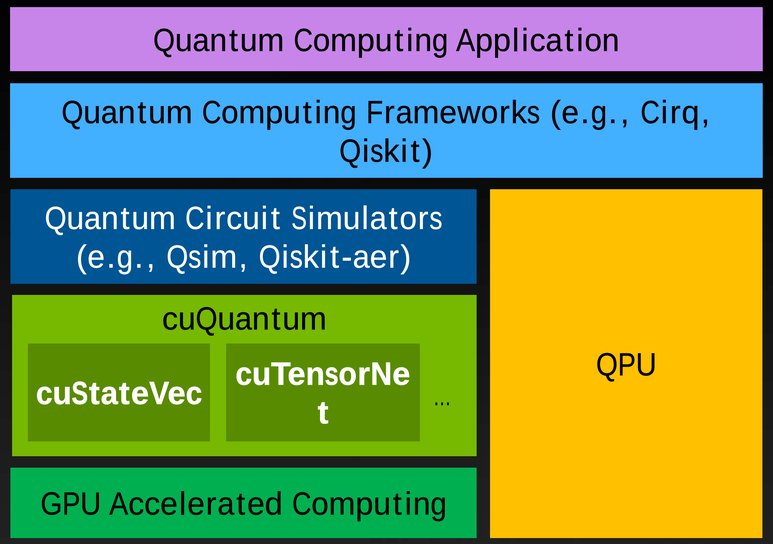
2022年5月サンプル予定

広がるコンピューティング需要

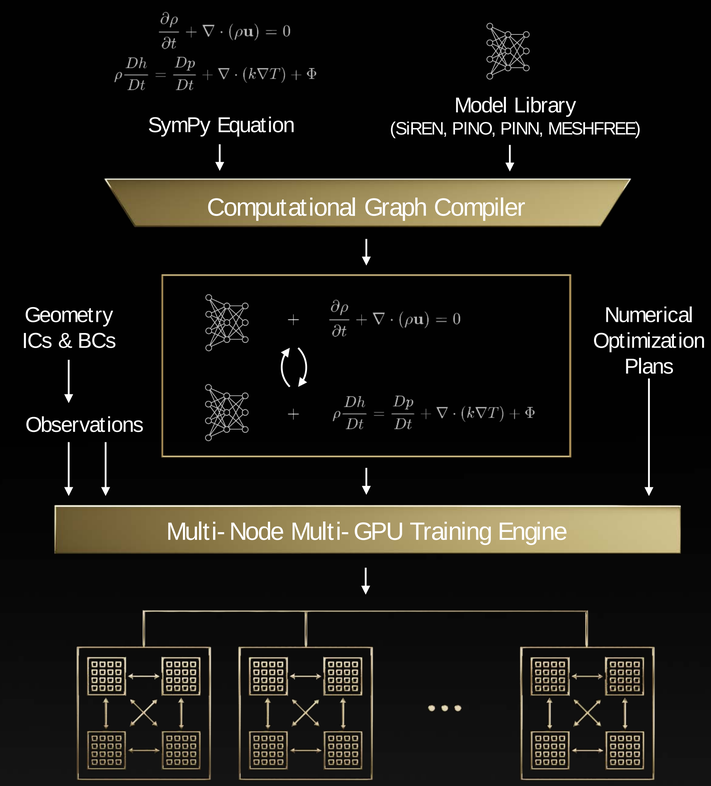
- cuNumeric



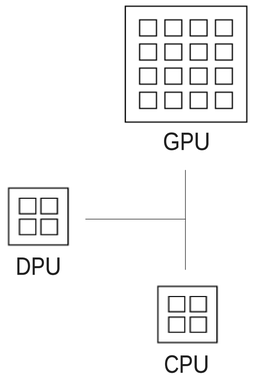
- cuQuantum



- Modulus

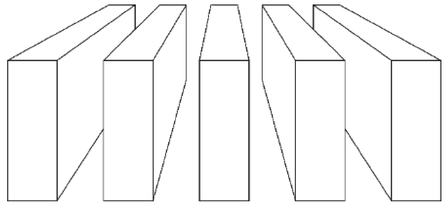


コンピューティング高速化の方針

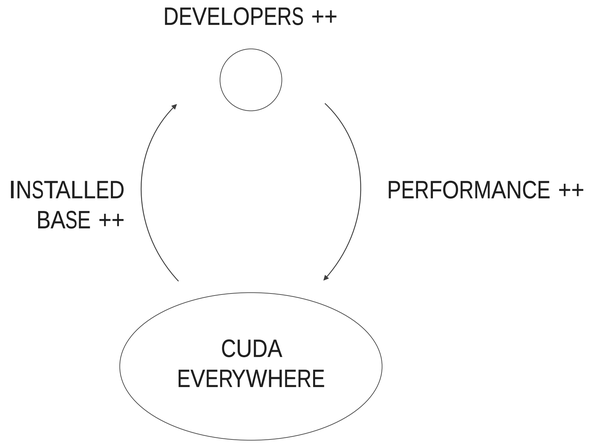


X-要素 高速化

フルスタック



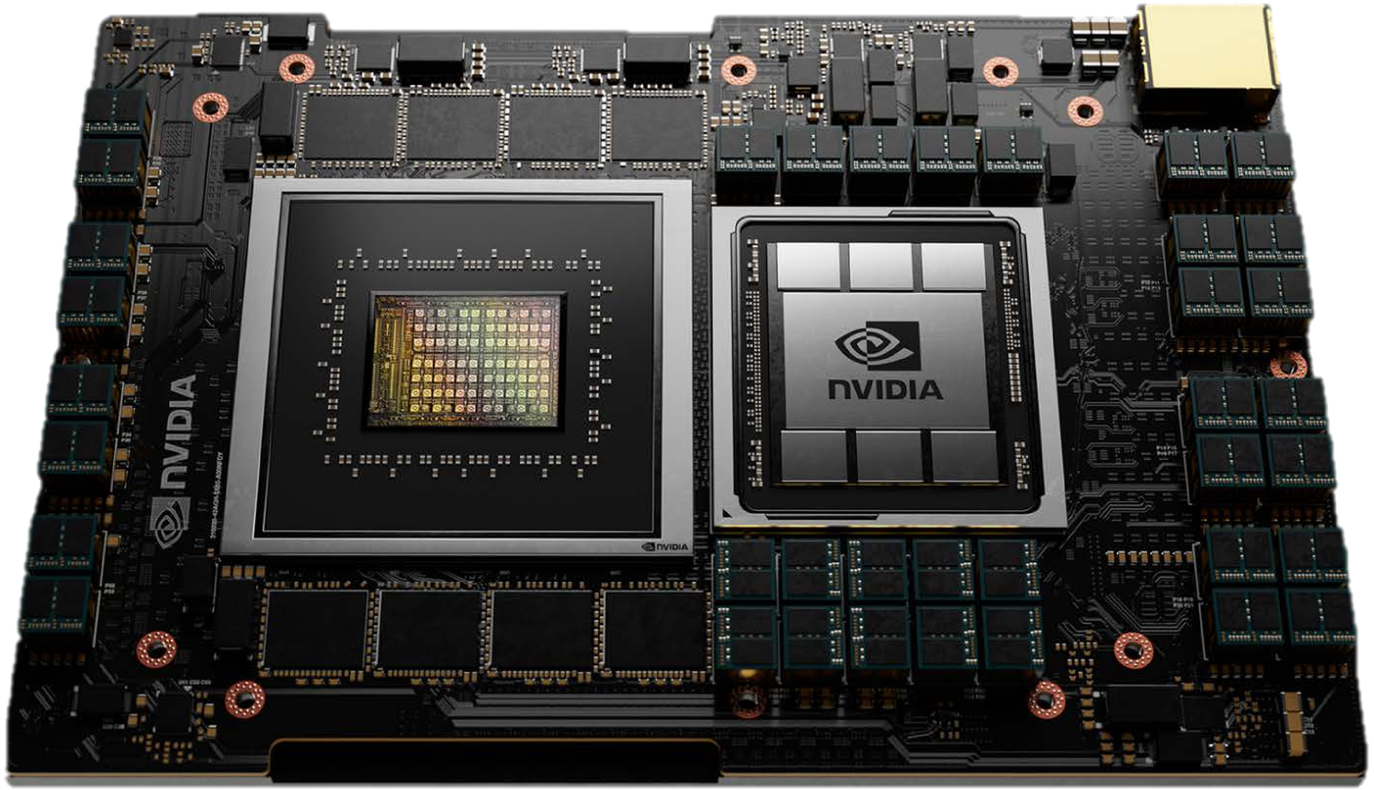
データセンタースケール



ワン・アーキテクチャー

NVIDIA GRACE

巨大スケールのAIおよびHPCアプリケーションに向け設計したブレークスルーCPU



最高速の相互接続

>900 GB/s キャッシュコヒーレント NVLink
CPU-GPU間通信 (14倍)
>600GB/s CPU-CPU間 (2倍)

最高のメモリバンド幅

500GB/s LPDDR5x ECC付
2倍以上のバンド幅
10倍のエネルギー効率

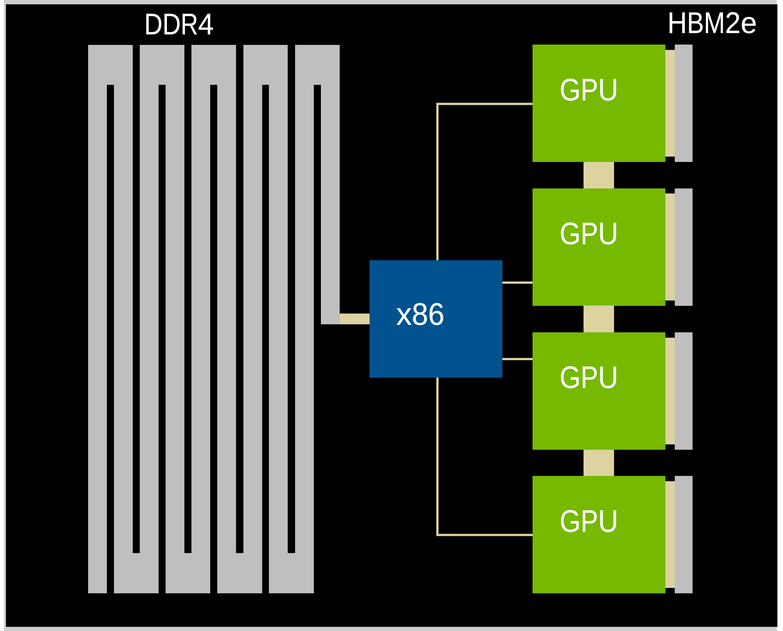
次世代 ARM NEOVERSE コア

>300 SPECrate2017_int_base est.
2023年出荷予定

テラバイトレベルに高速化されたコンピューティング

新たなワークロードに向けたアーキテクチャの革新

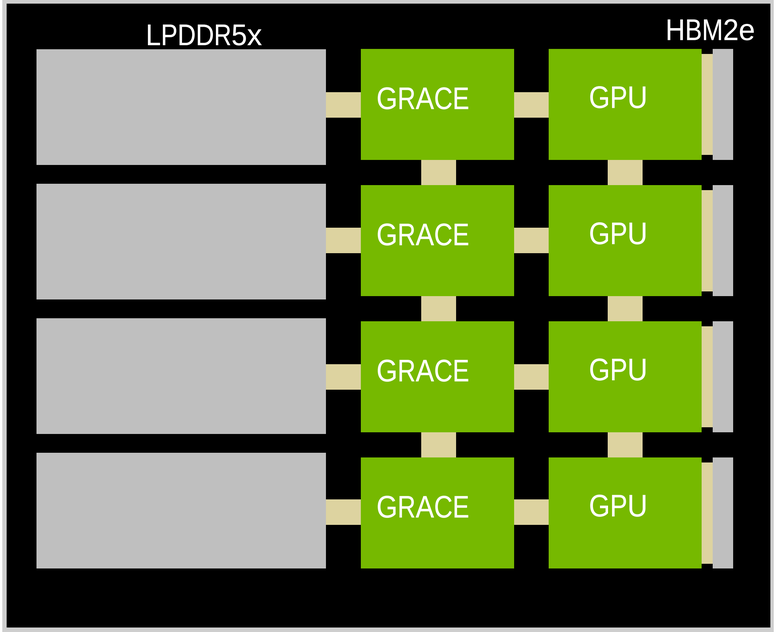
現行 x86 アーキテクチャ



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
メモリ-GPU間	**64**	**GB/sec**

2TBの転送に30秒

統合された CPU-GPU アーキテクチャ



GPU	8,000	GB/sec
CPU	500	GB/sec
NVLink	500	GB/sec
メモリ-GPU間	**2000**	**GB/sec**

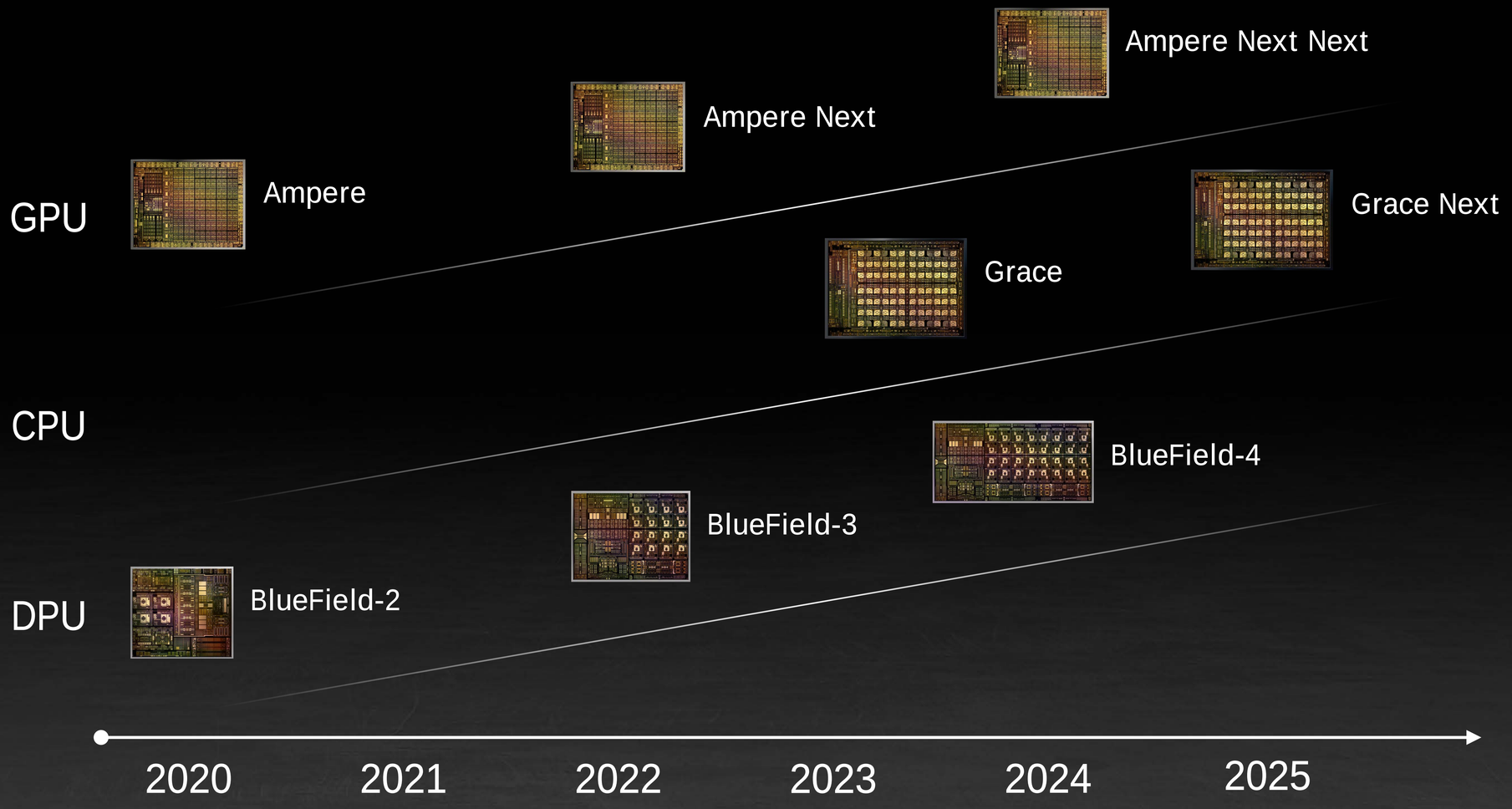
2TBの転送に1秒

3 DAYS FROM 1 MONTH
1Tモデルの学習ファインチューン

0.5Tモデルでのリアルタイム推論
シングルノードでのインタラクティブなNLP推論

Bandwidth claims rounded to nearest hundred for illustration.
Performance results based on projections on these configurations Grace : 8xGrace and 8xA100 with 4th Gen NVIDIA NVLink Connection between CPU and GPU and x86: DGX A100.
Training: 1 Month of training is Fine-Tuning a 1T parameter model on a large custom data set on 64xGrace+64xA100 compared to 8xDGXA100 (16xX86+64xA100)
Inference: 530B Parameter model on 8xGrace+8xA100 compared to DGXA100.

CPU/GPU/DPU のロードマップ




NVIDIA