サステナブルなコンピューティングのための取り組み

Dec 5,2024 – Hiroshi Aiko, Sr. Marketing Manager, NVIDIA.

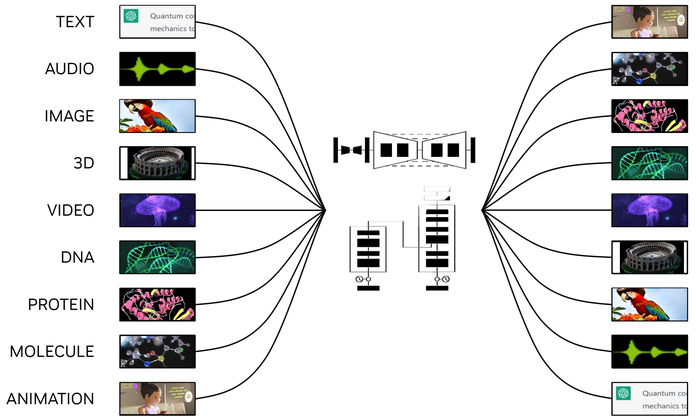
生成 AI の次の時代へ

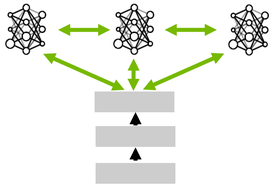
Image Classification

Transformer

Large Language Models (Transformer)

Large Language Models (Transformer)



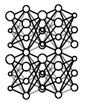


Google
Gemini

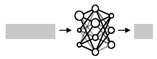
Meta
NLLB

MISTRAL AI_
Mixtral

Realtime
<50ms latency



Parameters
>10T

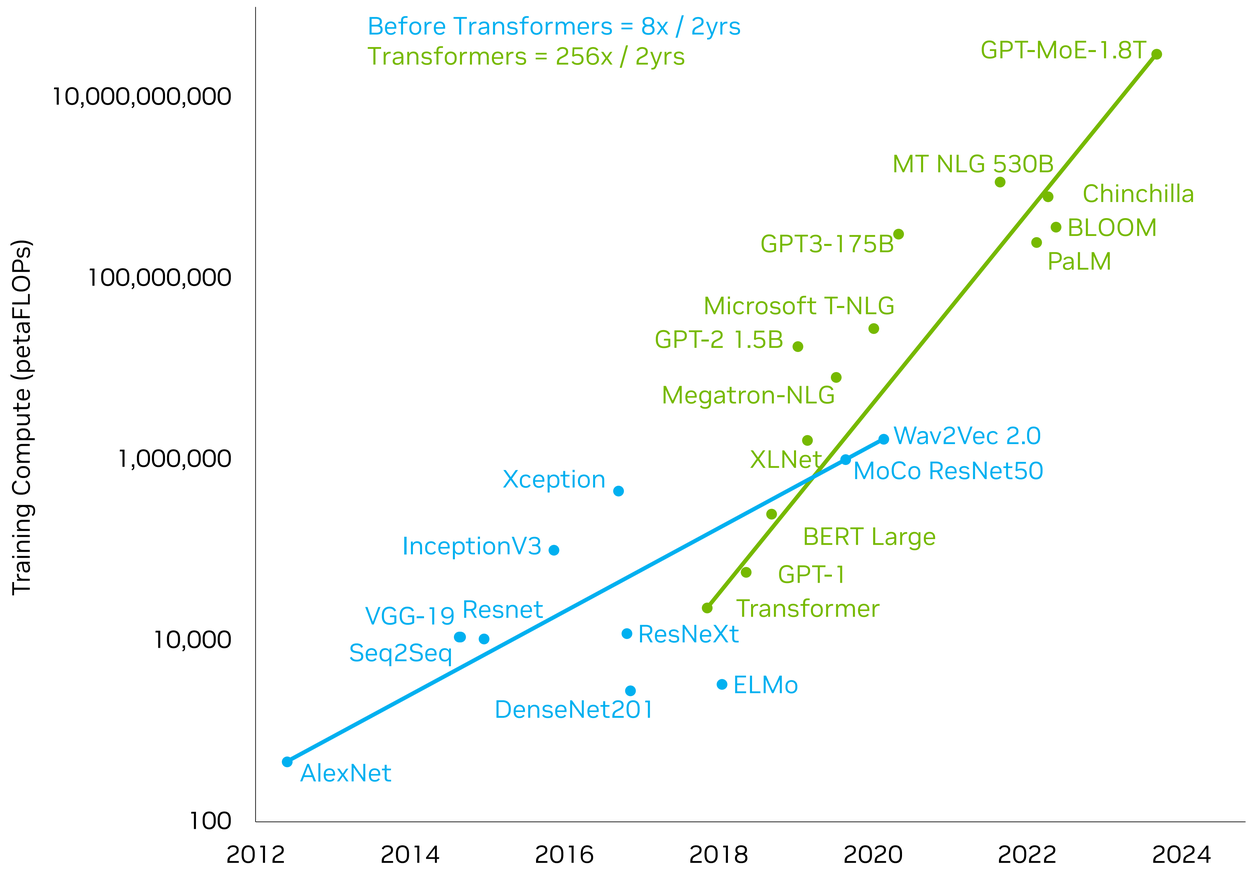


Sequence Length
>32K input

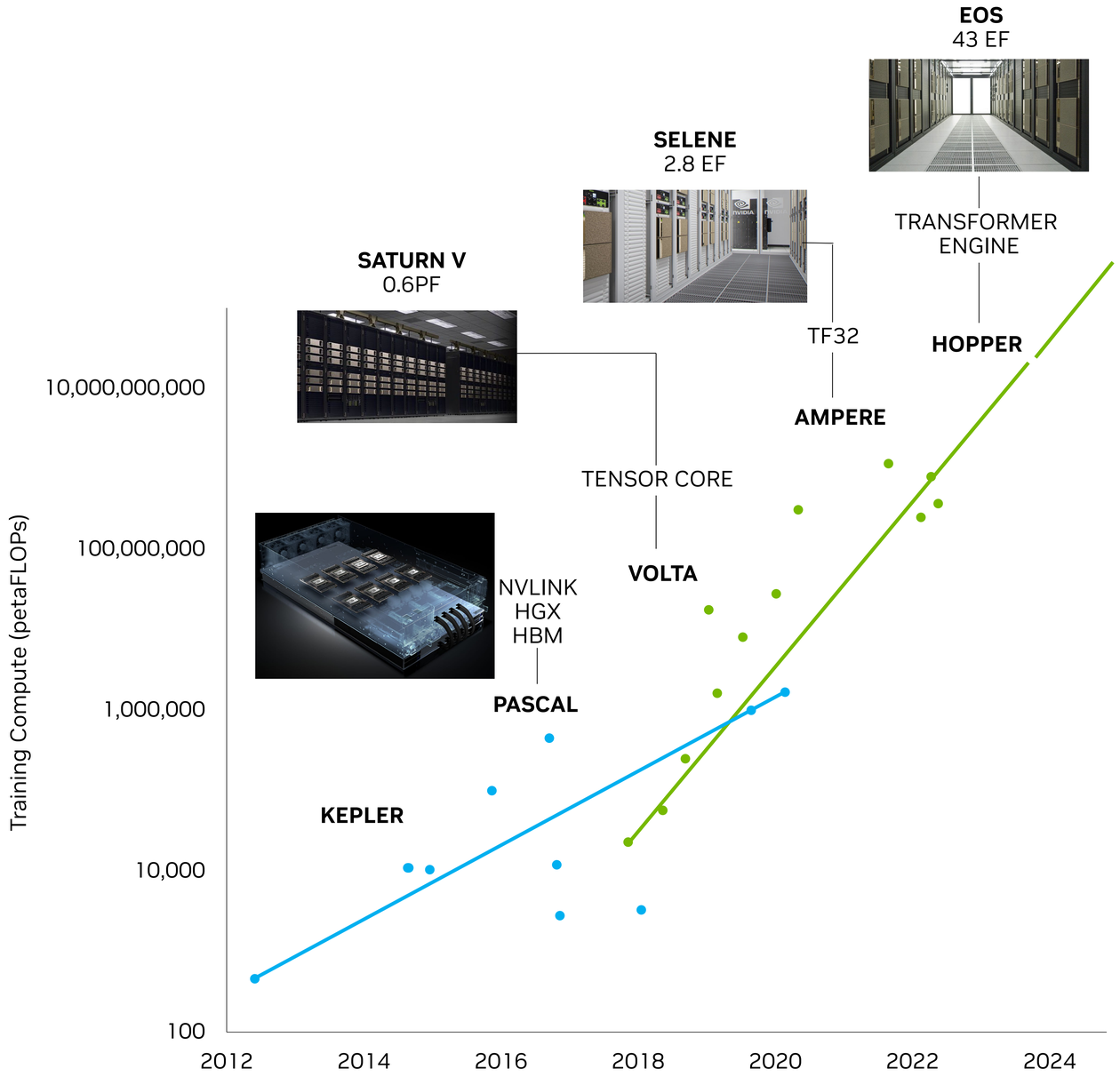
Labeled Datasets | Unlabeled Datasets | Generative AI | Multimodal Generative AI | Mixture of Experts (MoE) | Production GenAI Inference

NVIDIA

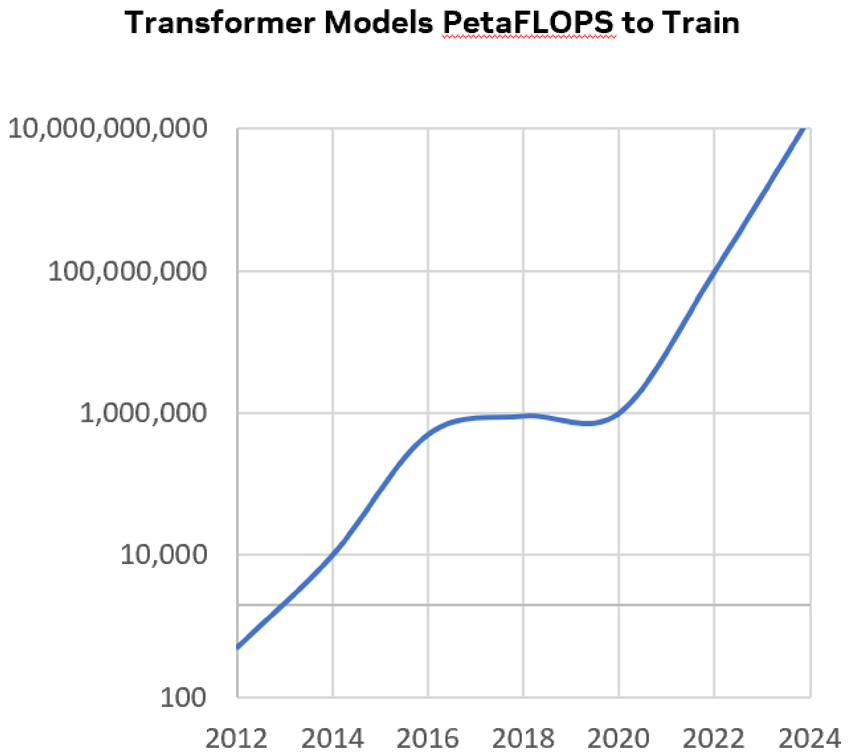
爆発的に増加する AI の計算要件



NVIDIA は AI 計算要件の爆発的な伸びを可能に



データセンターにおけるエネルギー使用量の爆発的増加



より多くのコンピューティングを要求する
モデルサイズ

Model sizes: Github AI Memory Wall, Arxiv papers | Data Center Electricity Usage: IEA report

AI と HPC において GPU は CPU より大幅なエネルギー効率向上を実現

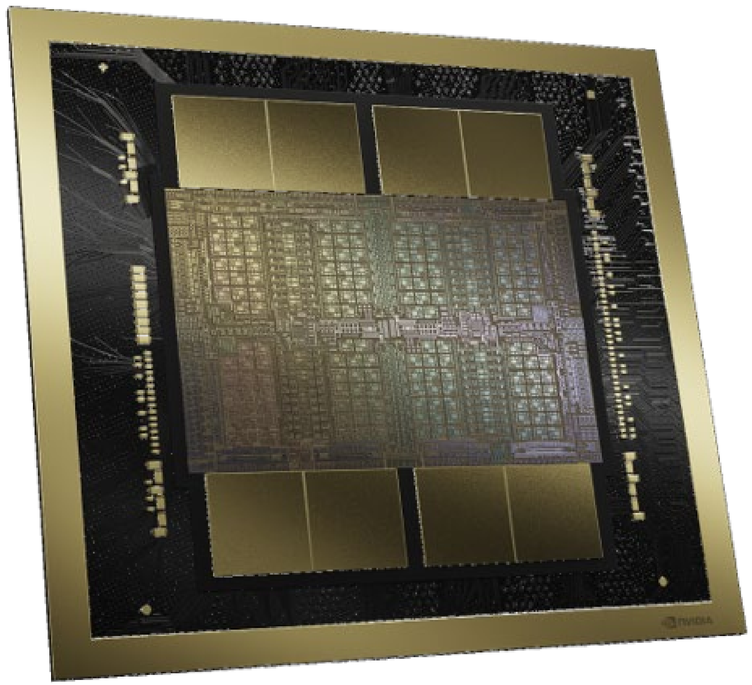
現在出荷中の NVIDIA GPU でデータセンターは今日からエネルギーを節約

NVIDIA Blackwell GPUは、AI や HPC のワークロードにおいて、CPU の20倍のエネルギー効率を発揮します。

AI、HPC、データ分析のワークロードにおいて、CPU のみサーバーが、すべて GPU に移行した場合に節約できる**年間エネルギー量**

計算集約型用途

AI　HPC　データ分析



GB200
20X
AI & HPC

40
TWh

25M
Metric Tons
CO_2

5M
家庭の電気使用量

に相当

NVIDIA GB200 GPU の優れたエネルギー効率

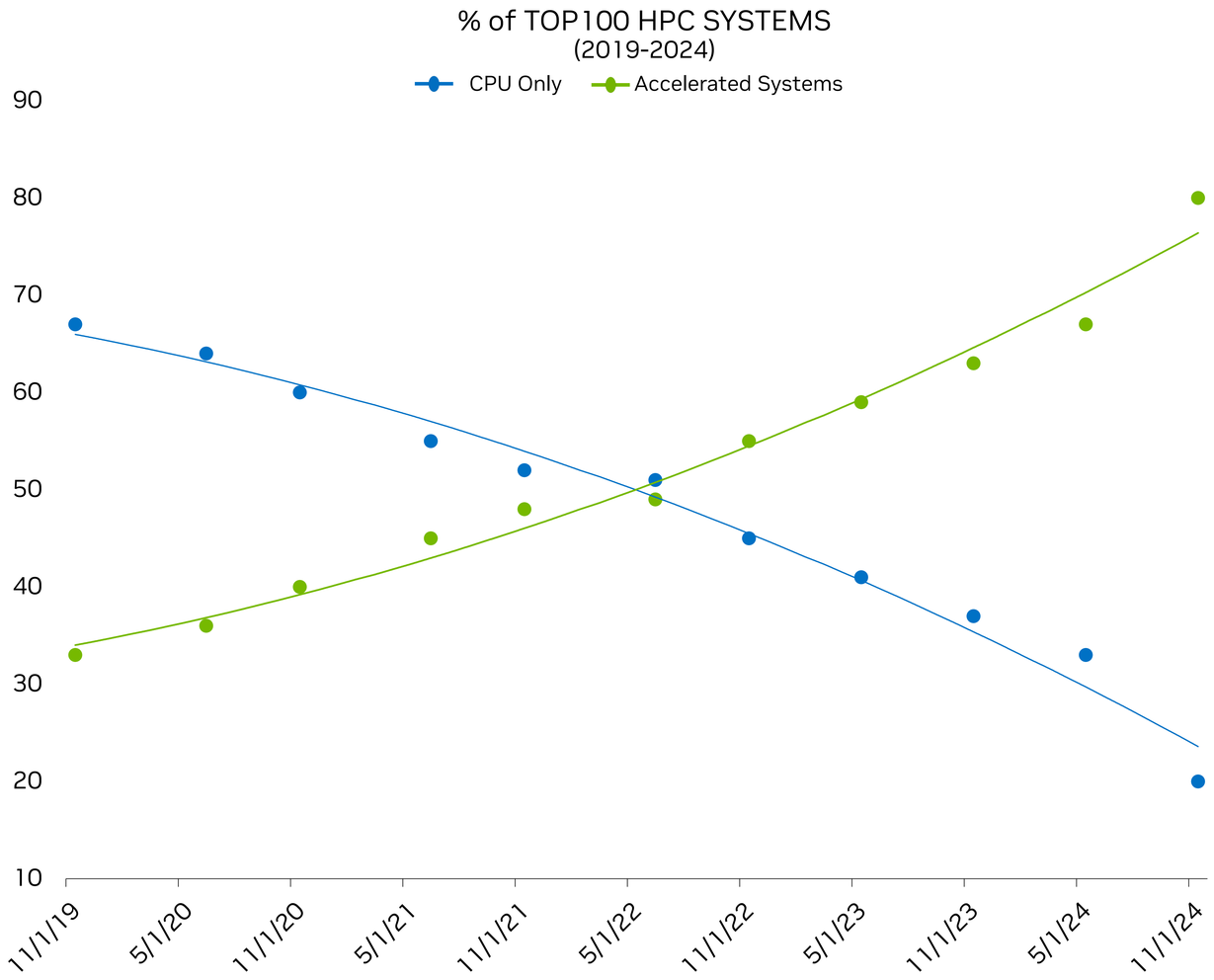
AI、HPC、データ分析ワークロードの、CPU から GPU への移行による大幅なコスト削減

Delivered Application Performance - Air Cooled GPUs | GB200 vs Sapphire Rapids on GPT-J large language model (LLM) inference | Carbon and Home equivalence calculated using: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results

NVIDIA

アクセラレーテッドコンピューティングがスーパーコンピューティングを変革

GPUがスーパーコンピューティング環境の主流に



Information available from top500.org

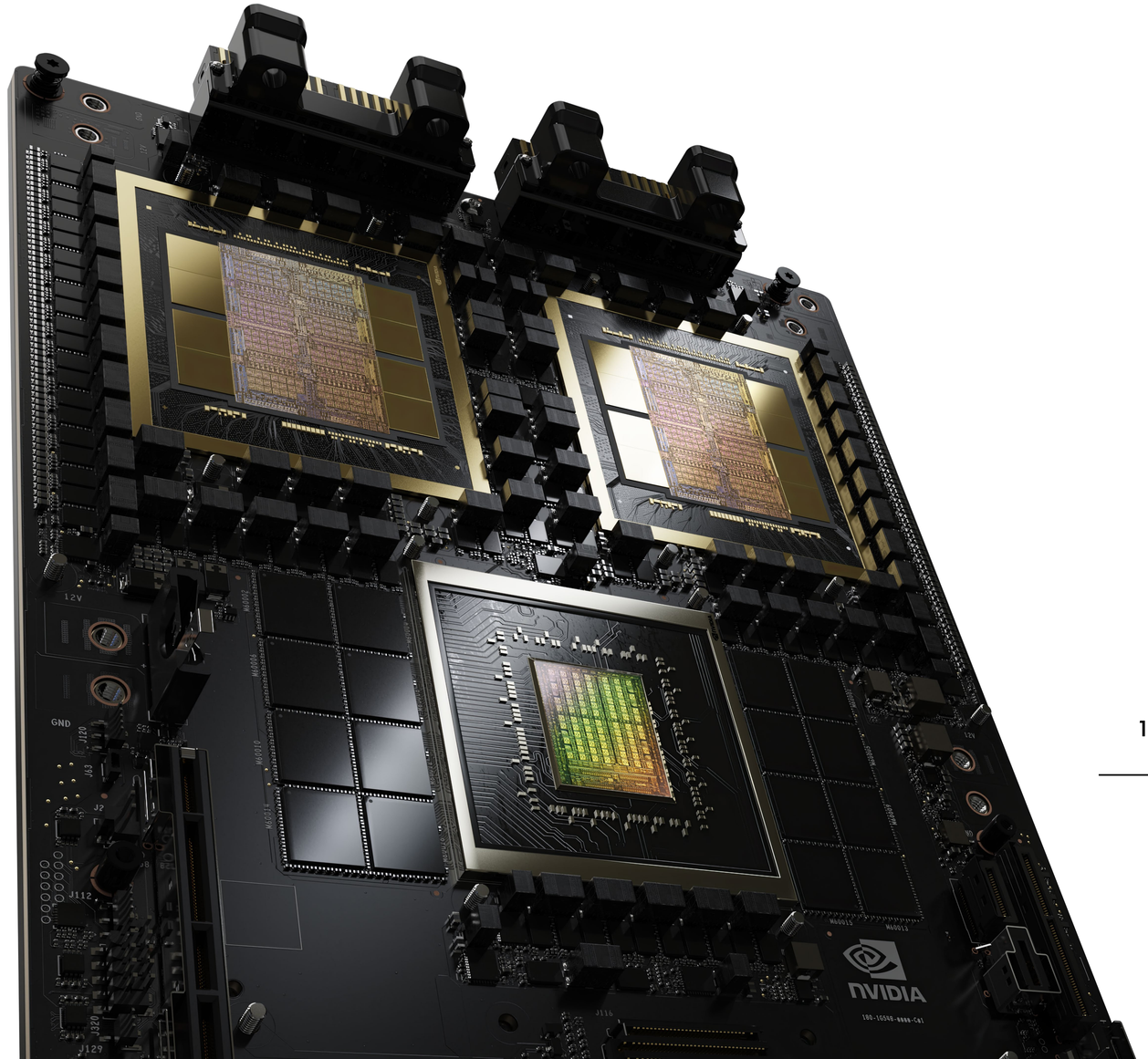
Green500 トップ10でのご採用

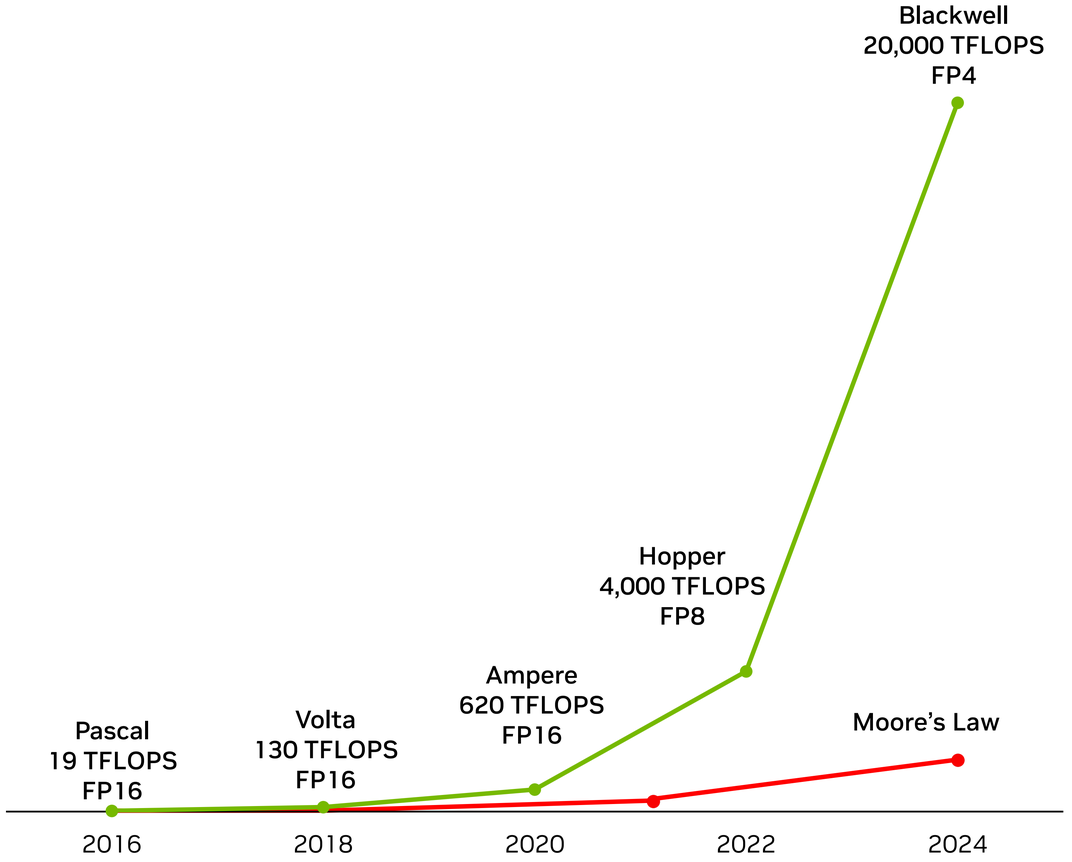
Green 500 上位 10 スパコン中 8 スパコンでのご採用 (2024年11月)

#	Name	Computer
1	JEDI	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, **NVIDIA GH200 Superchip**, Quad-Rail NVIDIA InfiniBand NDR200
2	ROMEO-2025	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, **NVIDIA GH200 Superchip**, Quad-Rail NVIDIA InfiniBand NDR200, Red Hat Enterprise Linux
3	Adastra 2	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, RHEL
4	Isambard-AI phase 1	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, **NVIDIA GH200 Superchip**, Slingshot-11
5	Capella	Lenovo ThinkSystem SD665-N V3, AMD EPYC 9334 32C 2.7GHz, **Nvidia H100 SXM5 94Gb**, Infiniband NDR200, AlmaLinux 9.4
6	JETI - JUPITER Exascale Transition Instrument	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, **NVIDIA GH200 Superchip**, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Linux and Modular Operating System
7	Helios GPU	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, **NVIDIA GH200 Superchip**, Slingshot-11
8	Henri	ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, **NVIDIA H100 80GB PCIe**, Infiniband HDR
9	HoreKa-Teal	ThinkSystem SD665-N V3, AMD EPYC 9354 32C 3.25GHz, **Nvidia H100 94Gb SXM5**, Infiniband NDR200
10	rzAdams	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS

NVIDIA Blackwell が飛躍的な演算能力を実現

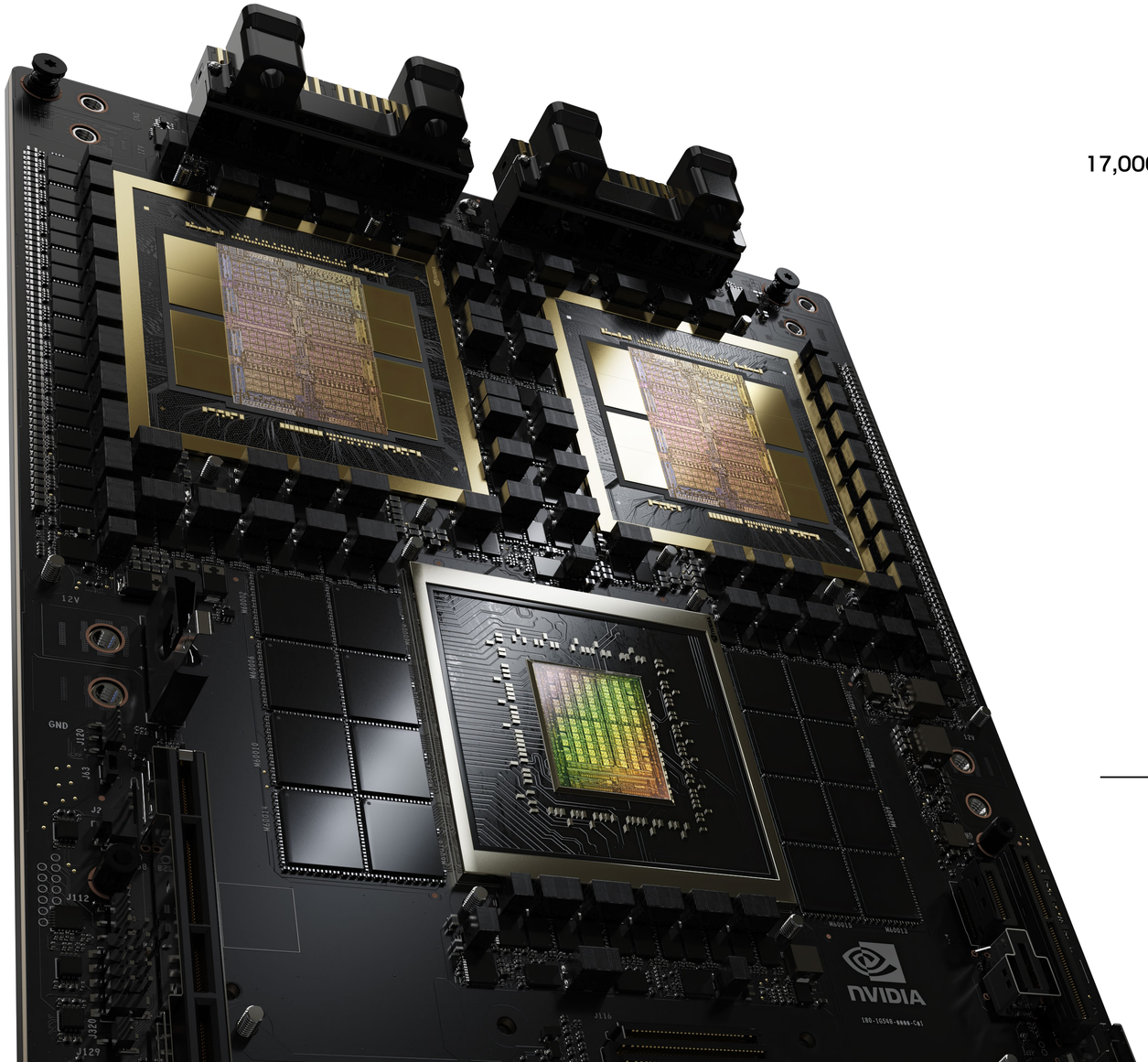
8年で1000倍の AI 演算能力

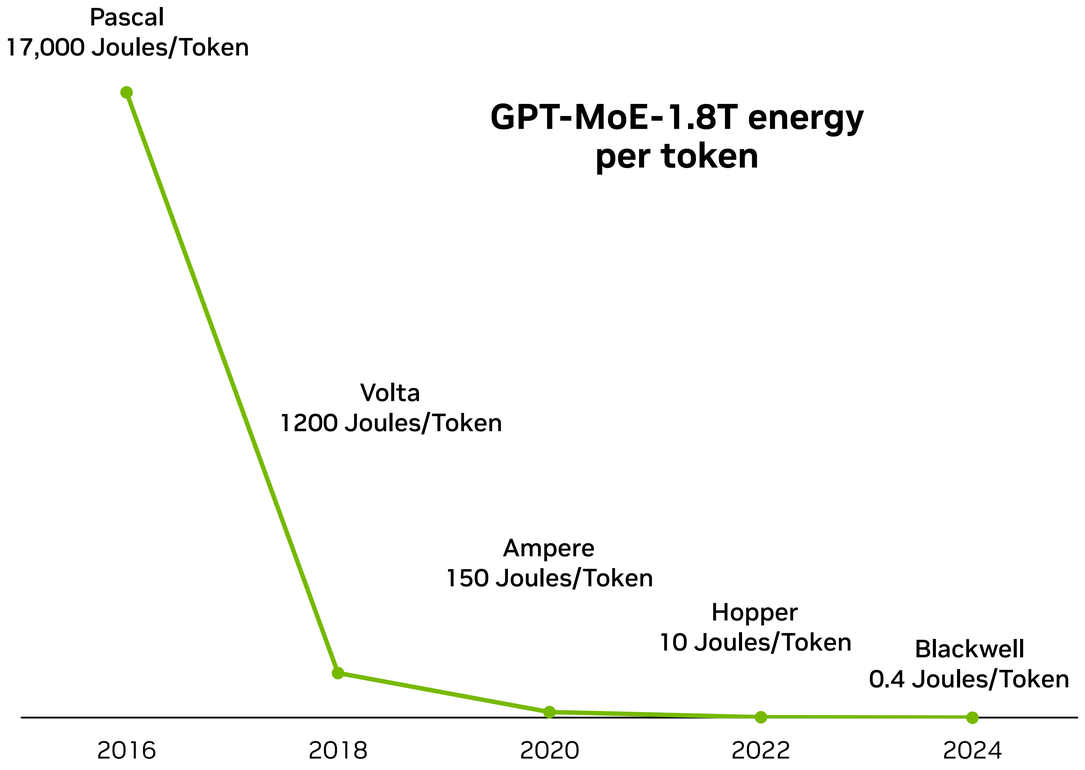




LLM推論のエネルギー効率は向上し続けている

トークンに必要なエネルギーは8年間で45,000分の1に減少





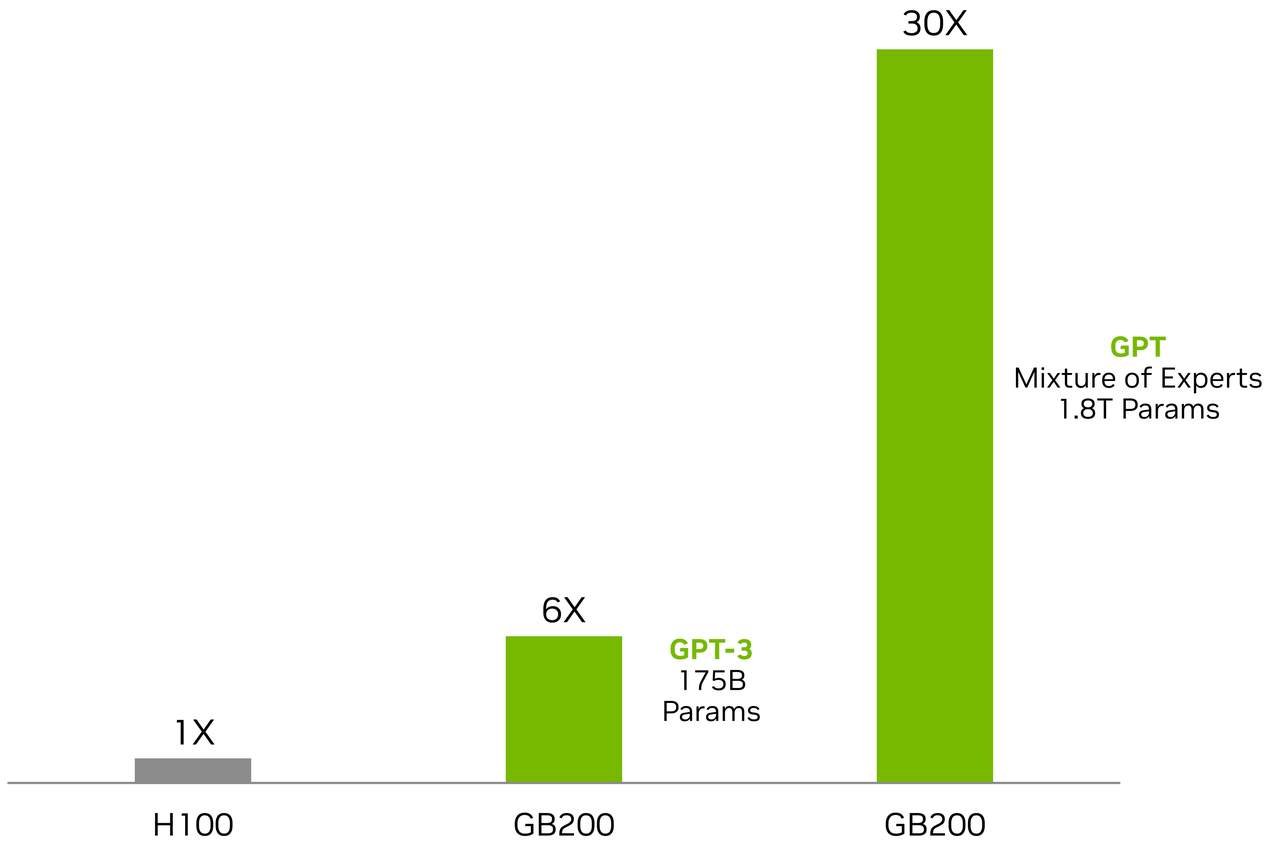
Blackwell は大規模モデルのパフォーマンスを向上

推論スループットとエネルギー効率は桁違いに

30X
Higher
Throughput

25X
Energy
Efficiency

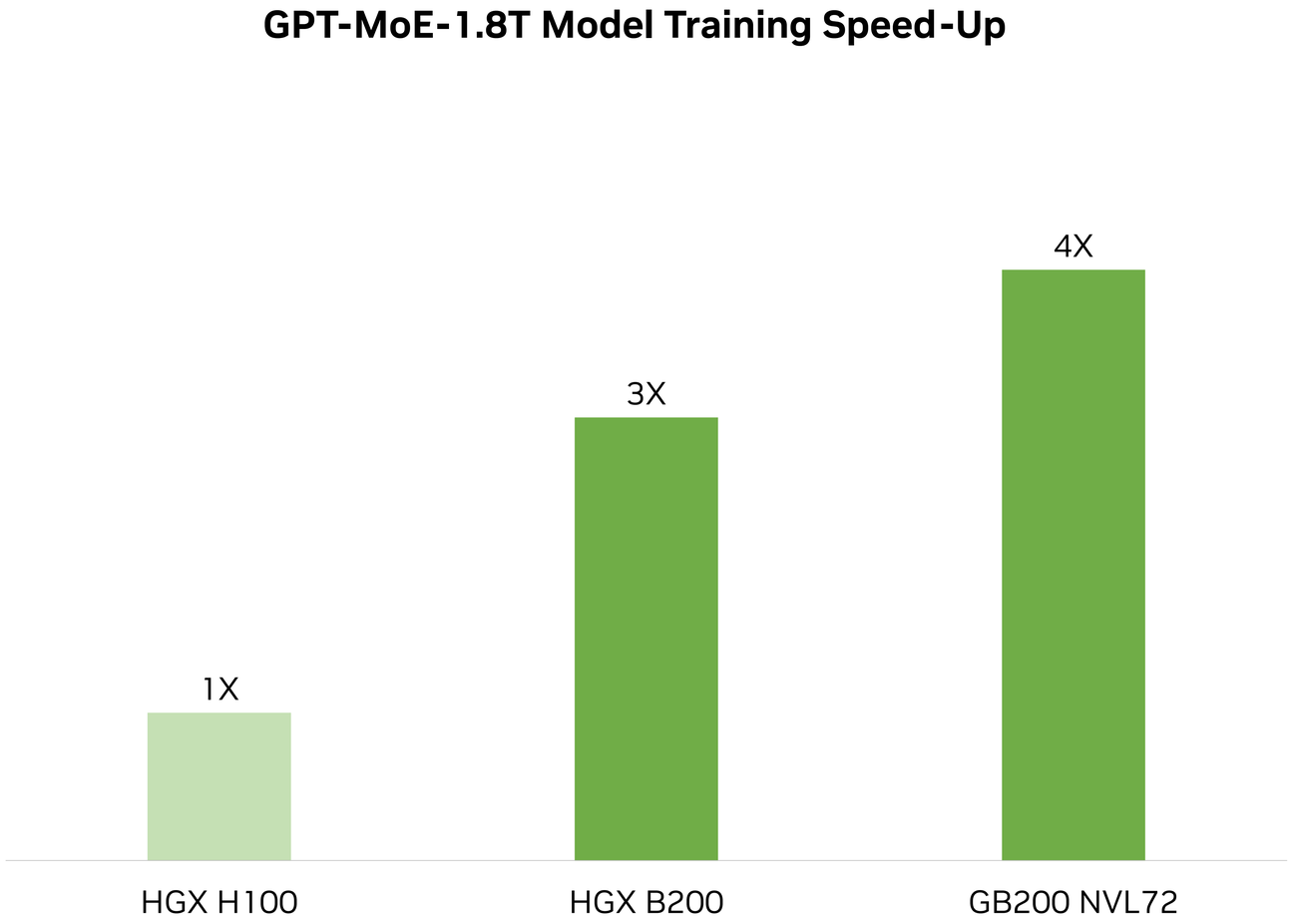
25X
Lower
TCO



Projected performance subject to change
Token-to-token latency (TTL) = 50 milliseconds (ms) real time | GPT-3 175B: First token latency (FTL) 2s; input sequence length = 2,048, output sequence length = 128, 4 HGX H100 air-cooled 400GB IB Network vs 2 GB200 Superchips liquid-cooled NVLink; per GPU performance comparison, | GPT-MoE-1.8T: FTL = 5s; input sequence length = 32,768, output sequence length = 1,024 , 8 HGX H100 air-cooled 400GB IB Network vs 18 GB200 Superchips liquid-cooled NVL36; per GPU performance comparison

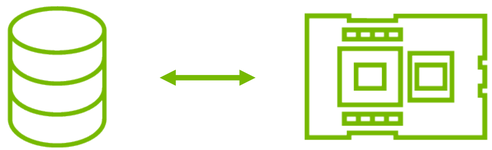
NVIDIA

AIトレーニングのパフォーマンス向上

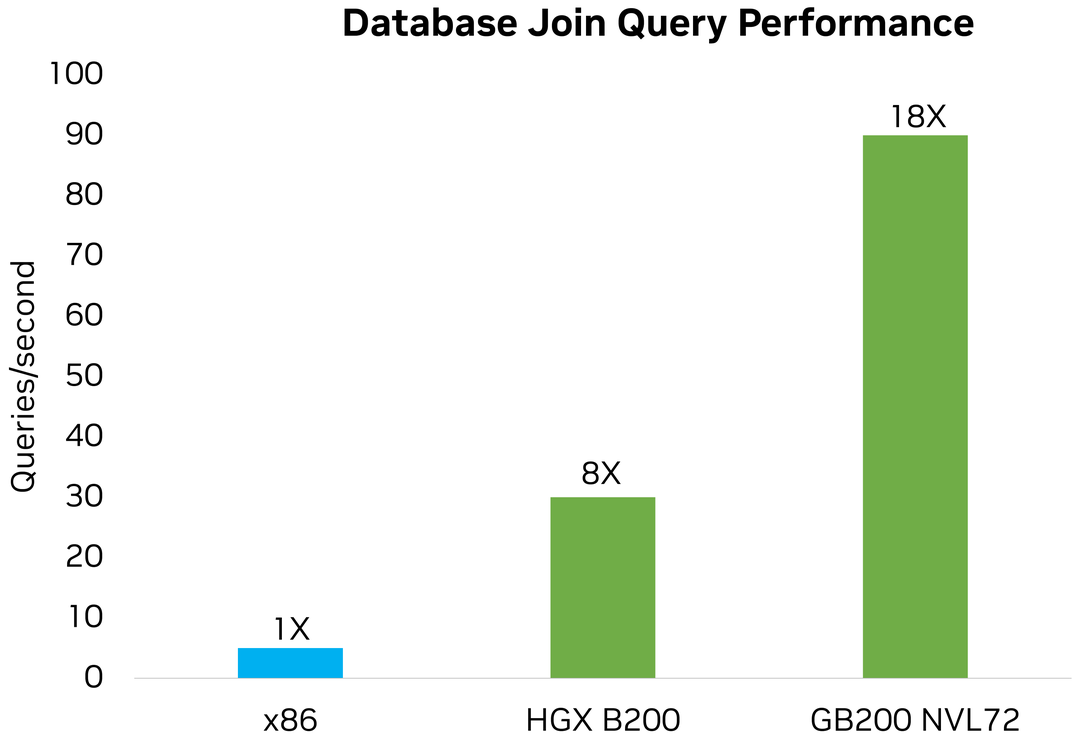


Projected performance subject to change. 32,768 GPU scale, 4,096 HGX H100 air-cooled cluster: 400G IB network, 456 GB200 NVL72 liquid-cooled cluster: 800G IB network

Decompression Engineでデータ処理を高速化



- サポートフォーマット: Deflate, Snappy, LZ4
- Grace CPUメモリと高速C2Cリンク - 大規模データベースのBlackwellへの迅速なアクセス



Projected performance subject to change. Database join query with Snappy / Deflate compression derived from TPC-H Q4 query. 1x x86, 1x H100 GPU, and 1x Blackwell GPU from Grace Blackwell GB200 Superchip.

DPU (Data Processing Unit) がデータセンターのエネルギー効率を高める

CPU から重要なインフラ処理をオフロードしてエネルギー消費を削減

インフラ オフロード

管理 **ストレージ** **セキュリティ** **ネットワーク**

NVIDIA Bluefield DPU

30%

CPU による OVS インフラ処理を DPU に移行した場合の**サーバ1台あたりのエネルギー削減量**

サーバ1台あたりの DPU 節電効果

9 TWh

すべての CPU インフラ運用を DPU に移行した場合に節約できる**年間エネルギー量**

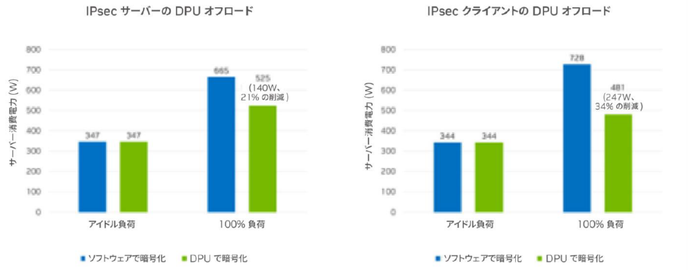
6.3M Metric Tons CO_2 **相当** **1.2M 家庭の電気使用量**

DPU による年間エネルギー節約額

Delivered Application Performance | Carbon and Home equivalence calculated using: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results

NVIDIA ホワイトペーパー「 DPU による電力仕様の効率化」より

図 5 および 6: IPsec の暗号化、復号化の実証試験により、BlueField-2 ハードウェアへのオフロードが IPsec 全体をソフトウェアで処理する場合と比べて顕著に省電力効果を示すことが判明。IPsec クライアントと IPsec サーバーの両方で顕著な省電力



予想どおり、サーバーの負荷が 0% (負荷なし = 暗号化トラフィックなし = オフロードするものなし) の場合に省電力効果は最小であり、負荷が 100% (大量の暗号化処理をオフロード) の場合に最大でした。つまり、DPU とサーバー効率化による最大の省電力効果は、各サーバーを可能な限り 100% に近い負荷で稼働させることで達成できます。この条件は、データセンターが一般的な戦略として打ち出す、必要なサーバーの台数を最小にし、ハードウェア アクセラレーションによるオフロードのメリットを最大にすることにも合致します。

表 3　　IPsec の暗号化を CPU から BlueField DPU にオフロードすることで得られた省電力効果[6]

100% 負荷時に IPsec を BlueField-2 にオフロード	**サーバーあたりの電力使用量 (削減)**	**サーバー 10,000 台を 3 年間運用した場合の電力コスト ($0.15/kWh で計算)**
IPsec サーバー、ソフトウェアで暗号化	665W	2,620 万ドル
IPsec サーバー、暗号化を DPU にオフロード	525W (140W、21% の削減)	2,070 万ドル (550 万ドルの削減)
IPsec クライアント、ソフトウェアで暗号化	728W	2,870 万ドル
IPsec クライアント、暗号化を DPU にオフロード	481W (247W、34% の削減)	1,900 万ドル (870 万ドルの削減)

持続可能なコンピューティングに向けて

- 近年のプロセッサ効率や性能の改善は性能要求に追いついていない
  - 結果として、データセンターやスパコンなどでの消費電力量は増加を続けている
- 非効率なコンピューティングによる無駄を削減
  - AI や HPC は、GPU をはじめとするアクセラレータの活用で何十倍ものエネルギー効率
  - AI 以外でも、暗号化や各種データ処理は専用のプロセッサ（ DPU など）を活用

ハードウェアやソフトウェアを含むあらゆるレイヤーにおいて
最適化や効率化を図ることでエネルギー効率が向上し
持続可能なコンピューティングにつながる