エヌビディア コンピューティング プラットフォーム最新情報

2021/12/8
エヌビディア合同会社

NVIDIA コンピューティング 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 | 医療診断 | 異常検知 | 品質保証 | 自動化 | マーケティング | シミュレーション | ++

NGC

ソフトウェア ハブ

HELM

コンテナイメージ

学習済みモデル

ソフトウェア開発キット

アプリケーション
フレームワーク

スマートシティ **Metropolis** | 対話型 AI **Riva** | 自動運転 **Drive** | レコメンデーション システム **Merlin** | ヘルスケア **Clara** | ++ …

開発者
ツールキット

ML & DATA ANALYTICS: RAPIDS, dmlc XGBoost

AI TRAINING & INFERENCE: TensorFlow, TensorRT, PYTORCH, mxnet

HIGH PERFORMANCE COMPUTING: *NVIDIA HPC SDK*

RENDERING & VISUALIZATION: *IndeX*, *OptiX*

高速化
ライブラリ

COMPUTE: *CUDA-X*

NETWORKING, STORAGE & SECURITY: *DOCA*, *MAGNUM IO*

運用

VIRTUAL GPU SOFTWARE

TRITON INFERENCE SERVER

FLEET COMMAND

NVIDIA GPU Operator

Red Hat

vmware

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

サーバー及びクラウド

DGX
HGX

ハイエンド

EGX

メインストリームとエッジ

クラウドのインスタンス

ハードウェア技術

GPU

NVSwitch

BlueField DPU

SMART NIC

Mellanox Switch

監視

DCGM | Grafana

Prometheus | UFM

SDK

NVIDIA cuQUANTUM DGX アプライアンス

最強の「古典」計算機でコンピューティングの未来を研究

Out-of-the-Box Optimized Stack for Cirq

Other Simulators in Development

cuQuantum SDK in Open Beta; Accelerate Popular Quantum Simulators from Google, IBM

LEADING QUANTUM SIMULATORS

Qiskit

Quantum AI

INDUSTRY PARTNERS

IONQ

PASQAL

PsiQuantum

RESEARCH COMMUNITY PARTNERS

Argonne NATIONAL LABORATORY

BERKELEY LAB Lawrence Berkeley National Laboratory

Caltech

JÜLICH Forschungszentrum

Los Alamos NATIONAL LABORATORY

MIT

OAK RIDGE National Laboratory

Pacific Northwest NATIONAL LABORATORY

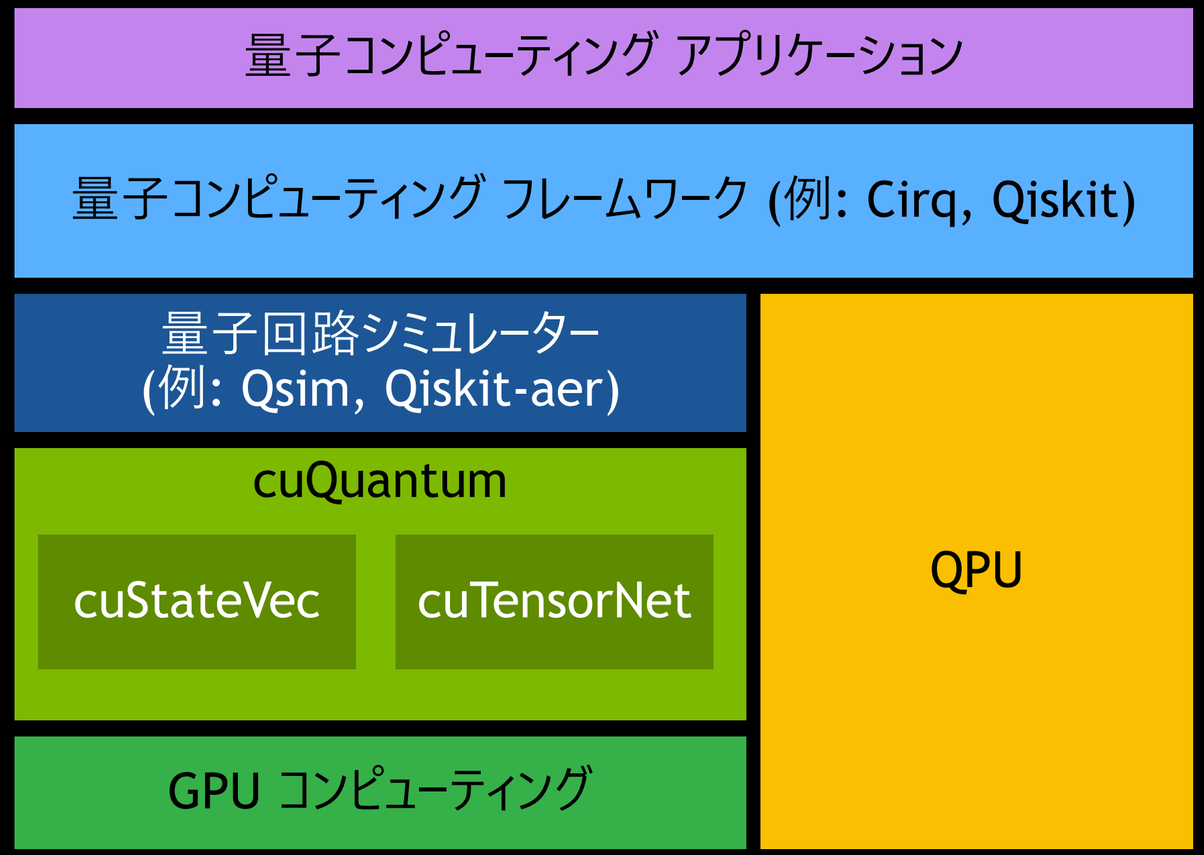
Appliance Available Q1 2022

cuQuantum Available Now for Download
developer.nvidia.com/cuquantum

cuQuantum とは？

- cuQuantum は量子コンピューティング向けの最適化されたライブラリやツールからなる SDK

- cuQuantum は以下のものでは**ありません**
  - 量子コンピューター
  - 量子コンピューティング フレームワーク
  - 量子回路シミュレーター



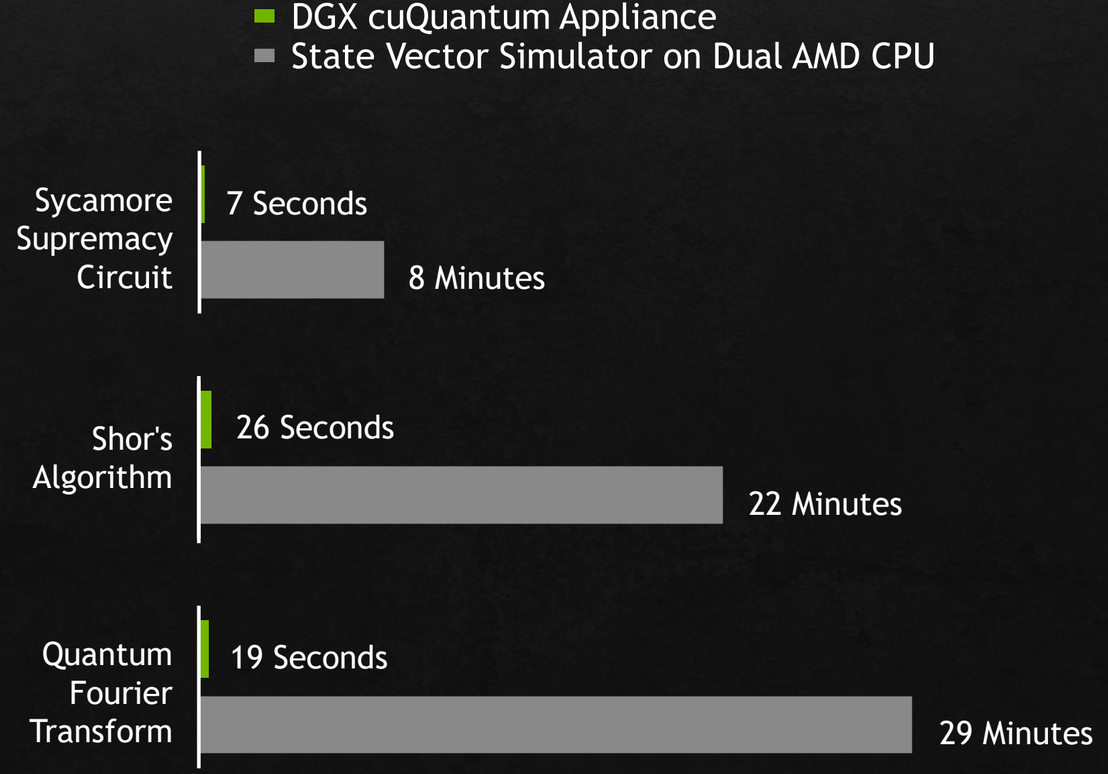
NVIDIA cuQUANTUM
DGX アプライアンス

最強の「古典」計算機でコンピューティングの未来を研究

Out-of-the-Box Optimized Stack for Cirq

Other Simulators in Development

cuQuantum SDK in Open Beta;
Accelerate Popular Quantum
Simulators from Google, IBM



Appliance Available Q1 2022

cuQuantum Available Now for Download
developer.nvidia.com/cuquantum

NVIDIA cuNUMERIC

PyData / NumPy エコシステムを GPU で高速化

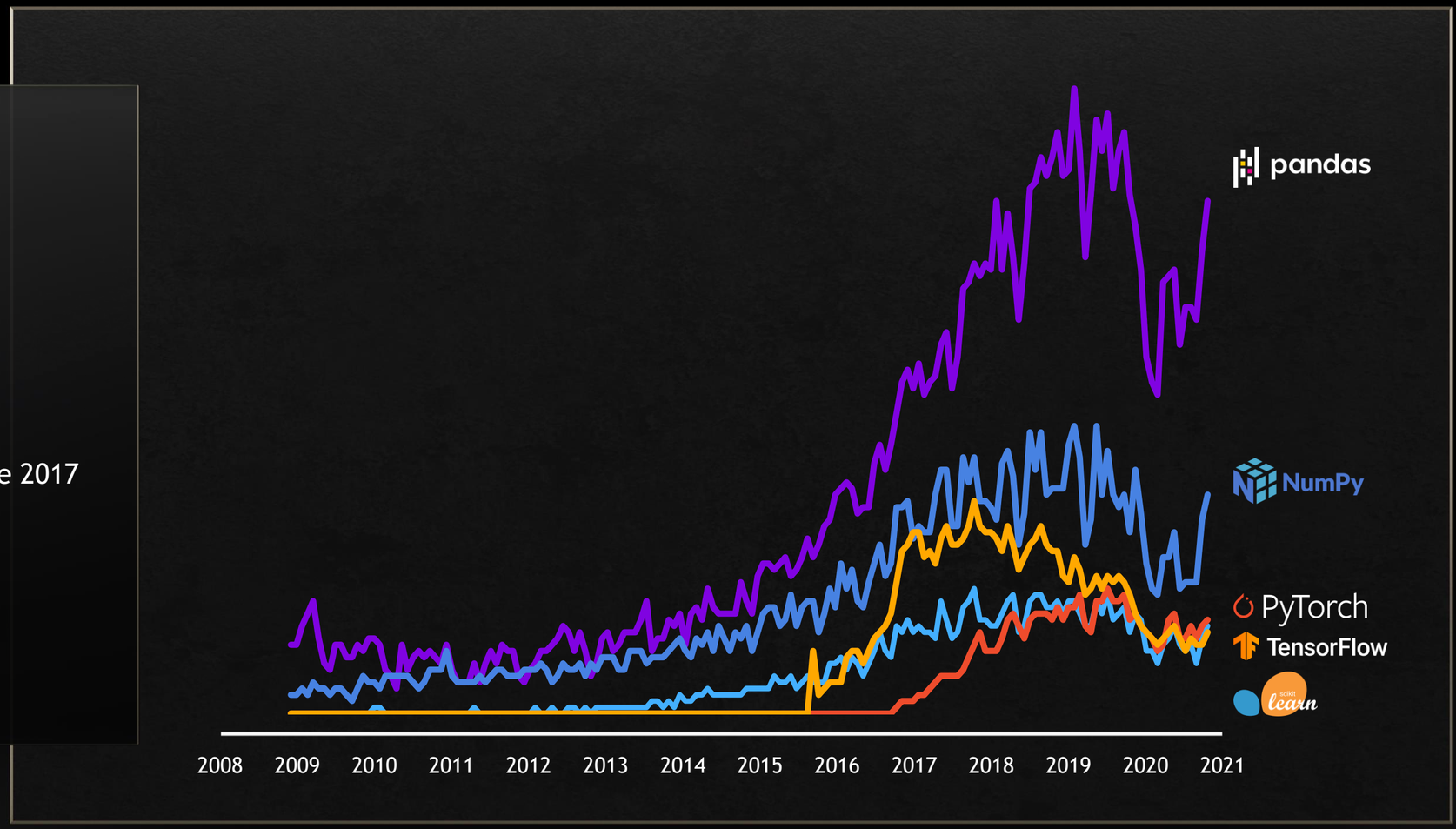
Python Used by 20 Million Data Scientists, Researchers, and Scientists

NumPy Downloaded 122,000,000 Times Since 2017

NumPy Used by 790,000 Projects in GitHub

NumPy is the Foundation of Pandas, SciPy, and Scikit-Learn

https://github.com/nv-legate/cunumeric



NVIDIA cuNUMERIC

PyData / NumPy エコシステムを GPU で高速化

Transparently Accelerates and Scales NumPy Workflows

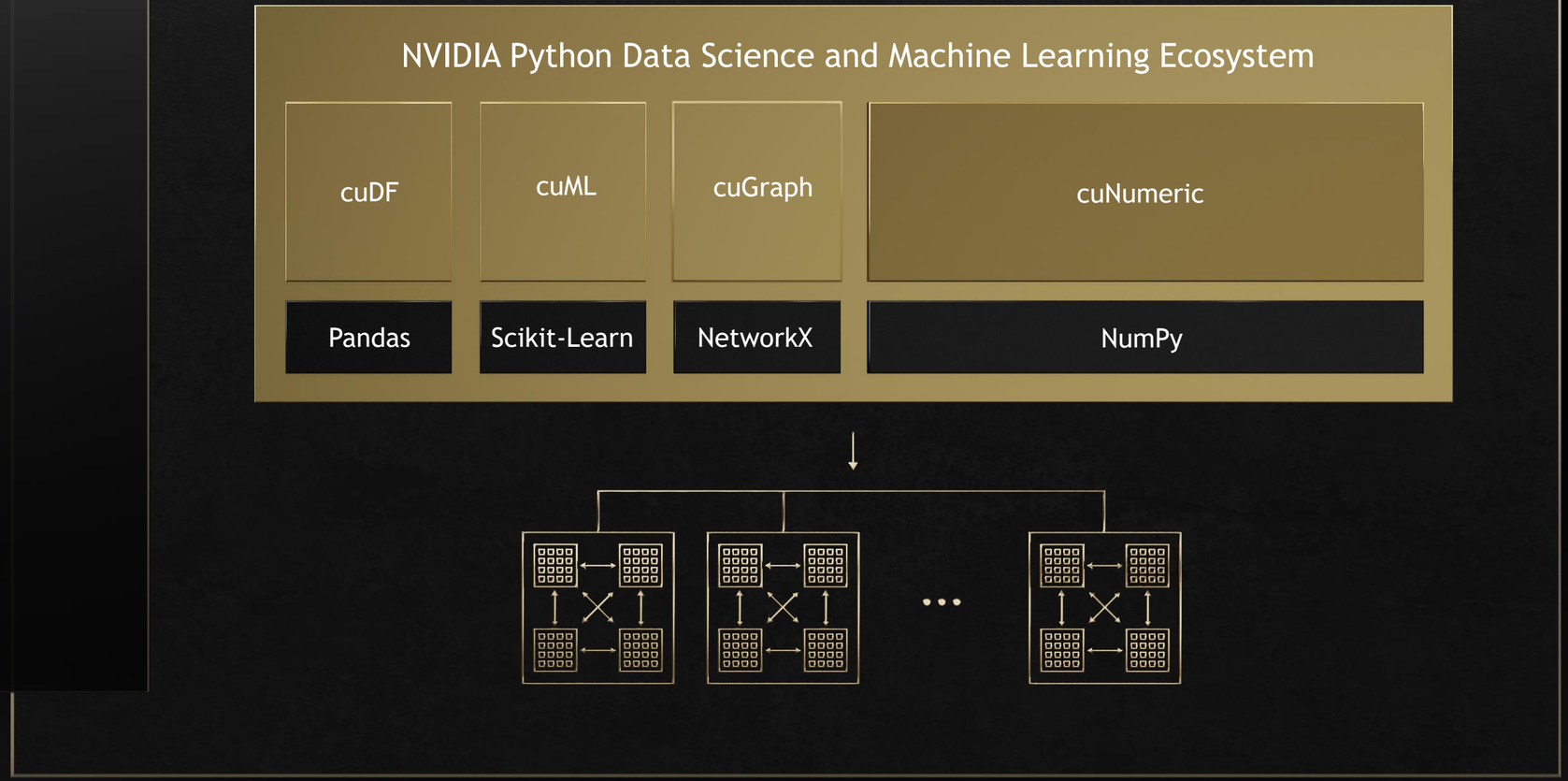
Zero Code Changes

Automatic Parallelism and Acceleration for Multi-GPU, Multi-Node Systems

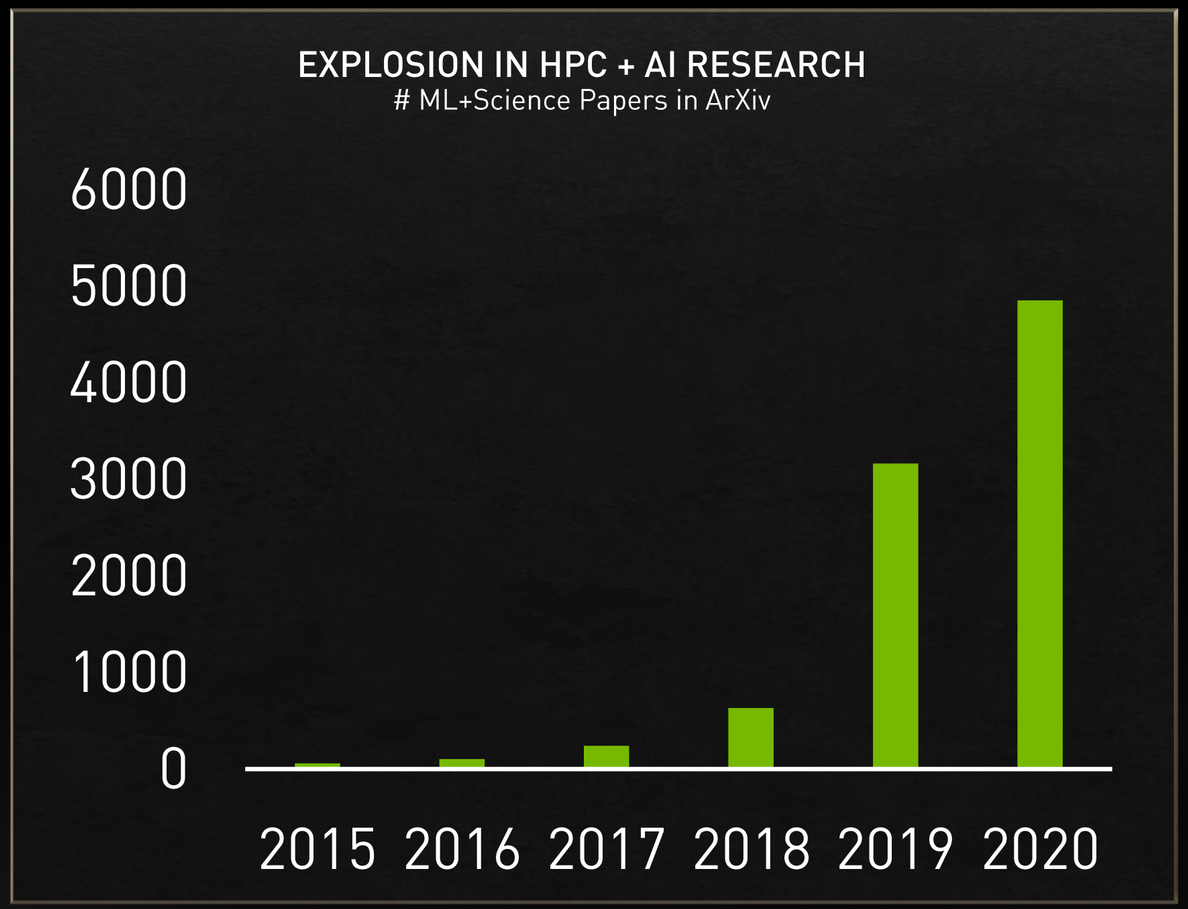
Scales to 1,000s of GPUs

Available Now on GitHub and Conda

https://github.com/nv-legate/cunumeric



PHYSICS-ML TURBOCHARGES SCIENCE



Physics-Informed Neural Operator for Learning Partial Differential Equations

Zongyi Li*, Hongkai Zheng*, Nikola Kovachki, David Jin, Haoxuan Chen,
Burigede Liu, Kamyar Azizzadenesheli, Anima Anandkumar

October 28, 2021

Abstract

Machine learning methods have recently shown promise in solving partial differential equations (PDEs). They can be classified into two broad categories: solution function approximation and operator learning. The Physics-Informed Neural Network (PINN) is an example of the former while the Fourier neural operator (FNO) is an example of the latter. Both these approaches have shortcomings. The optimization in PINN is challenging and prone to failure, especially on multi-scale dynamic systems. FNO does not suffer from this optimization issue since it carries out supervised learning on a given dataset, but obtaining such data may be too expensive or infeasible. In this work, we propose the physics-informed neural operator (PINO), where we combine the operating-learning and function-optimization frameworks, and this improves convergence rates and accuracy over both PINN and FNO models. In the operator-learning phase, PINO learns the solution operator over multiple instances of the parametric PDE family. In the test-time optimization phase, PINO optimizes the pre-trained operator ansatz for the querying instance of the PDE. Experiments show PINO outperforms previous ML methods on many popular PDE families while retaining the extraordinary speed-up of FNO compared to solvers. In particular, PINO accurately solves long temporal transient flows and Kolmogorov flows, while PINN and other methods fail to converge.

NVIDIA MODULUS

物理法則に基づいたニューラル シミュレーション
フレームワーク

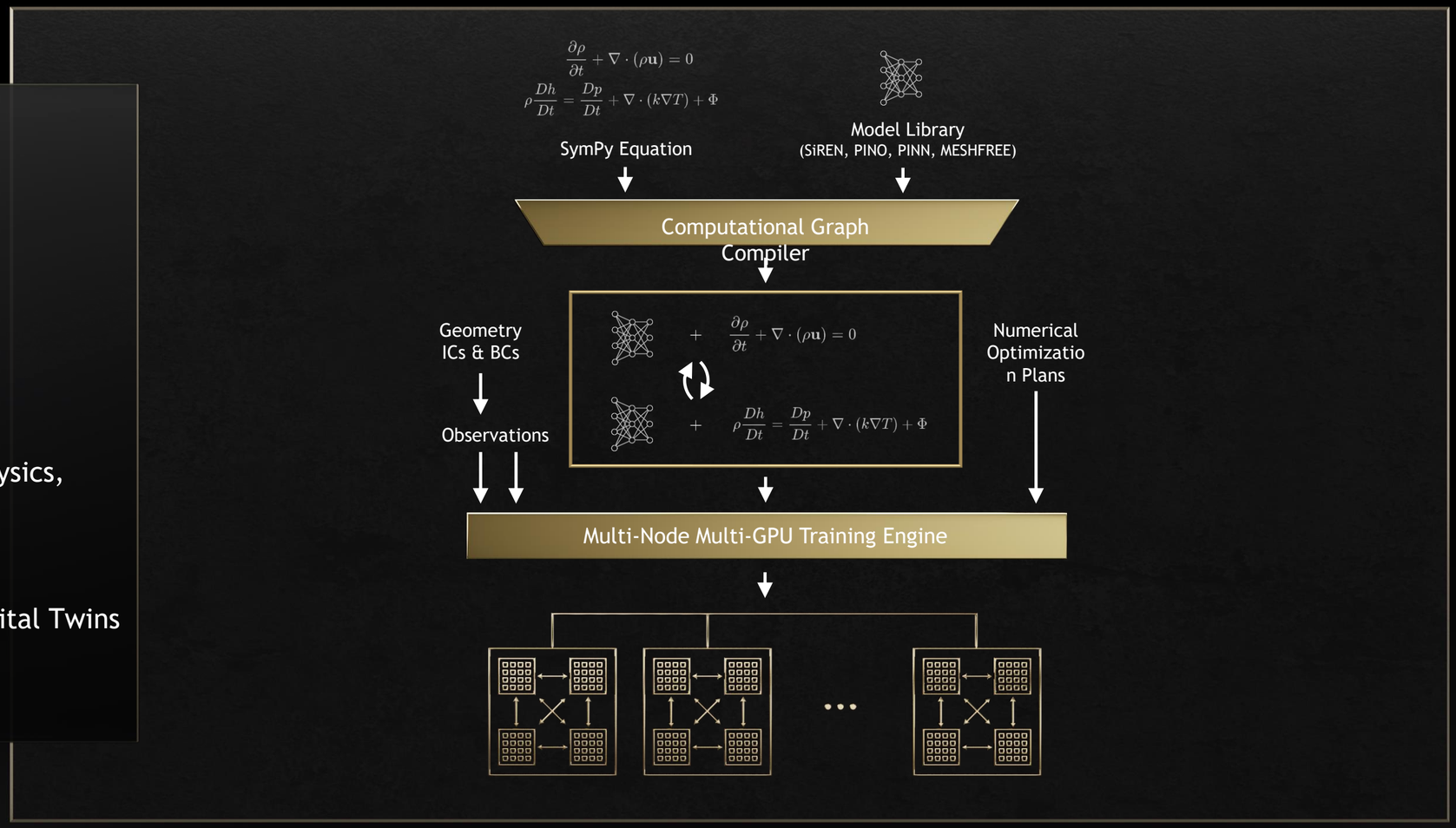
物理法則に基づいた機械学習モデルを
開発するためのフレームワーク

Train Physics-ML Models Using Governing Physics, Simulation, and Observed Data

Multi-GPU, Multi-Node Training

1,000-100,000X Speed Models – Ideal for Digital Twins

Available Now
developer.nvidia.com/modulus



GPU

データセンター向け GPU 製品

	A100	A30	A2	A40	A16	A100X	A30X
Design	Highest Perf Compute	Mainstream Compute	Entry-Level Small Footprint	High Perf Graphics	High Density Virtual Desktop	High Perf Converged Accelerator	Mainstream Converged Accelerator
Max Power	300W	165W	40-60W	300W	250W	300W	230W
Form Factor	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x8 PCIe Gen4 1 Slot LP	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge
GPU Memory	80GB HBM2e	24GB HBM2	16GB GDDR6	48GB GDDR6	4x 16GB GDDR6	80GB HBM2e	24GB HBM2e
Multi-Instance GPU (MIG)	Up to 7	Up to 4	-	-	-	Up to 7	Up to 4
Media Acceleration	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV decode)	1 Video Encoder 2 Video Decoder (+AV1 decode)	4 Video Encoder 8 Video Decoder (+AV1 decode)	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder
Ray Tracing	-		Yes	Yes	Yes	-	
Fast FP64	Yes		-	-	-	Yes	
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)		Good	Best	Better	For in-situ visualization (no NVIDIA vPC or RTX vWS)	
vGPU	Yes		Yes*	Yes	Yes	Yes*	
Hardware Root of Trust	Yes		Yes	Yes	Yes	Yes	
Integrated DPU	-	-	-	-	-	BlueField-2	
Server Availability	In Production	In Production	Q1 ’22	In Production	In Production	Q1 ‘22	

**Coming soon*

NVIDIA A2

NVIDIA A2

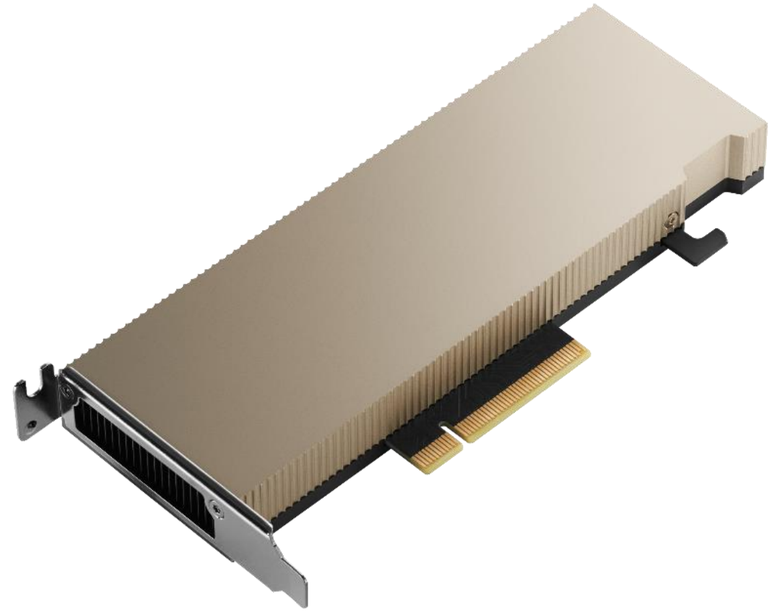
様々なサーバーに搭載可能なエントリーレベル GPU

Compact, Entry-Level Inference
Single slot LP, lower power – fits any server
& optimal for thermally constrained systems

Latest Ampere Architecture Features
3rd gen Tensor cores, 2nd gen RT cores, Secure RoT

Higher Intelligent Video Analytics (IVA) Performance
1.3X better performance vs T4

Up to 20X Higher Performance versus CPU
Speedups for AI inference and cloud gaming

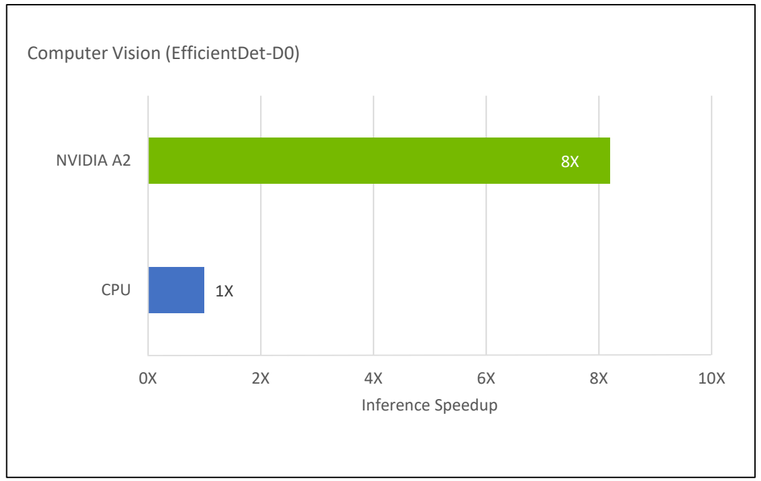


**Server availability 1-3 months later*

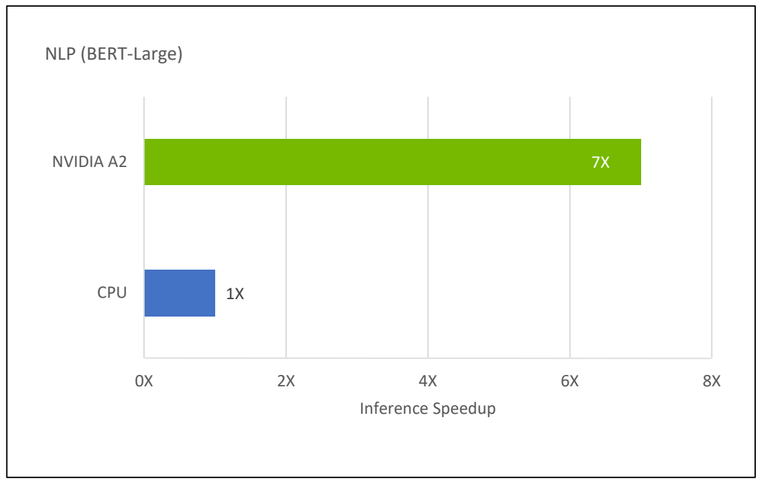
A2 が推論を高速化

CPU サーバーに対し最大 20 倍の高速化

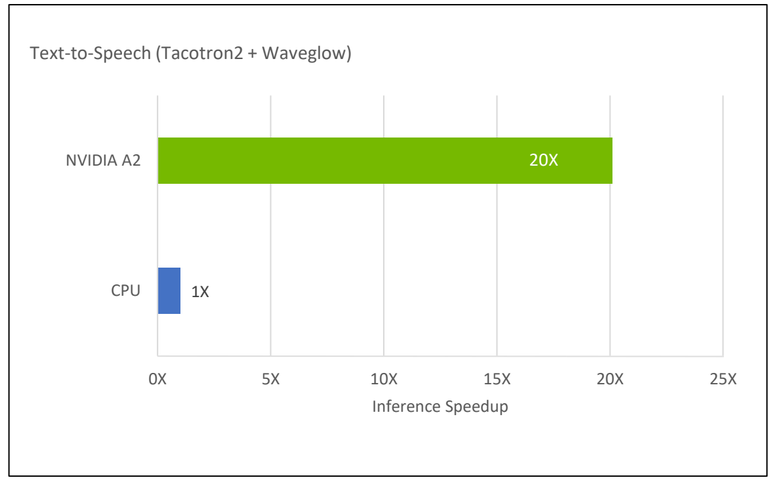
Computer Vision



Natural Language Processing



Text-to-Speech



Comparisons of one NVIDIA A2 Tensor Core GPU versus a dual-socket Xeon Gold 6330N CPU

System Config: [CPU: HPE DL380 Gen10 Plus, 2S Xeon Gold 6330N @2.2GHz, 512GB DDR4]
Computer Vision: EfficientDet-D0 (COCO, 512x512) | TensorRT 8.2, Precision: INT8, BS:8 (GPU) | OpenVINO 2021.4, Precision: INT8, BS:8 (CPU)
NLP: BERT-Large (Sequence length: 384, SQuAD: v1.1) | TensorRT 8.2, Precision: INT8, BS:1 (GPU) | OpenVINO 2021.4, Precision: INT8, BS:1 (CPU)
Text-to-Speech: Tacotron2 + Waveglow E2E pipeline (input length: 128) | PyTorch 1.9, Precision: FP16, BS:1 (GPU) | PyTorch 1.9, Precision: FP32, BS:1 (CPU)

NVIDIA A100X と NVIDIA A30X

DPU と GPU のコンバージド アクセラレーター

A100 / 30 Tensor Core GPU

BlueField-2 DPU

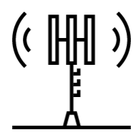
ConnectX-6 Dx

8 ARM A72 Cores

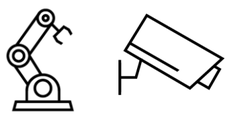
Integrated PCIe Gen4 Switch

2 slot FHFL

300 / 230 W

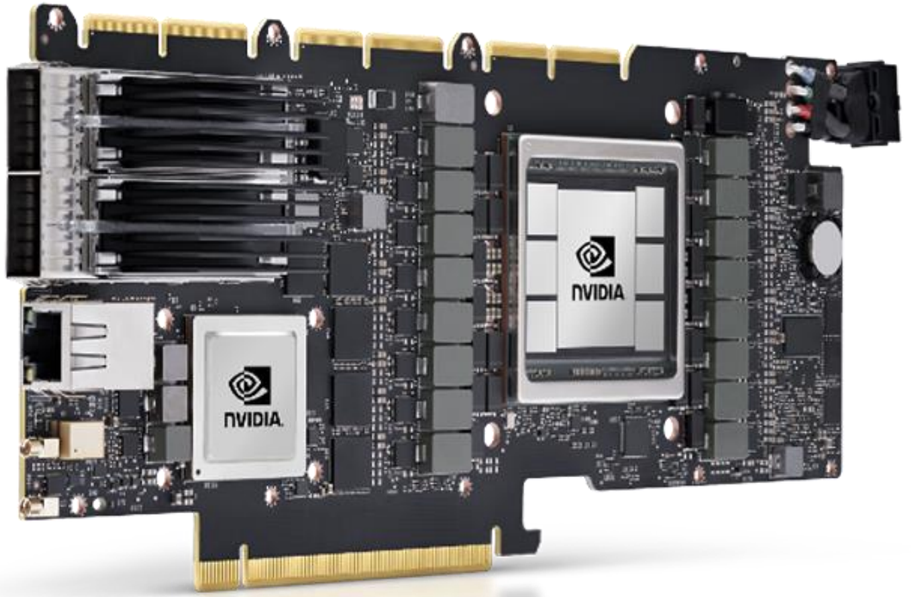


5G vRAN



AI on 5G

AI-Based Security

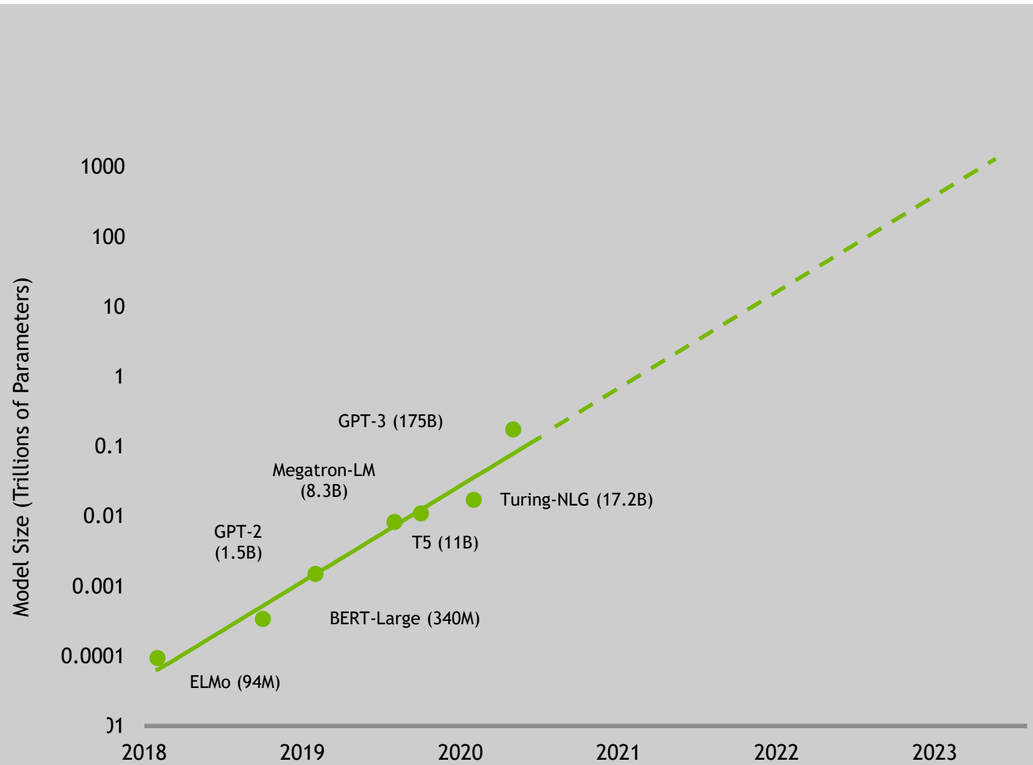


GRACE CPU

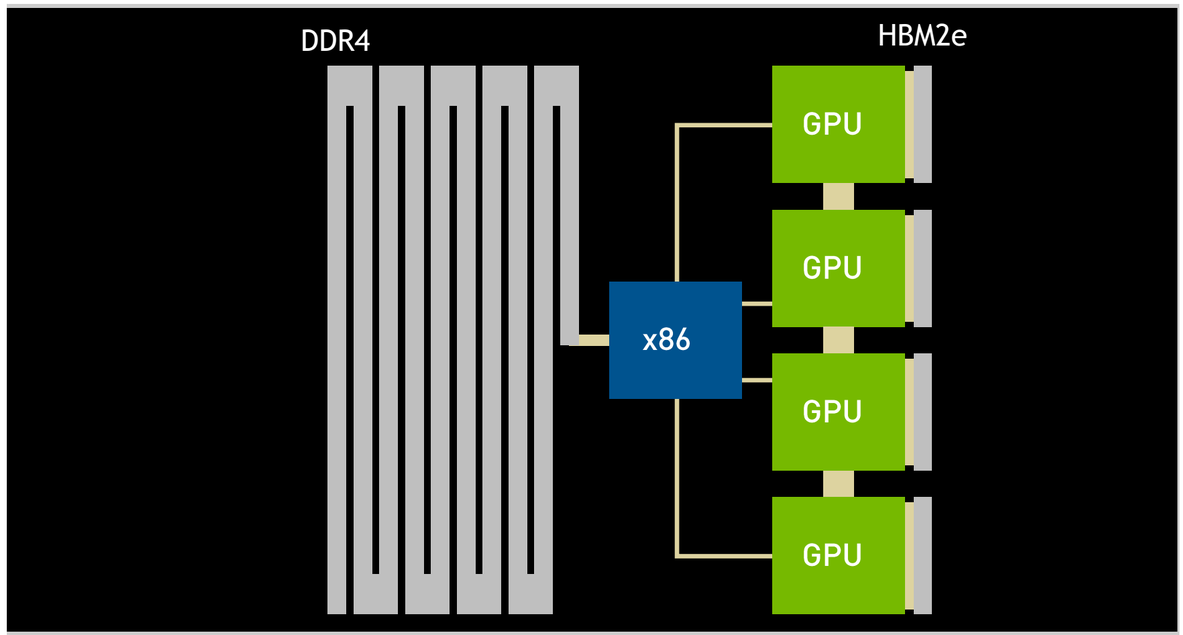
既存アーキテクチャの限界を超える巨大なモデル

新たなアーキテクチャが必要

100 TRILLION PARAMETER MODELS BY 2023



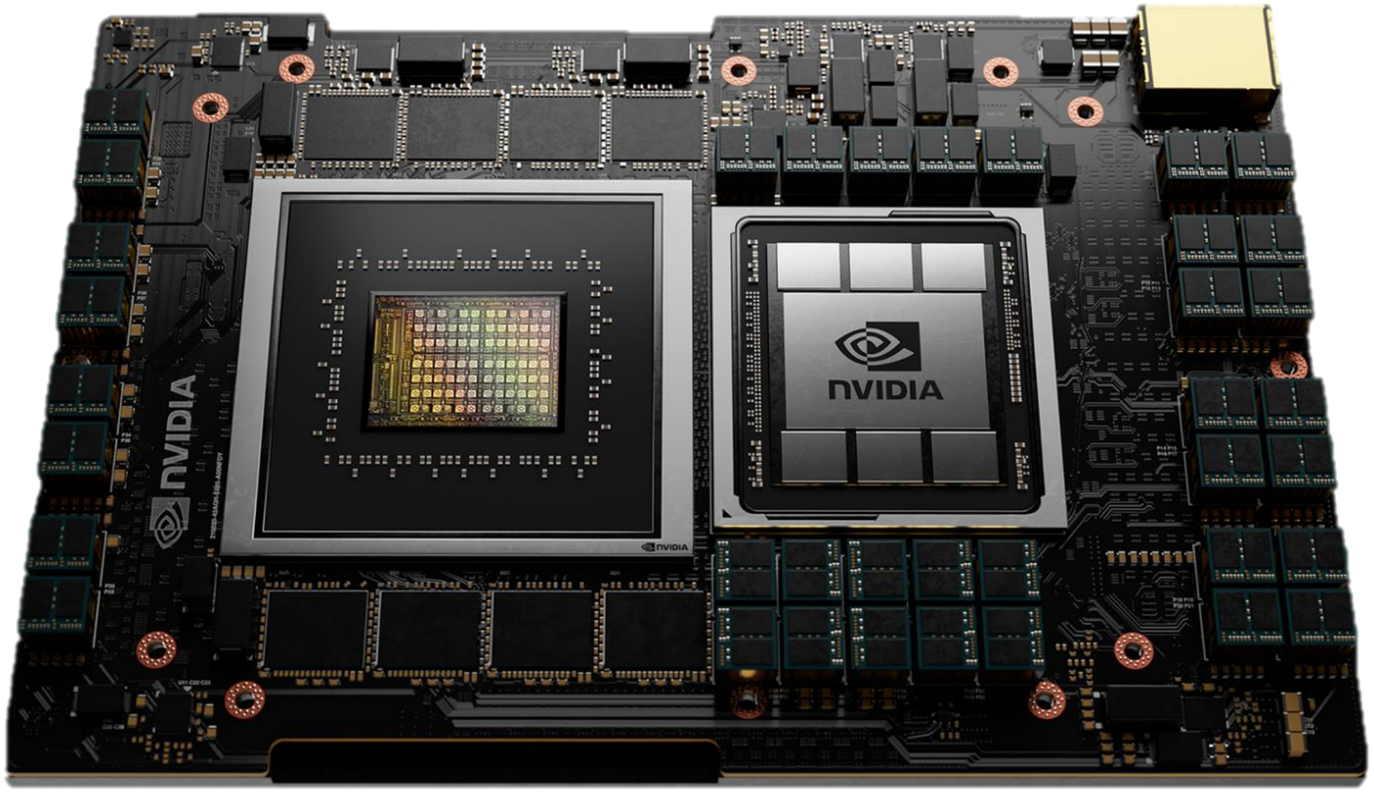
System Bandwidth Bottleneck



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
Mem-to-GPU	**64**	**GB/sec**

NVIDIA GRACE

大規模な AI と HPC アプリケーション向けに設計される画期的な CPU



FASTEST INTERCONNECTS
>900 GB/s Cache Coherent NVLink CPU To GPU (14x)
>600GB/s CPU To CPU (2x)

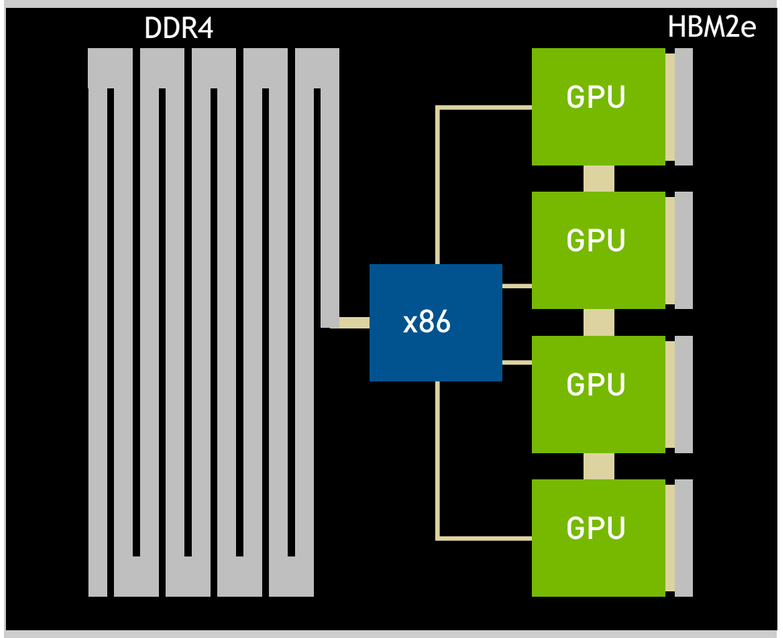
HIGHEST MEMORY BANDWIDTH
>500GB/s LPDDR5x w/ ECC
>2x Higher B/W
10x Higher Energy Efficiency

NEXT GENERATION ARM NEOVERSE CORES
>300 SPECrate2017_int_base est.
Availability 2023

テラバイト級のアクセラレーテッド コンピューティング

新たなワークロードに向けて進化するアーキテクチャ

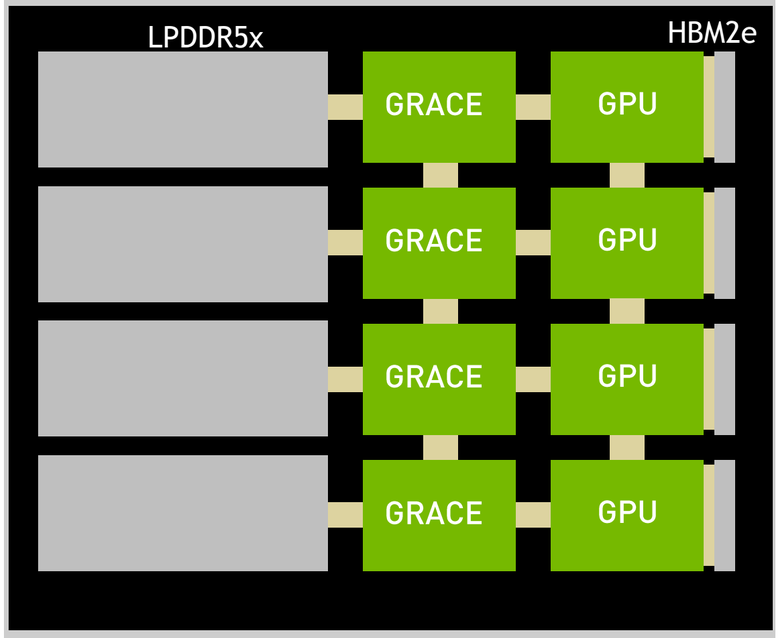
CURRENT x86 ARCHITECTURE



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
Mem-to-GPU	**64**	**GB/sec**

Transfer 2TB in 30 secs

INTEGRATED CPU-GPU ARCHITECTURE



GPU	8,000	GB/sec
CPU	500	GB/sec
NVLink	500	GB/sec
Mem-to-GPU	**2000**	**GB/sec**

Transfer 2TB in 1 secs

3 DAYS FROM 1 MONTH

Fine-Tune Training of 1T Model

REAL-TIME INFERENCE ON 0.5T MODEL

Interactive Single Node NLP Inference

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.

NVIDIA

世界最速の AI スーパーコンピューター

20 Exaflops of AI

Accelerated w/ NVIDIA Grace CPU and NVIDIA GPU

HPC and AI For Scientific and Commercial Apps

Advance Weather, Climate, and Material Science

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Hewlett Packard
Enterprise

NVIDIA

NETWORKING

IN-NETWORK がスーパーコンピューティングを飛躍的に加速

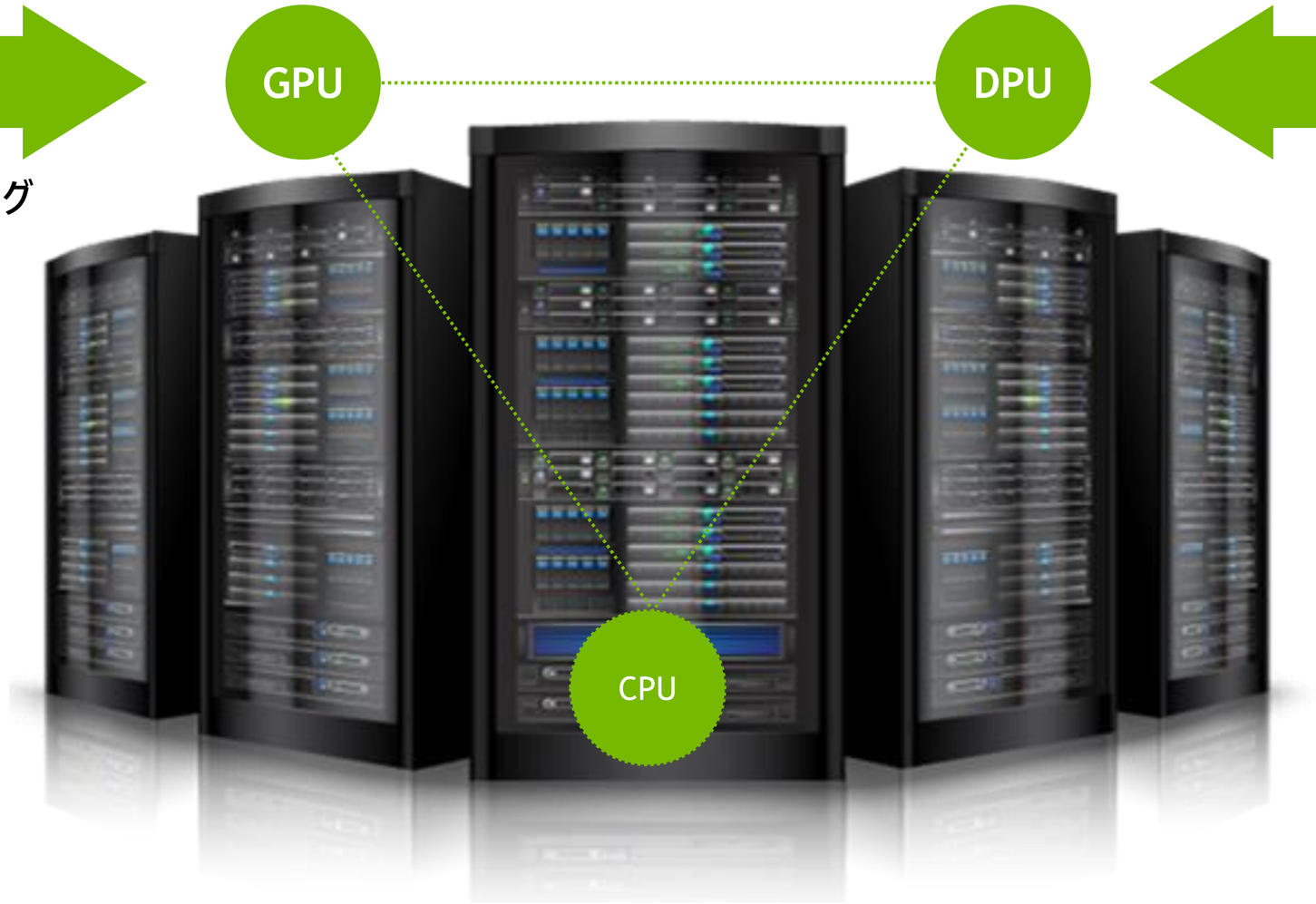
ソフトウェア デファインド、ハードウェア アクセラレーション、InfiniBand ネットワーク

AI & ML

アクセラレーテッド コンピューティング

GPU で加速された AI & 機械学習

すべてのAIワークロードが加速



Data Processing Unit

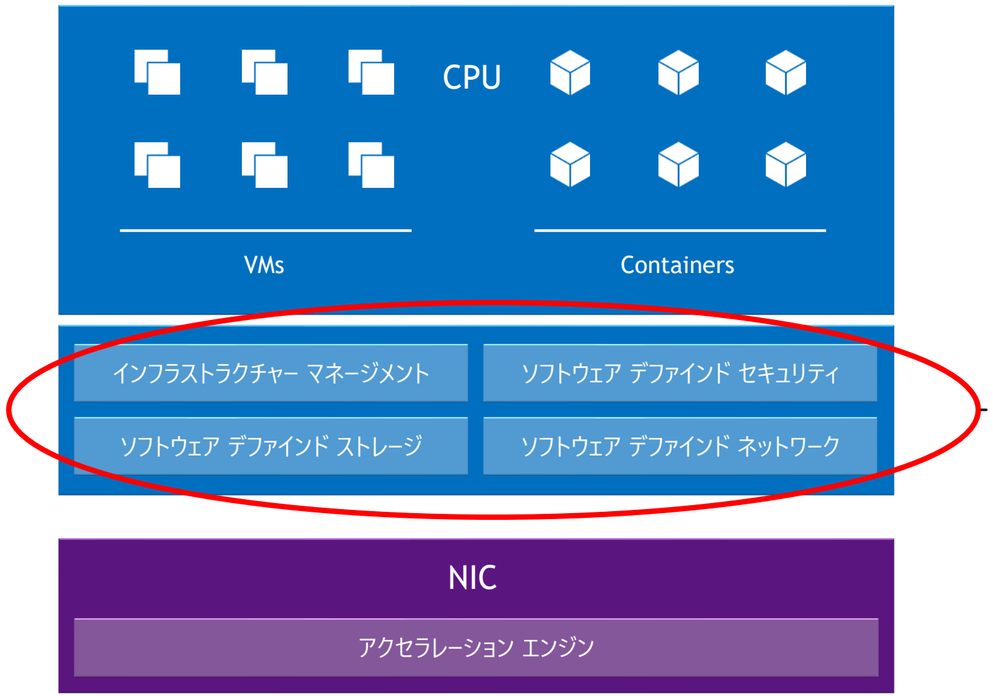
ソフトウェア デファインド、ハードウェア アクセラレーション

DPUはデータ集約型タスクを加速

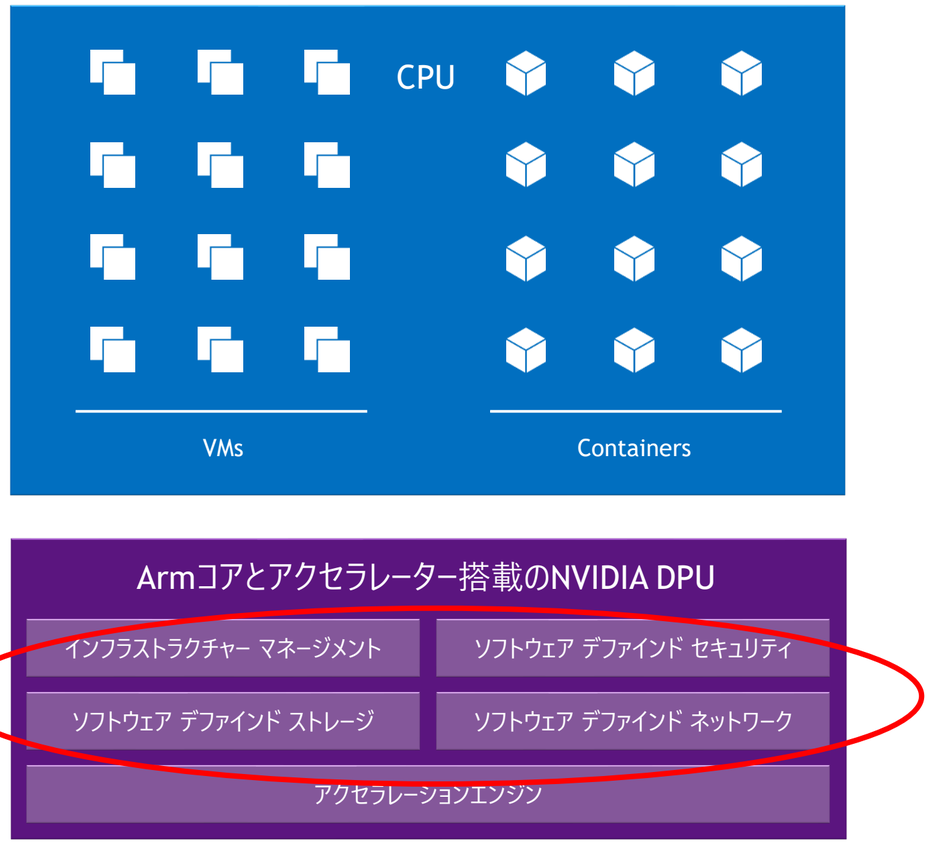
ネットワーク、セキュリティ、ストレージ

DATA PROCESSING UNIT

データセンター インフラストラクチャをソフトウェア デファインド, ハードウェア アクセラレーションでチップにオフロード



手動インフラストラクチャ管理| セキュリティ アプライアンス| ストレージシステム| 静的ネットワーク| マイクロサービス|エンドツーエンド接続| ストレージアクセス| ゼロトラストセキュリティ



NVIDIA BLUEFIELD-2 DATA PROCESSING UNIT

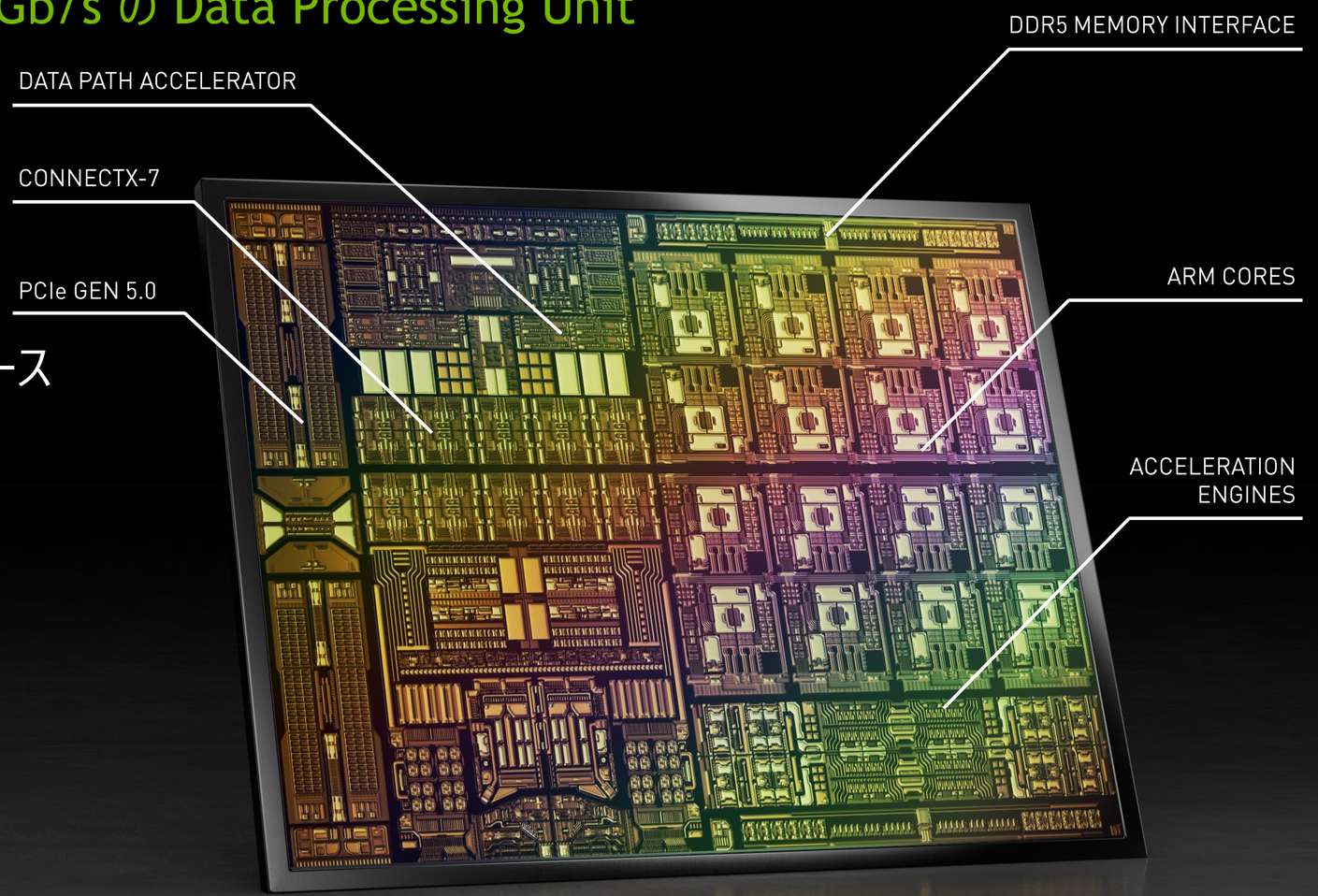
データセンター インフラストラクチャをチップにオフロード

- 最大 200Gb/s Ethernet and InfiniBand, PAM4/NRZ
- ConnectX-6 Dx を内蔵
- 8 Arm A72 CPUs subsystem – up to 2.75GHz
  - 8MB L2 cache, 6MB L3 cache in 4 Tiles
  - 非常に高速で低遅延なインターコネクト
- 外部、内部を統合するPCIe switch,16x Gen4.0
  - PCIeルートコンプレックスモードまたはエンドポイントモード
- Single DDR4 メモリチャネル
- 1GbE 外部管理ポート
- セキュリティ、ストレージ、ネットワークを加速

NVIDIA BLUEFIELD-3 DPU

最速 400Gb/s の Data Processing Unit

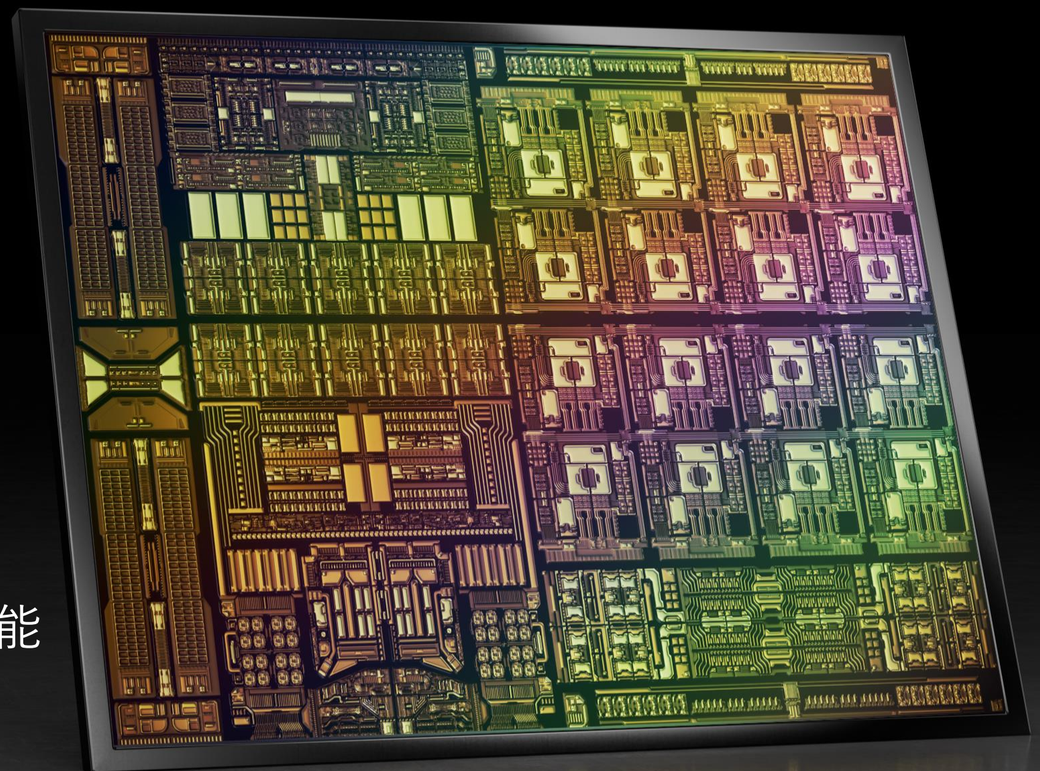
- 220億トランジスタ
- 400Gb/s のEthernet版 InfiniBand版をリリース
- 400Gb/s 暗号化アクセラレーション
- X86 Core 300個と同等
- 18M IOP/s ストレージ性能
- DDR5 Memory



NVIDIA BLUEFIELD-3 DPU

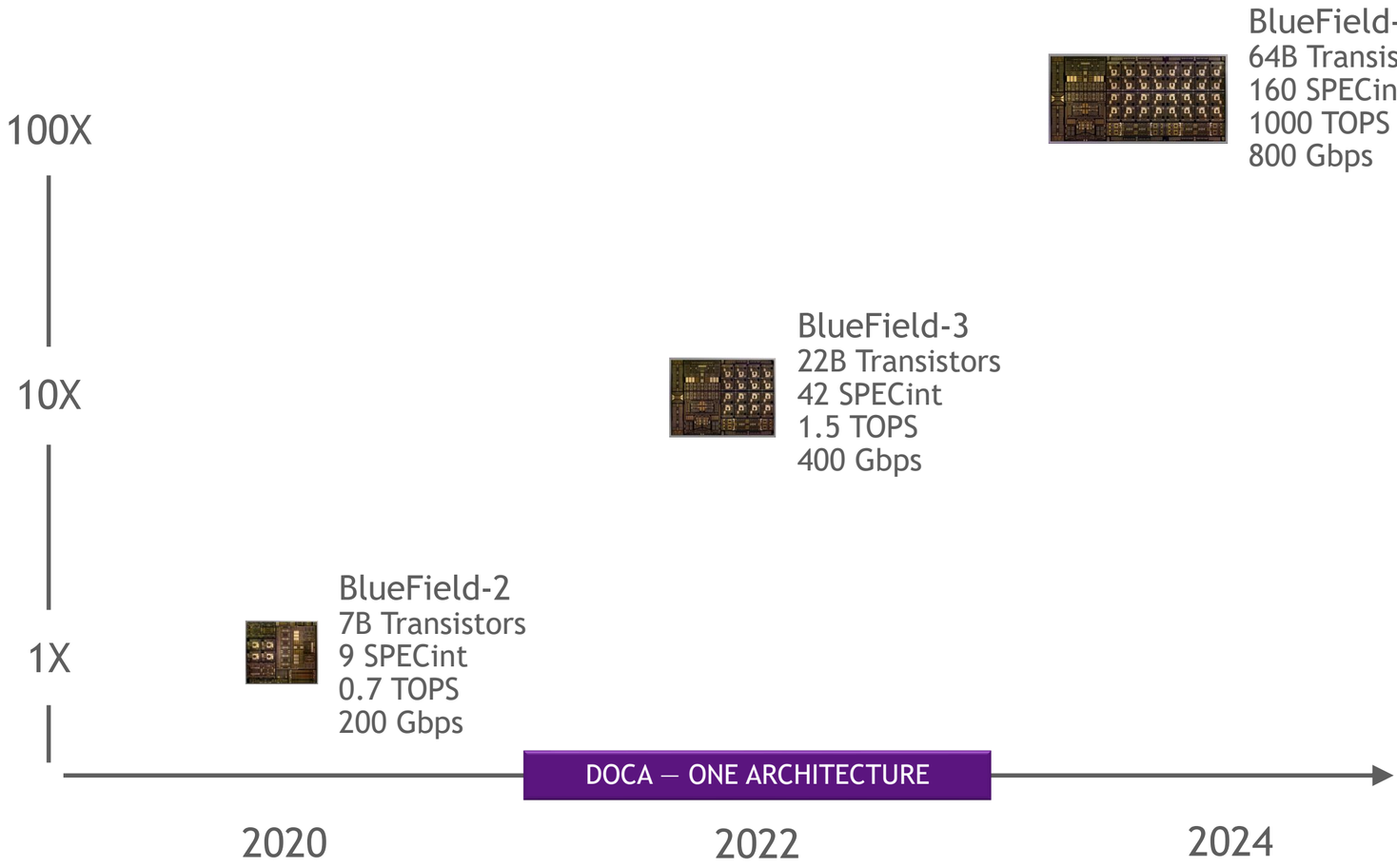
最速 400Gb/s の Data Processing Unit

- データセンター インフラストラクチャーのオフロードと高速化
- ホストのアプリケーションワークロードから
  管理機能、制御機能を完全分離
- 強力な CPU – 16x Arm A78 Coresを搭載
- データパスアクセレラレータ – 16x Cores, 256 Threads
- ネットワーク, ストレージ,セキュリティを400 Gb/sで処理可能



NVIDIA DPU ロードマップ

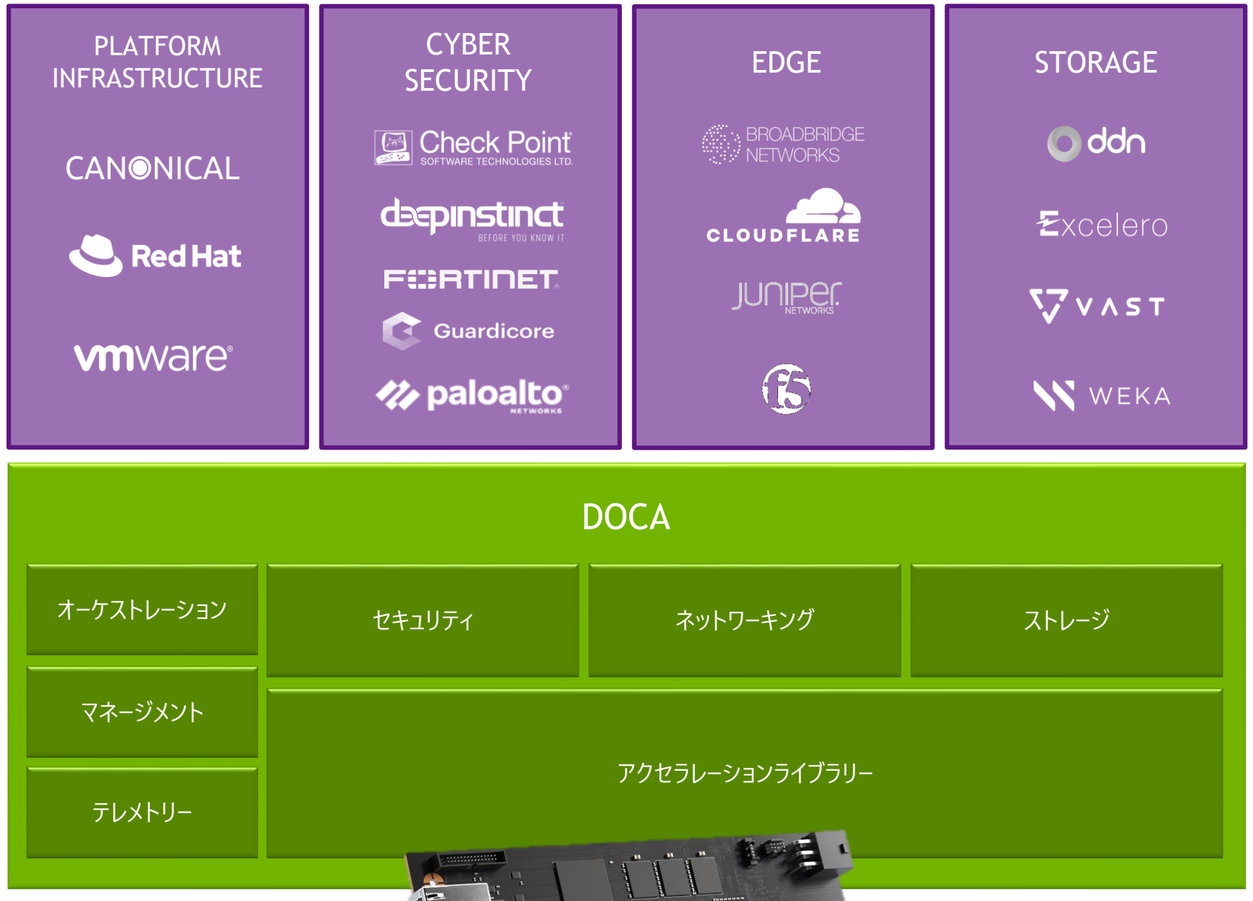
データセンターインフラ処理の飛躍的な成長

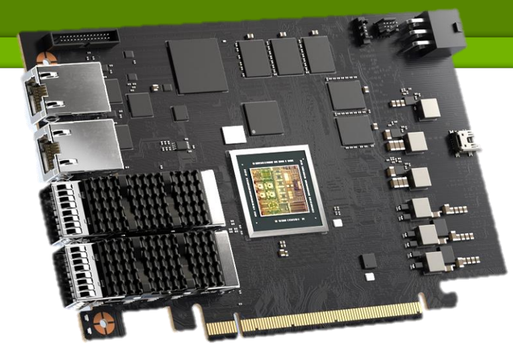


NVIDIA DOCA

広範な DPU パートナーエコシステムの実現

- BlueField DPU のソフトウェア アプリケーション フレームワーク
- GPU には CUDA、DPU には DOCA
- 開発したソフト、習得技術は将来の DPU 上でも動作が可能
- 認定リファレンス アプリケーション, API & パートナー ソリューション
- 業界およびワークロード全体にわたる豊富なパートナー エコシステム





NVIDIA DOCA – クラウド データセンターの SDK

ソフトウェアで定義されハードウェアで高速化されるインフラとアプリケーション

108
New DOCA APIs

1400
DOCA Developers

Services

DOCA 1.2
Zero Trust Security Framework

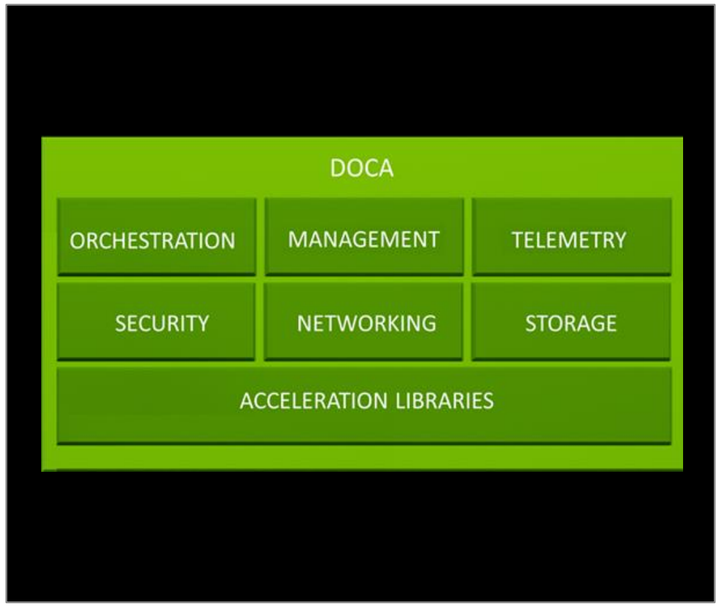
Load Balancers	Deep Packet Inspection	Intrusion Detection
Firewall	Telemetry	Authentication and Security Groups

DOCA 1.0
Accelerated Secure Bare Metal Cloud

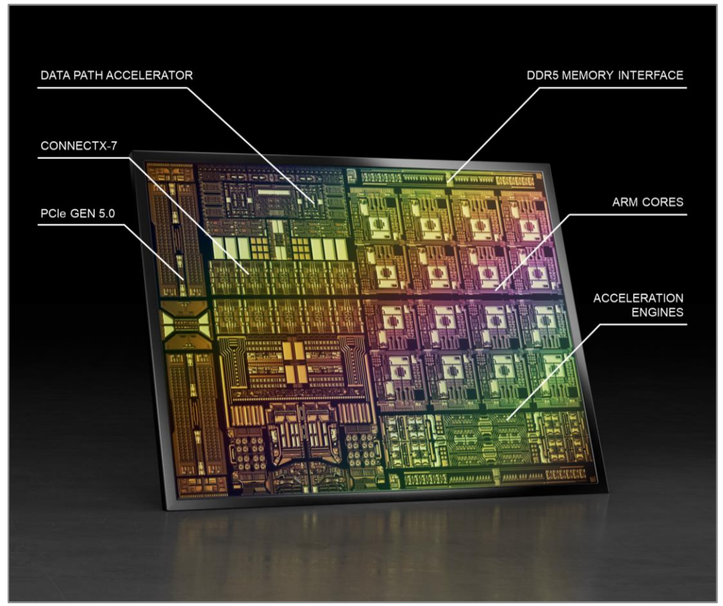
Software Defined Networking	Crypto	Storage Acceleration
RegEx	De/Compression	RDMA

NVIDIA BlueField DPUs

これからの HPC と AI を支えるネットワーク技術



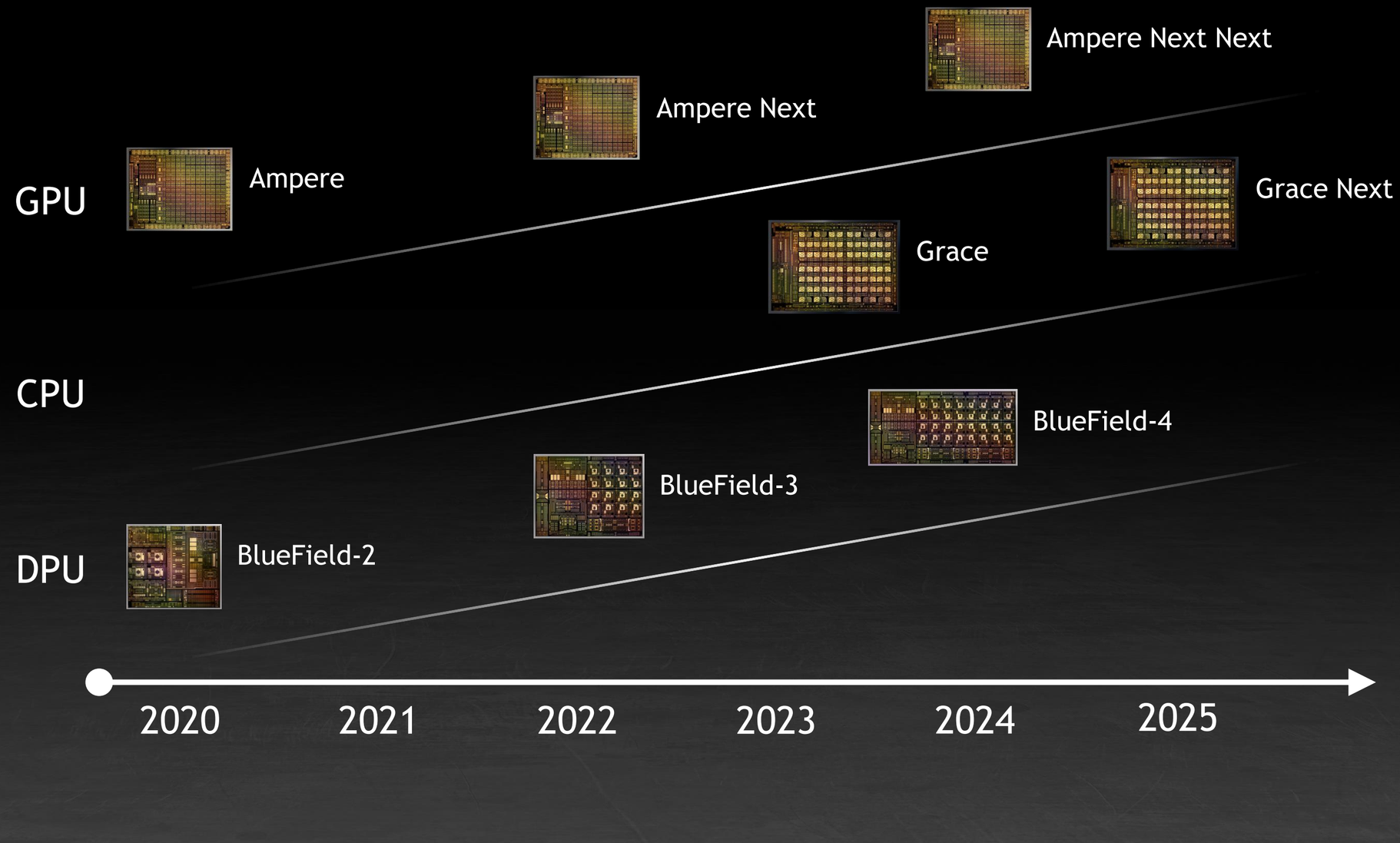
DOCA のパートナーエコシステム



次世代の 400Gbps DPU - Bluefield-3

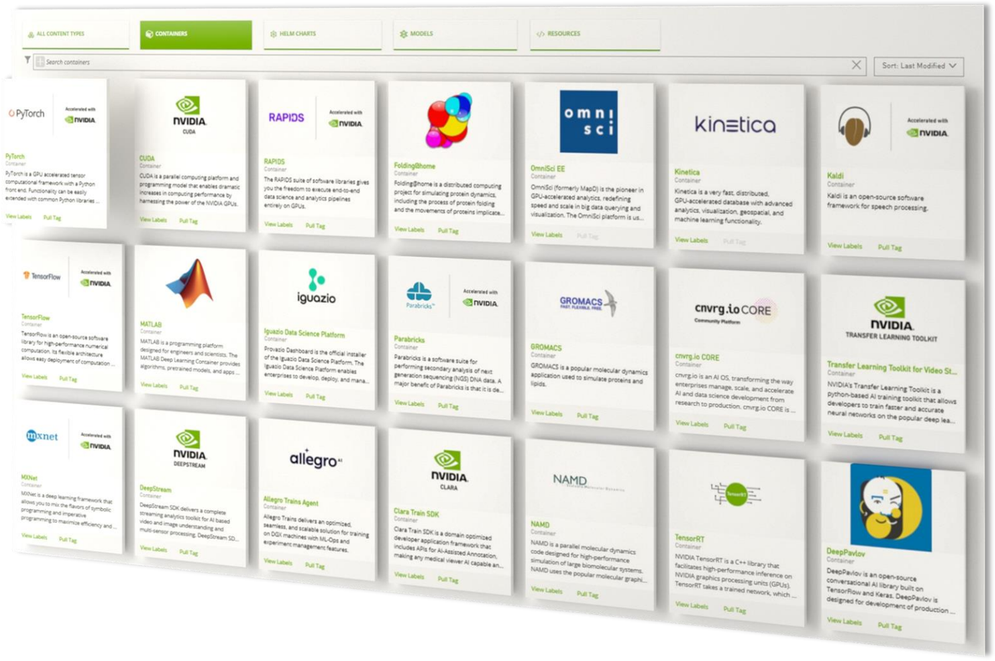
NVIDIA Quantum-2
NDR 400G InfiniBand

CPU/GPU/DPU のロードマップ



NGC

NGC コンテナが AI 開発を効率化



企業ユース向けの品質

セキュリティ検査

信頼性テスト

エンタープライズ サポート

パフォーマンス最適化

スケーラブル

月例更新

最適化による性能向上

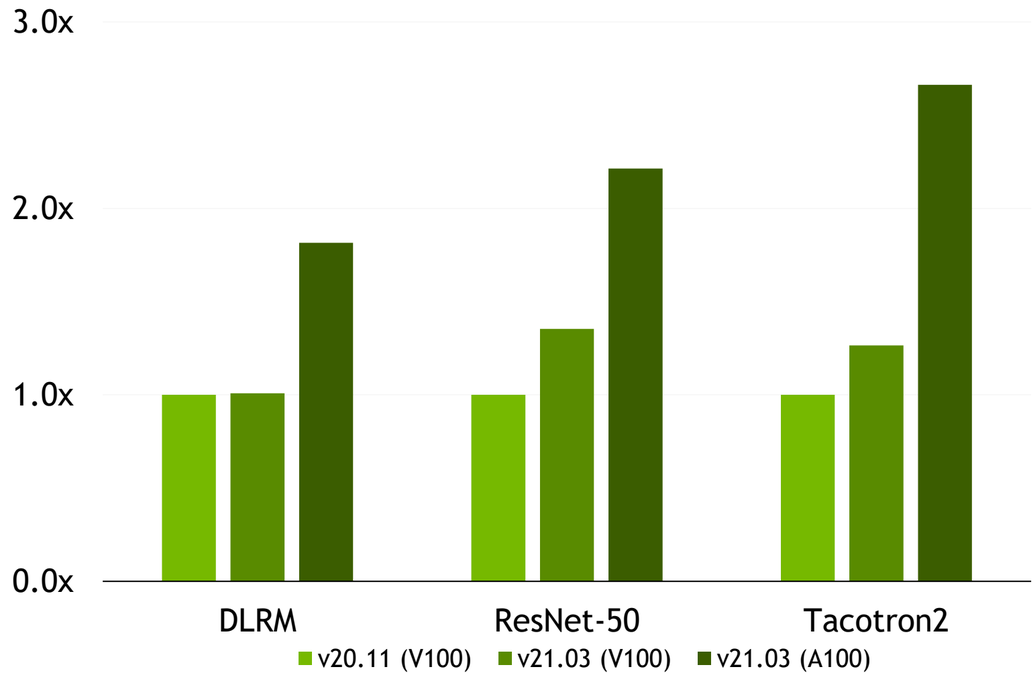
様々な環境で利用可能

Docker | cri-o | containerd | Singularity

ベアメタル、仮想マシン、Kubernetes

オンプレミス、クラウド、エッジ環境

常に最高の性能を



企業ユース向けの品質

セキュリティ検査

信頼性テスト

エンタープライズ サポート

パフォーマンス最適化

スケーラブル

月例更新

最適化による性能向上

様々な環境で利用可能

Docker | cri-o | containerd | Singularity

ベアメタル、仮想マシン、Kubernetes

オンプレミス、クラウド、エッジ環境

Tacotron2 Training with PyTorch container on 8x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size: 104 (V100). 128 (A100). DLRM training with PyTorch on 1x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size: 32,768. ResNet-50 v1.5 training with TensorFlow on 8x V100 (32GB) & A100 (40GB). Mixed Precision. Batch size: 256

NGC プライベート レジストリ

安全で迅速なコラボレーションを促進

Build and Secure
- コンテナイメージのセキュリティを確保するために既知の脆弱性に対する自動的なスキャンを実施

Share
- ユーザー自身が作成したコンテナイメージや Helm チャート、学習済みモデル、モデルスクリプトを組織内で安全に共有

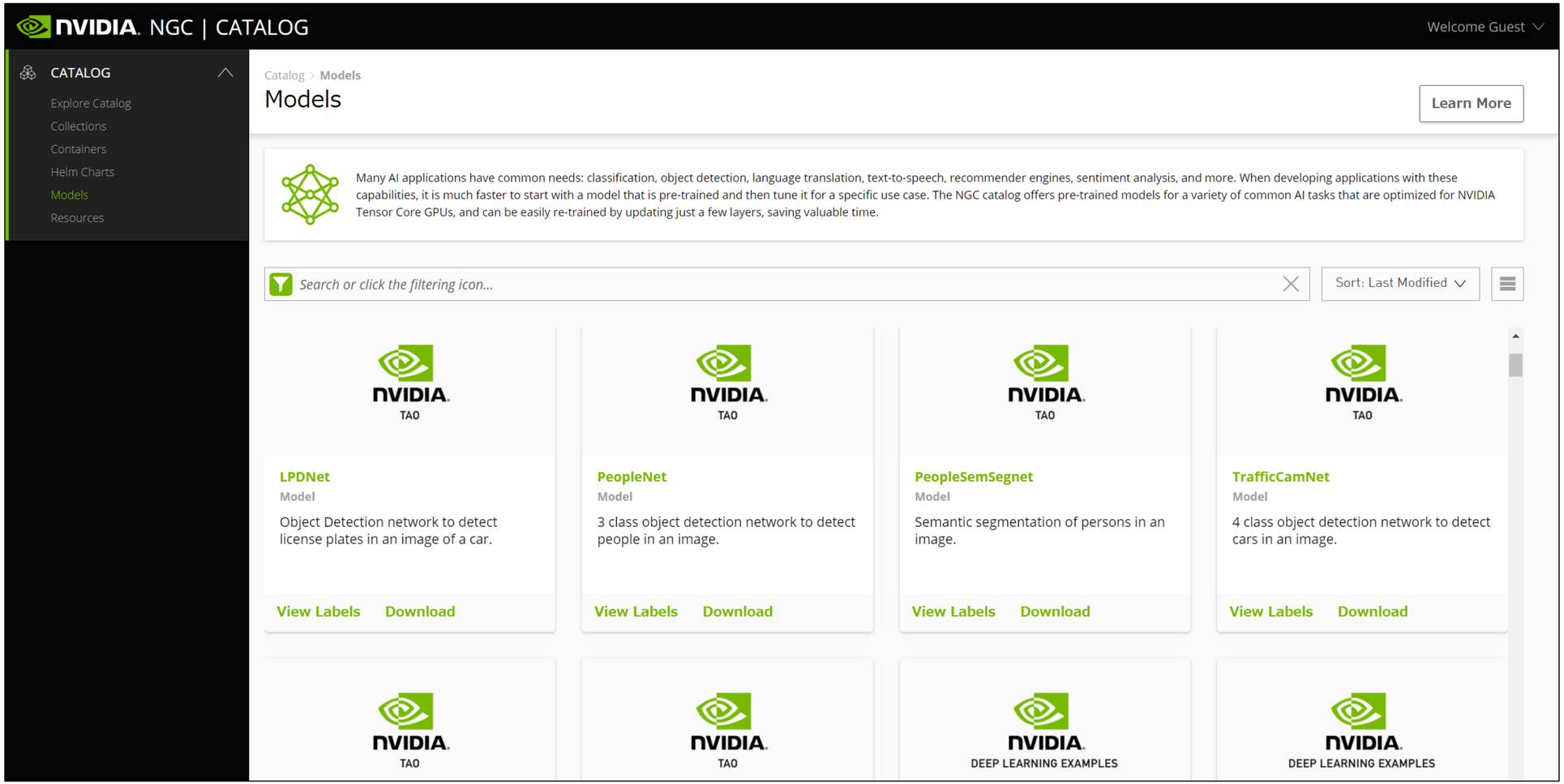
Manage
- モデルのバージョン管理が可能

Protect your IP
- 安全なストレージ、ユーザーとチームの管理、API キーの管理

[プライベートレジストリのドキュメントはこちら]

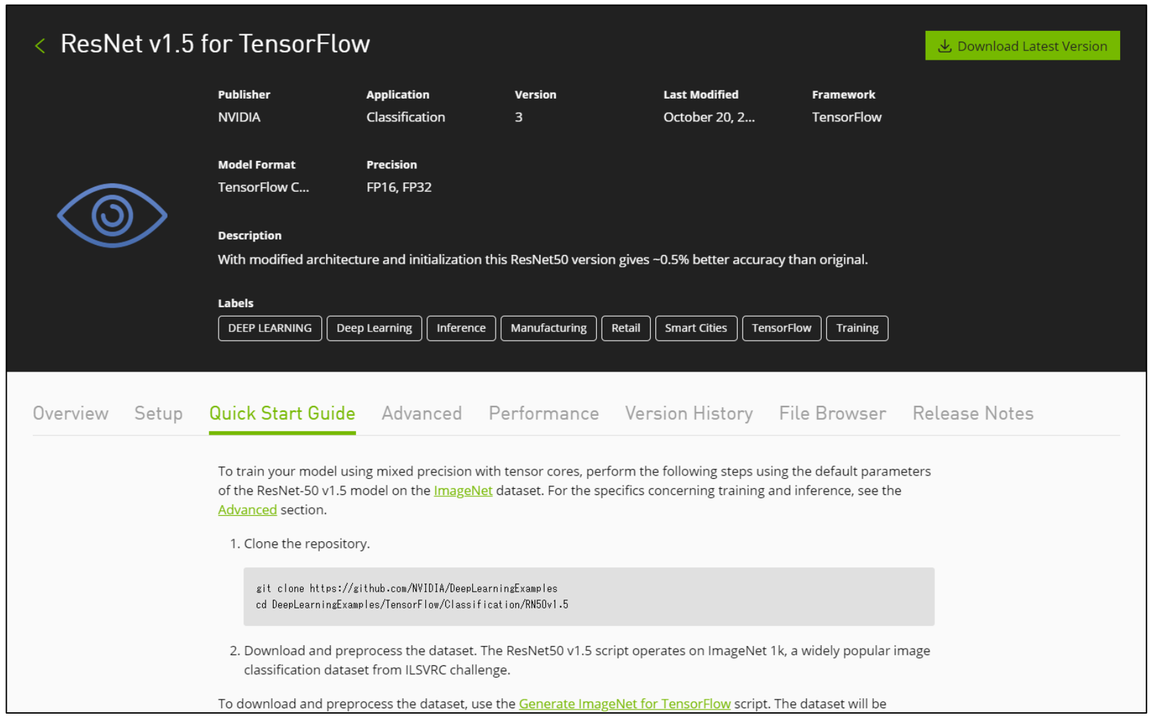
DGX, EGX, NGC サポートサービスのユーザー

トレーニング済みモデル



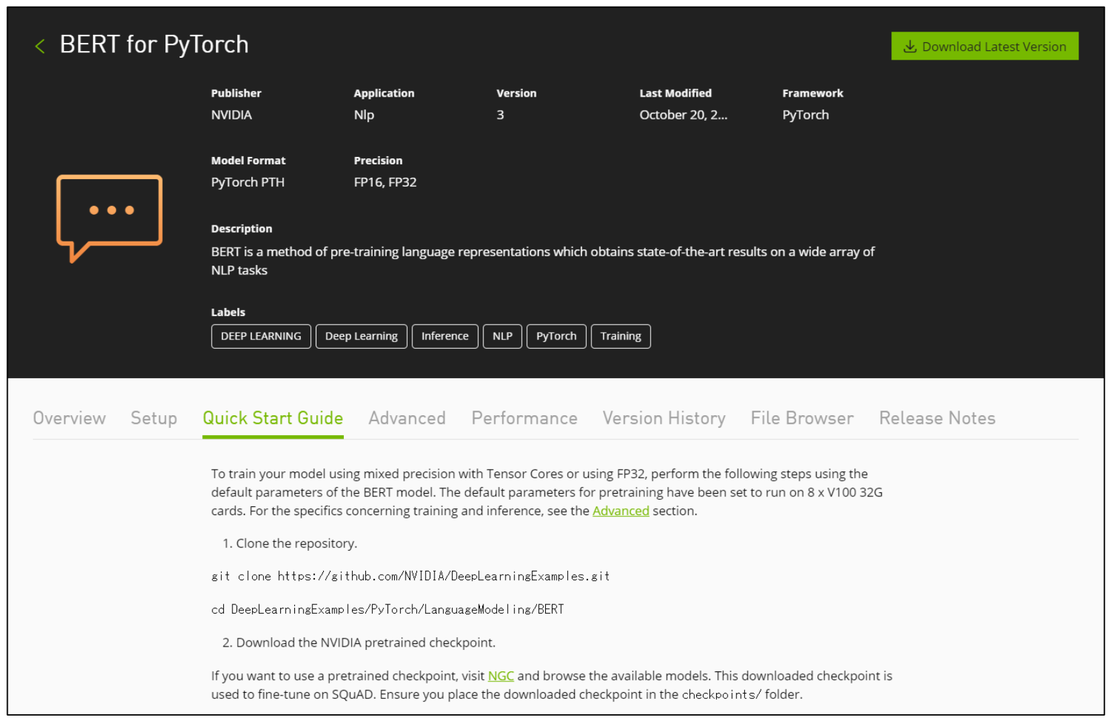
NGC のモデルスクリプトを活用

ResNet-50 for TensorFlow



https://ngc.nvidia.com/catalog/model-scripts/nvidia:resnet_50_v1_5_for_tensorflow/quickStartGuide

BERT for PyTorch



https://ngc.nvidia.com/catalog/model-scripts/nvidia:bert_for_pytorch/quickStartGuide

NVIDIA コンピューティング 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 | 医療診断 | 異常検知 | 品質保証 | 自動化 | マーケティング | シミュレーション | ++

NGC

ソフトウェア ハブ

HELM

コンテナイメージ

学習済みモデル

ソフトウェア開発キット

アプリケーション
フレームワーク

スマートシティ **Metropolis** | 対話型 AI **Riva** | 自動運転 **Drive** | レコメンデーション システム **Merlin** | ヘルスケア **Clara** | ++ …

開発者
ツールキット

ML & DATA ANALYTICS: RAPIDS, dmlc XGBoost

AI TRAINING & INFERENCE: TensorFlow, TensorRT, PYTORCH, mxnet

HIGH PERFORMANCE COMPUTING: *NVIDIA HPC SDK*

RENDERING & VISUALIZATION: *IndeX*, *OptiX*

高速化
ライブラリ

COMPUTE: *CUDA-X*

NETWORKING, STORAGE & SECURITY: *DOCA*, *MAGNUM IO*

運用

VIRTUAL GPU SOFTWARE

TRITON INFERENCE SERVER

FLEET COMMAND

NVIDIA GPU Operator

Red Hat

vmware

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

サーバー及びクラウド

DGX
HGX

ハイエンド

EGX

メインストリームとエッジ

クラウドのインスタンス

ハードウェア技術

GPU

NVSwitch

BlueField DPU

SMART NIC

Mellanox Switch

監視

DCGM | Grafana

Prometheus | UFM

NVIDIA