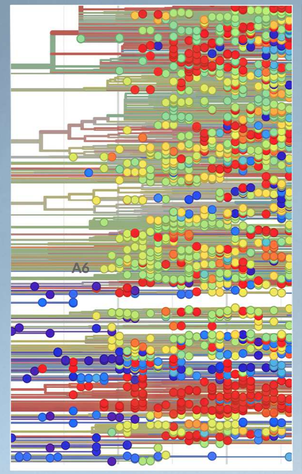
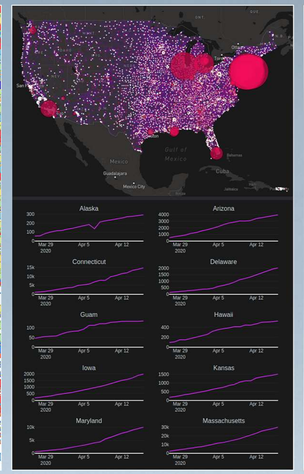

NVIDIA


THANK YOU
HEROES

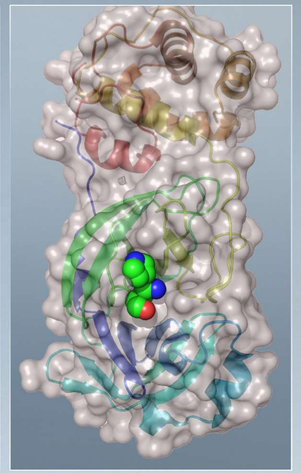
# NVIDIA の COVID-19 との戦い



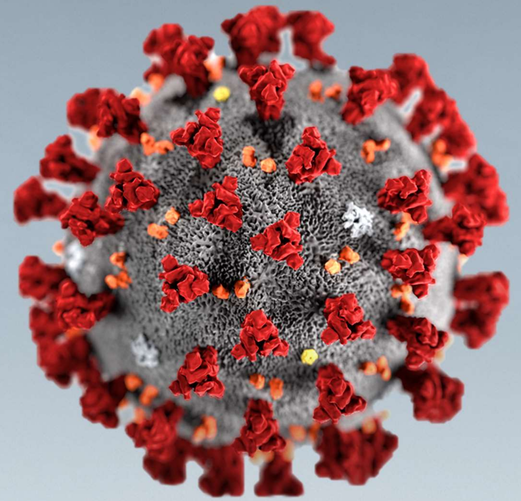

Oxford Nanopore
ウイルスのゲノムを
7 時間で解析



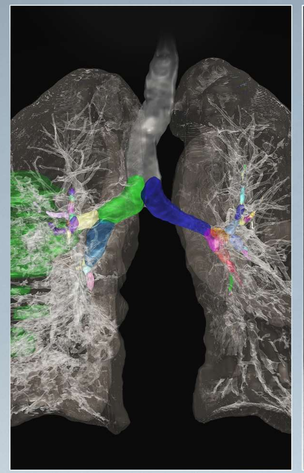

Plotly, NVIDIA
リアルタイムの
感染率分析



ORNL, Scripps
10 億の薬物化合物の
スクリーニングを
1 年から 1 日に短縮

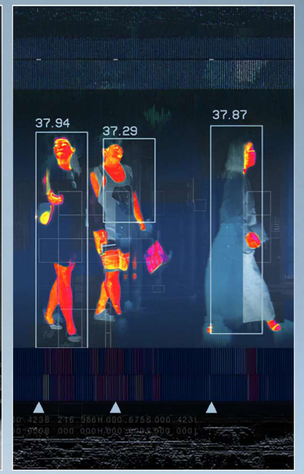


Structura, NIH, UT Austin
CryoSPARC
初のウイルス スパイク
プロテインの 3 次元構造



NIH, NVIDIA
AI による COVID-19 の分類

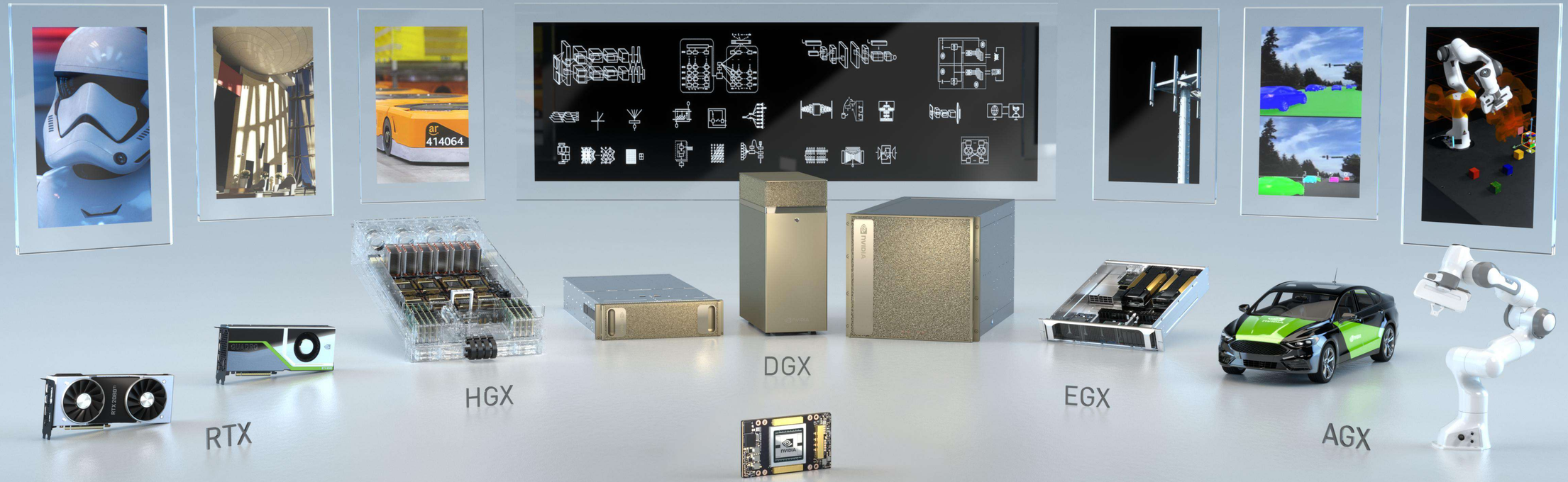

Kiwibot
ロボットによる医療品の配達



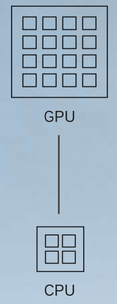

Whiteboard Coordinator
AI による体温上昇の
スクリーニング システム

封じ込め — 緩和 — 治療 — 追跡と監視

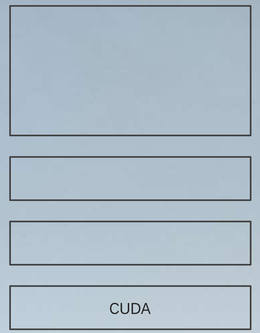
# 私たちの時代のダ ヴィンチのためのコンピューティング



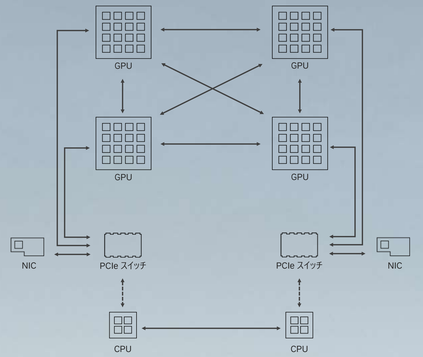
# アクセラレーテッド コンピューティングの 25 年



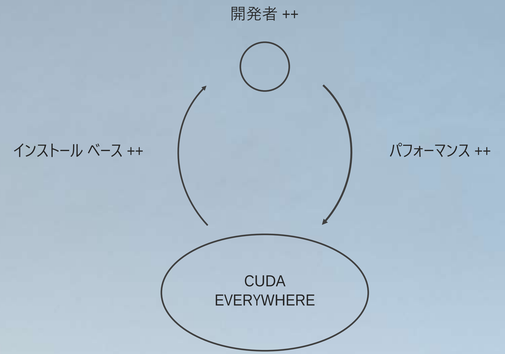

X-FACTOR スピードアップ




フル スタック

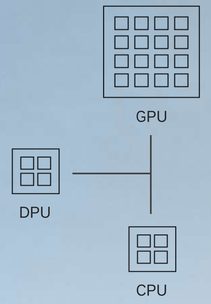



システム



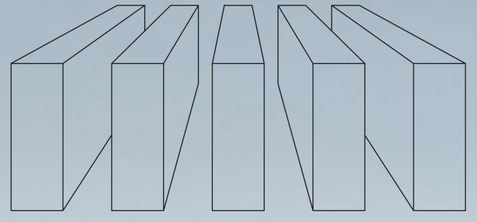

1 つのアーキテクチャ

# データセンター規模のアクセラレーテッド コンピューティング

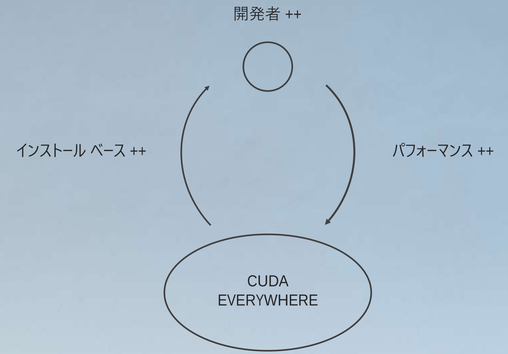



X-FACTOR スピードアップ

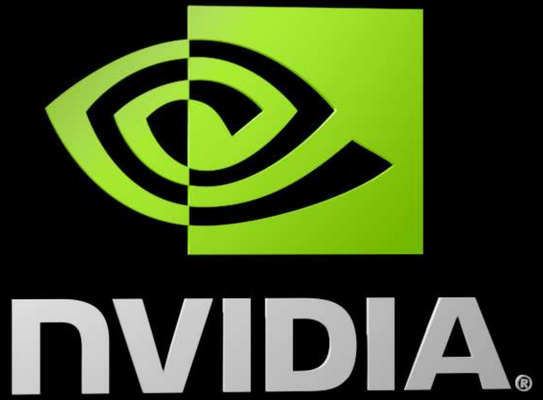

フル スタック



システム

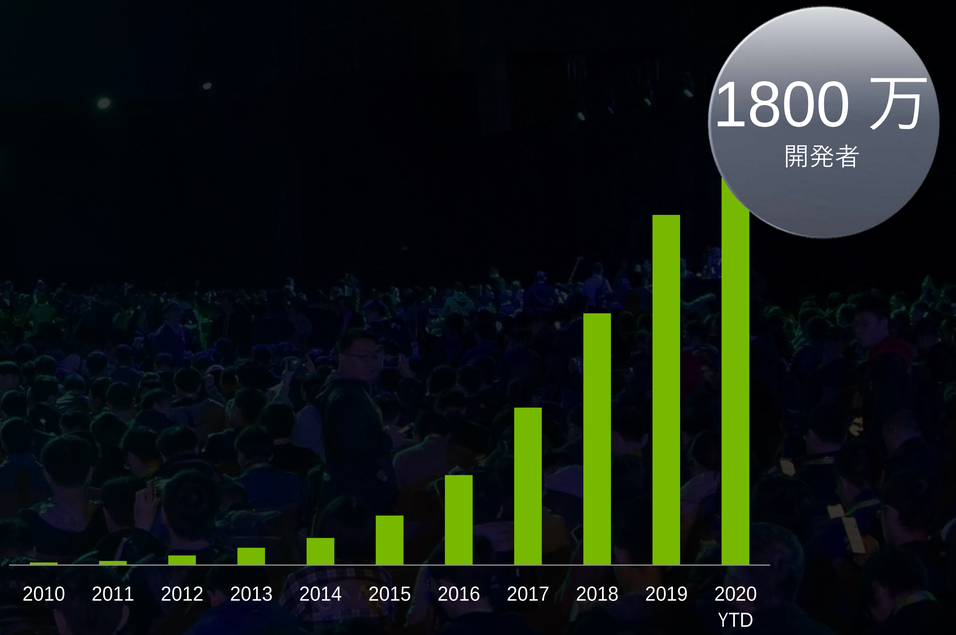



1 つのアーキテクチャ


NVIDIA


Mellanox
TECHNOLOGIES

# 開発者にとって素晴らしい年

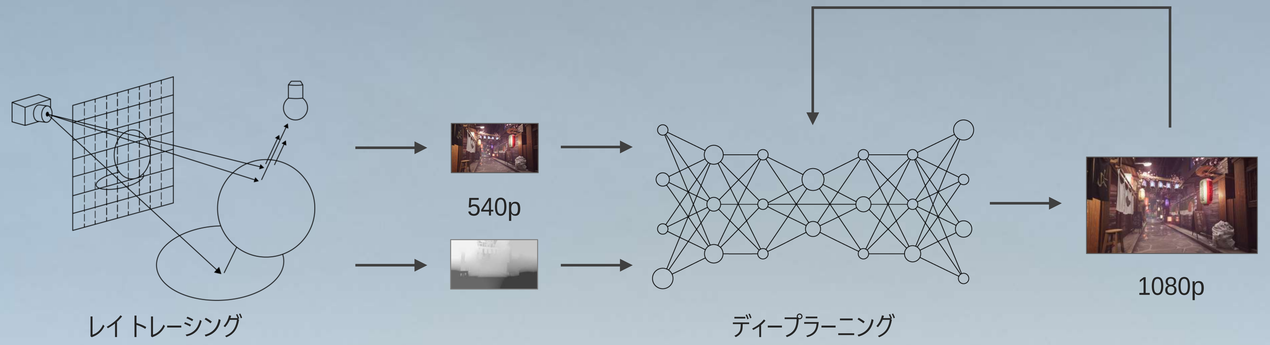



2018 年 8 月 13 日、SIGGRAPH で NVIDIA は RTX を発表、
コンピュータ グラフィックスの新時代の幕が上がる 。

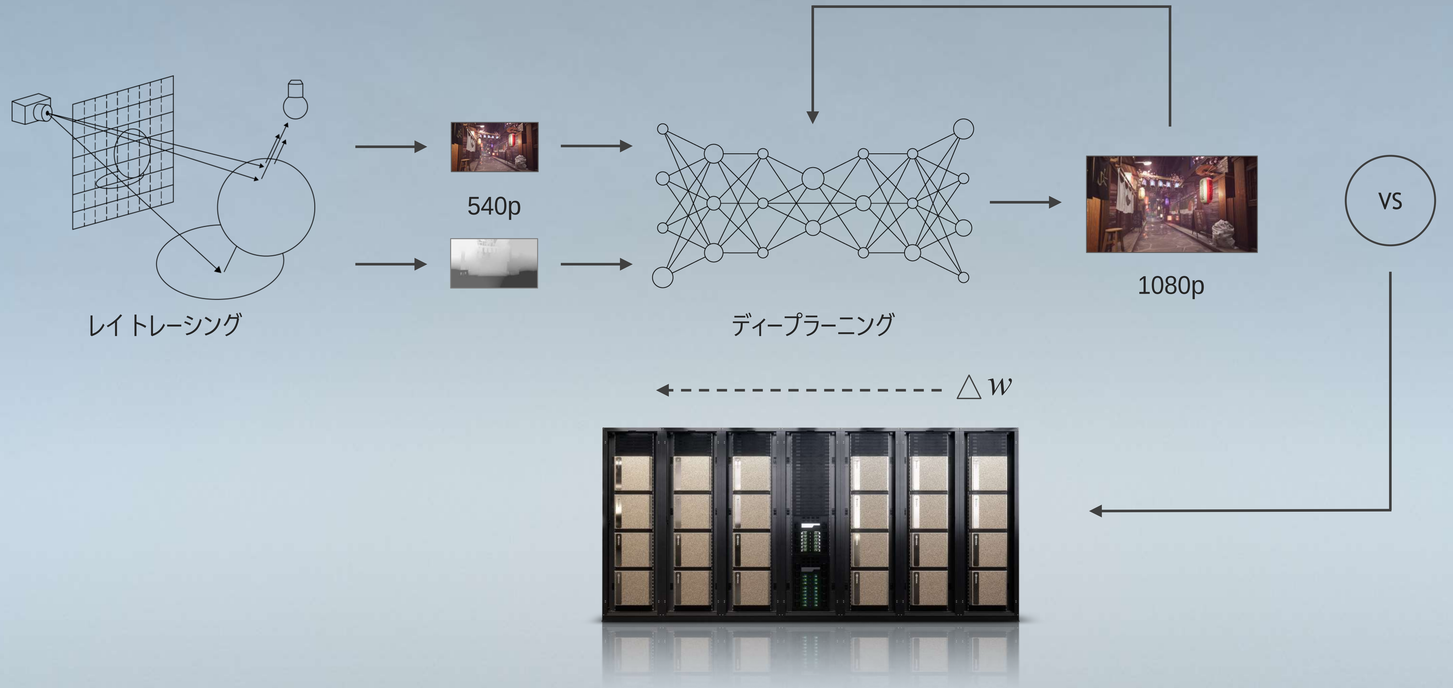
「30 年以上前から約束されていたリアルタイム レイ トレーシングが、
予想より 10 年早くやって来ました。」

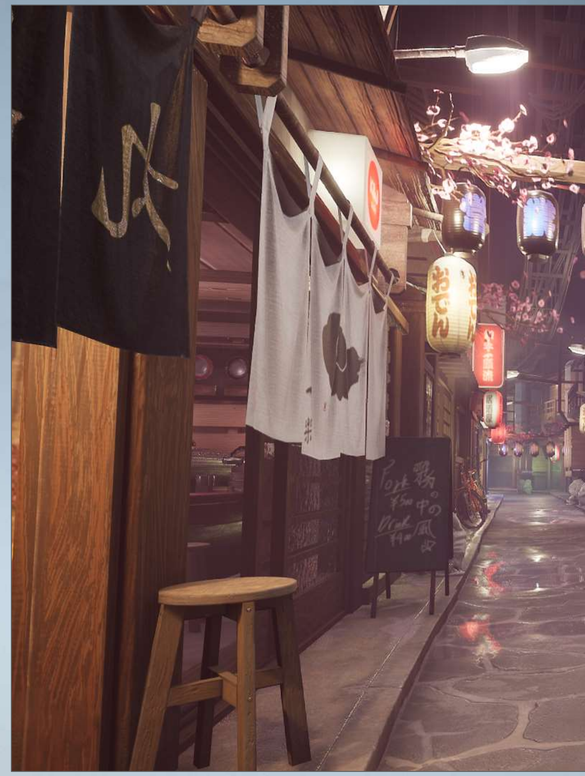
– Jon Peddie, JPR

# NVIDIA RTX
# コンピュータ グラフィックスの新時代 — レイ トレーシングと AI



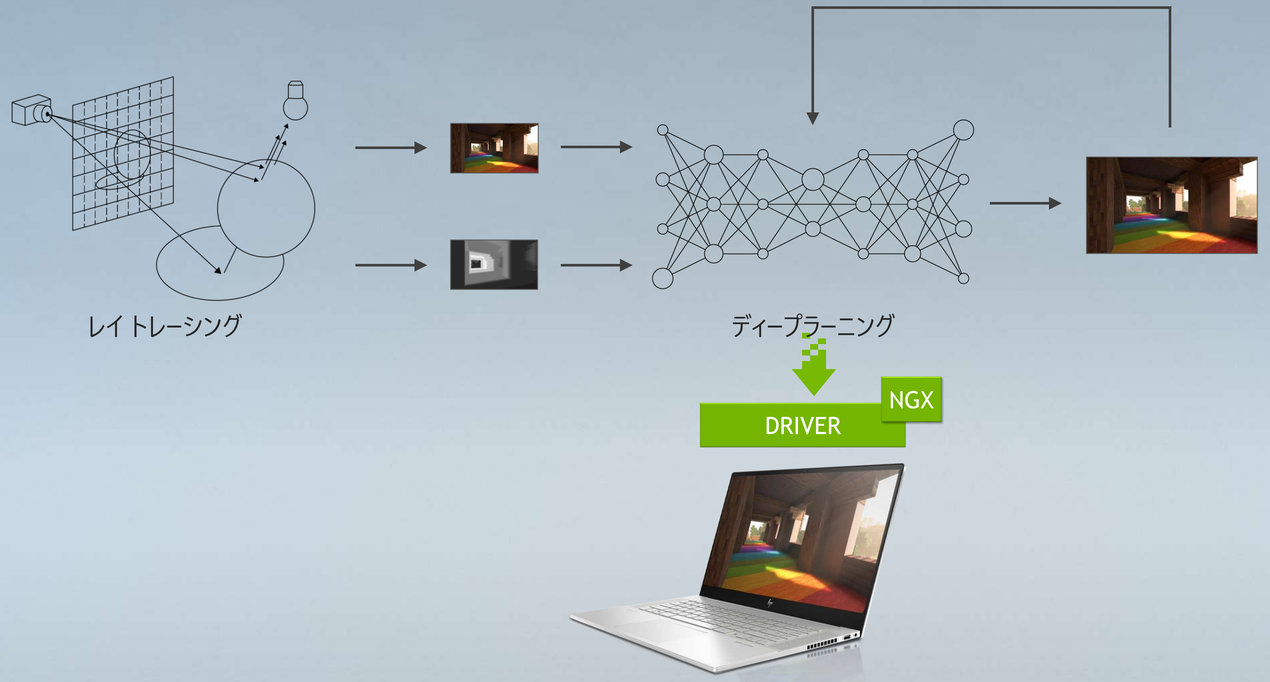
# NVIDIA RTX
# コンピュータ グラフィックスの新時代 — レイ トレーシングと AI



スーパーコンピューターでレンダーされた 16K Ground Truth

# NVIDIA RTX
# コンピュータ グラフィックスの新時代 — レイ トレーシングと AI




Ground
Truth 16K


Native
720p


DLSS 1.0
720p > 1080p


DLSS 2.0
720p > 1080p


Native
1080p


Ground
Truth 16K


Native
540p


DLSS 2.0
540p > 1080p

MINECRAFT


MOJANG

「ゲームチェンジャー」

– Digital Foundry

「なんてゴージャスなんだ」

– PCWorld

「素晴らしいに他ならない」

– IGN

「見事だ」

– PC Gamer

「驚いて開いた口が塞がらない」

– Trusted Reviews

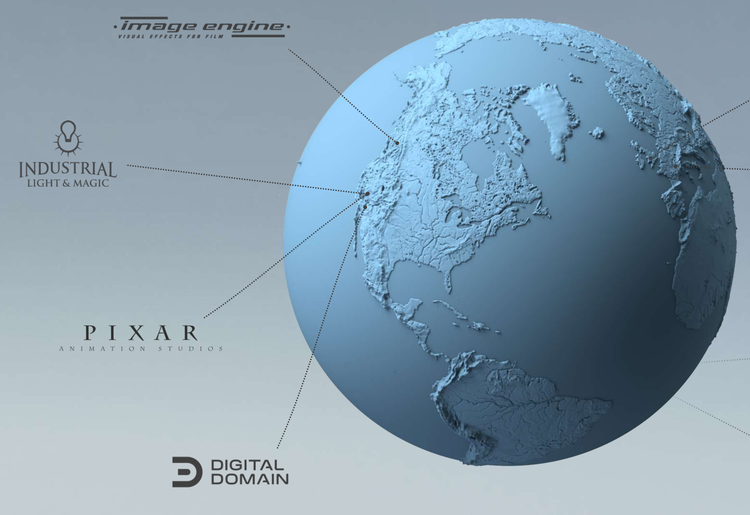
# 3D は非常に複雑

さまざまなツールと巨大なデータ セット

多様な専門家による大規模なチーム

複数のロケーションとスタジオ

高価

# NVIDIA OMNIVERSE

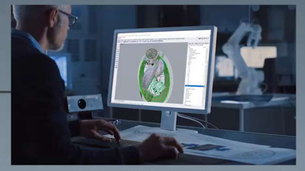
デザイン ワークフローのコラボレーション プラットフォーム

USD (Universal Scene Description) ベースに構築

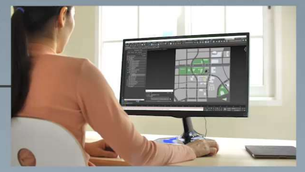
マテリアルと物理をインタラクティブ レンダラーに搭載

PC と Linux に対応、ストリーミング端末には Mac、Android が対応

Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS

RHINO



MAX

REVIT

AR

Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS


HOORKA
OIL COLOUR

# 発表:
# リモート コラボレーションに最適化された NVIDIA RTX サーバー

Omniverse によるデザイン ワークフロー コラボレーション
完全なレイ トレーシングによるグローバル イルミネーションに対応したインタラクティブ プロダクション レンダリング
デザインとシミュレーションで検証された Quadro 仮想ワークステーション

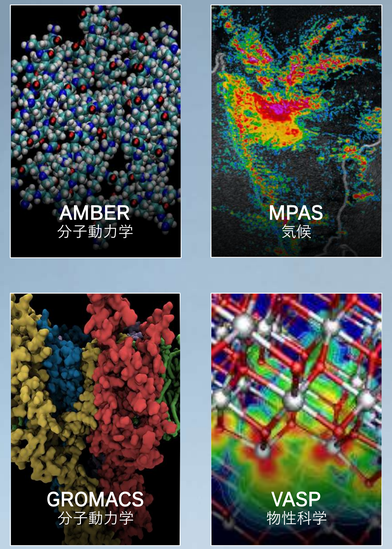
NVIDIA RTX **サーバーは出荷を開始**

BOXX DELL Technologies

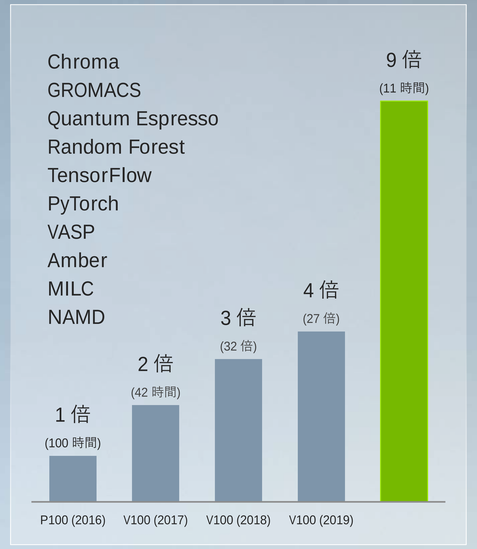
Hewlett Packard Enterprise SUPERMICRO

Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS

# NVIDIA HPC



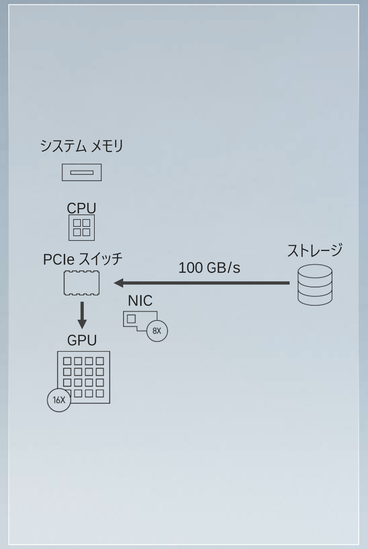

700 以上のアプリが
CUDA で高速化




9 倍のパフォーマンス
過去 4 年で

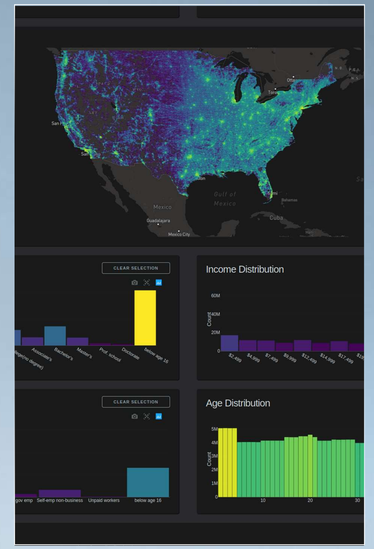
HPE Apollo70
Marvell ThunderX
ARM CPU

CUDA ON ARM



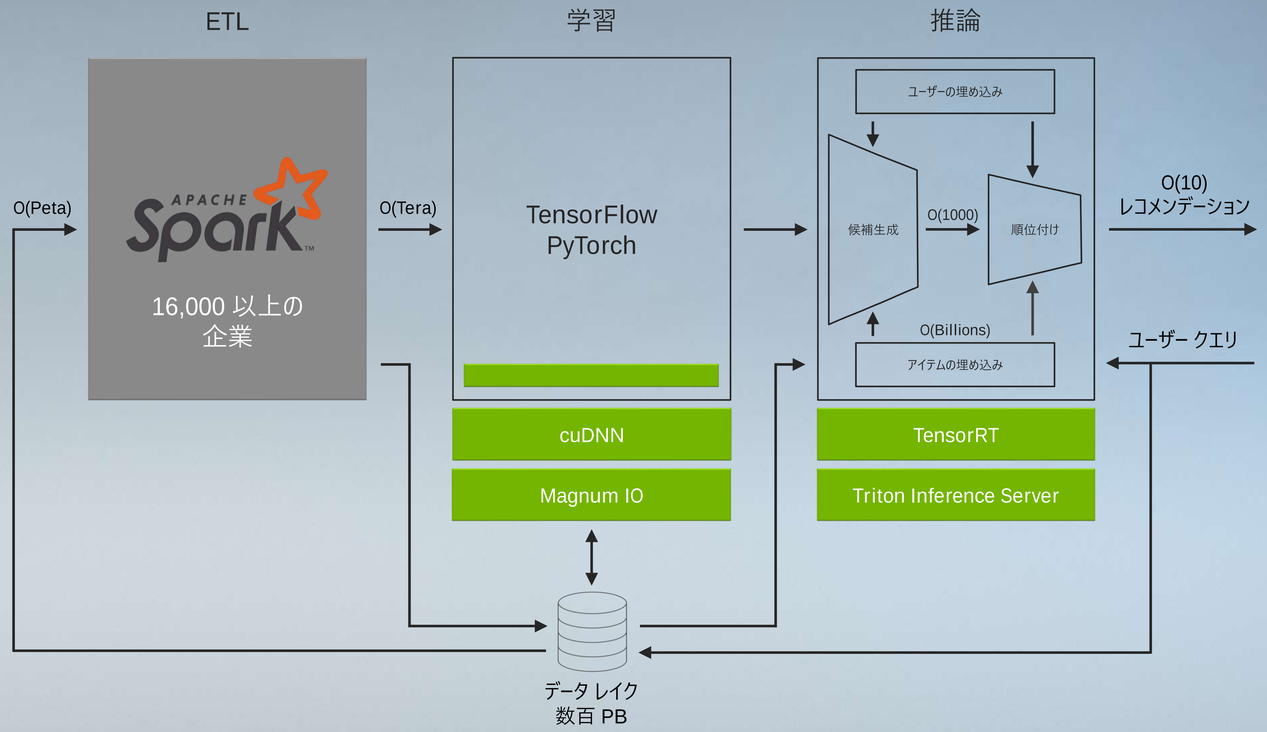

NVIDIA MAGNUM-IO
I/O アクセラレーション

NVIDIA PARABRICKS
ゲノミクス

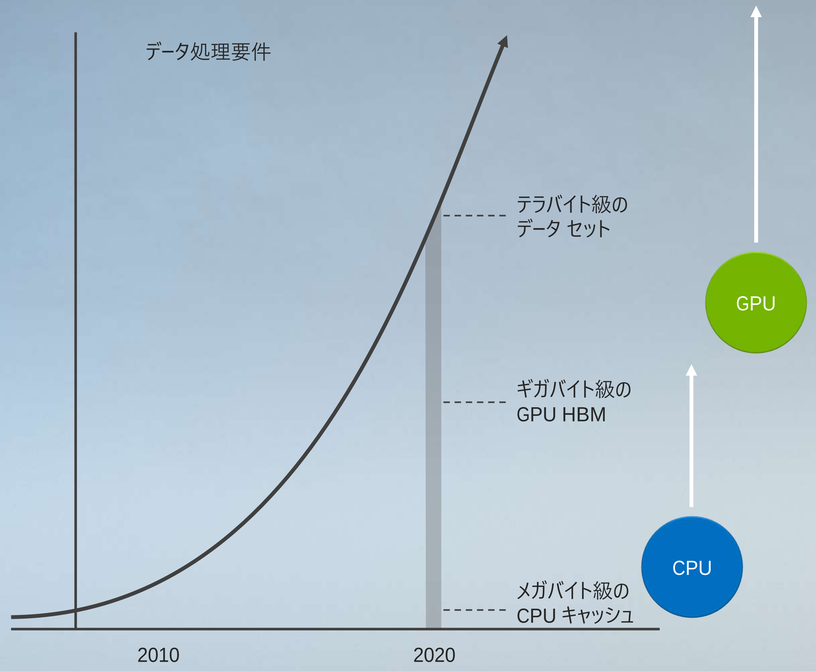


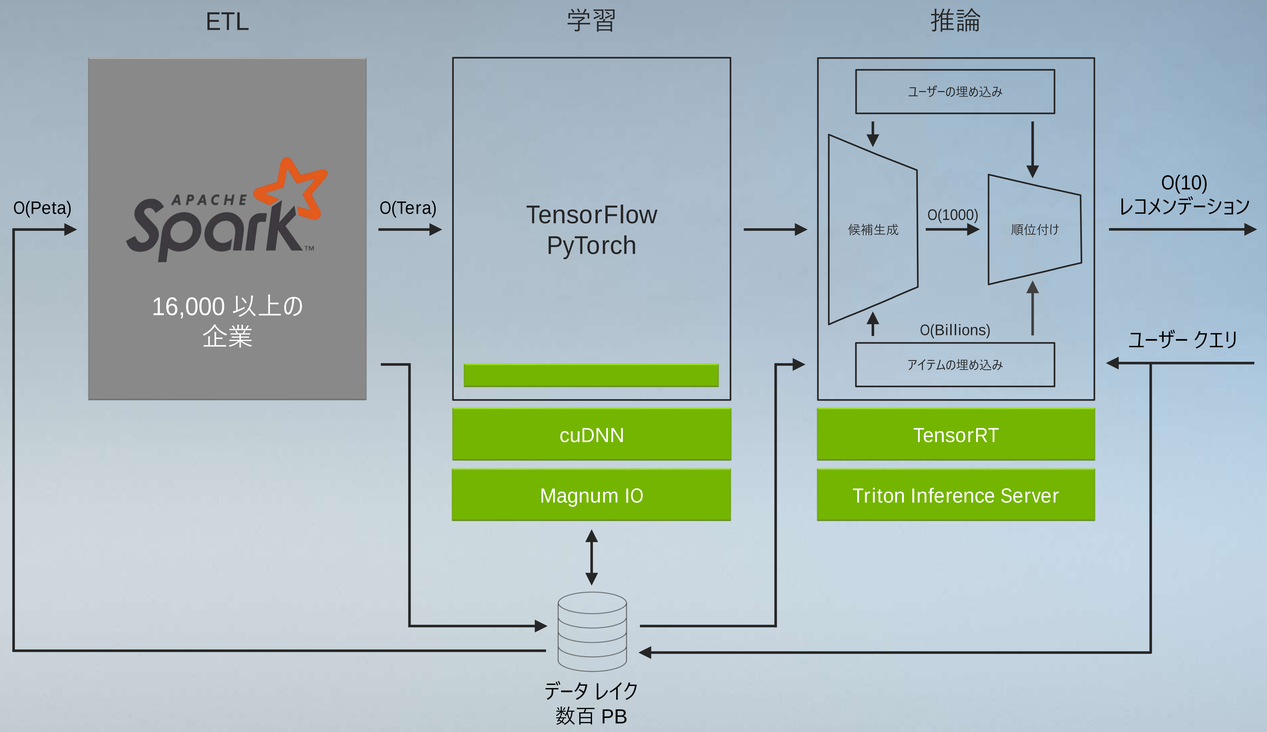

NVIDIA RAPIDS
データ分析

# 機械学習パイプラインは HPC の課題

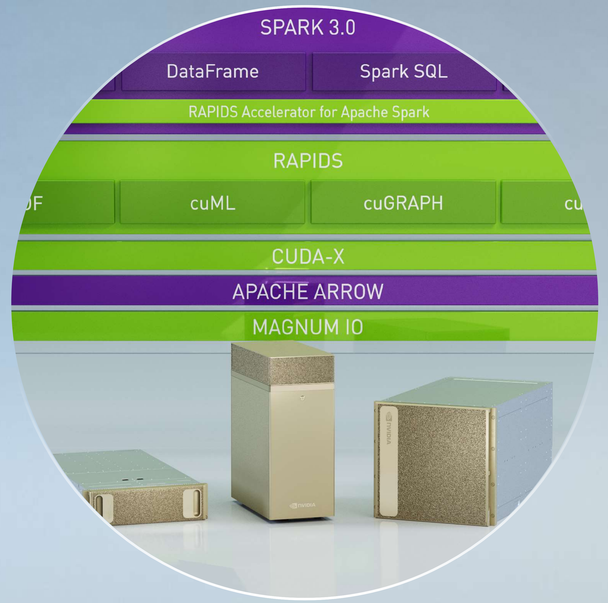


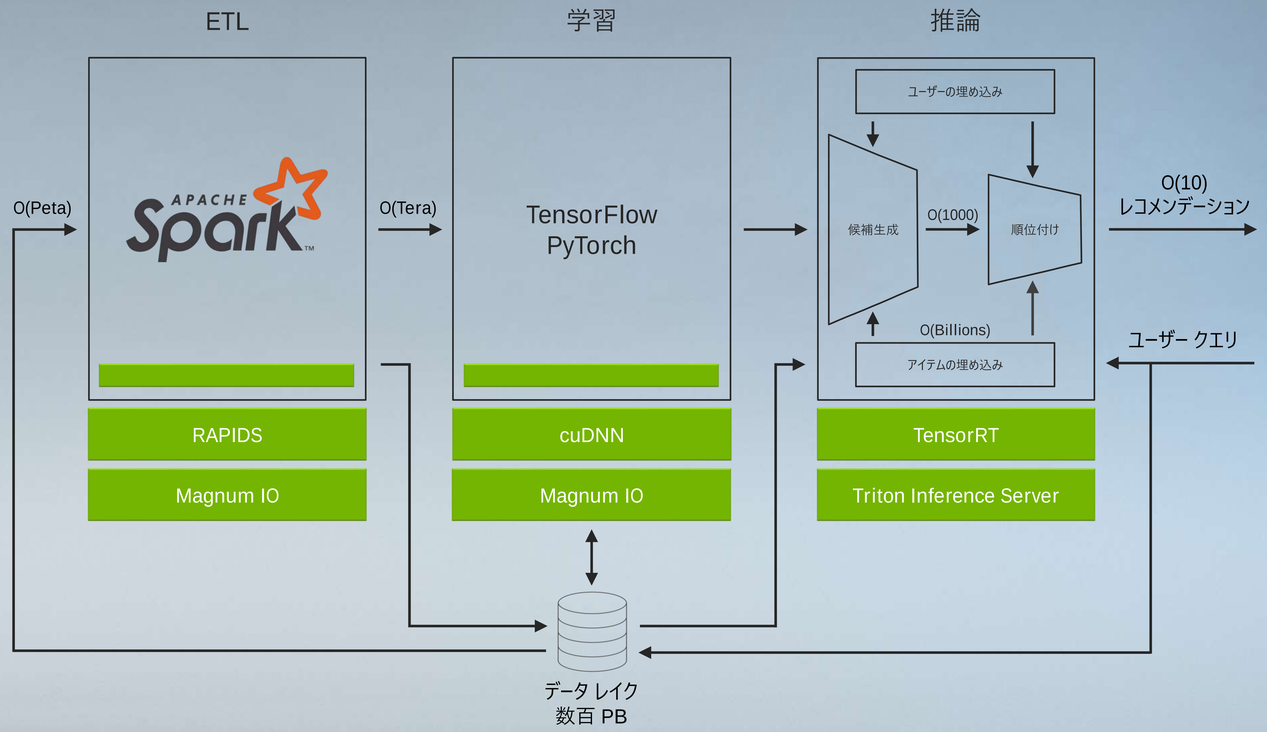
# 機械学習のデータが飛躍的に増大



# 発表:
# NVIDIA が SPARK 3.0 を高速化

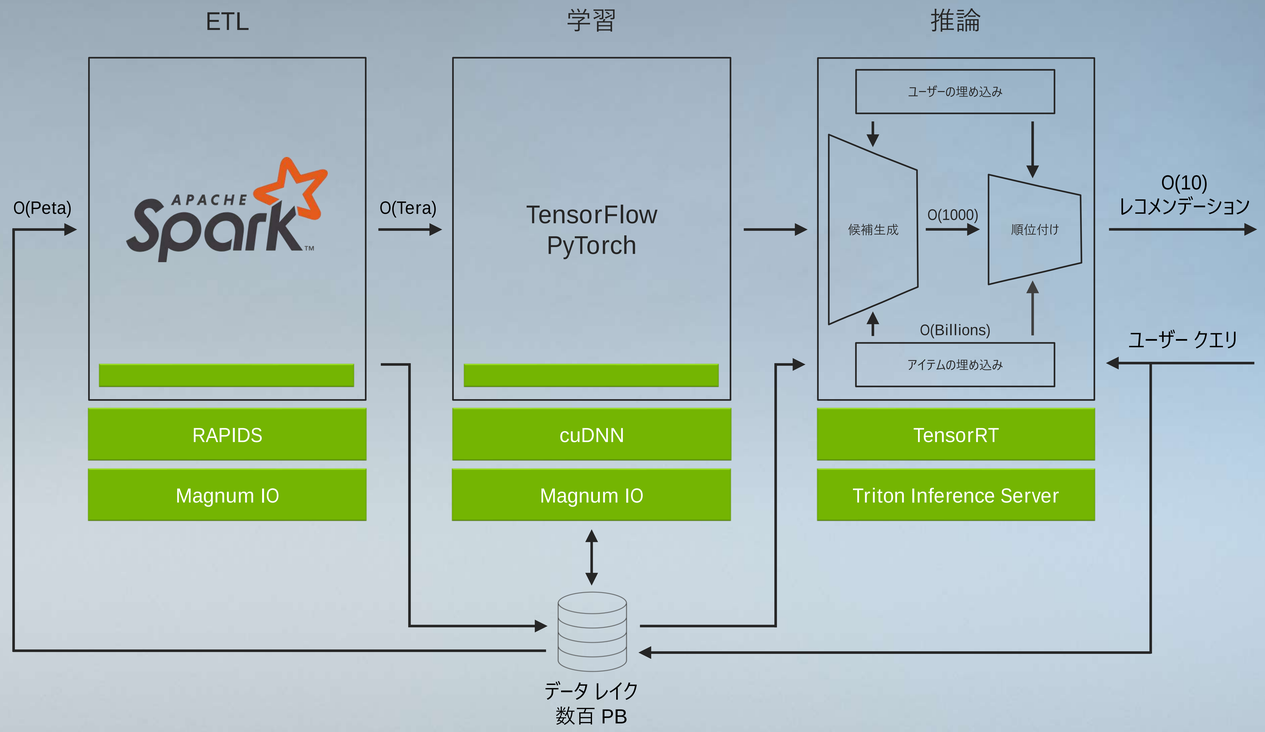


# SPARK 3.0 は最先端技術の基に構築
# RAPIDS が ETL ベンチマークを打ち破る


TPCx-BB @ SF 10K で 17 GB/秒のスループット

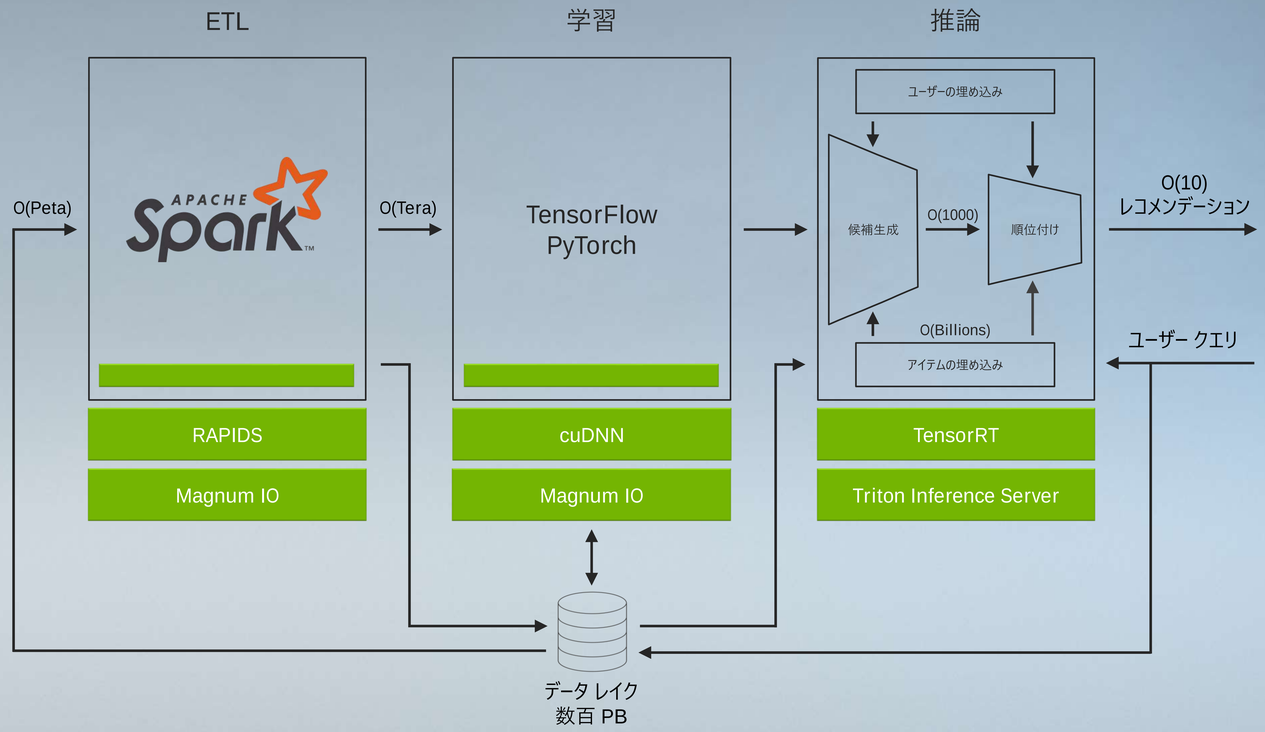
$1M | 18 2U CPU システム | 2 ラック | 16 kW



# SPARK 3.0 は最先端技術の基に構築
# RAPIDS が ETL ベンチマークを打ち破る


RAPIDS を実装した TPCx-BB @ SF 10K で 163 GB/秒のスループット

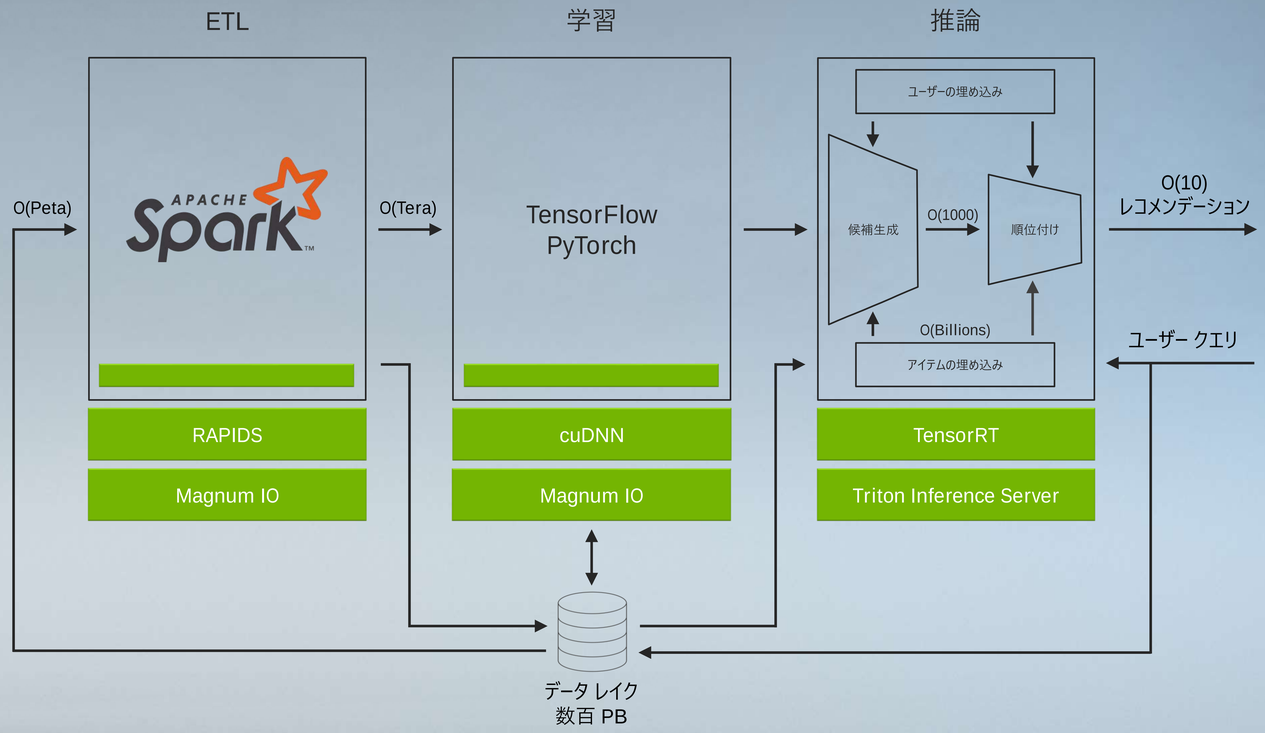
$2M | 16 台の DGX-1 | 2 ラック | 56 kW



# SPARK 3.0 は最先端技術の基に構築
# RAPIDS が ETL ベンチマークを打ち破る


RAPIDS を実装した TPCx-BB @ SF 10K で 163 GB/秒のスループット

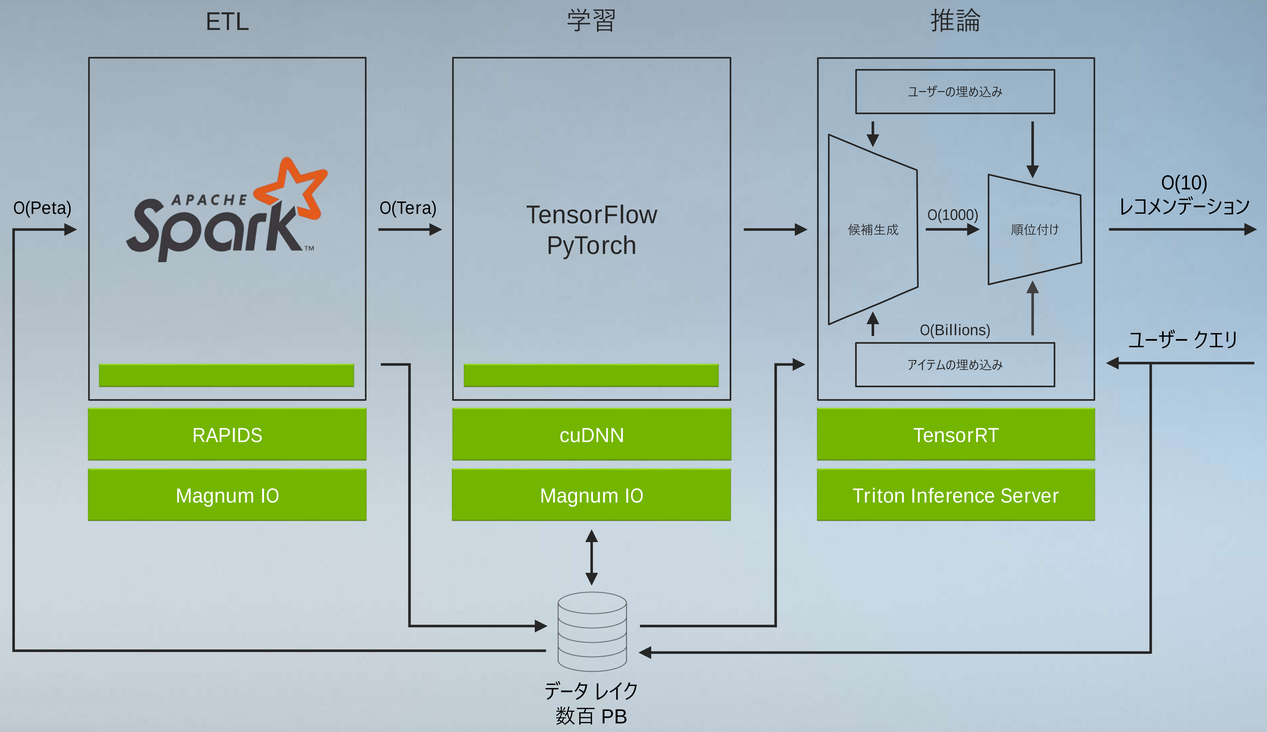
$2M | 16 台の DGX-1 | 2 ラック | 56 kW



# SPARK 3.0 は最先端技術の基に構築
# RAPIDS が ETL ベンチマークを打ち破る


TPCx-BB @ SF 10K で 163 GB/秒相当のスループット

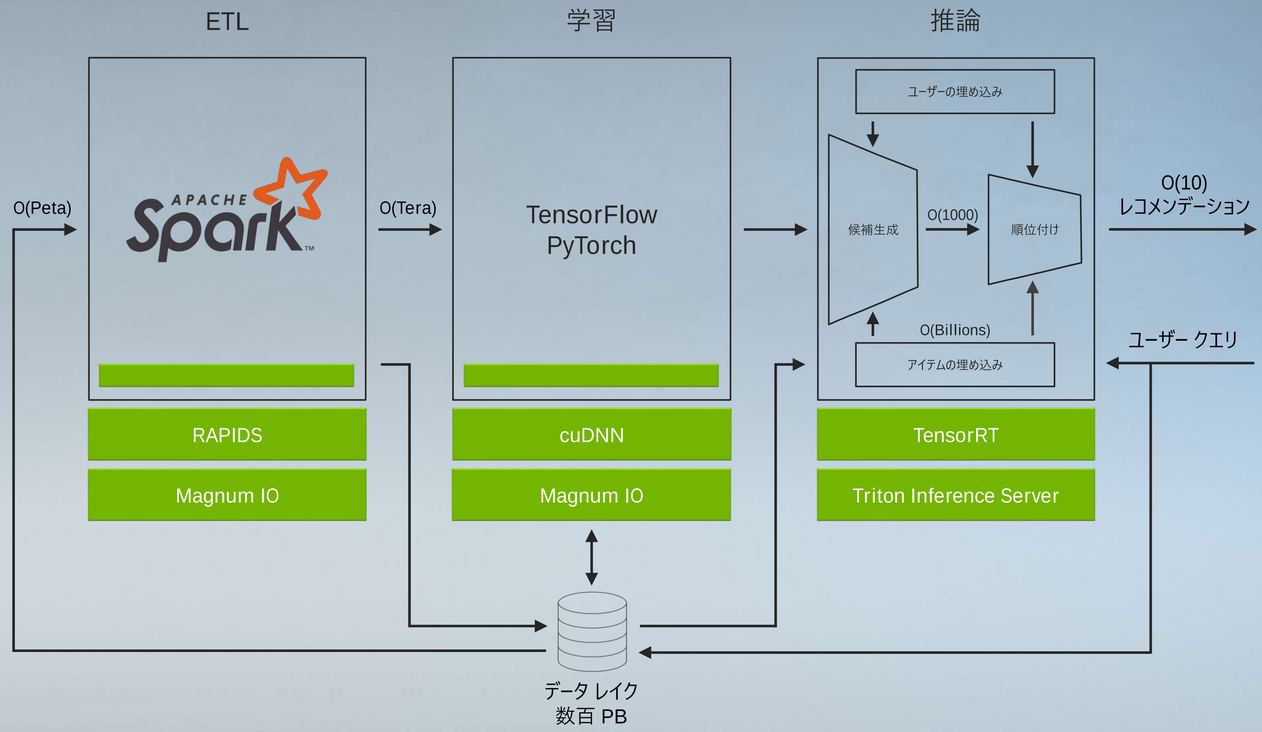
$10M | 167 台の 2U CPU サーバー | 11 ラック | 140 kW



# SPARK 3.0 は最先端技術の基に構築
# RAPIDS が ETL ベンチマークを打ち破る


RAPIDS を実装した TPCx-BB @ SF 10K で 163 GB/秒のスループット

$2M | 16 台の DGX-1 | 2 ラック | 56 kW

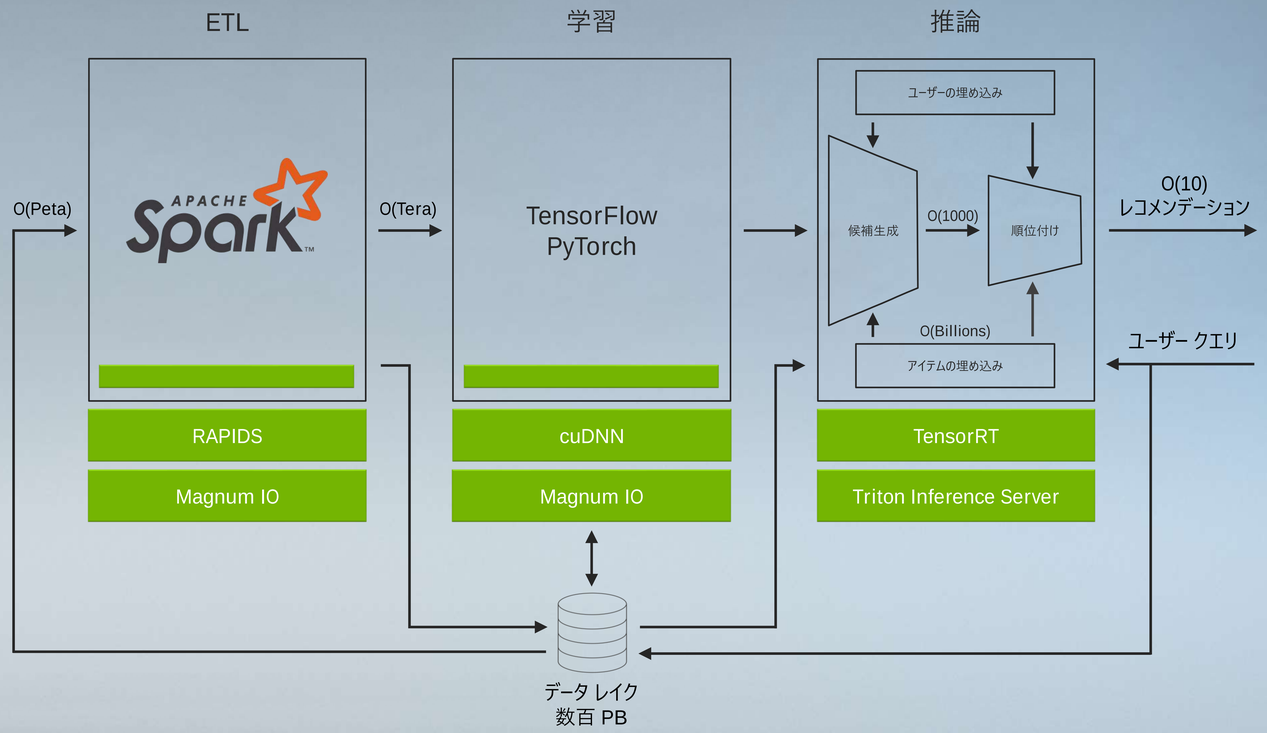


# 発表:
# DATABRICKS が NVIDIA で高速化

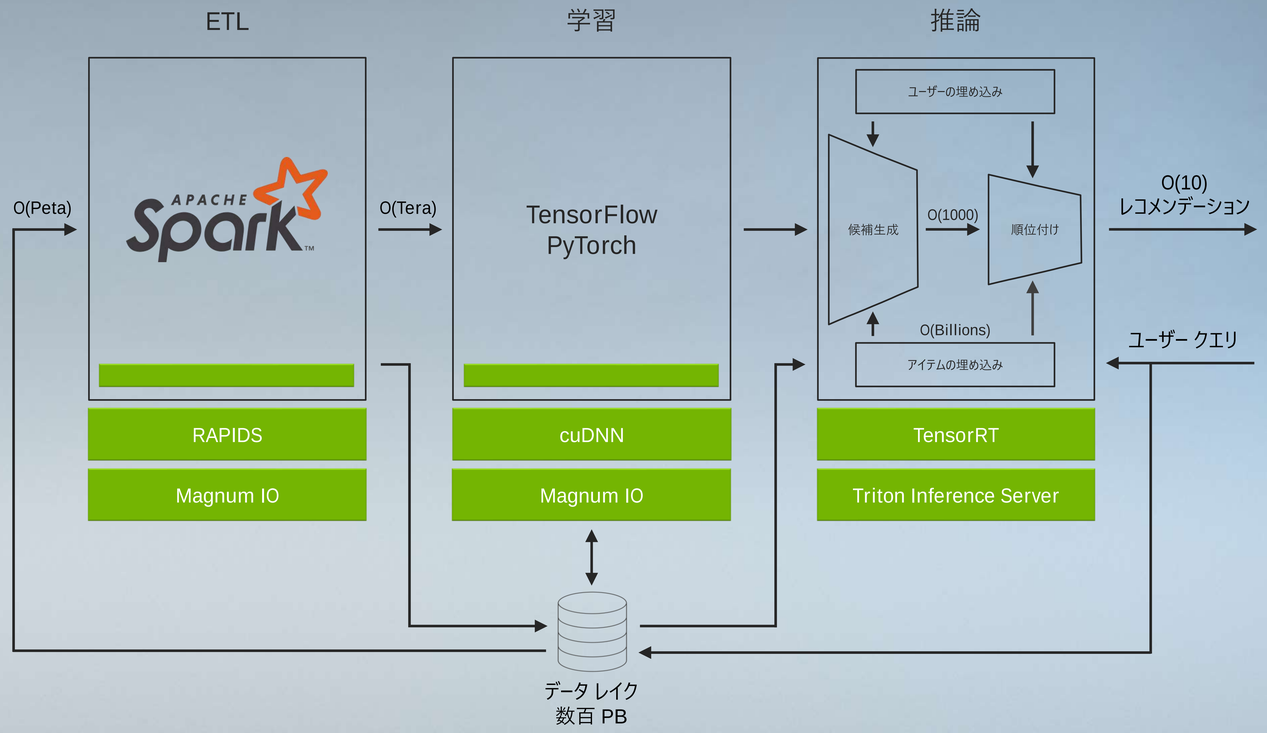
databricks

「これらの貢献は、Apache Spark 3.0 と Databricks を使用でデータ パイプライン、モデルの学習とスコアリングを高速化し、より多くのブレイクスルーと洞察の獲得につながります。」

— Matei Zaharia, original creator of Apache Spark and chief technologist at Databricks

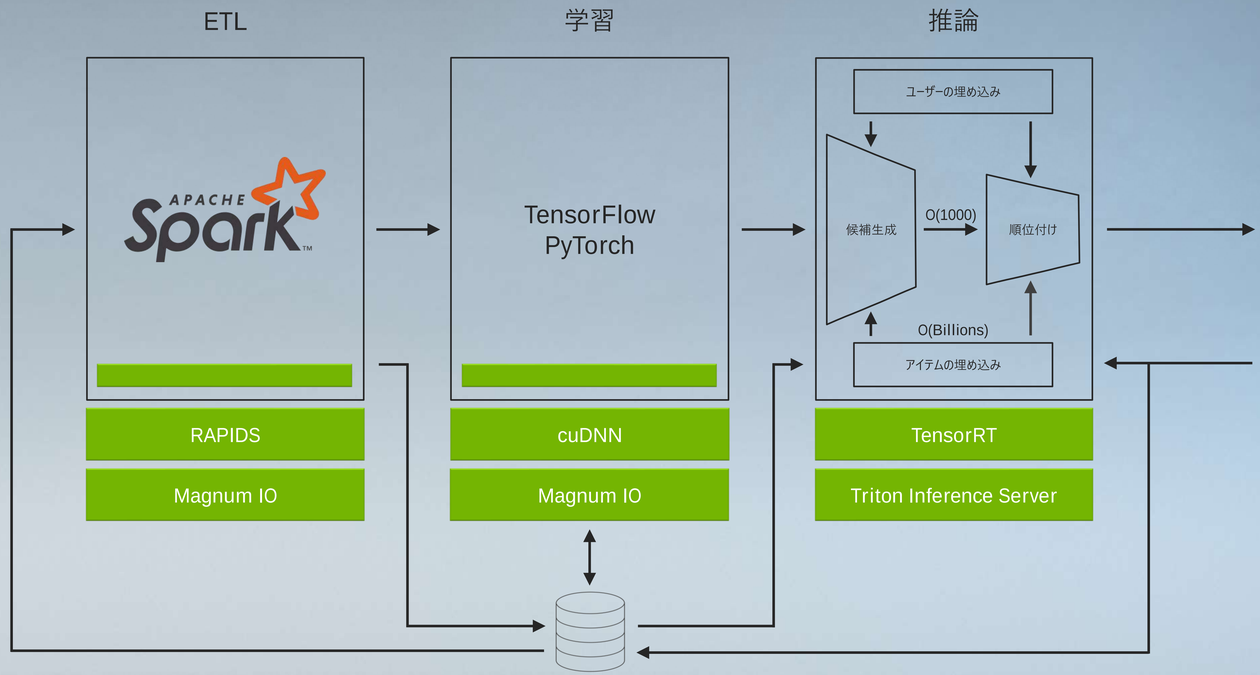


# 発表:
# クラウド分析プラットフォームが NVIDIA で高速化

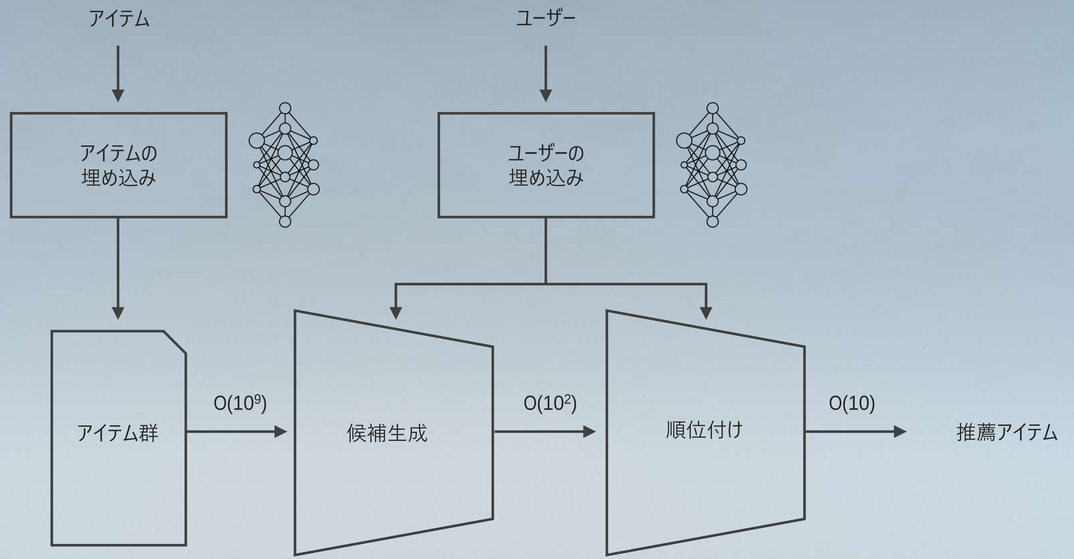



# NVIDIA AI

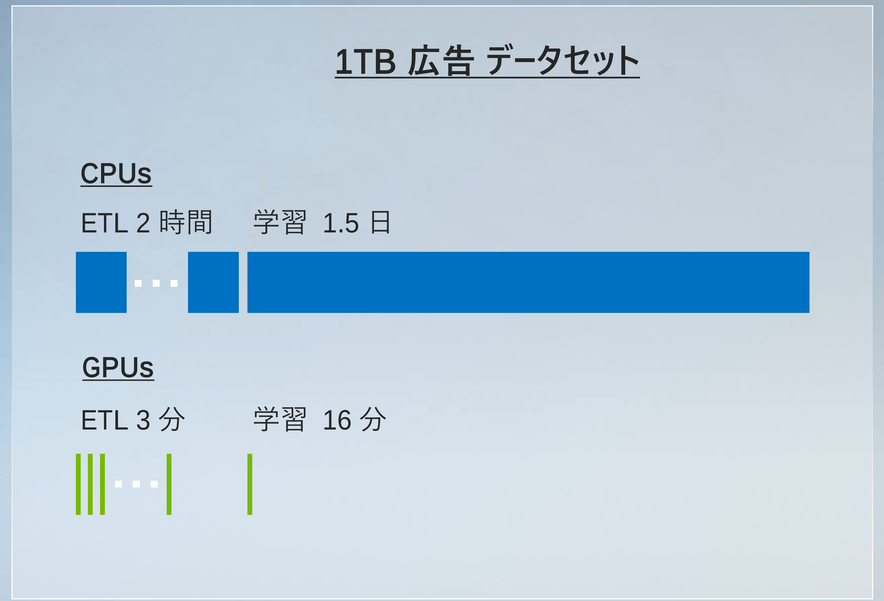

Amazon SageMaker
Azure Machine Learning
databricks
Google Cloud AI
Google Cloud Dataproc
DASK
XGBOOST
TensorFlow
PyTorch
RAPIDS
cuDF
cuML
cuGRAPH
cuDNN
CUDA-X
APACHE ARROW
MAGNUM IO

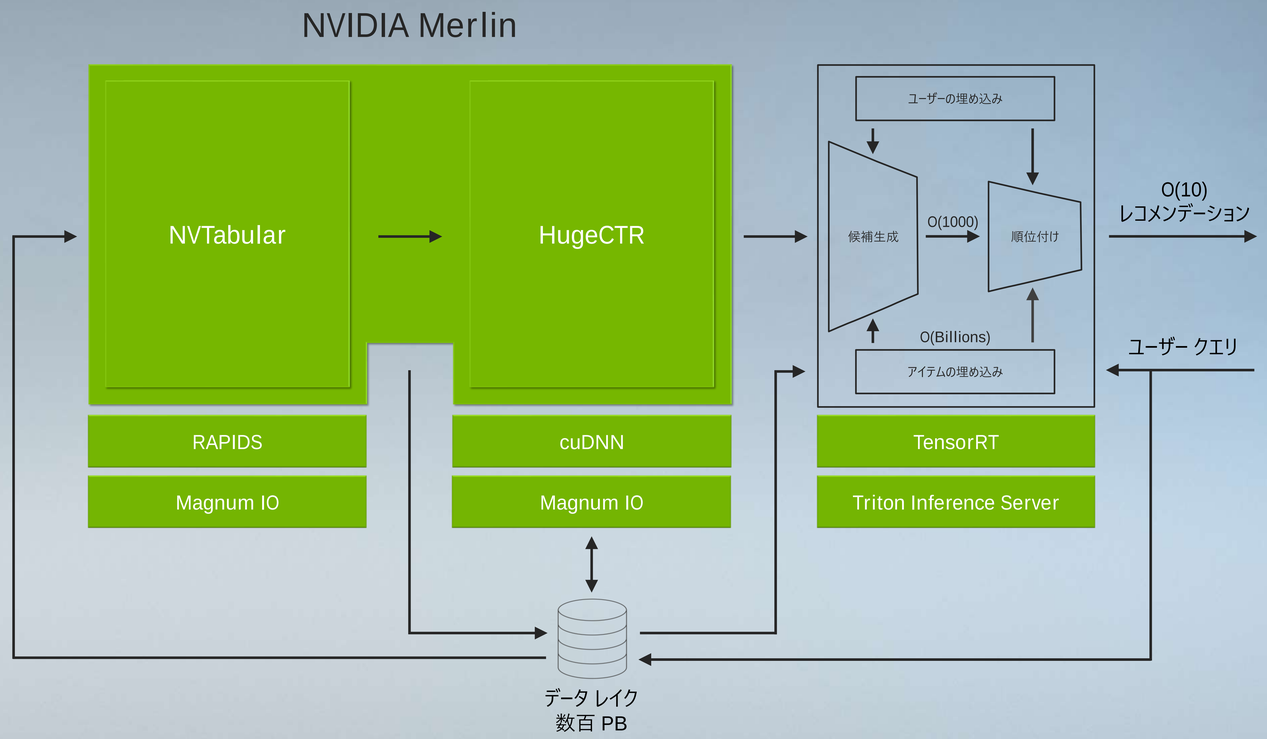

ETL
学習
推論
APACHE Spark
TensorFlow
PyTorch
ユーザーの埋め込み
候補生成
O(1000)
順位付け
O(Billions)
アイテムの埋め込み
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton Inference Server

# レコメンダー システムはパーソナライズされたインターネットのエンジン



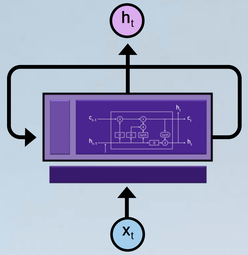
# 発表:
# NVIDIA MERLIN — ディープ レコメンダー アプリケーション フレームワーク



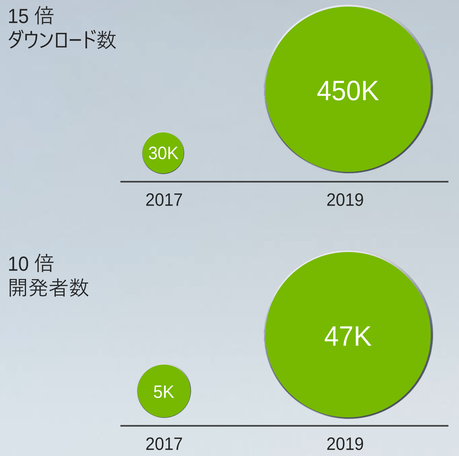



# NVIDIA AI 推論
# 過去 2 年で TENSORRT のダウンロードが 15 倍

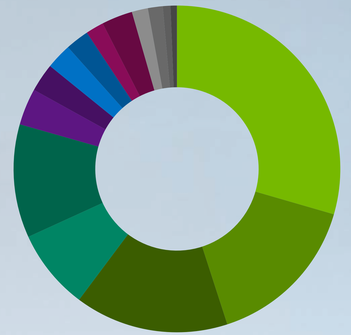
TENSORRT 7




CNN、Transformer、RNN
1000 以上の最適化されたカーネル
自動混合精度 FP32、FP16、INT8

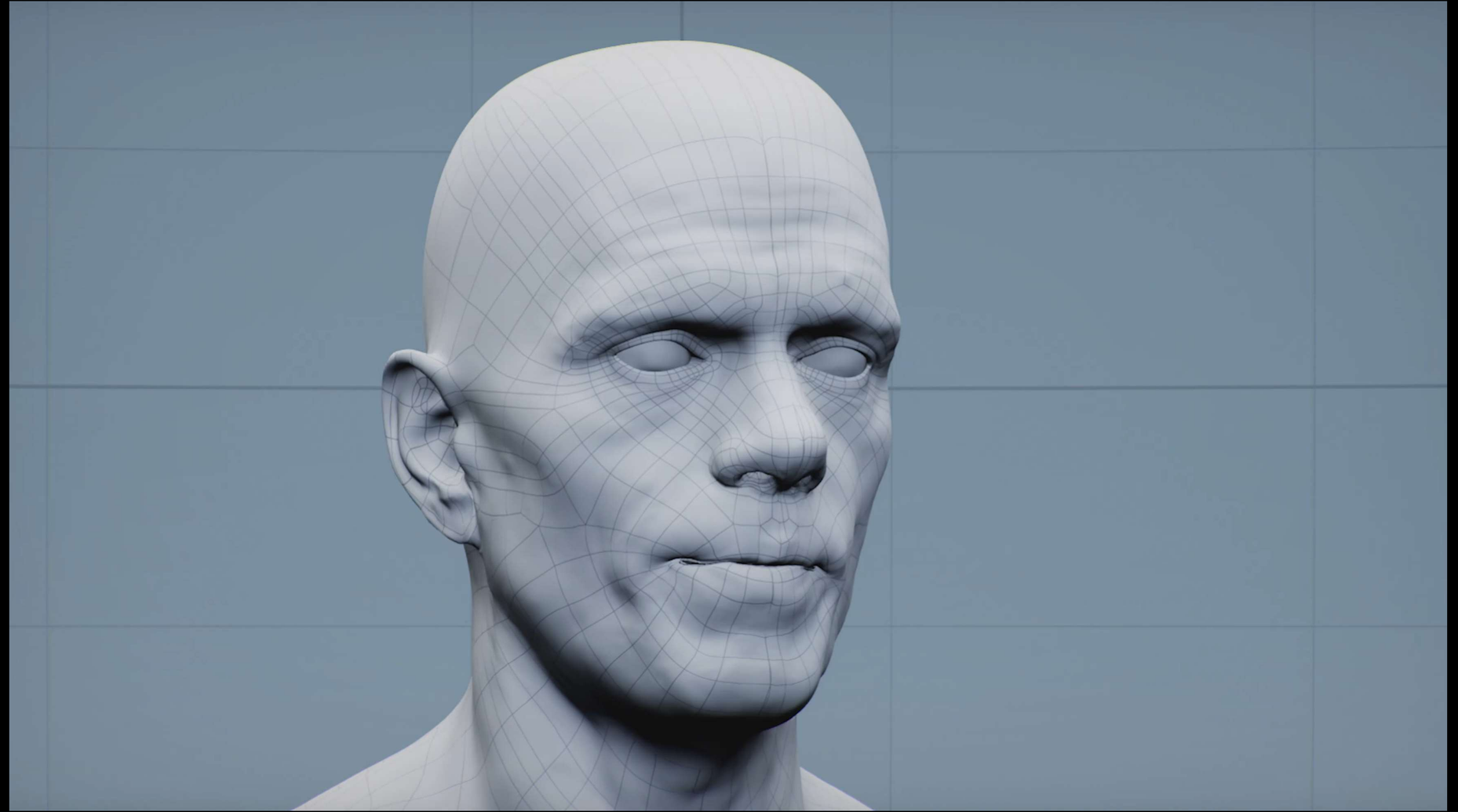



Alibaba.com
aws
Baidu 百度
ByteDance
criteo
ebay
Flipkart
Google Cloud
Microsoft
美团
NAVER
ORACLE Cloud Infrastructure
Tencent 腾讯



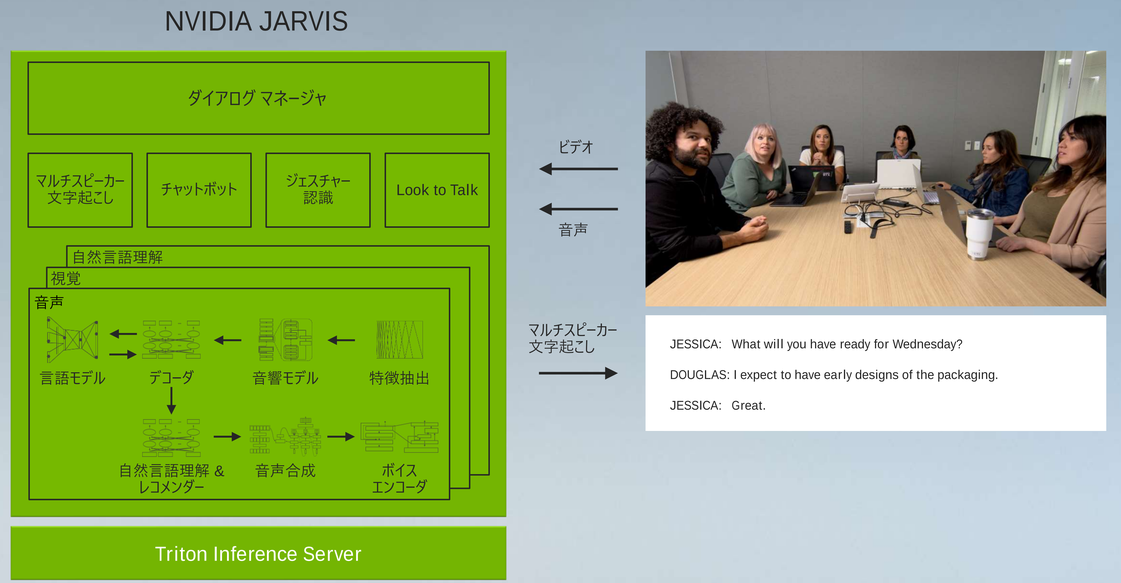
- Software
- IT Services
- Other
- Healthcare & Life Sciences
- Manufacturing
- Public Sector
- Consulting Services
- Research / Higher Ed
- Automotive
- Internet / Telecom
- Hardware / Semiconductor
- Cloud Services
- Financial Services
- Energy / Oil & Gas

# 発表:
# NVIDIA JARVIS —マルチモーダル 対話型 AI サービス フレームワーク

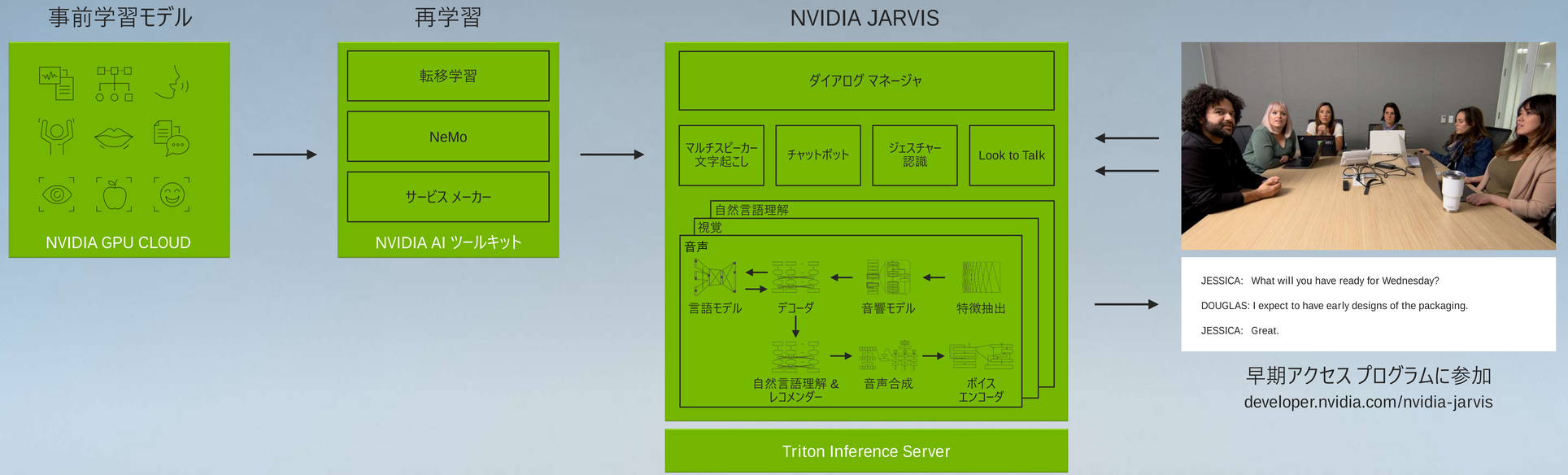


# JARVIS デモ
# MISTY — インタラクティブ 3D チャットボット

# 発表:
# NVIDIA JARVIS —マルチモーダル 対話型 AI サービス フレームワーク



# 発表:
# NVIDIA JARVIS —マルチモーダル 対話型 AI サービス フレームワーク




早期アクセス プログラムに参加
developer.nvidia.com/nvidia-jarvis

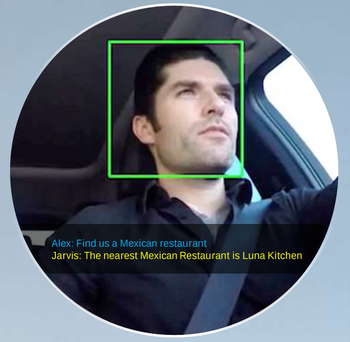
# 対話型 AI が産業を変革中

ビデオ会議
字幕、翻訳、文字起こし
1 日あたり 2 億回の会議

コール センター
1 日あたり 5 億コール

スマート スピーカー
年間 1 億 5000 台販売

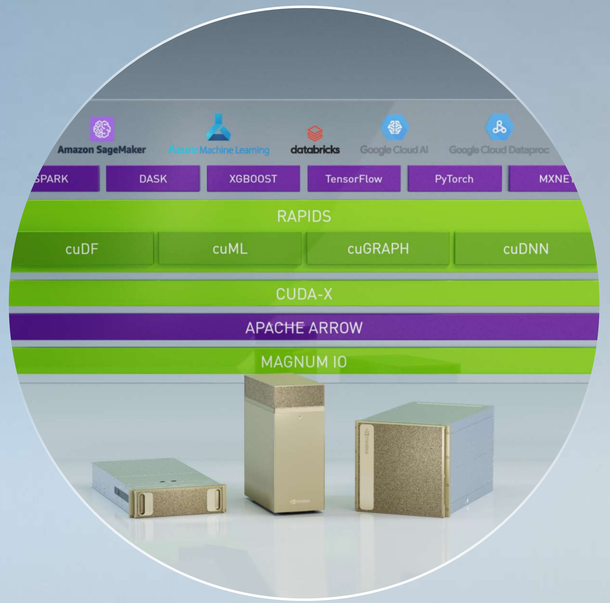
リテール アシスタント
1200 万の小売店

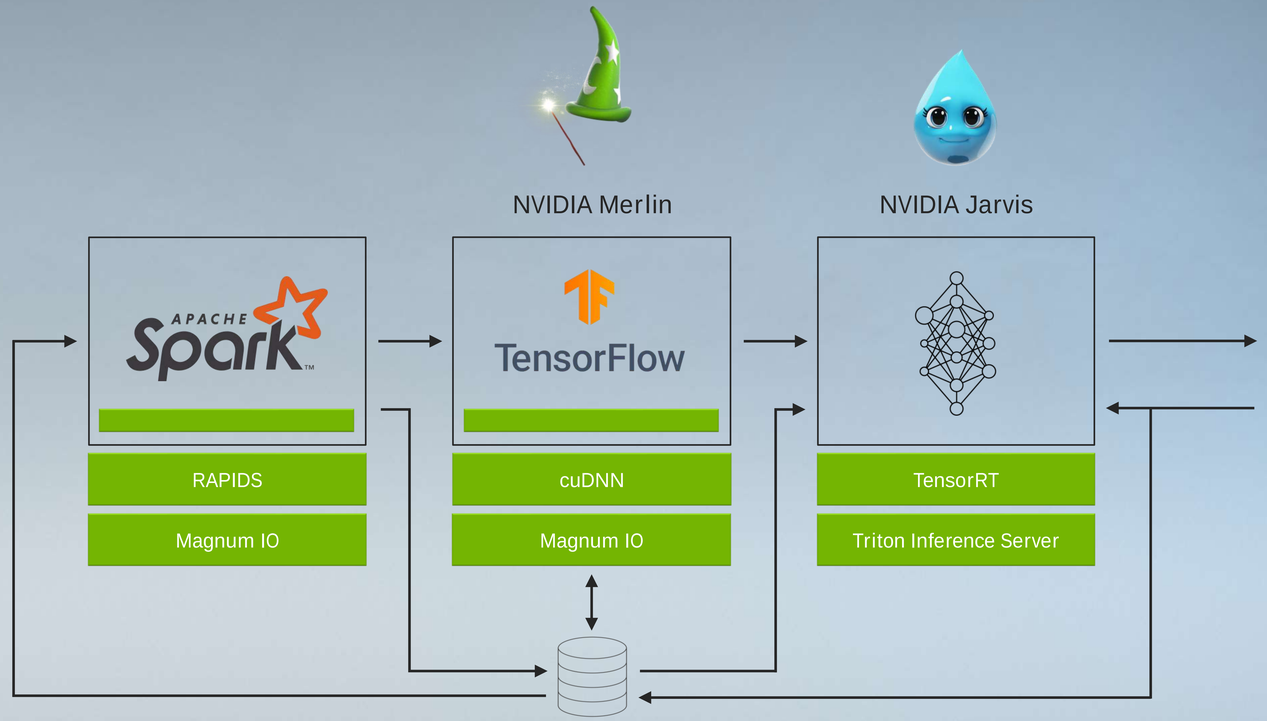



車内アシスタント
年間 750 万台の新車

KENSHO + S&P Global · Microsoft · NUANCE · Square · voca.ai · fast.ai · KALDI · ESPnet · spaCy · Alibaba Cloud · aws · Baidu 百度 · Google Cloud · Microsoft Azure · ORACLE Cloud Infrastructure · Tencent Cloud

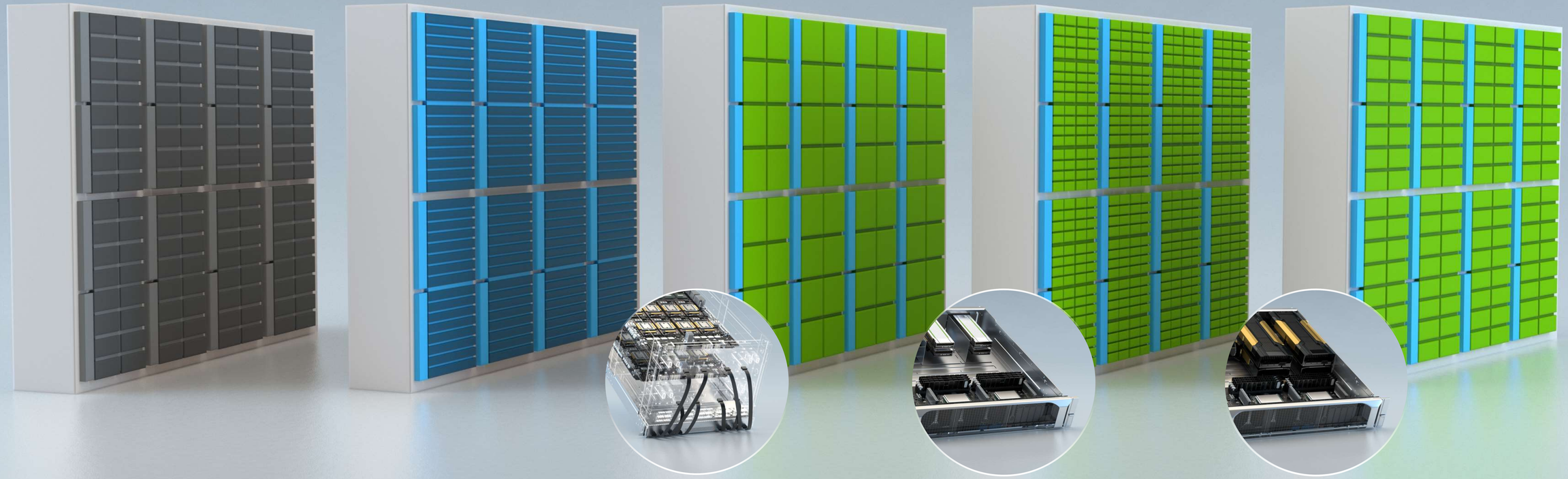
# NVIDIA AI


Amazon SageMaker
Azure Machine Learning
databricks
Google Cloud AI
Google Cloud Dataproc
DASK
XGBOOST
TensorFlow
PyTorch
RAPIDS
cuDF
cuML
cuGRAPH
cuDNN
CUDA-X
APACHE ARROW
MAGNUM IO

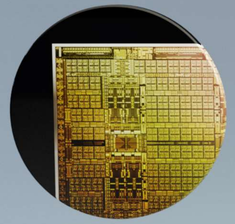

NVIDIA Merlin
NVIDIA Jarvis
APACHE Spark
TensorFlow
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton Inference Server

# 現代のクラウド データ センター

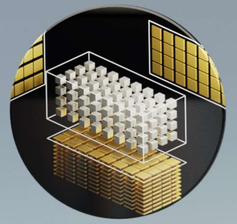
多様なアプリケーション | スケールアップとスケールアウトのワークロード | 飽くなき需要



# 発表: NVIDIA A100

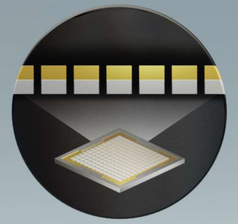


540 億トランジスタ

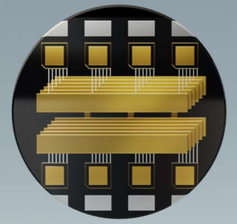


第 3 世代 TENSOR コア

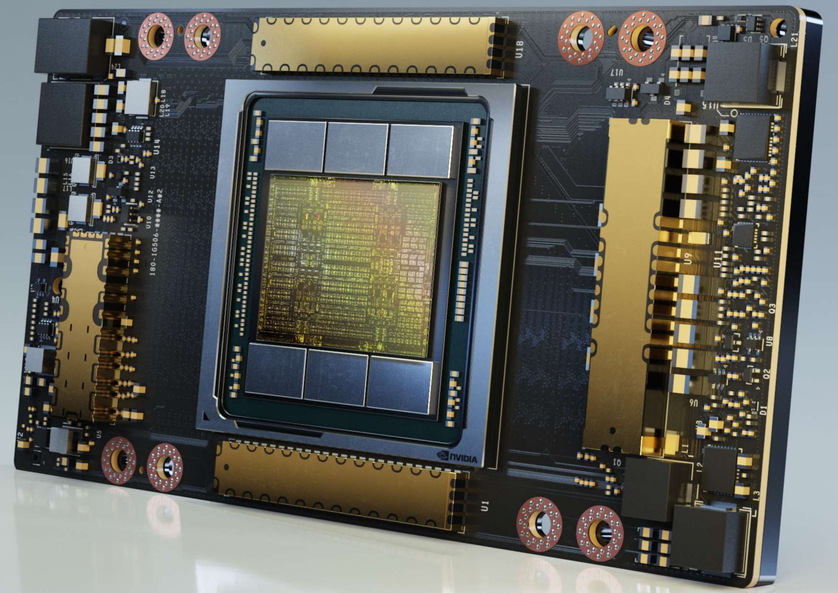
スパーシティ アクセラレーション



MIG

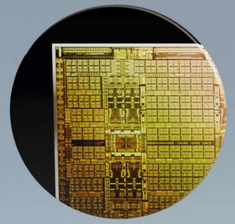


第 3 世代 NVLINK & NVSWITCH

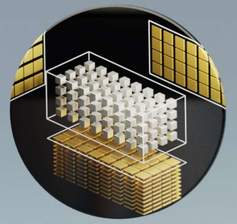


# 発表: NVIDIA A100

TSMC 7nm | HBM2 — 1.6 テラバイト / 秒 | 3D チップ スタック

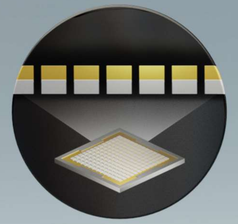


540 億トランジスタ

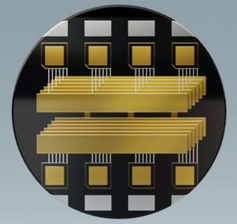


第 3 世代 TENSOR コア

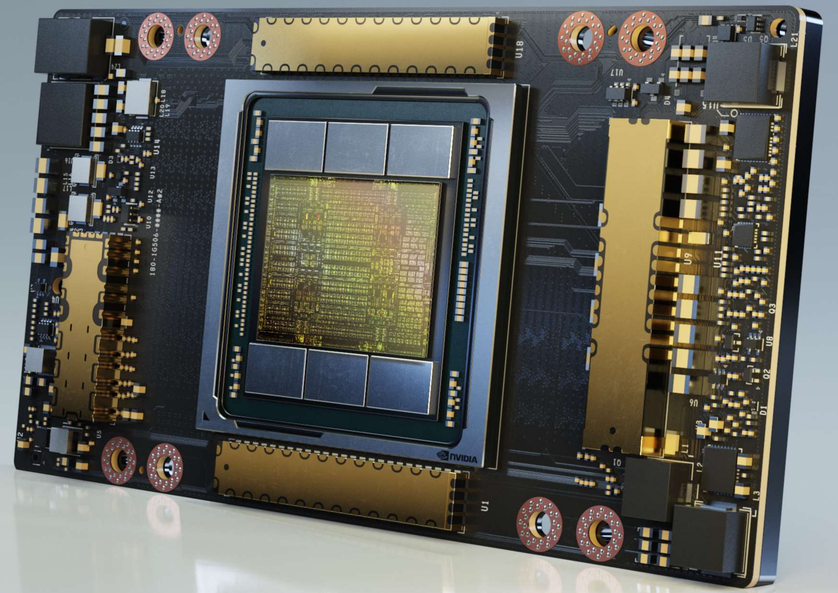
スパーシティ アクセラレーション



MIG

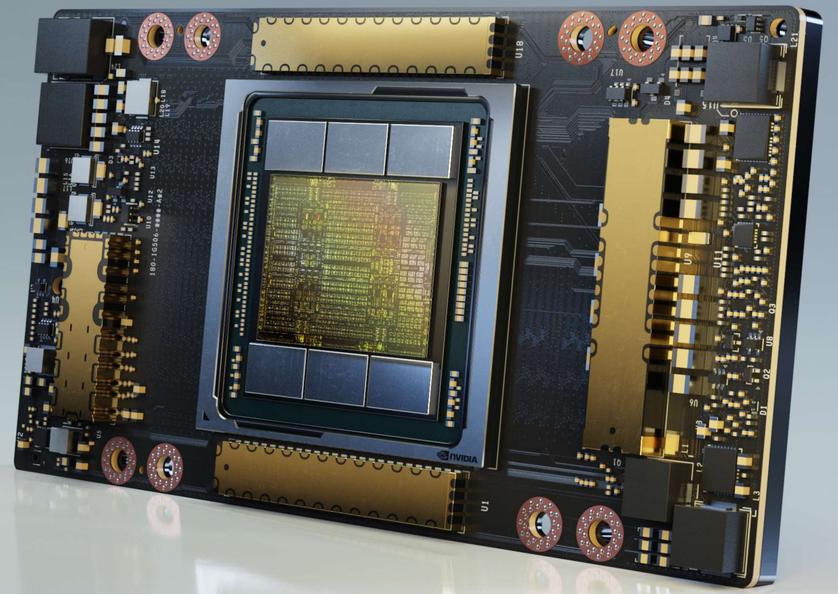


第 3 世代 NVLINK & NVSWITCH



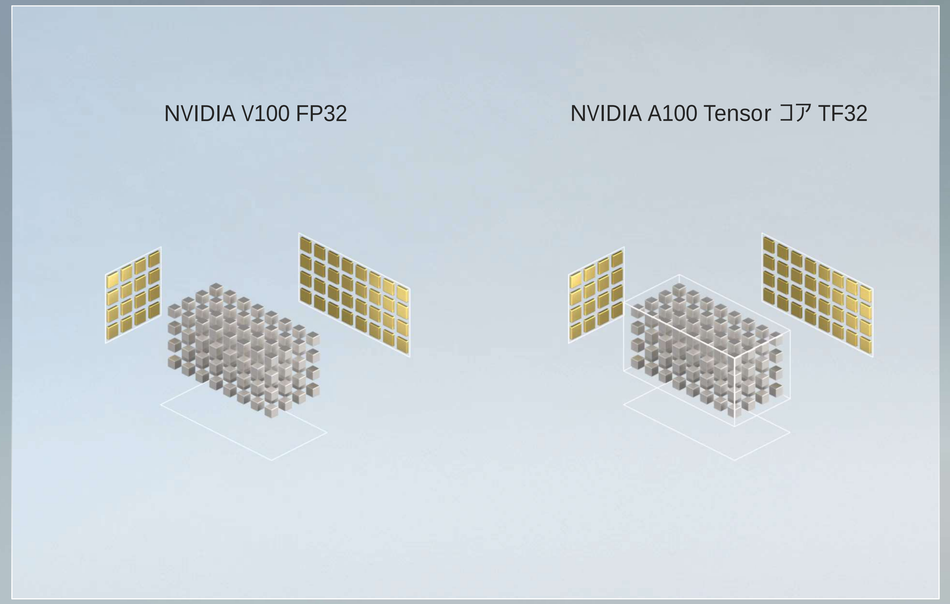
# 新しい TF32 TENSOR コア

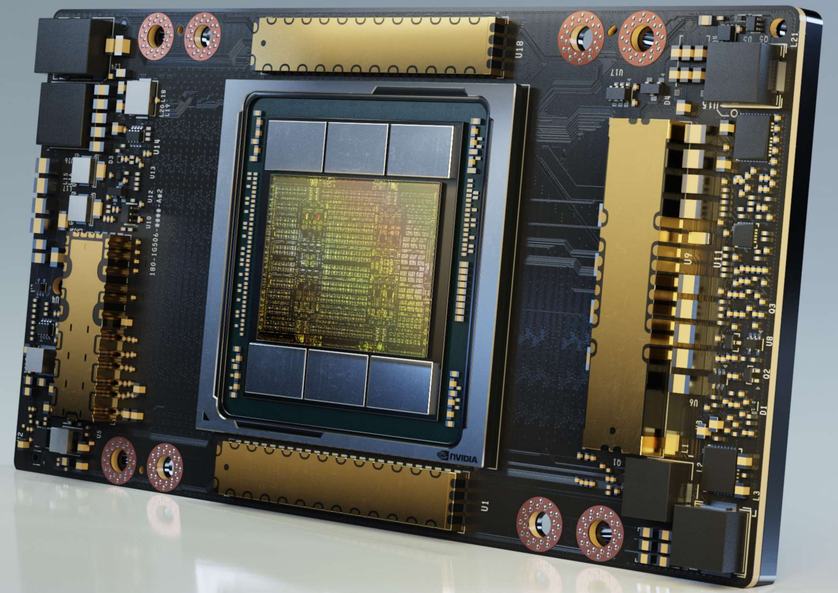
FP32 の範囲と FP16 の精度 | FP32 の入力と FP32 でのアキュミュレーション | コードの変更なしで学習を高速化

# 新しい TF32 TENSOR コア

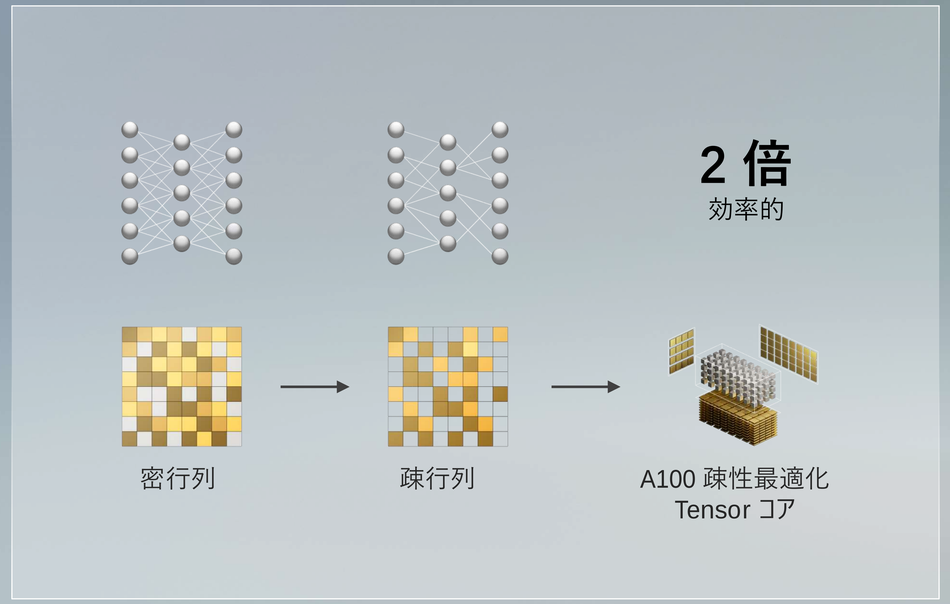
FP32 の範囲と FP16 の精度 | FP32 の入力と FP32 でのアキュミュレーション | コードの変更なしで学習を高速化

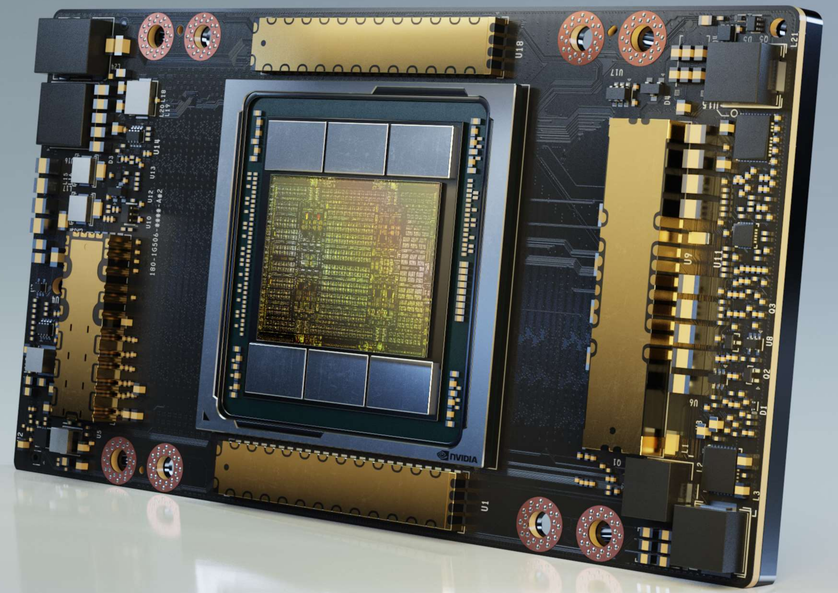


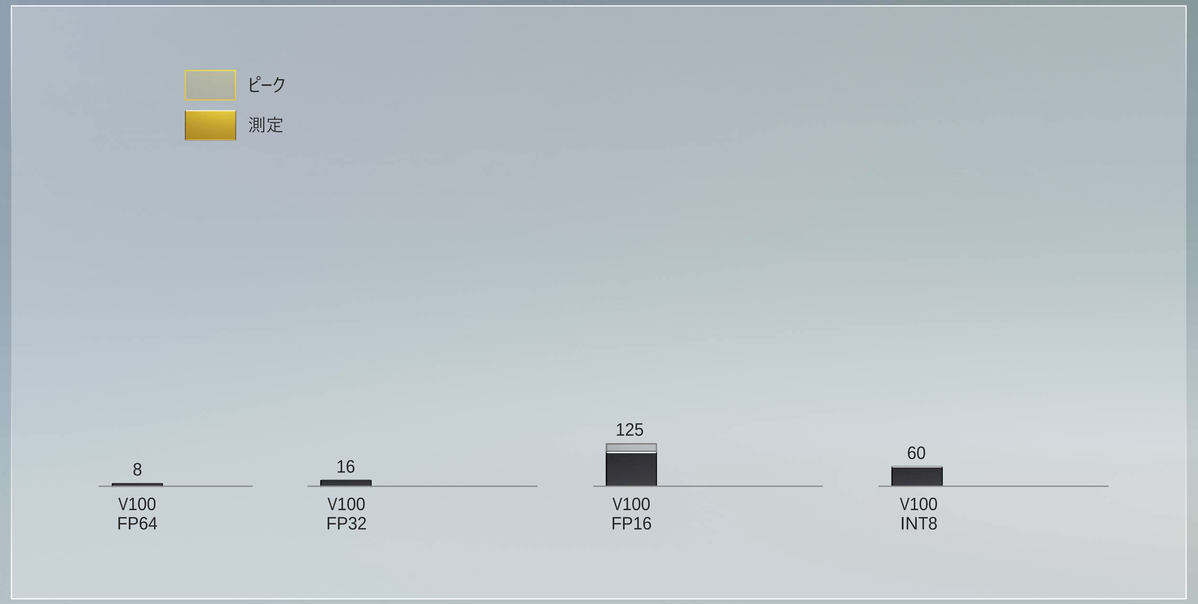
# 疎行列向けの新しい TENSOR コア アクセラレーション

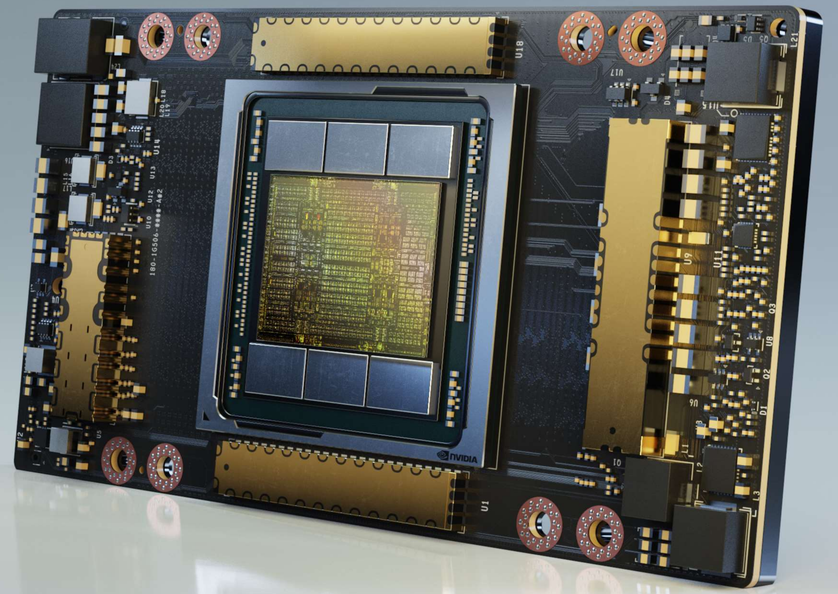
疎な AI Tensor 命令を最適化| 2 倍高速な実行 | TF32、FP16、BFLOAT16、INT8、INT4 で対応



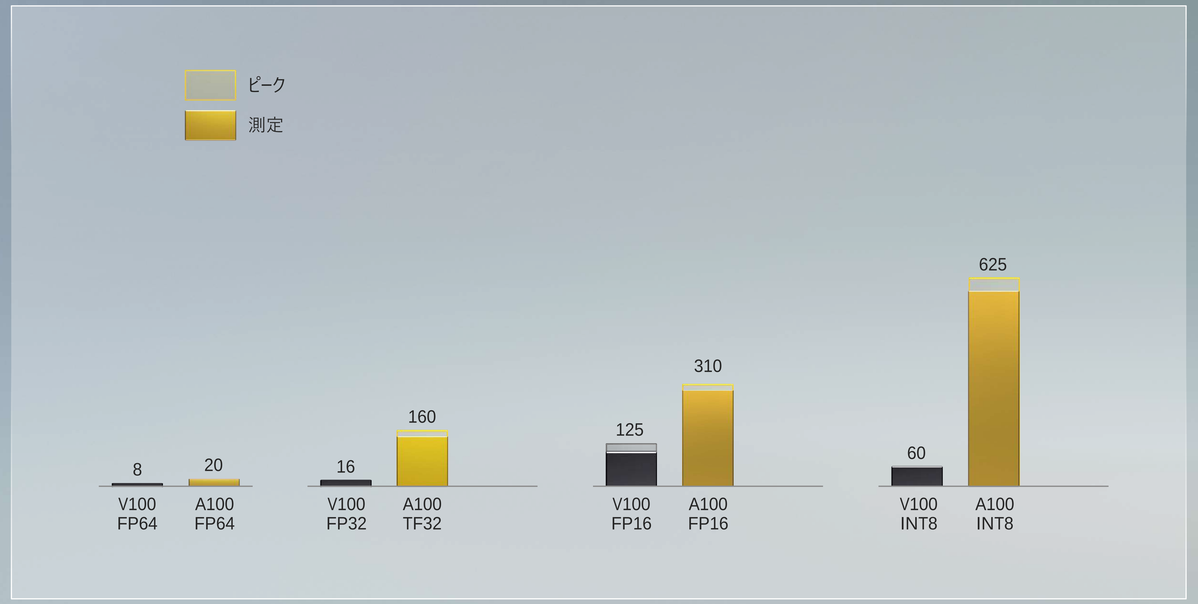



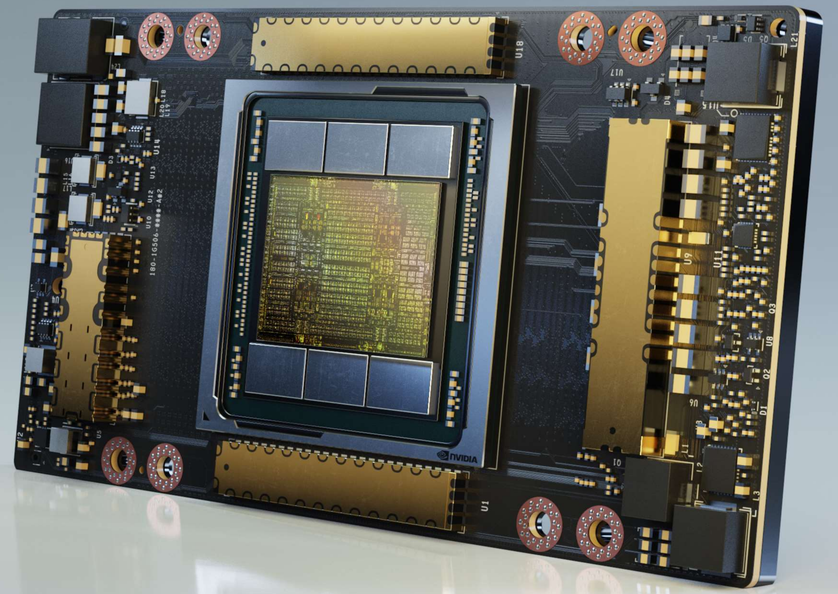
# 発表: NVIDIA A100
# 最大の世代間の飛躍



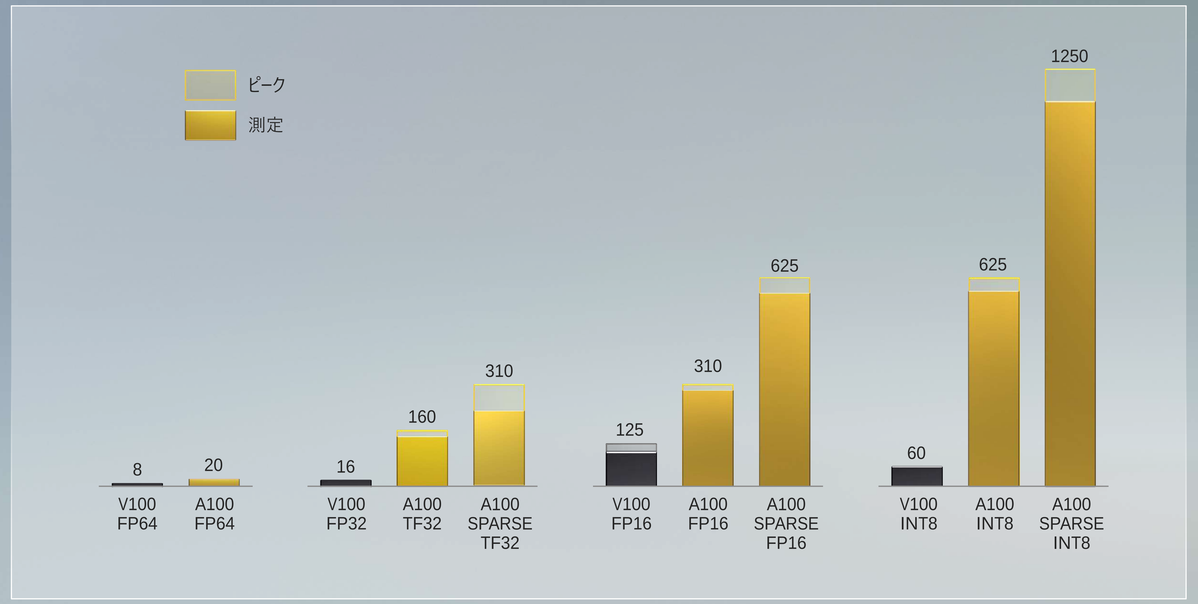



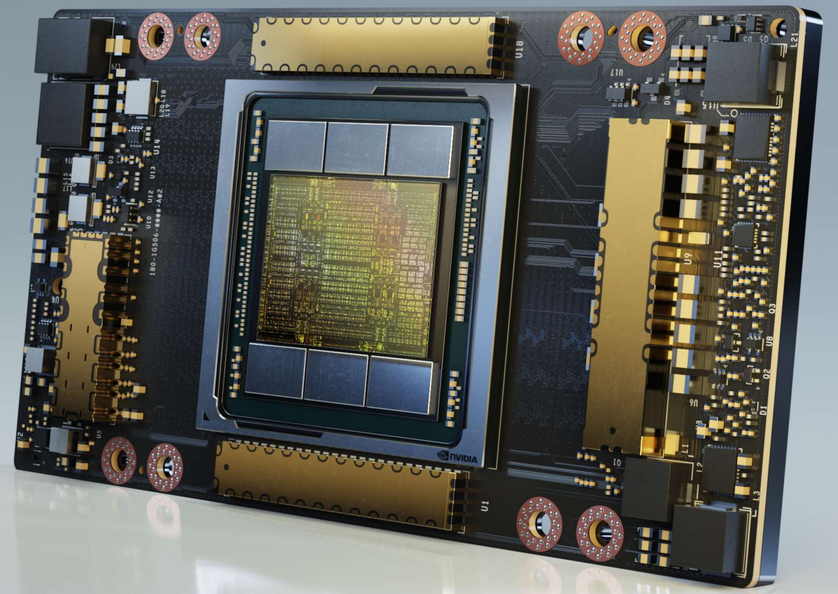
# 発表: NVIDIA A100
# 最大の世代間の飛躍



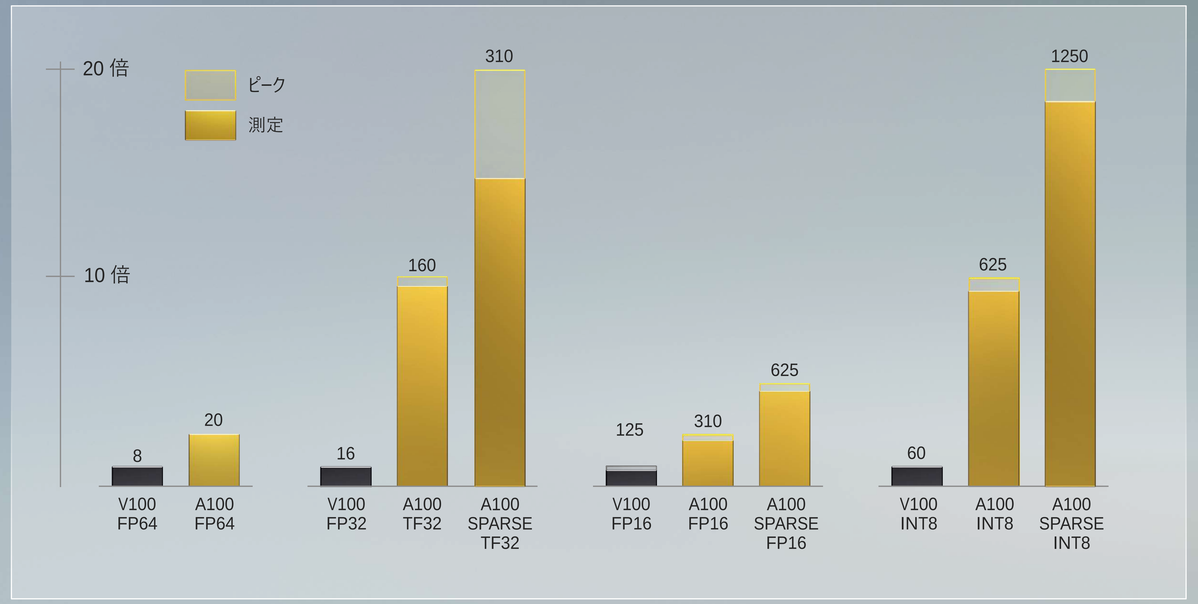



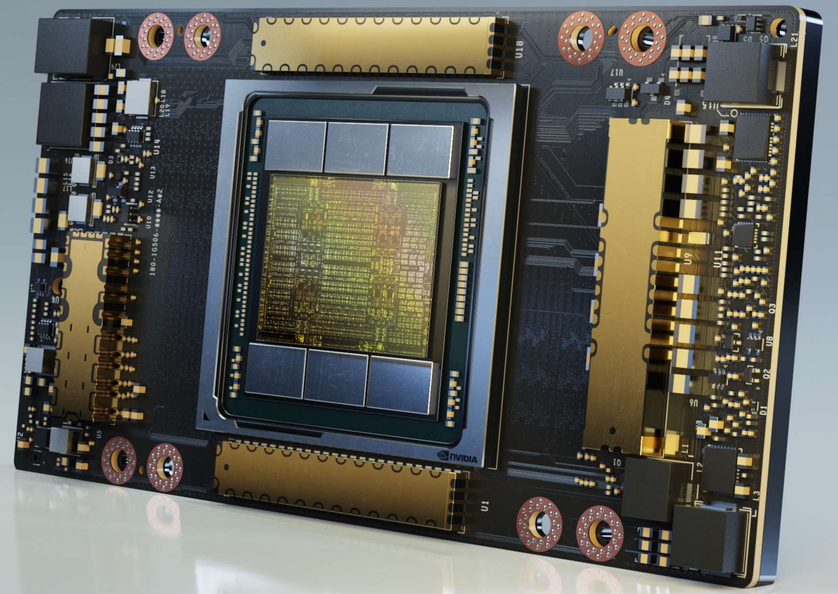
# 発表: NVIDIA A100
# 最大の世代間の飛躍



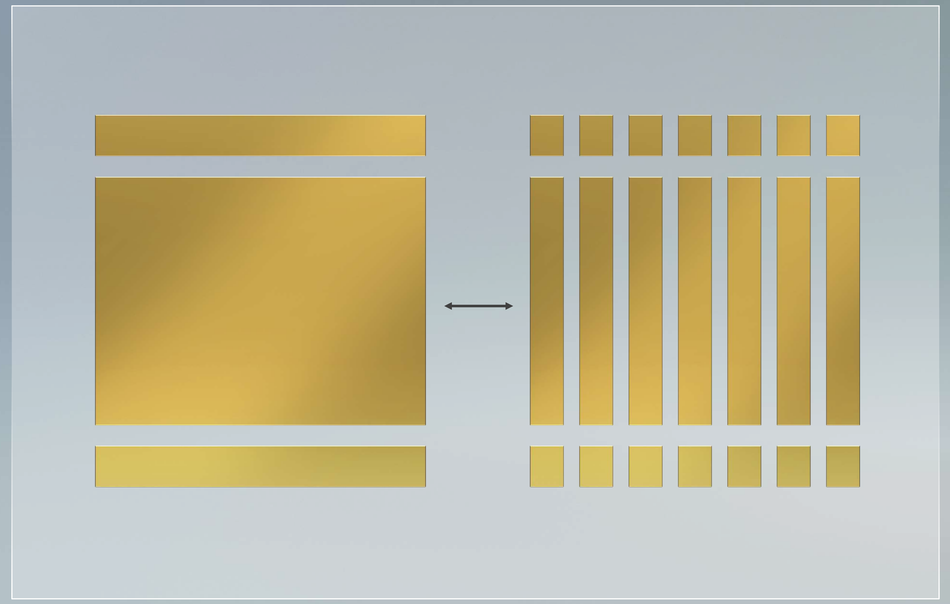
# 発表: NVIDIA A100
# 最大の世代間の飛躍

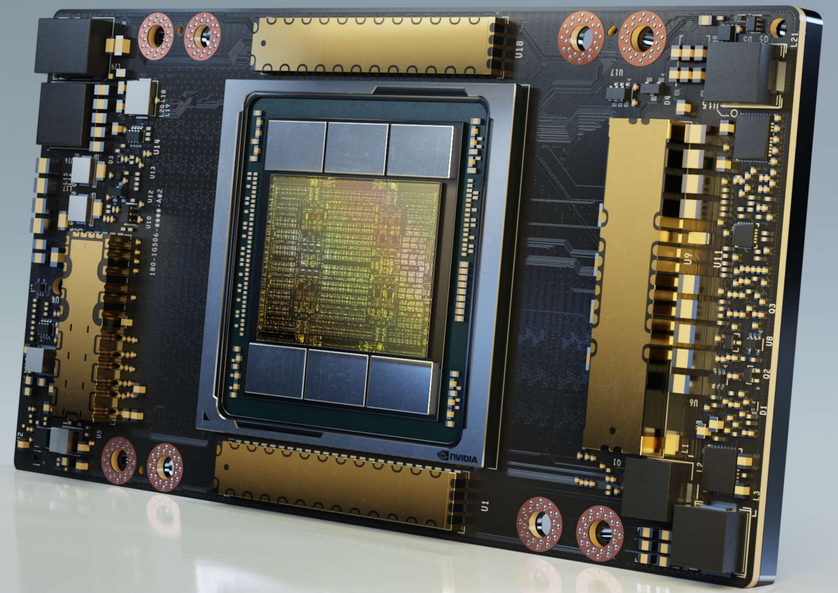


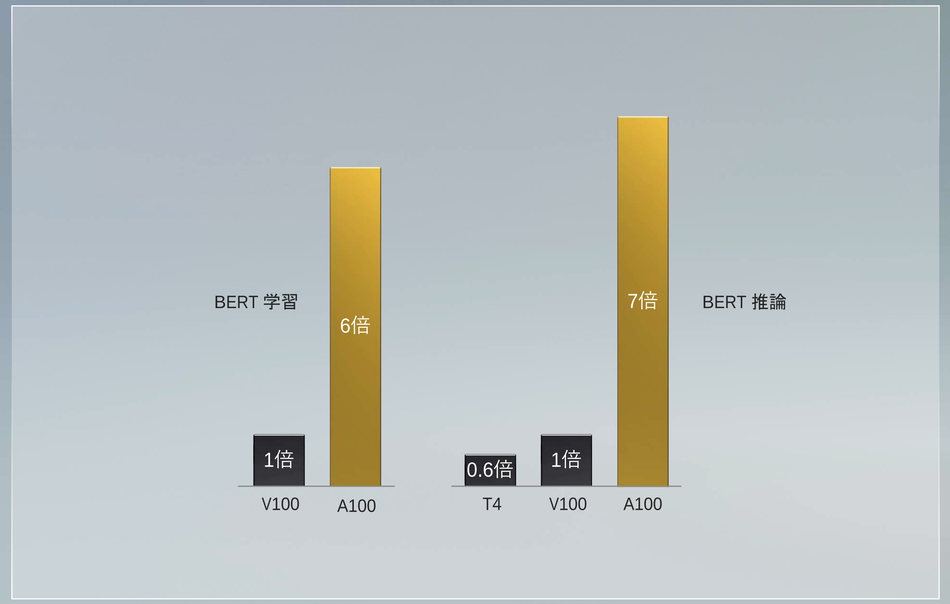
# エラスティック GPU コンピューティングのための新しいマルチインスタンス GPU

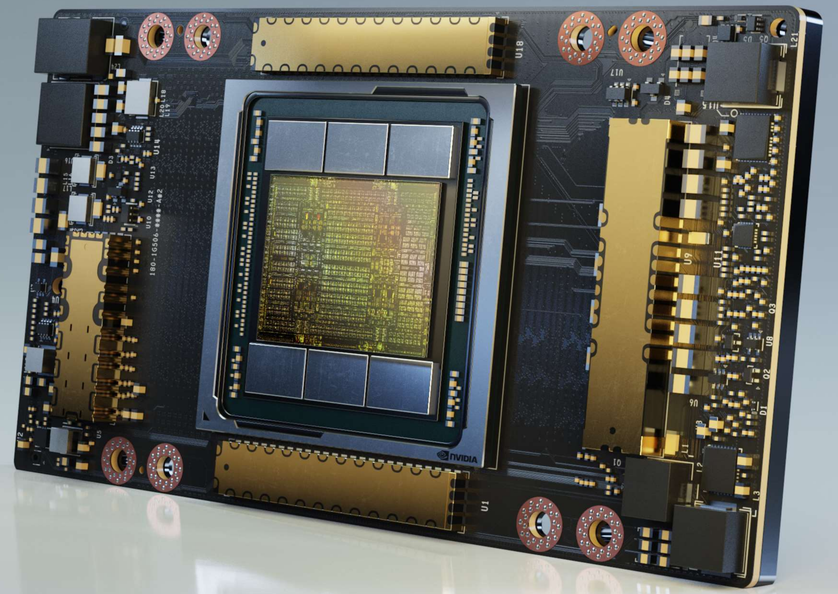
GPU あたり複数のインスタンスを同時に動かし V100 の 7 倍高いスループット



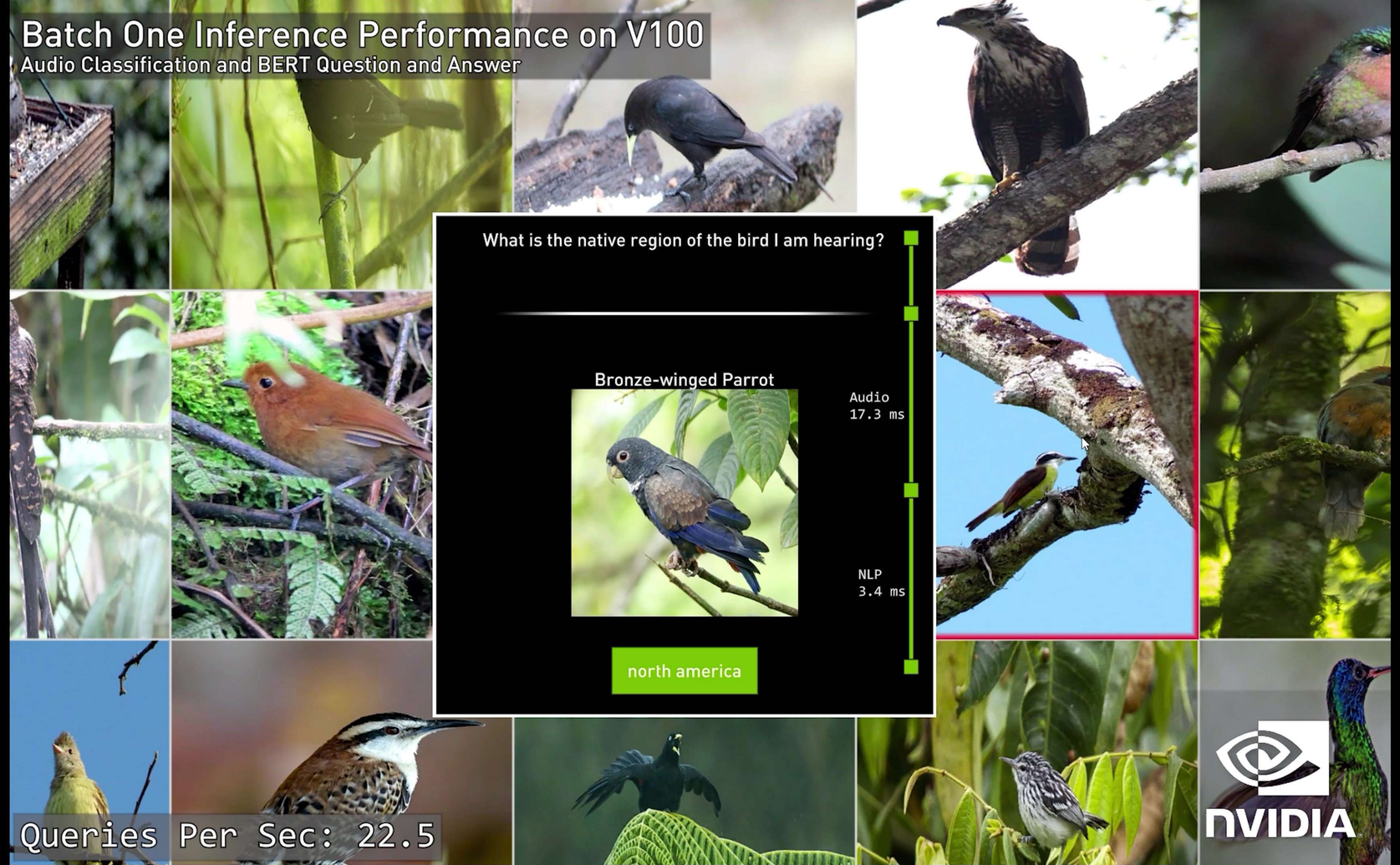


# 学習と推論のアクセラレーションを統合

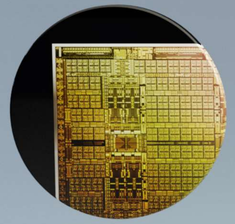


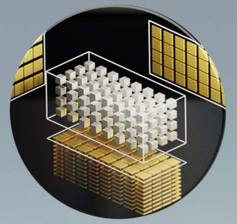
BERT 事前学習のスループット、Pytorch を使用、(2/3)Phase 1 と (1/3)Phase 2 を含む | Phase 1 Seq Len = 128、Phase 2 Seq Len = 512 V100: FP32 精度を使用した 8 基の V100 を搭載した DGX-1 サーバー A100: TF32 精度を使用した 8　基の A100 を搭載した DGX A100 サーバー | BERT Larget 推論 | T4、V100: TRT 7.1、精度 = FP16、バッチサイズ = 256 | A100 MIG: リリース前の TRT、バッチサイズ = 94、精度 = INT8、疎性高速化を有効


Batch One Inference Performance on V100
Audio Classification and BERT Question and Answer
What is the native region of the bird I am hearing?
Bronze-winged Parrot
Audio
17.3 ms
NLP
3.4 ms
north america
Queries Per Sec: 22.5
NVIDIA

# 発表: NVIDIA A100

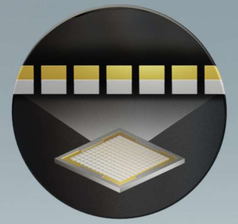


540 億トランジスタ

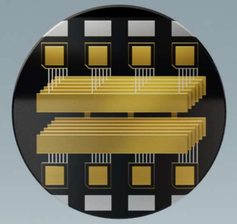


第 3 世代 TENSOR コア

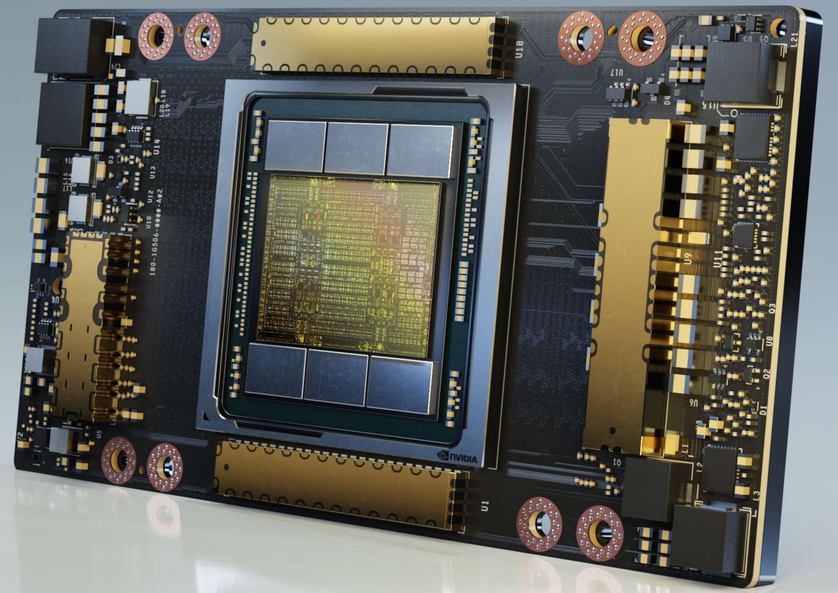
スパーシティ アクセラレーション



MIG



第 3 世代 NVLINK & NVSWITCH



# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

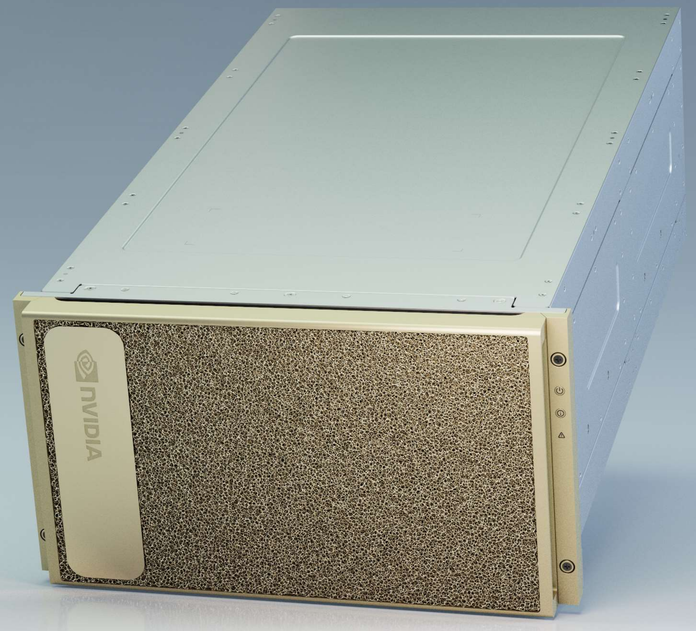
## シングル ノードで 5 PetaFLOPS の性能

エンドツーエンドのデータ サイエンスと AI のための統合システム

完全にアクセラレーションされたスタック — Spark 3.0、RAPIDS、TensorFlow、PyTorch、Triton

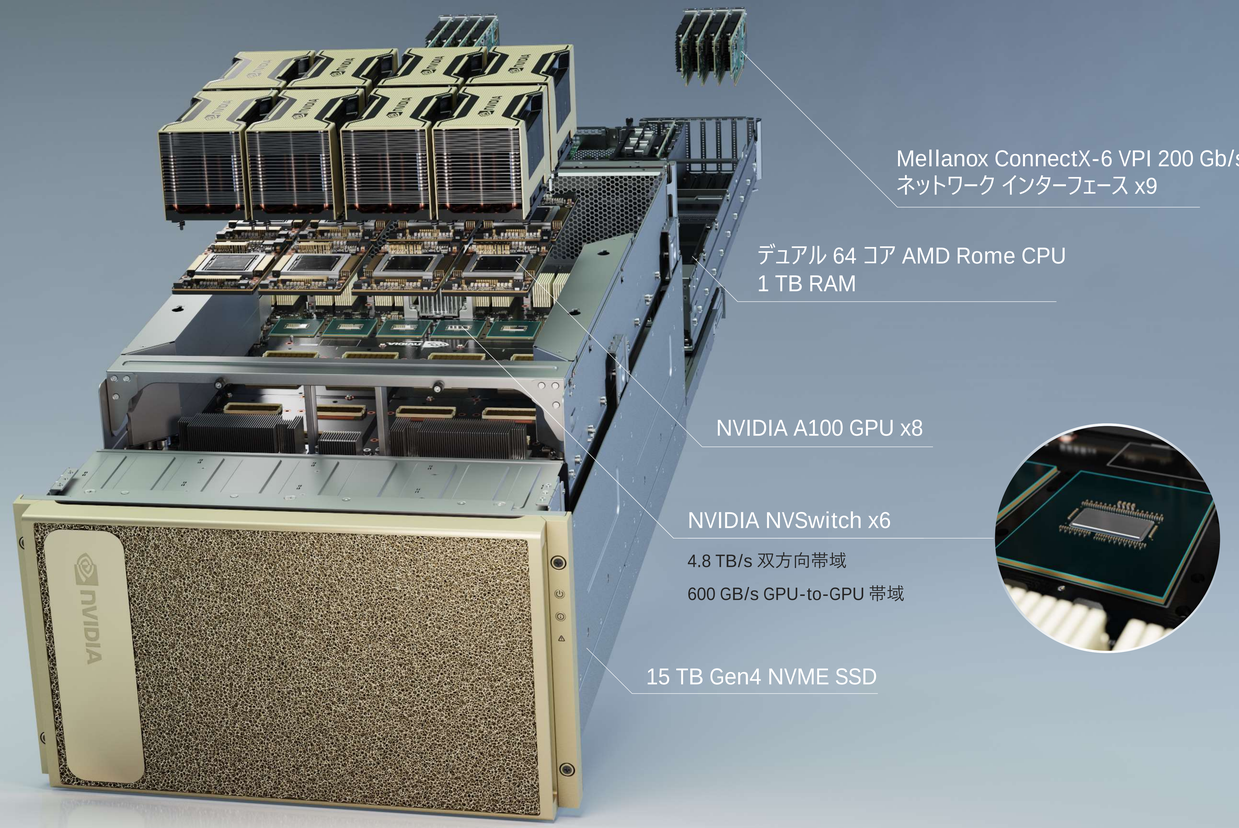
エラスティックなスケールアップまたはスケールアウト コンピューティング

Mellanox ネットワーキングによる高いスケーラビリティ



# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

シングル ノードで 5 PetaFLOPS の性能

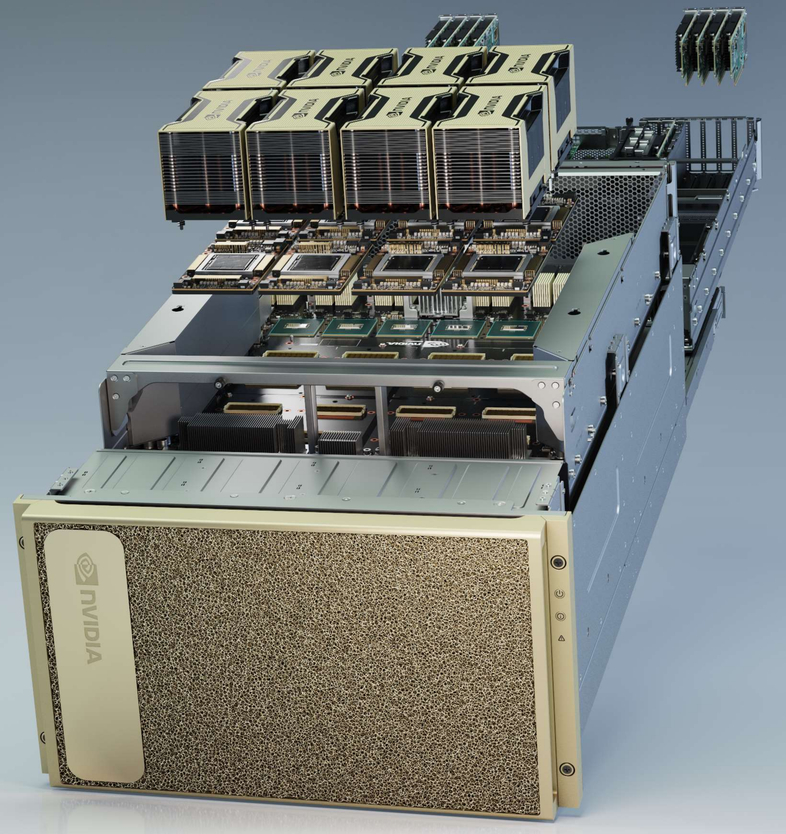


# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

シングル ノードで 5 PetaFLOPS の性能

| | | |
|---|---|---|
| INT8 | 10 | PetaOPS ピーク |
| FP16 | 5 | PFLOPS ピーク |
| TF32 | 2.5 | PFLOPS ピーク |
| FP64 | 156 | TFLOPS ピーク |

Tensor コア + スパーシティ アクセラレーション



# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

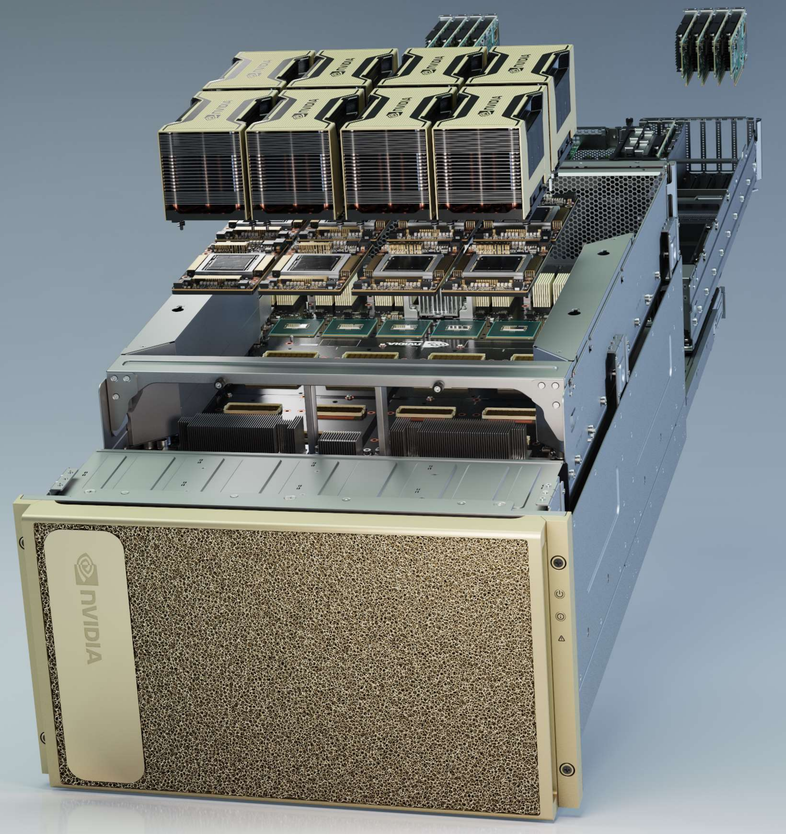
シングル ノードで 5 PetaFLOPS の性能

**150 倍** AI 演算

**40 倍** メモリ帯域幅

**40 倍** IO 帯域幅

ハイエンド CPU サーバーとの比較



# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

シングル ノードで 5 PetaFLOPS の性能

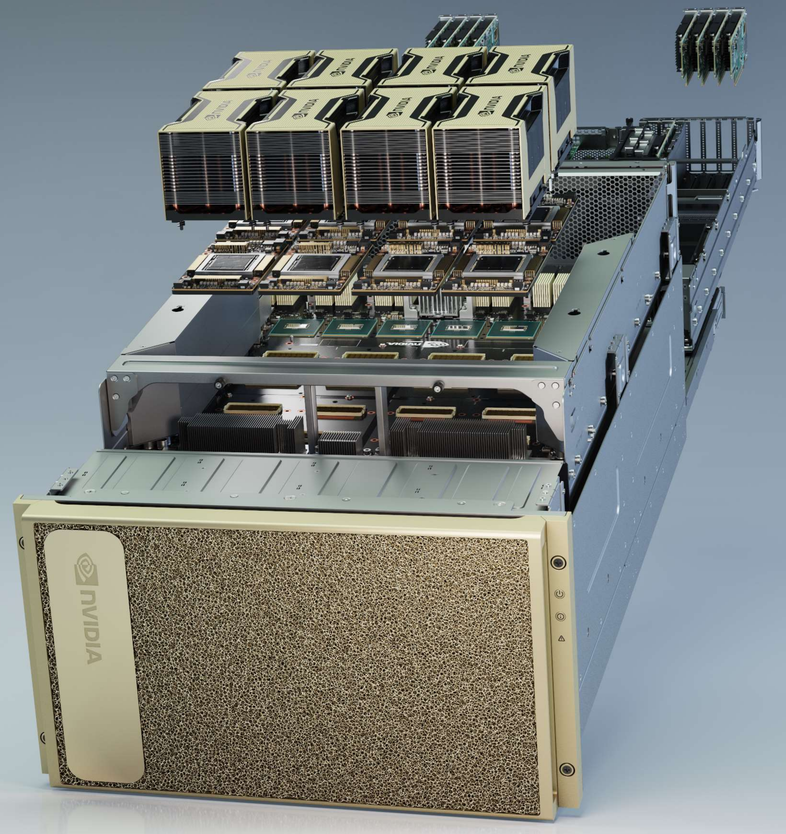
**150 倍** AI 演算

**40 倍** メモリ帯域幅

**40 倍** IO 帯域幅

ハイエンド CPU サーバーとの比較

**199,000 ドルで販売中**



# 発表:
# NVIDIA A100 ライトハウス カスタマー

業界のリーダーが選ぶ エラスティック データ センター アクセラレータ

## クラウド

Alibaba Cloud

aws

BAIDU AI CLOUD

Google Cloud

Microsoft Azure

ORACLE Cloud Infrastructure

Tencent Cloud

## システム

Atos

CRAY a Hewlett Packard Enterprise company

DELL Technologies

FUJITSU

GIGABYTE

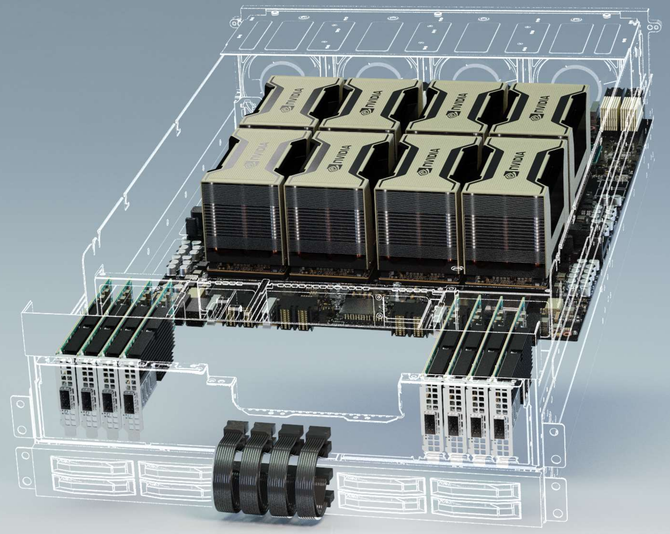
Hewlett Packard Enterprise

inspur

Lenovo

QCT

SUPERMICRO



# 今日の AI データ センター

50 台の DGX-1 システム : AI 学習向け

600 台の CPU システム : AI 推論向け

1,100 万ドル

25 ラック

630 kW

# DGX A100 AI

5 台の DGX A100 システム : AI 学習と推論向け

100 万ドル

1 ラック

28 kW

# 今日の AI データ センター

50 台の DGX-1 システム : AI 学習向け

600 台の CPU システム : AI 推論向け

1,100 万ドル

25 ラック

630 kW

# DGX A100 AI

5 台の DGX A100 システム : AI 学習と推論向け

100 万ドル

1 ラック

28 kW

# ページランクの事例

Common Crawl データ セット
2.6TB グラフ – 1280 億エッジ

3,000 CPU サーバー – 105 ラック

# ページランクの事例

Common Crawl データ セット
2.6TB グラフ – 1280 億エッジ

外部 NVLINK で接続された 4 台の DGX A100

# ページランクの事例

Common Crawl データ セット
2.6TB グラフ - 1280 億エッジ

3,000 CPU サーバー – 105 ラック

# ページランクの事例

Common Crawl データ セット
2.6TB グラフ - 1280 億エッジ

外部 NVLINK で接続された 4 台の DGX A100

# 発表:
# NVIDIA DGX A100 SUPERPOD

140 台の DGX A100 システム (1,120 基の A100)
170 台の Mellanox Quantum 200G InfiniBand スイッチ
280 Tb/s ネットワーク ファブリック - 15km の光ケーブル
4 PB のオールフラッシュ ネットワーク ストレージ
700 PFLOPS の AI 性能
3 週間以内で構築

# NVIDIA が SATURNV を拡張

拡張前
1,800 台の DGX システム
1.8 ExaFLOPS

4 セットの DGX SuperPOD を追加
560 台の DGX A100 = 2.8 ExaFLOPS

4.6 ExaFLOPS 総演算性能

# 発表:
# NVIDIA DGX A100
# 第 3 世代 統合 AI システム

シングル ノードで 5 PetaFLOPS の性能

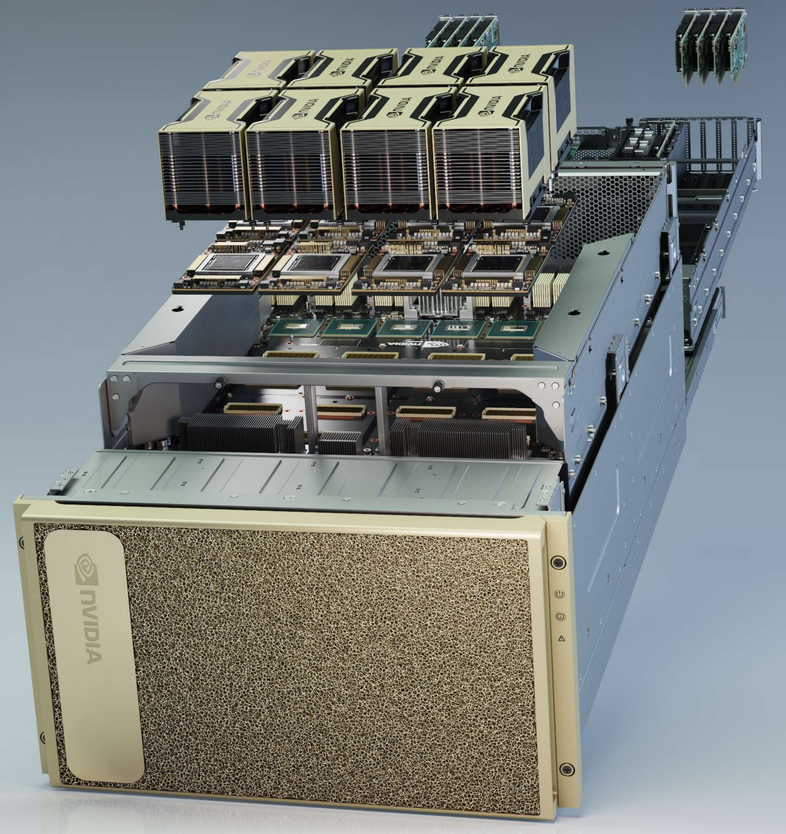
**150 倍** AI 演算

**40 倍** メモリ帯域幅

**40 倍** IO 帯域幅

ハイエンド CPU サーバーとの比較

**199,000 ドルで販売中**



# SMART EVERYTHING 革命


常時オン | 即座に感知-推論-行動 | 遠隔 | 数兆個

# 発表:
# MELLANOX CX6 DX 搭載 NVIDIA EGX A100

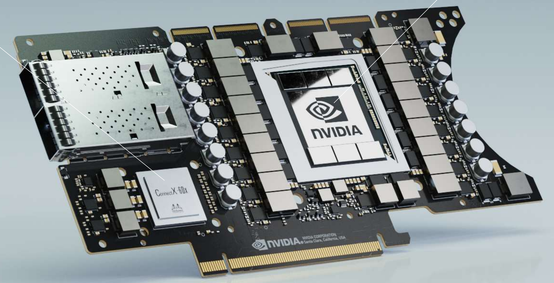
## NVIDIA Mellanox ConnectX-6 DX

Dual 100 Gb/s Ethernet または InfiniBand

ラインスピード TLS/IPSec 暗号化エンジン

通信事業者向けタイムトリガ方式転送技術:
Time Triggered Transmission Tech for Telco
(5T for 5G)

ASAP$^2$ SR-IOV と VirtIO オフロード



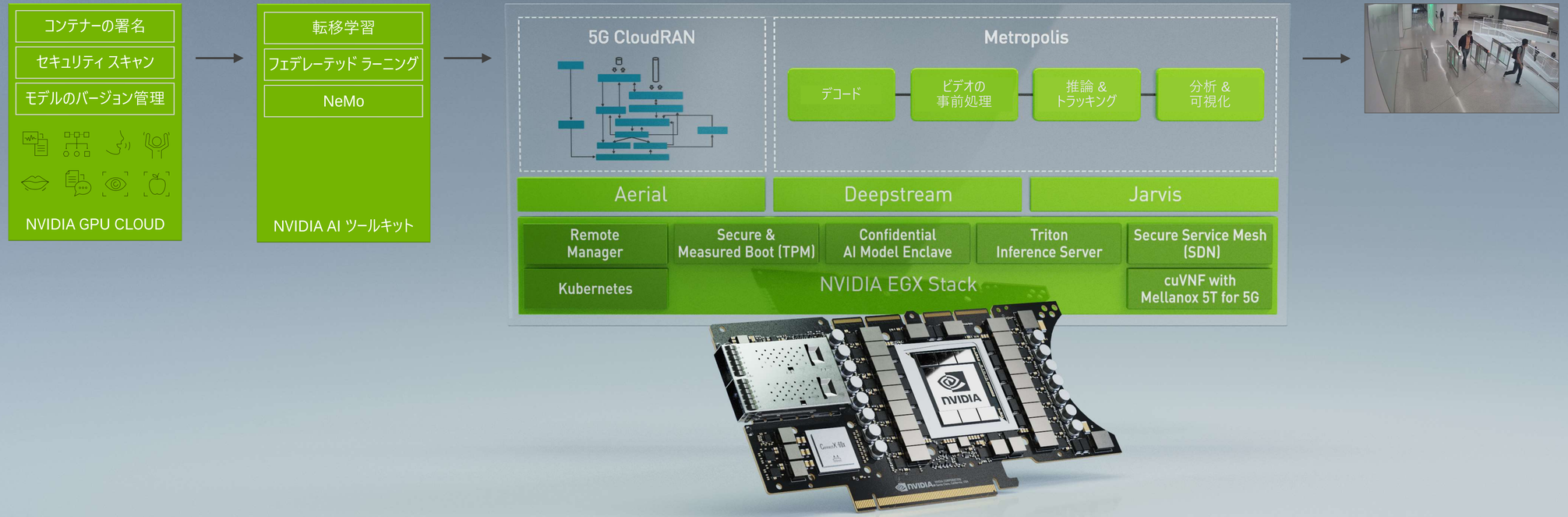

## NVIDIA Ampere GPU

第 3 世代 Tensor コア

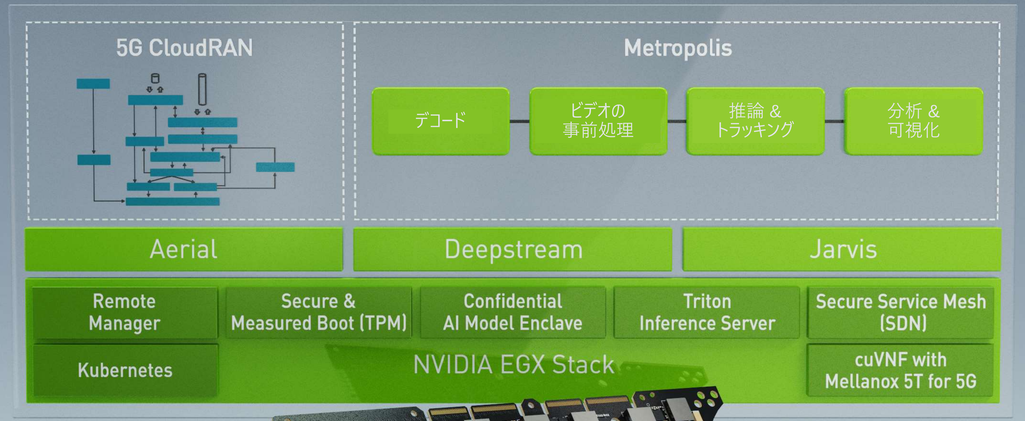
対話型 AI 向けの新しいセキュリティエンジン

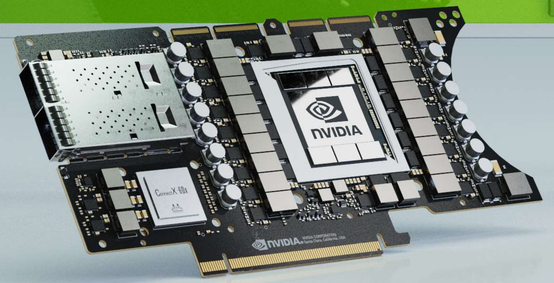
セキュアで認証済みのブート

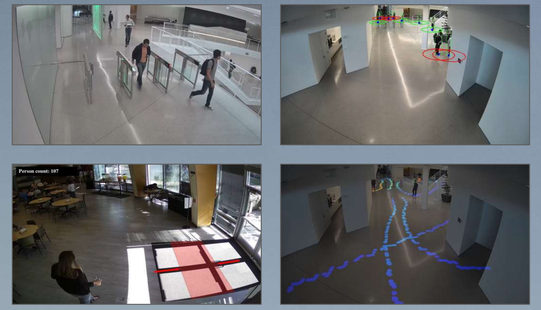
# 発表:
# NVIDIA EGX エッジ AI プラットフォーム



# NVIDIA EGX で動作する
# METROPOLIS ビデオ AI と AERIAL 5G

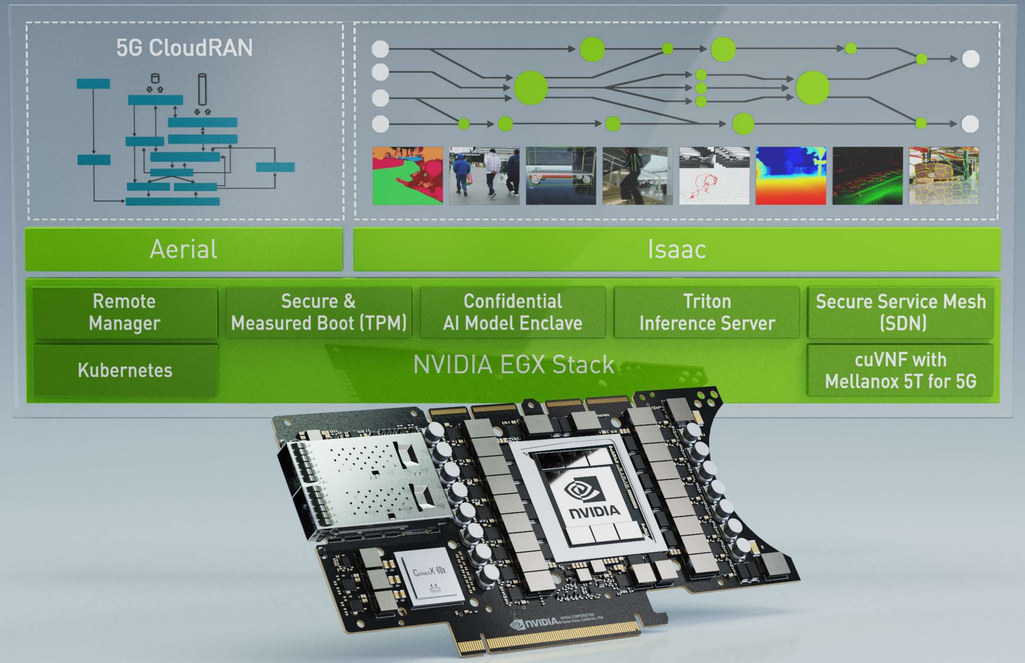



# NVIDIA EGX で動作する
# ISAAC ロボティクス ファクトリと AERIAL 5G

実際の工場




バーチャル工場
デジタル ツイン


KEEP CLEAR

# 発表:
# BMW が NVIDIA ISAAC ロボティクスを採用

**“The Power of Choice”**

40 以上の BMW 車種、車あたり 100 オプション

99% の注文がカスタマイズ/ユニーク

$2^{100}$ 種類の可能な構成

**“Raw Parts In, Parts Trays Out”**

3,000 万のパーツが毎日到着

1.800 社のサプライヤーから、31 の工場

23 万のパーツ ナンバー

**“Just in Time, Just in Sequence”**

ラインあたり最大 10 台

56 秒ごとに新しい車を生産

SplitBot

PickBot

PlaceBot

Smart Transport Robot (STR)

SortBot

# NVIDIA EGX エコシステム

## 5G & CloudRAN

ALTRAN ERICSSON MAVENIR Viasat

## セキュリティ & ネットワーキング

CUMULUS FORTINET JUNIPER nuagenetworks From Nokia

## インフラストラクチャ

IBM Red Hat vmware WIND

## システム

Atos DELL Technologies Hewlett Packard Enterprise Lenovo

## クラウド

Baidu 百度 Microsoft


## 対話型 AI

gnani.ai Intelligent Voice KENSHO S&P Global voca.ai

## インテリジェント ビデオ分析

AIFI DEEPVISION MALONG IRONYUN SAFR from RYAM

## ロボティクス

BMW DENSO FANUC FOXCONN KOMATSU

## 医療

CHOOCH DawnLight GE Healthcare UCSF WHITEBOARD COORDINATOR

## 工業

Micron MUSASHI SAMSUNG SEAGATE tsmc

## 業界のリーダー

goli P&G SoftBank UNITED STATES POSTAL SERVICE Walmart

# ORIN と AMPERE 搭載の NVIDIA DRIVE
# 5W から 2,000 TOPS まで — ひとつのプログラマブル アーキテクチャ

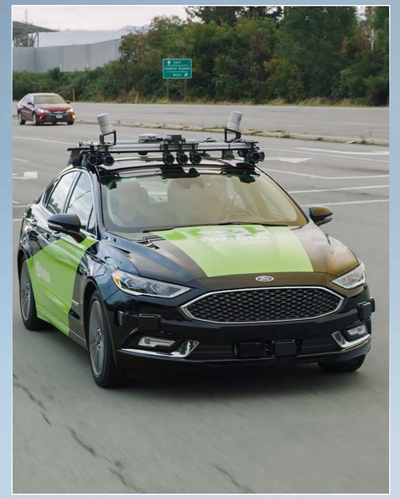
ADAS
Windshield NCAP
10 TOPS、5W

L2+
Autopilot
200 TOPS、45W

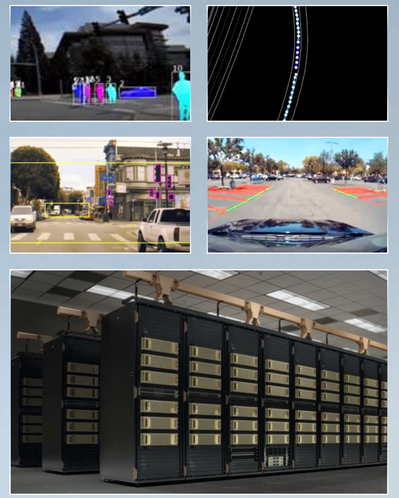
L5
Robotaxi
2,000 TOPS、800W

# NVIDIA DRIVE — ソフトウェアデファインド AV プラットフォーム

データ収集



モデルの学習



シミュレーション

DRIVE AV

DRIVE IX

DRIVE RC

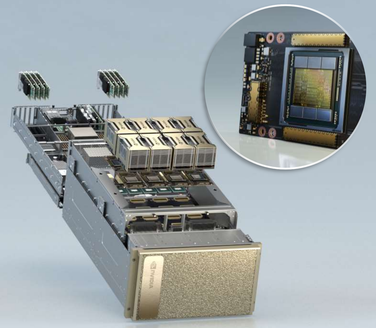
# NVIDIA DRIVE グローバル エコシステム

| 乗用車 | トラック | サプライヤー | 移動サービス | スタートアップ | ソフトウェア | マッピング | シミュレーション |
|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

# GTC 2020 発表

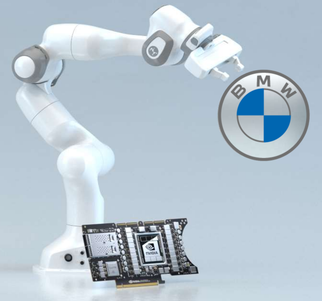

データセンタースケール
コンピューティング

Omniverse RTX サーバー



A100 と DGX A100


NVIDIA AI




EGX と ISAAC


NVIDIA