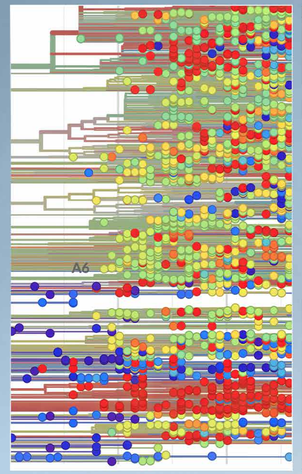
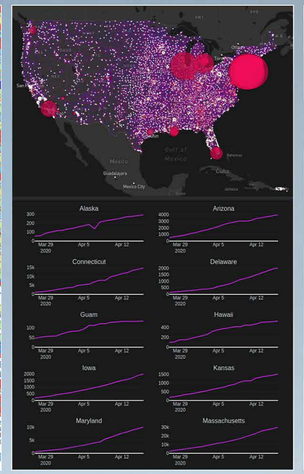

NVIDIA


THANK YOU
HEROES

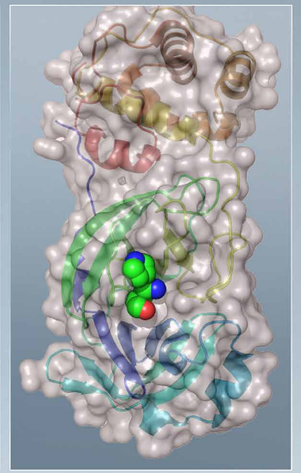
NVIDIA 抗击新冠肺炎 (COVID-19)



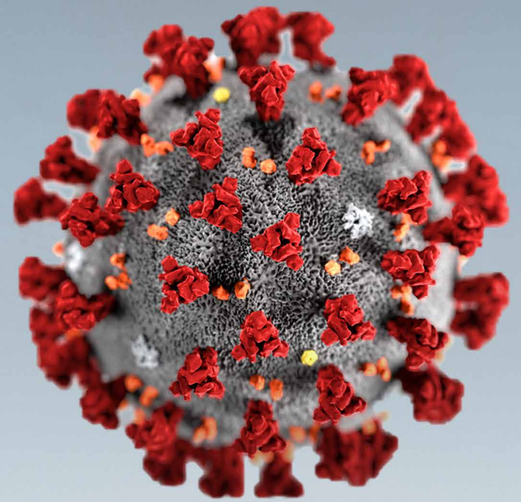
Oxford Nanopore
在 7 小时内完成
病毒基因测序



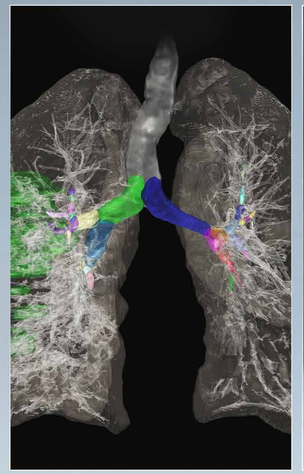
Plotly，NVIDIA
实时感染率分析



ORNL、Scripps
在一天内筛查
10 亿个药物化合物
而以前需要一年时间

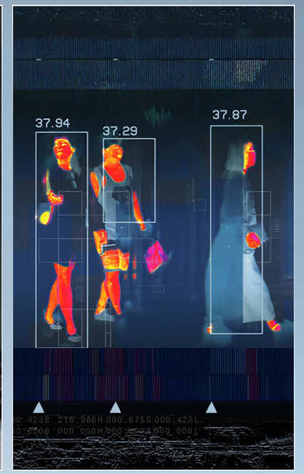


美国国立卫生研究院 (NIH) Structura
团队，德州大学奥斯丁分校
CryoSPARC
第 1 个病毒刺突蛋白三维结构



NIH，NVIDIA
AI COVID-19 分类

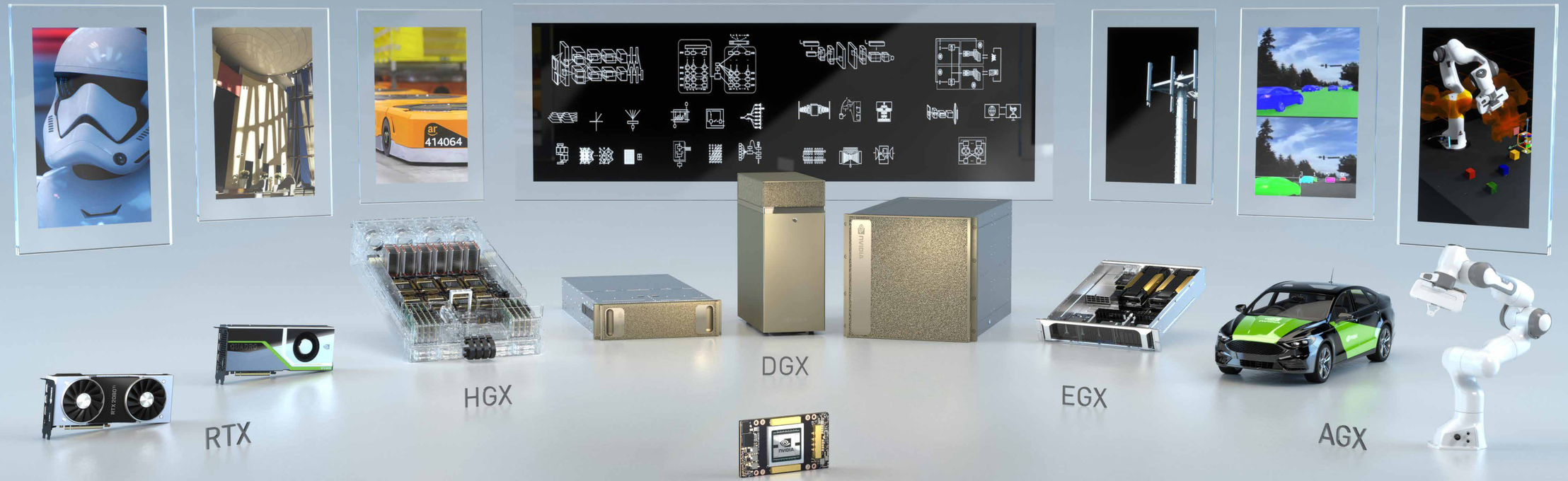
Kiwibot
通过机器人递送医疗物资

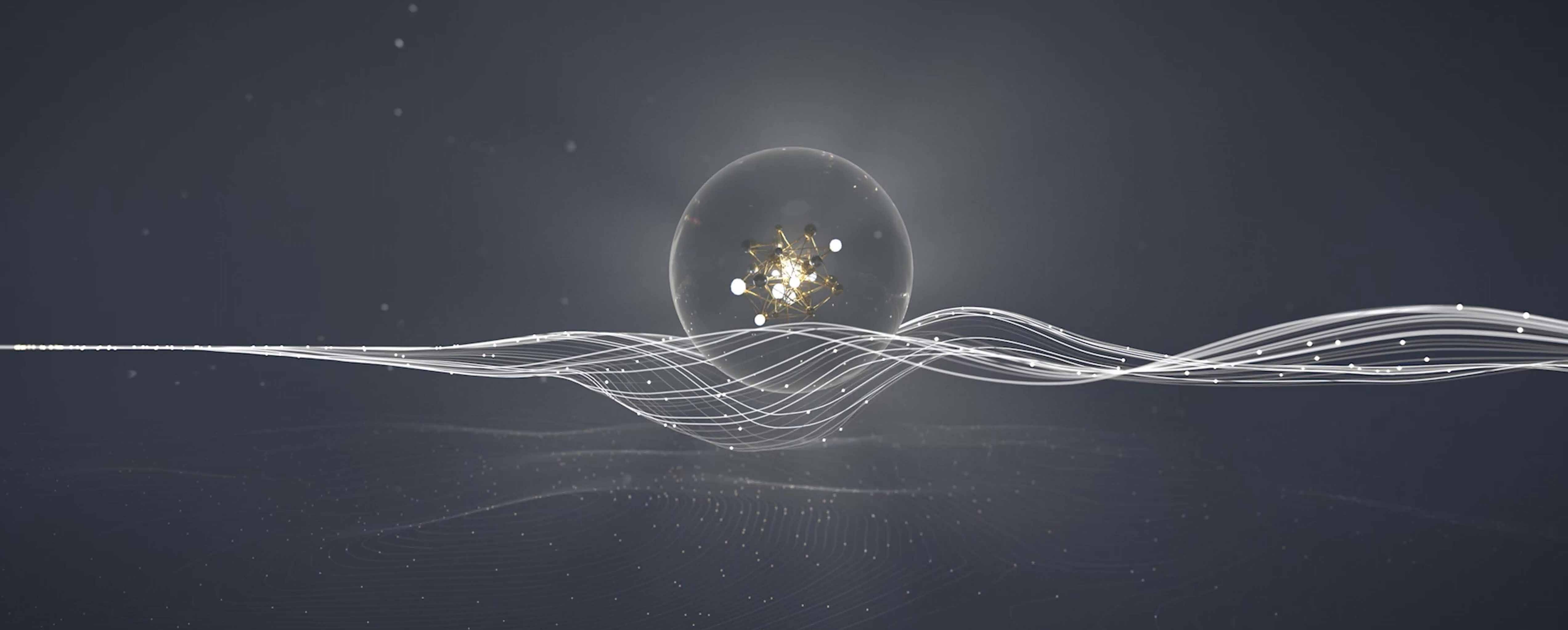


Whiteboard Coordinator
AI 体温升高筛查系统

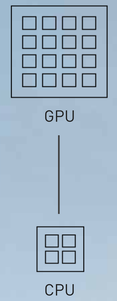
控制 — 缓解 — 治疗 — 跟踪和监控

为当代的达芬奇提供出色计算性能

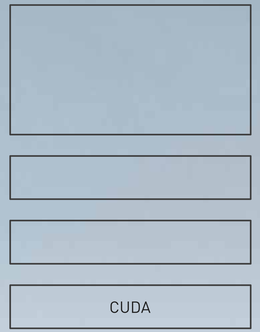




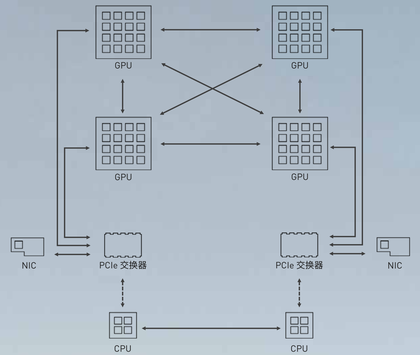
在加速计算领域深耕 25 年



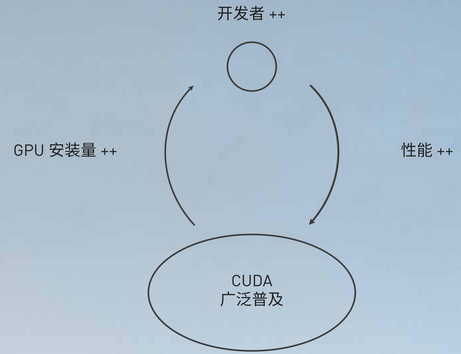
X 倍提速



完整加速技术栈

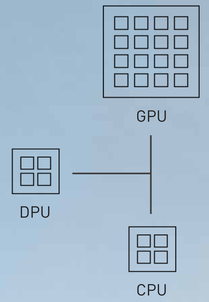


系统

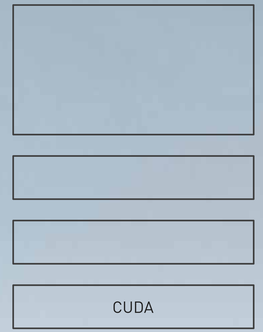


同一个架构

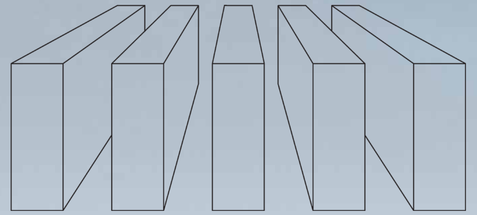
数据中心级加速计算



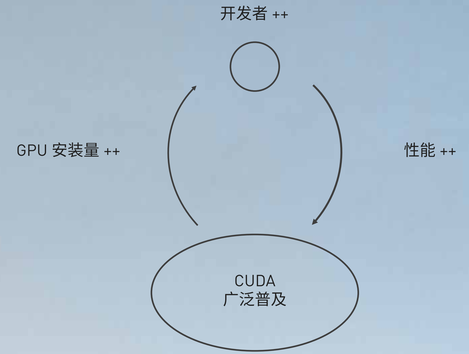
X 倍提速



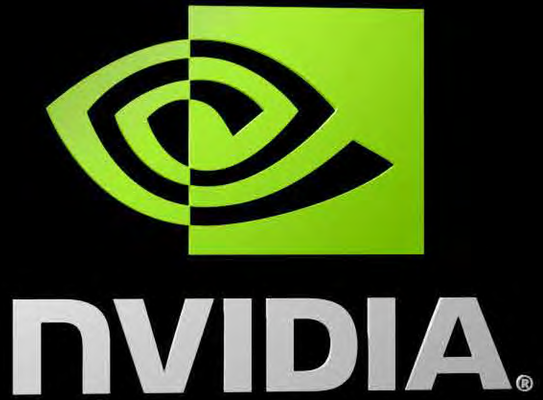
完整加速技术栈



数据中心规模

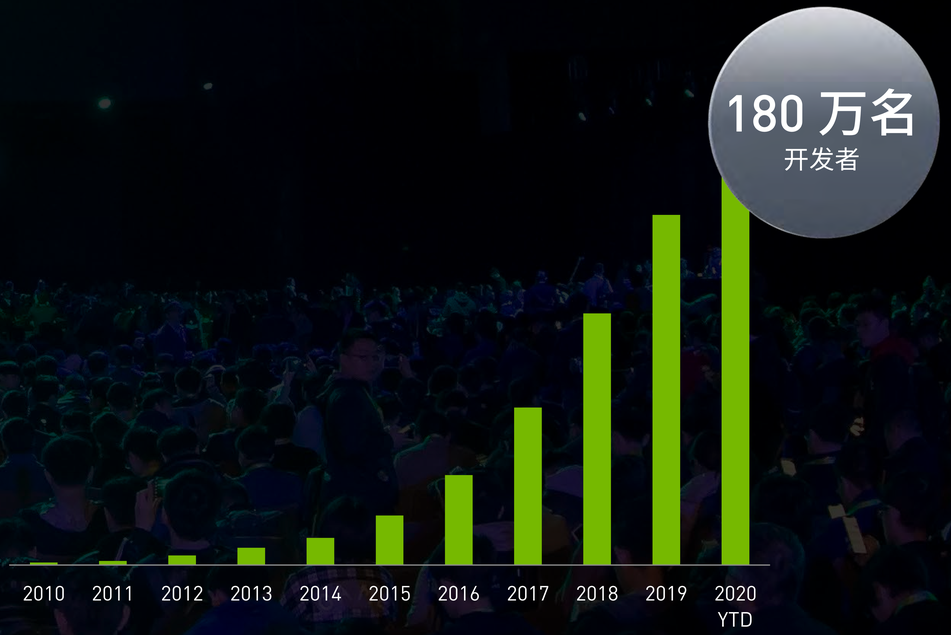


同一个架构


NVIDIA


Mellanox
TECHNOLOGIES

对于我们的开发者而言，这是美妙的一年

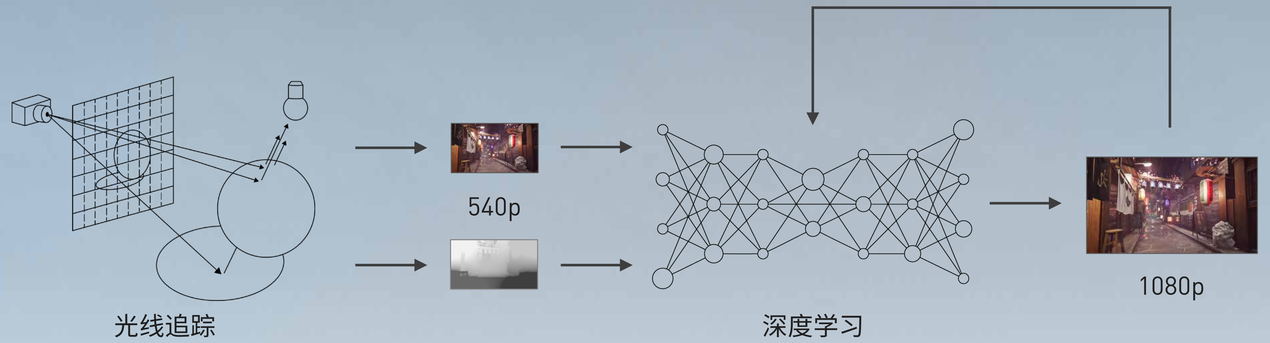


2018 年 8 月 13 日，NVIDIA 在 SIGGRAPH 大会上发布了 RTX，开创了计算机图形技术的新时代。

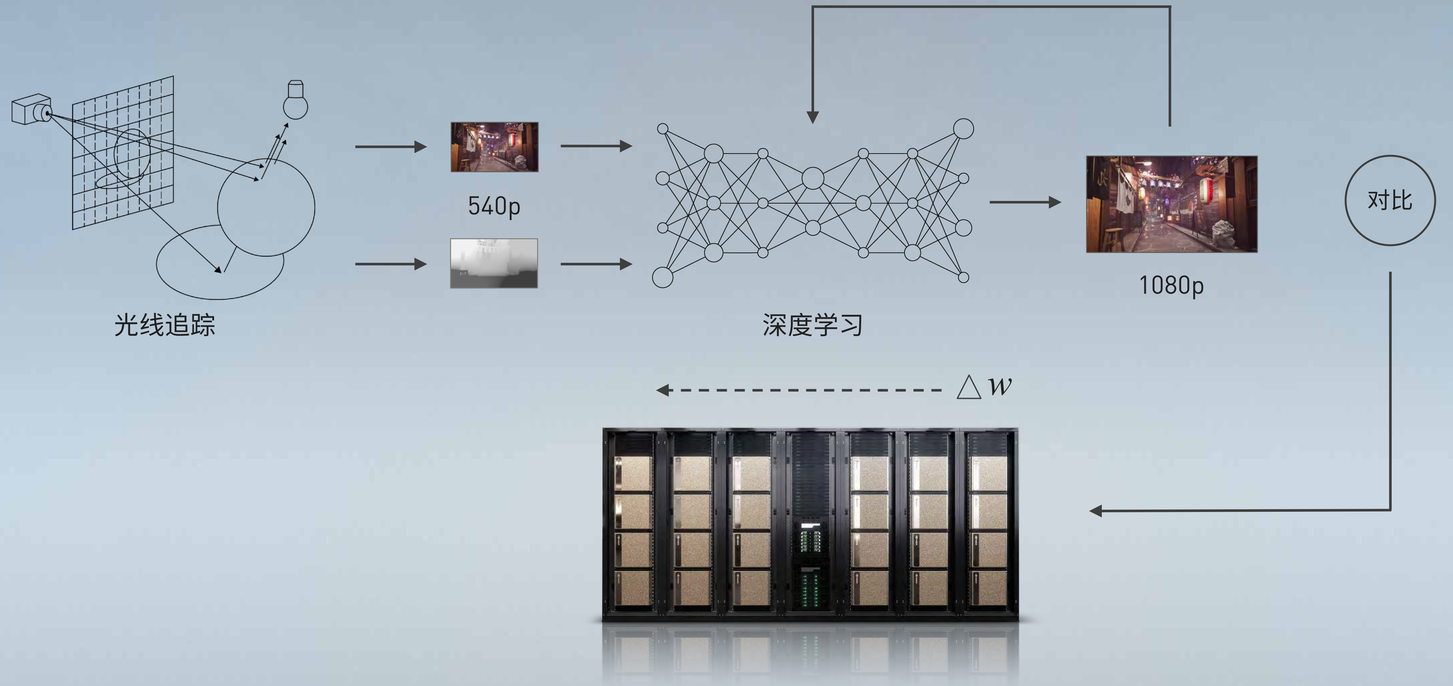
“我们在过去 30 余年间一直承诺提供的实时光线追踪技术终于问世 – 比预期提前了十年。”

– Jon Peddie，JPR

NVIDIA RTX
计算机图形技术的新时代 – 光线追踪与 AI

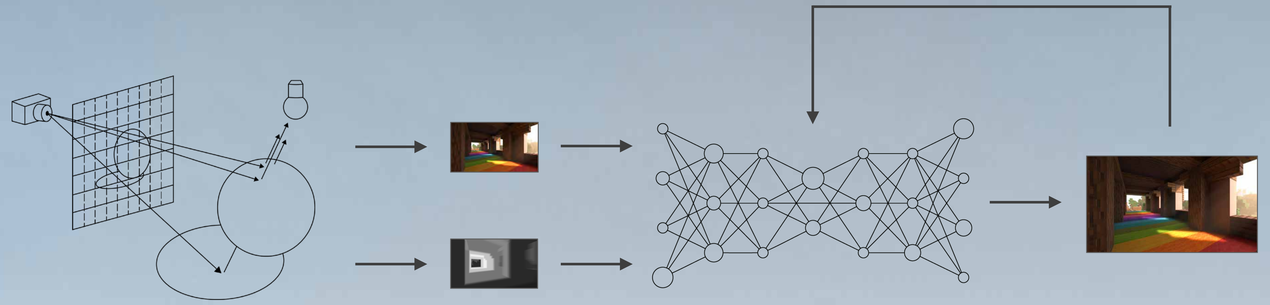


NVIDIA RTX
计算机图形技术的新时代 – 光线追踪与 AI



由超级计算机渲染 – 16K 真值

NVIDIA RTX
计算机图形技术的新时代 – 光线追踪与 AI




真值 16K


原生 720p


DLSS 1.0
720p > 1080p


DLSS 2.0
720p > 1080p

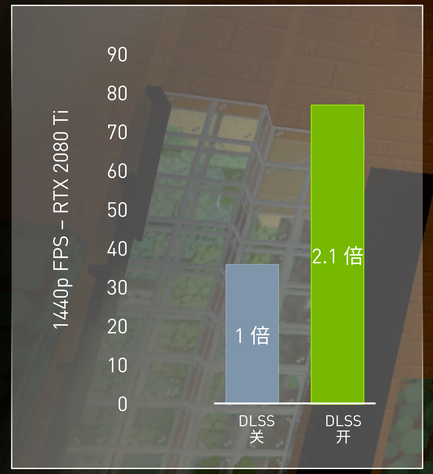

原生 1080p


真值 16K


原生 540p


DLSS 2.0
540p > 1080p

MINECRAFT



MOJANG

“变革者”

– Digital Foundry

“这真的太华丽了，让人叹为观止”

– PCWorld

“简直是太棒了”

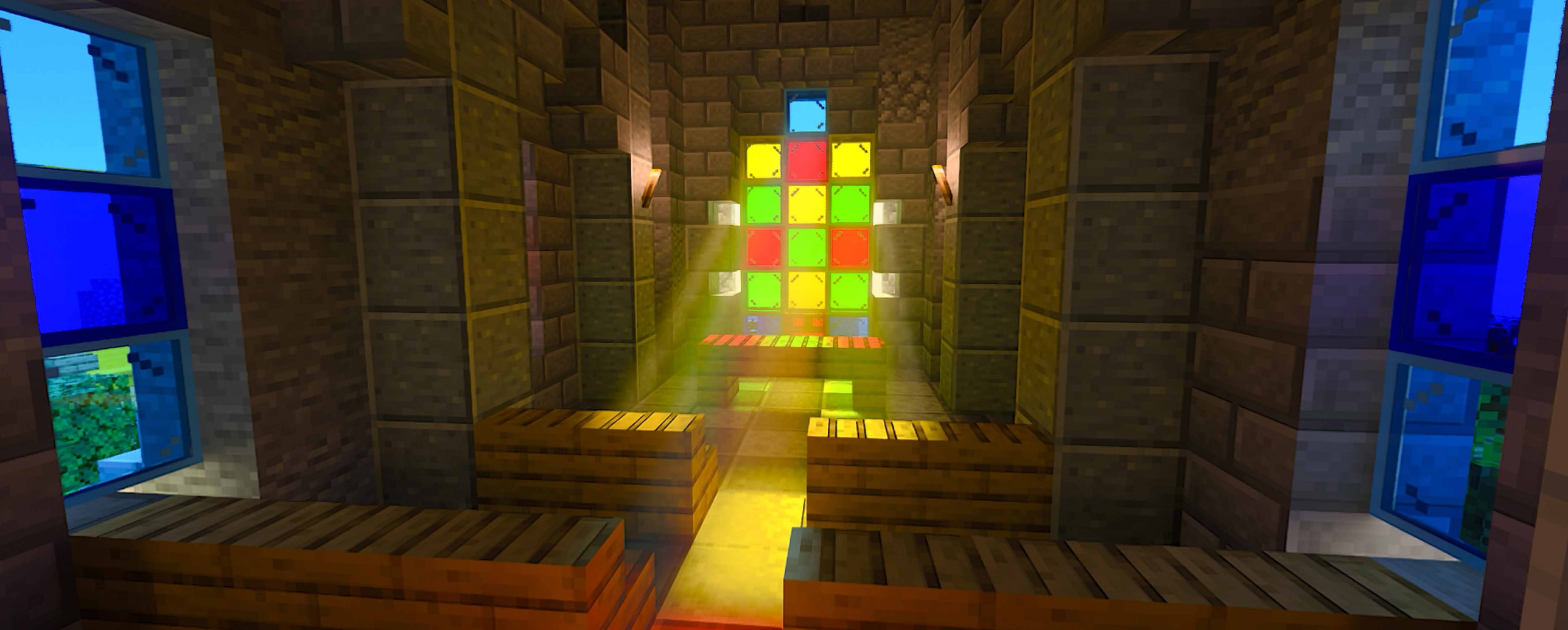
– IGN

“令人惊艳”

– PC Gamer

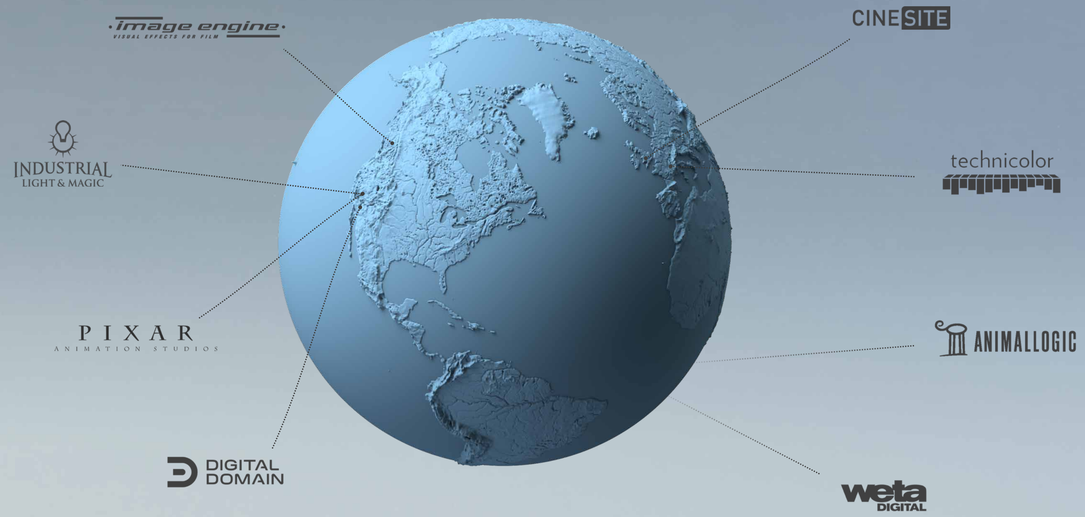
“令人瞠目”

– Trusted Reviews



3D 技术非常复杂

不同的工具和无比庞大的数据集
由各类技术专家组成的大型团队
涉及多个地点和工作室费用高昂



NVIDIA OMNIVERSE

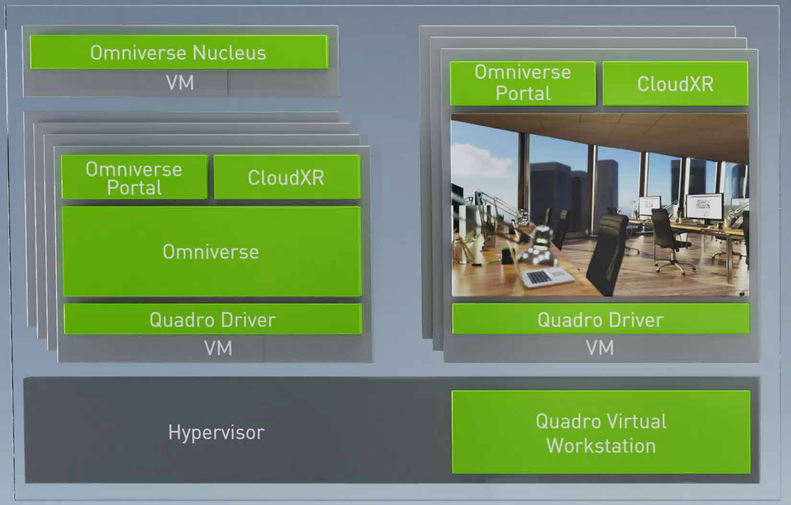
设计工作流程协作平台

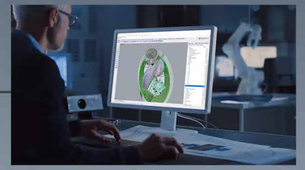
基于通用场景描述 (USD) 打造

具有材质和物理效果的内置交互式渲染器

支持 PC 和 Linux；适用于 Mac、Android 的流式传输客户端

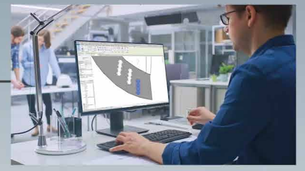
丽泽 SOHO，北京
由 ZAHA HADID ARCHITECTS 设计





RHINO

MAX



REVIT

AR

丽泽 SOHO，北京
由 ZAHA HADID ARCHITECTS 设计


HOORKA
OIL COLOUR

宣布推出
针对远程协作优化的
NVIDIA RTX 服务器

通过 Omniverse 开展设计工作流程协作
支持全光线追踪全局照明的交互式生产渲染
经过设计和模拟验证的 Quadro 虚拟工作站

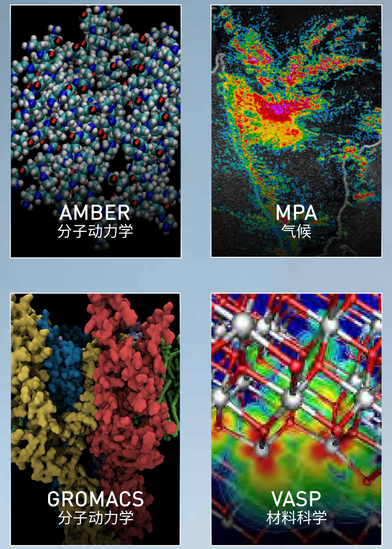
NVIDIA RTX 服务器现已发售

BOXX DELL Technologies

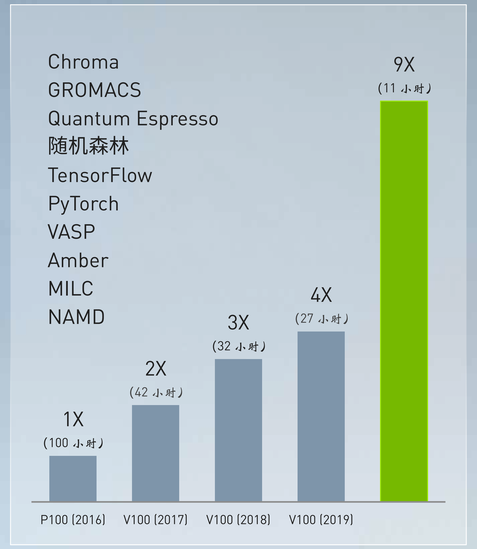
Hewlett Packard Enterprise SUPERMICRO

丽泽 SOHO，北京
由 ZAHA HADID ARCHITECTS 设计

NVIDIA HPC



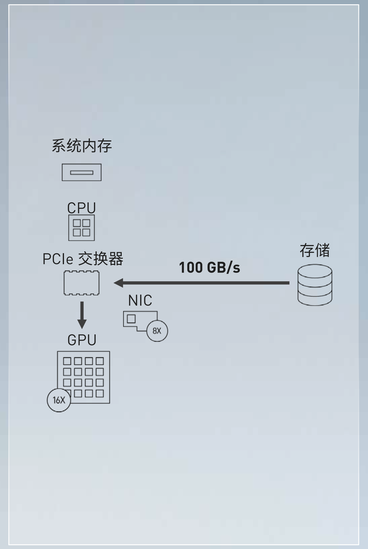
700 多款 CUDA 加速应用程序



在 4 年内实现
9 倍性能提升

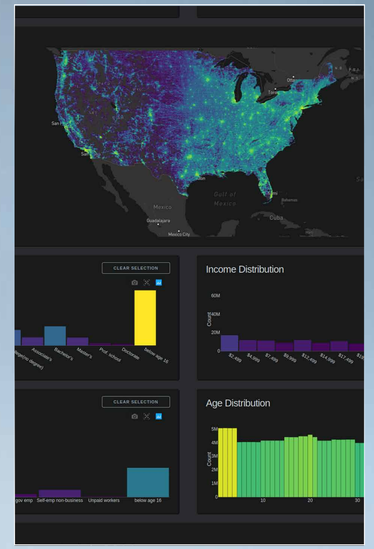
HPE Apollo70
Marvell ThunderX
ARM CPU

基于 ARM 架构的 CUDA



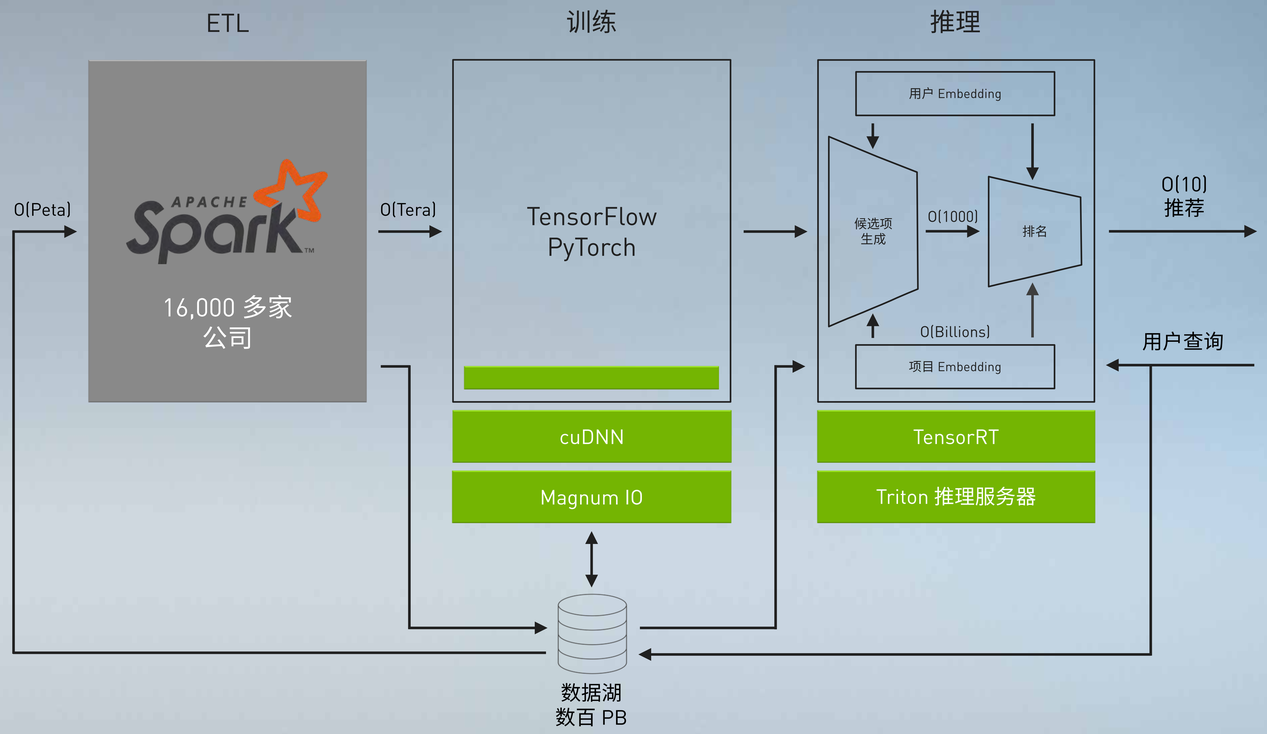
NVIDIA MAGNUM-IO
I/O 加速

NVIDIA PARABRICKS
基因组学

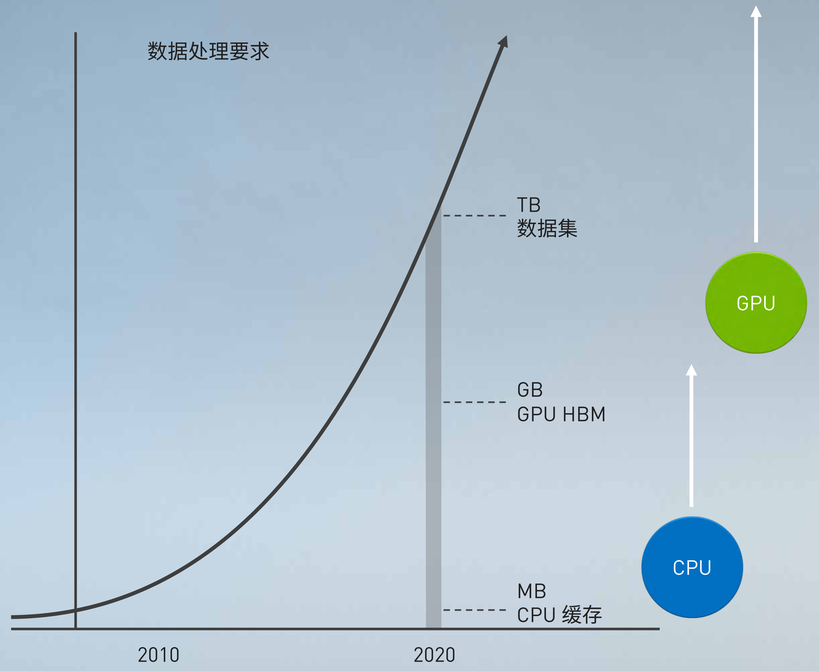


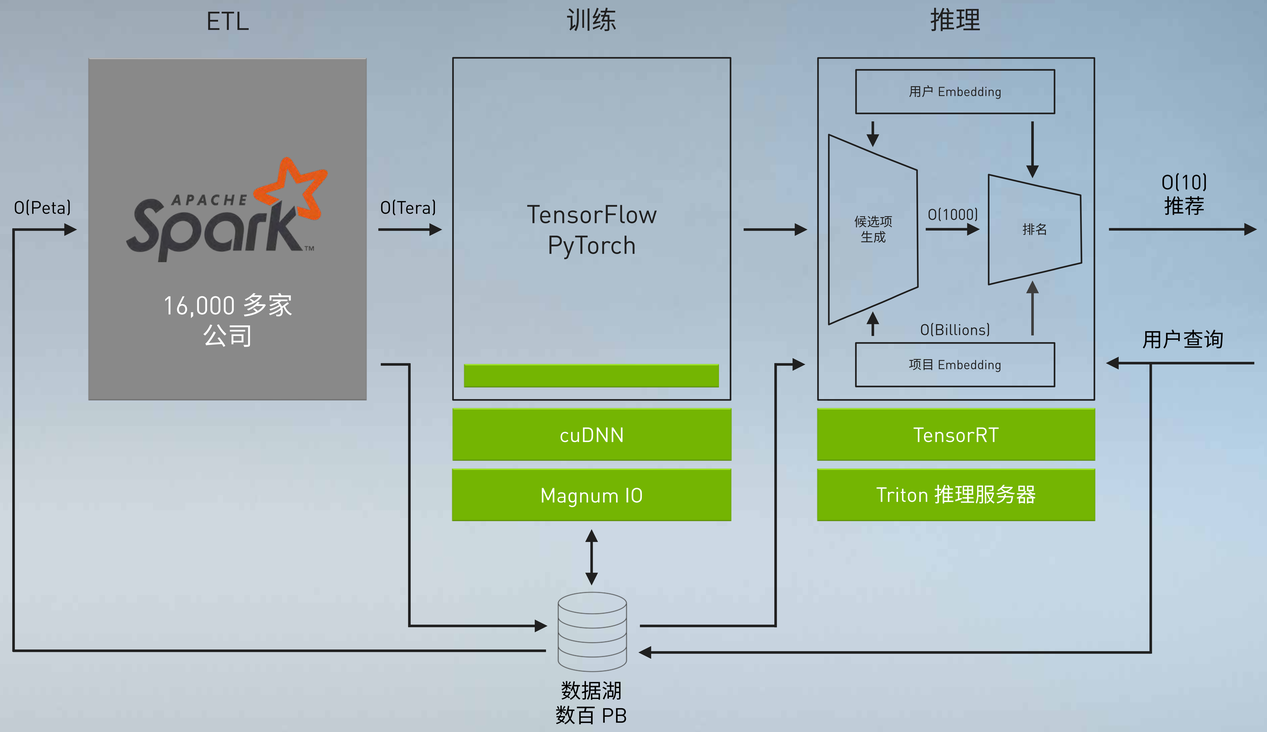
NVIDIA RAPIDS
数据分析

机器学习流程是一项 HPC 挑战

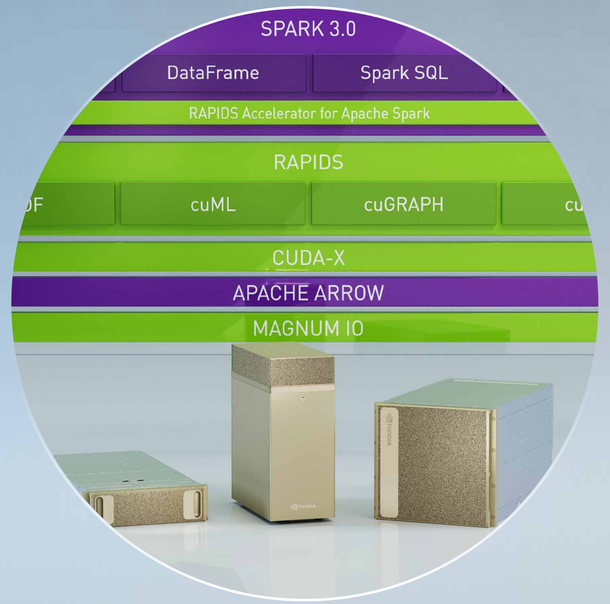


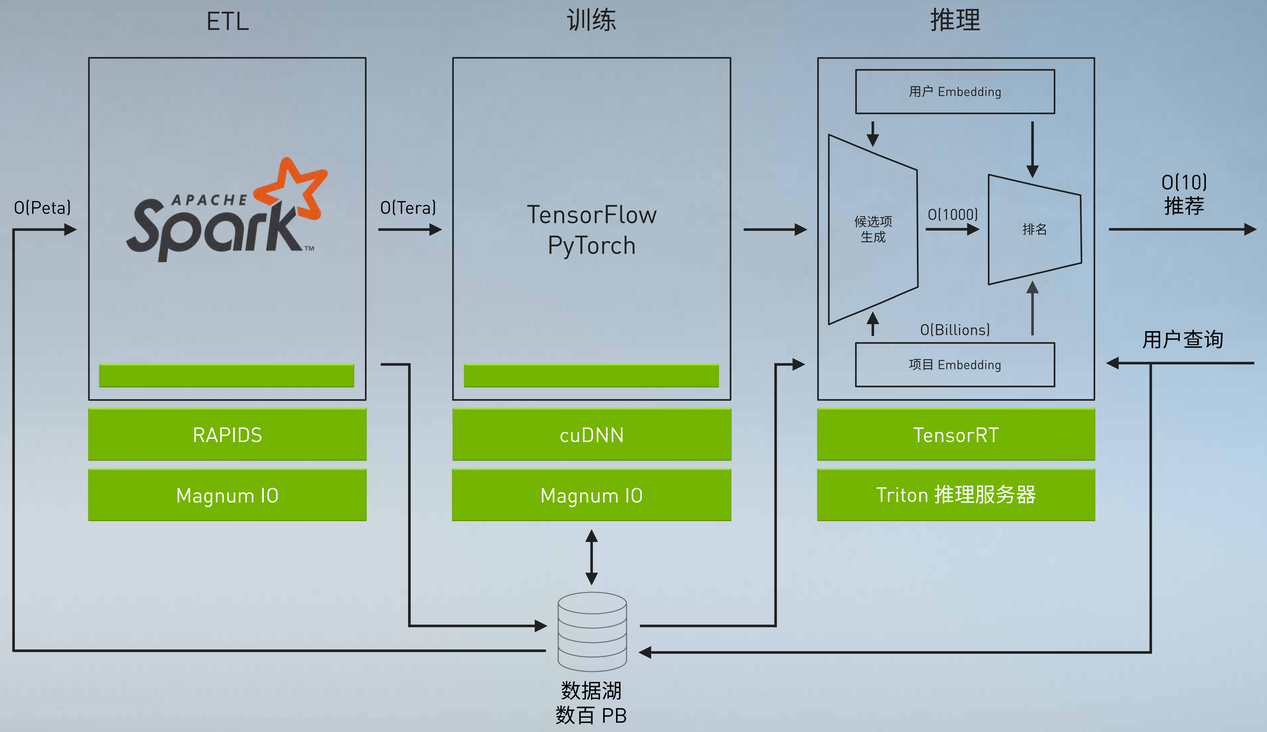
机器学习促使数据呈指数级增长





宣布
NVIDIA 助力加速 SPARK 3.0

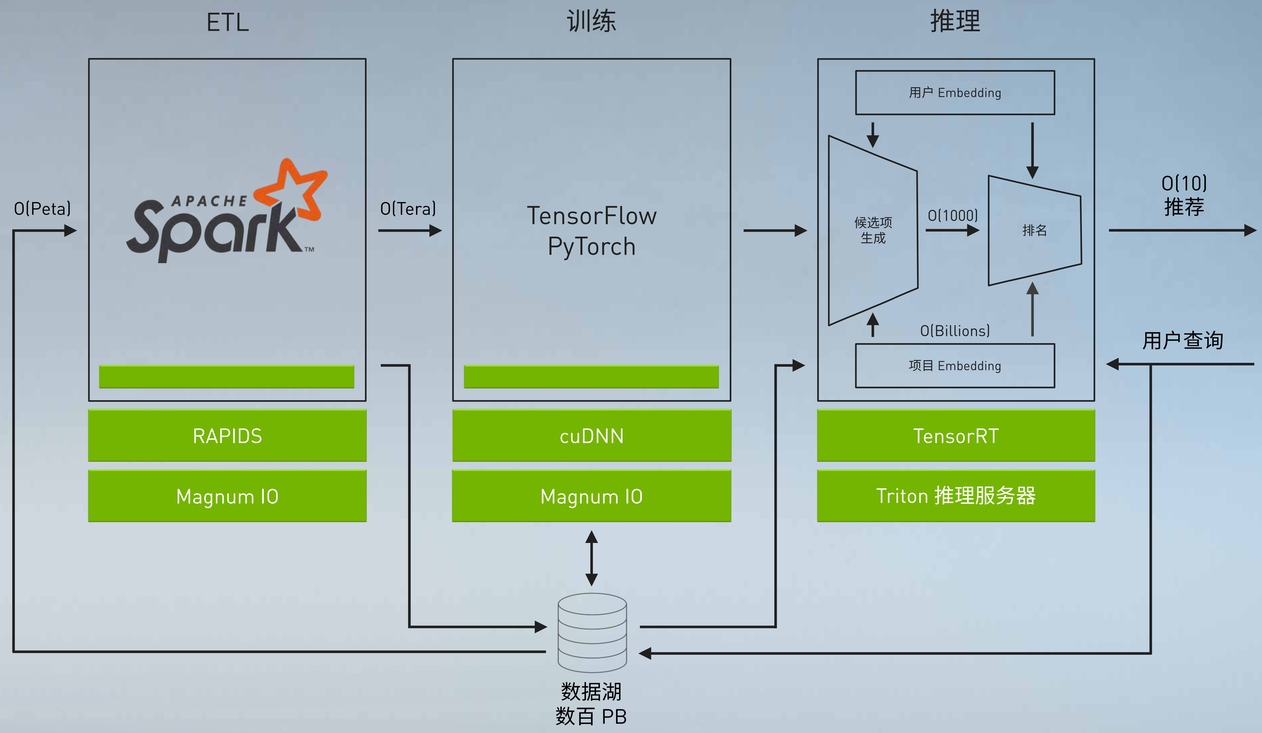




SPARK 3.0 基于先进的基础技术构建
RAPIDS 打破 ETL 基准测试记录

在 TPCx-BB @ SF 10K 条件下达到 17 GB/s 的吞吐量

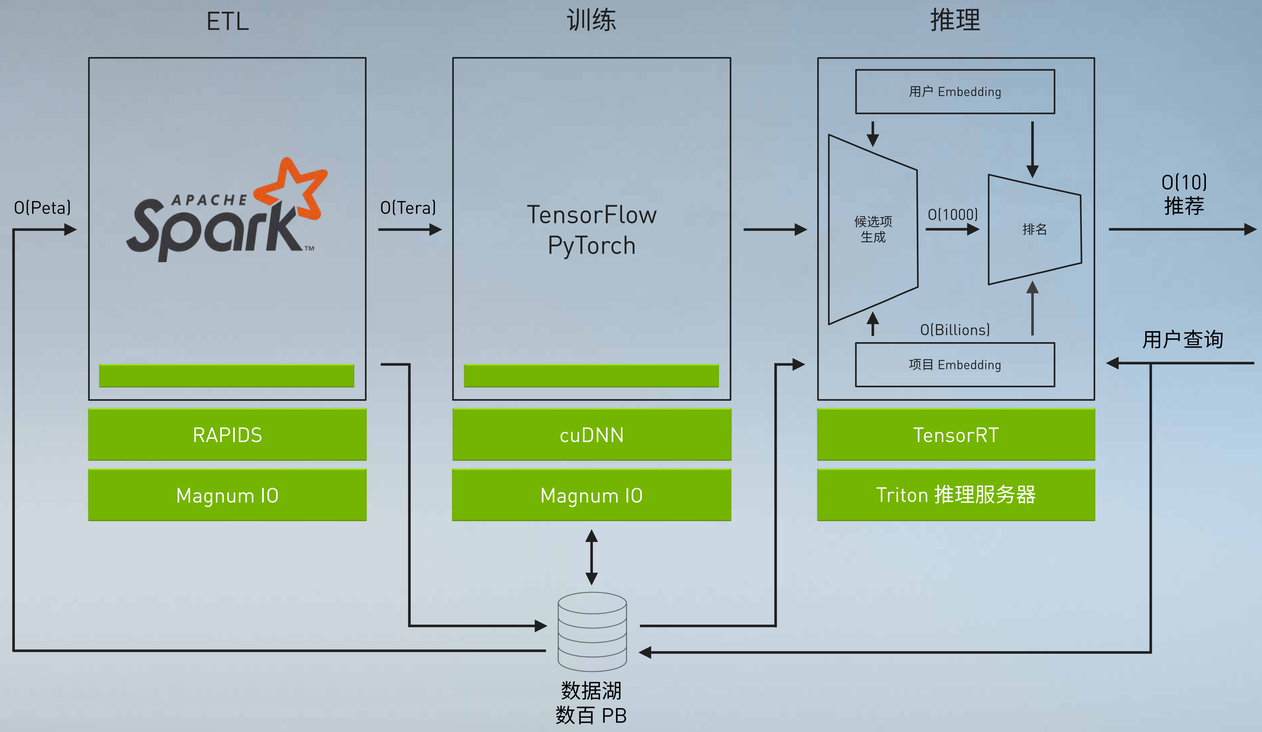
100 万美元 | 18 个 2U CPU 服务器 | 2 个机架 | 16 kW



SPARK 3.0 基于先进的基础技术构建
RAPIDS 打破 ETL 基准测试记录

在 TPCx-BB @ SF 10K 条件下使用 RAPIDS 可达到 163 GB/s 的吞吐量

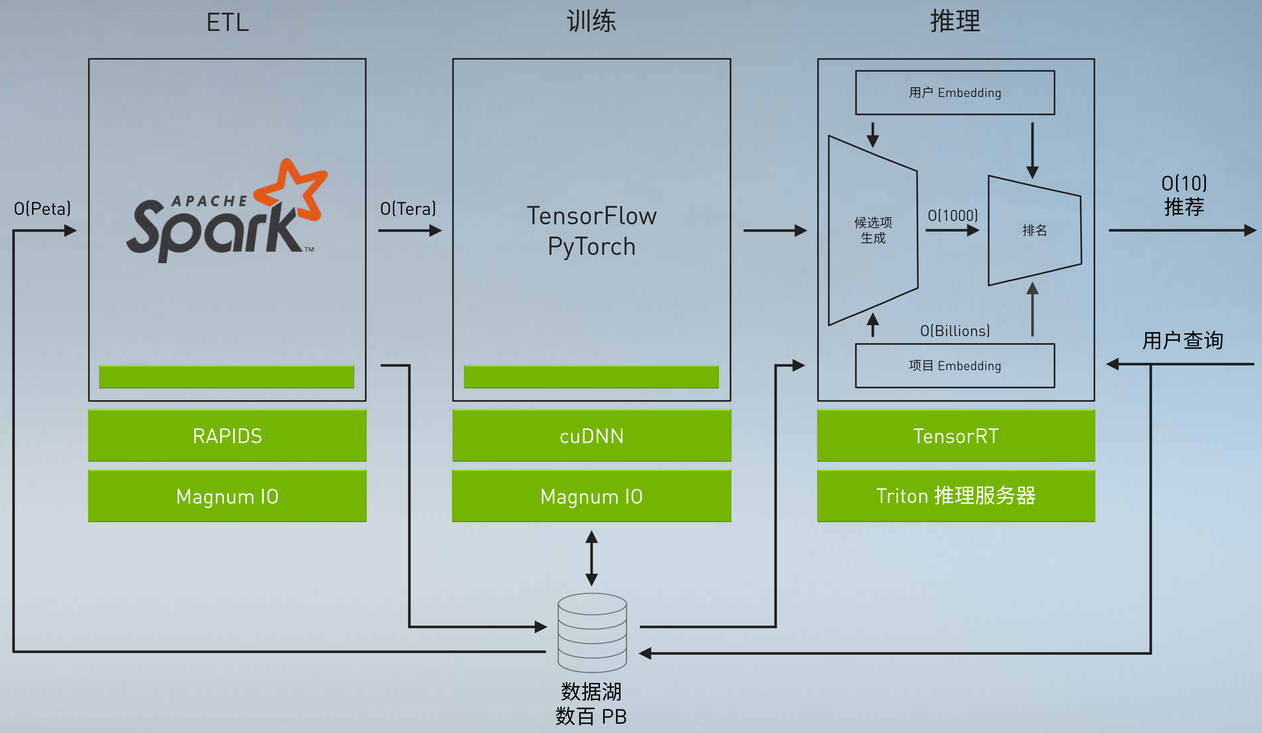
200 万美元 | 16 个 DGX-1 | 2 个机架 | 56 kW



SPARK 3.0 基于先进的基础技术构建
RAPIDS 打破 ETL 基准测试记录

在 TPCx-BB @ SF 10K 条件下使用 RAPIDS 可达到 163 GB/s 的吞吐量

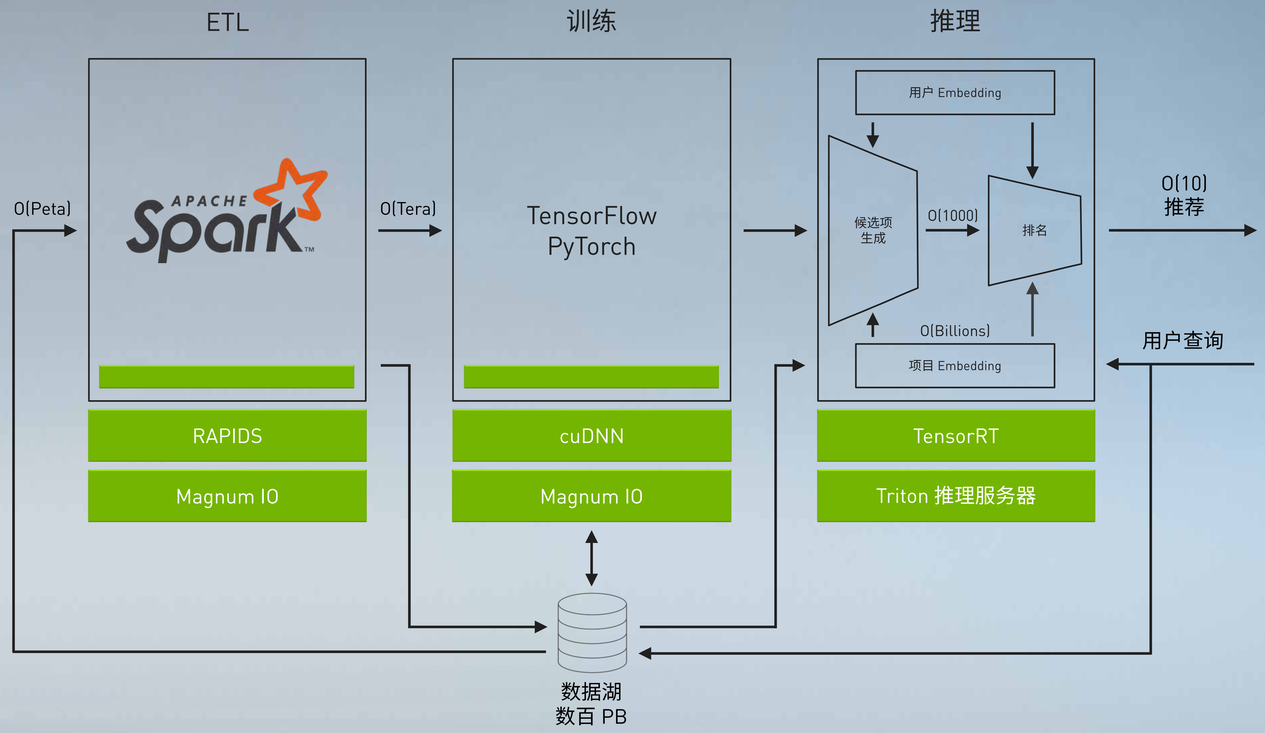
200 万美元 | 16 个 DGX-1 | 2 个机架 | 56 kW



SPARK 3.0 基于先进的基础技术构建
RAPIDS 打破 ETL 基准测试记录

在 TPCx-BB @ SF 10K 条件下实现相当于 163 GB/s 的吞吐量

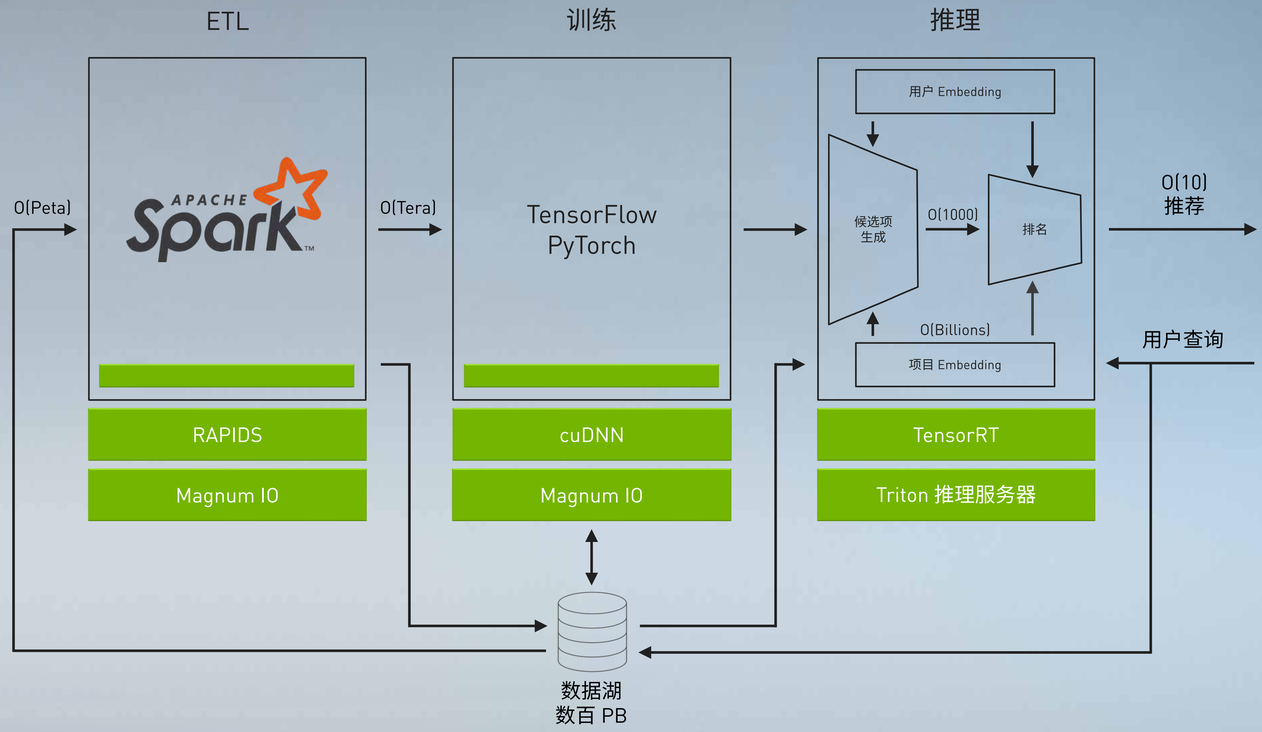
1000 万美元 | 167 个 2U CPU 服务器 | 11 个机架 | 140 kW



SPARK 3.0 基于先进的基础技术构建
RAPIDS 打破 ETL 基准测试记录

在 TPCx-BB @ SF 10K 条件下使用 RAPIDS 可达到 163 GB/s 的吞吐量

200 万美元 | 16 个 DGX-1 | 2 个机架 | 56 kW

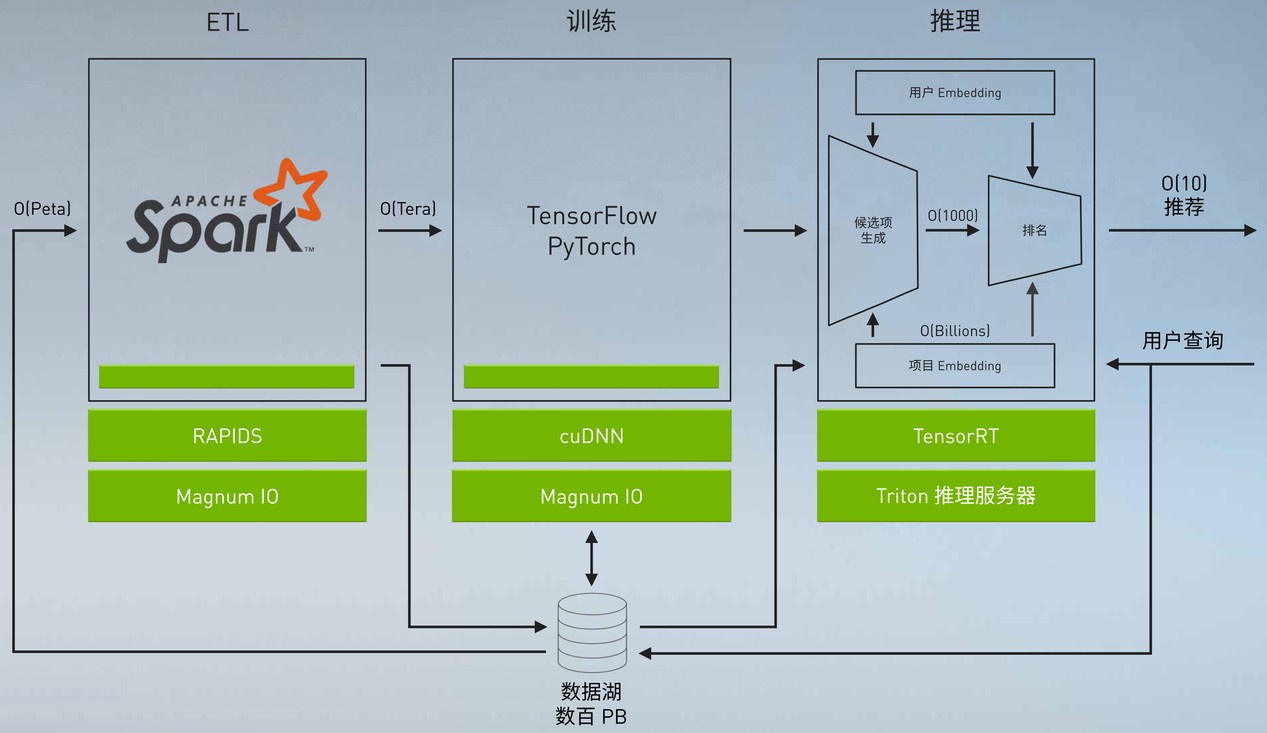


宣布
NVIDIA 助力加速 DATABRICKS

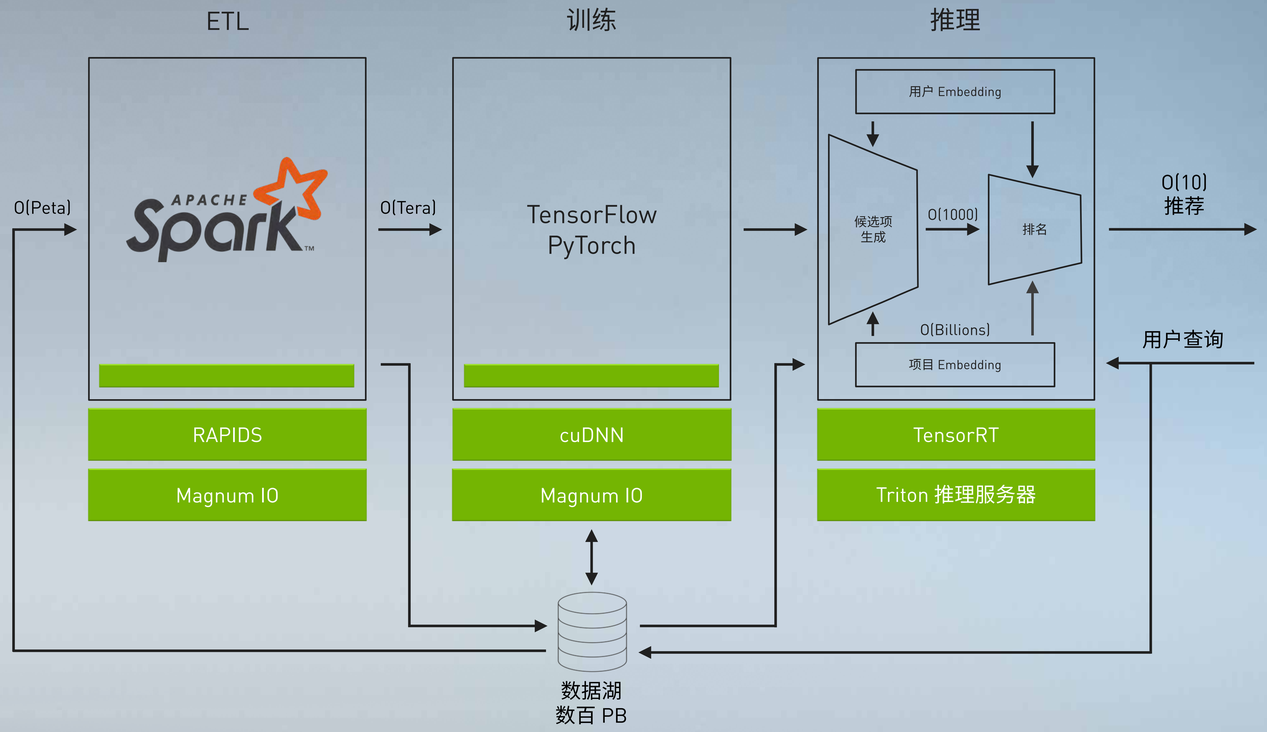
databricks

“这些贡献将带来更高速的数据管道、模型训练和评分，助力在 Apache Spark 3.0 和 Databricks 中取得更多突破、发掘更多见解。”

– Matei Zaharia，Apache Spark 的原始创建者及 Databricks 首席技术专家

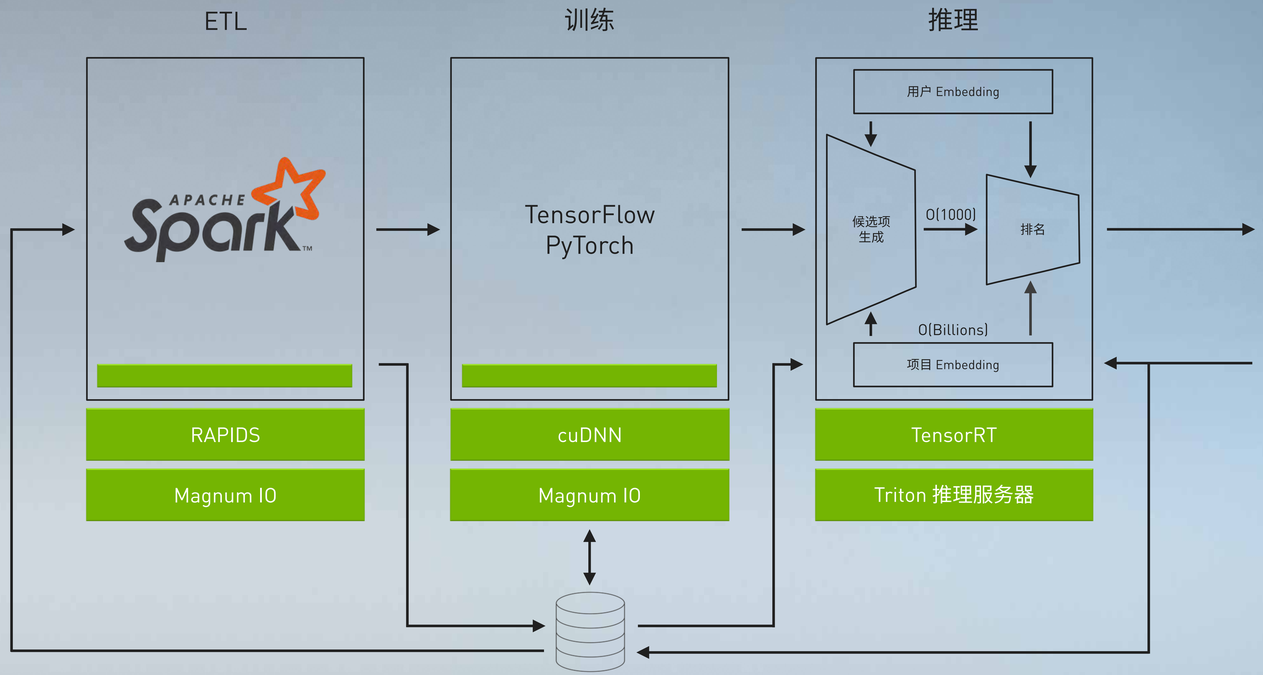


宣布
NVIDIA 助力加速云分析平台

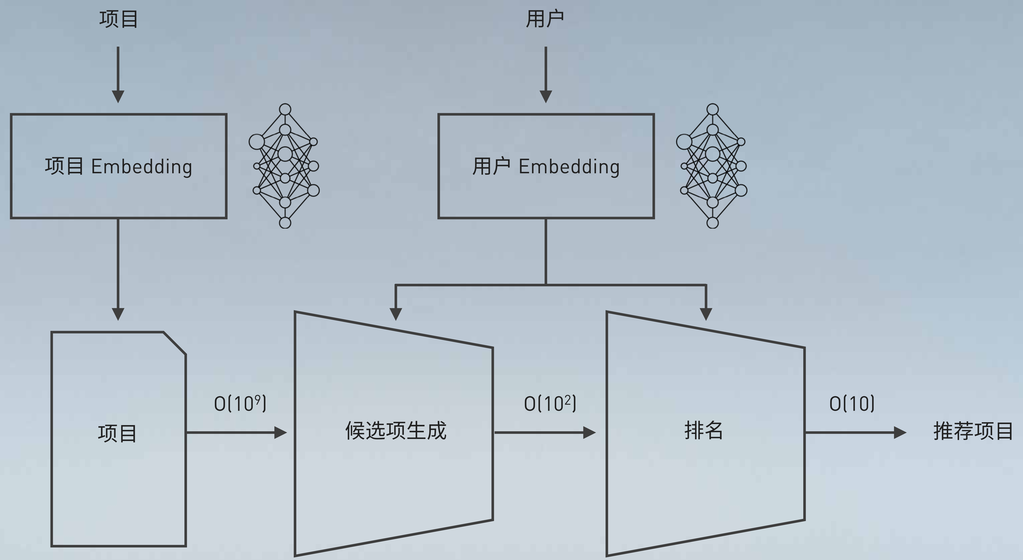


NVIDIA AI


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


ETL
训练
推理
APACHE Spark
TensorFlow
PyTorch
用户 Embedding
候选项
生成
O(1000)
排名
O(Billions)
项目 Embedding
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton 推理服务器

推荐系统是打造个性化互联网的引擎



宣布推出
深度推荐应用程序框架 NVIDIA MERLIN

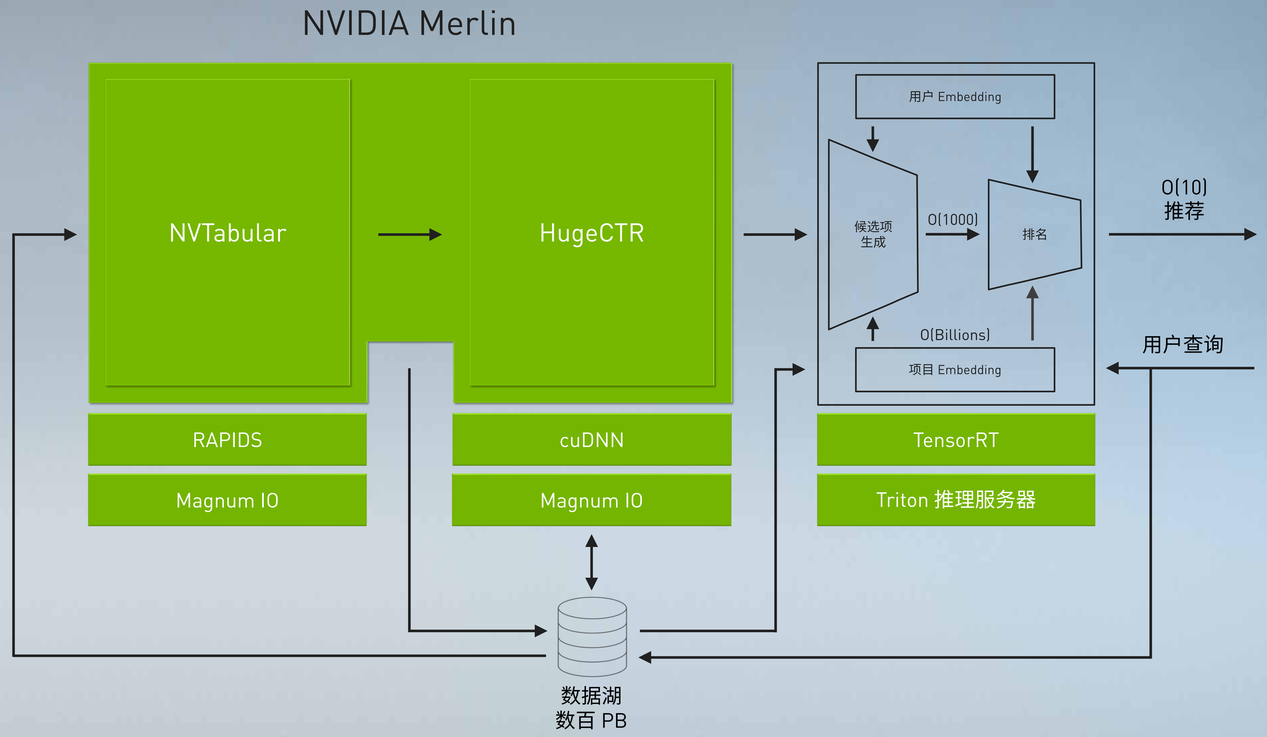
1 TB 广告数据集

CPU

ETL：2 小时　训练：1.5 天

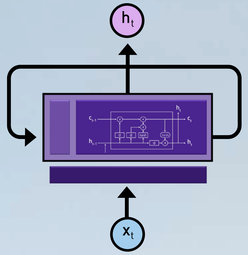
GPU

ETL：3 分钟　训练：16 分钟

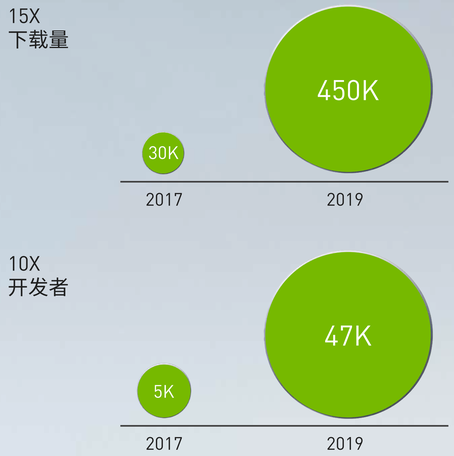


NVIDIA AI 推理
2 年内 TENSORRT 下载量达到 15 倍

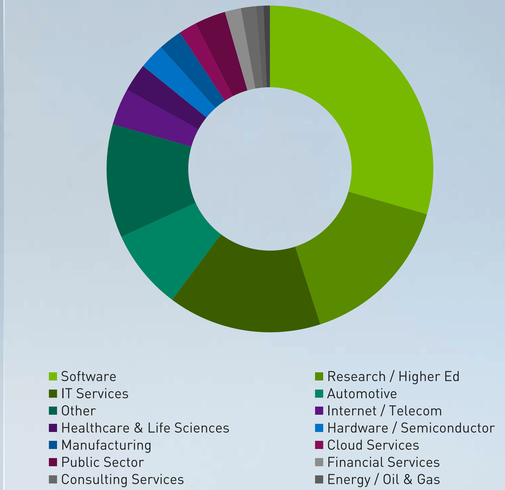
TENSORRT 7



CNN、Transformer、RNN
1000 多个经优化的内核
自动混合精度 FP32、FP16 和 INT8



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

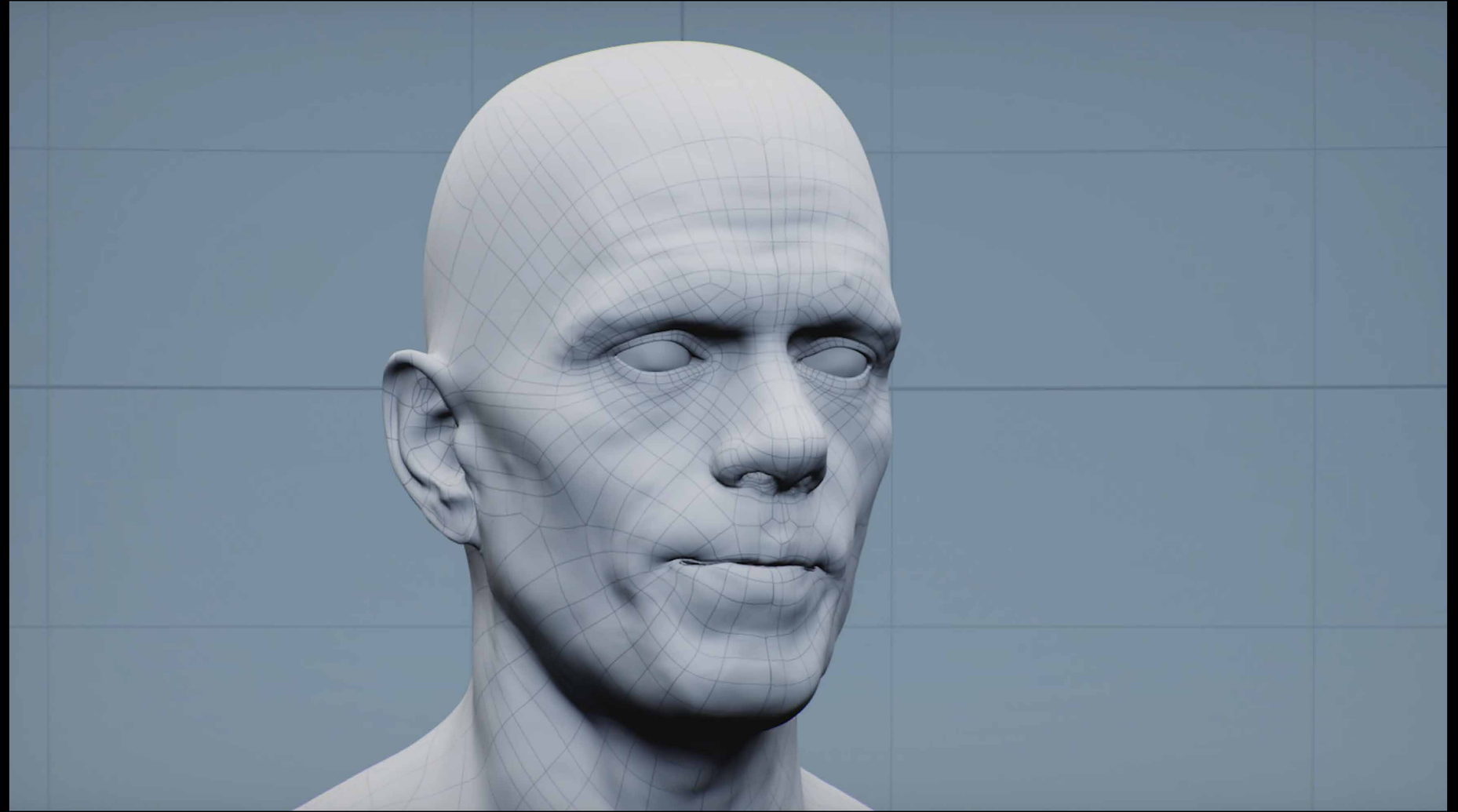


宣布推出
多模对话式 AI 服务框架 NVIDIA JARVIS

ASR NLU TTS A2F

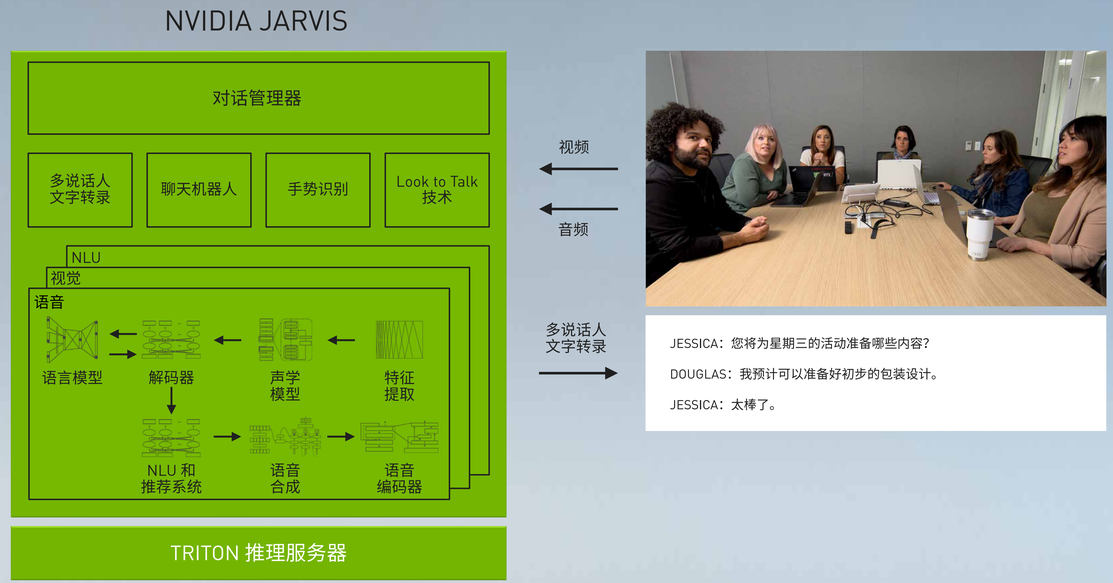
Jarvis

Omniverse



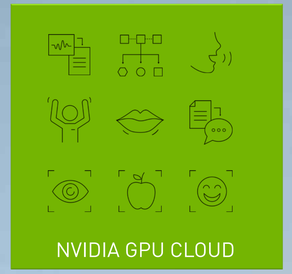
JARVIS 演示
MISTY – 交互式 3D 聊天机器人

宣布推出
多模对话式 AI 服务框架 NVIDIA JARVIS



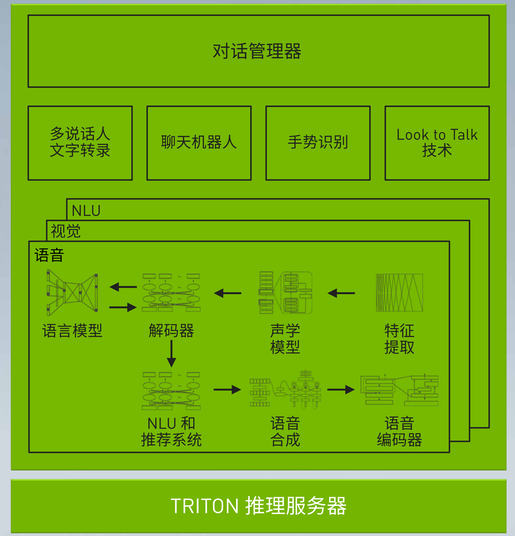
宣布推出
多模对话式 AI 服务框架 NVIDIA JARVIS

经过预先训练的模型



重新训练

NVIDIA JARVIS



JESSICA：您将为星期三的活动准备哪些内容？

DOUGLAS：我预计可以准备好初步的包装设计。

JESSICA：太棒了。

参与抢先体验计划

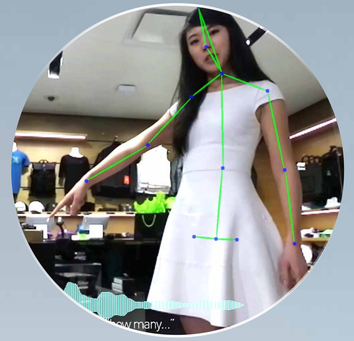
developer.nvidia.com/nvidia-jarvis

对话式 AI 正在各行各业中掀起变革

视频会议
字幕、翻译、文字转录
每天 2 亿场会议

呼叫中心
每天 5 亿通电话

智能音箱
每年售出 1.5 亿台



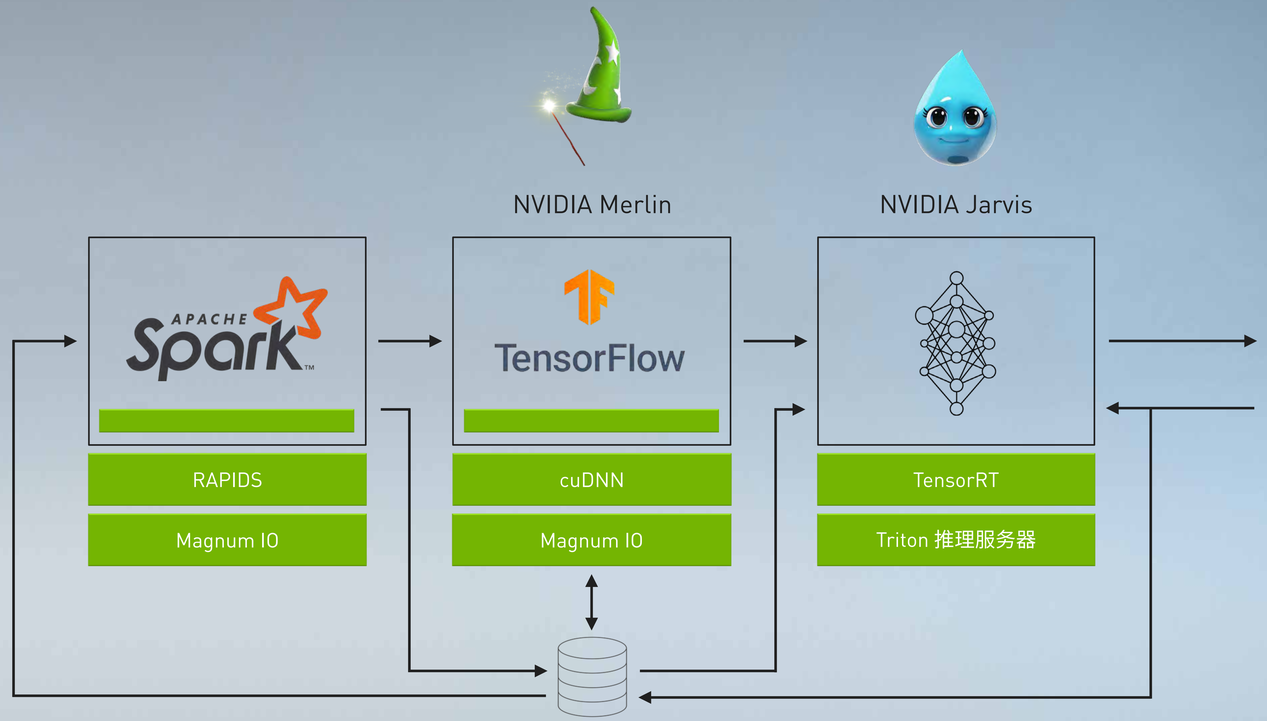
零售助理
1200 万家零售门店

车载助理
每年 7500 万台新车

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

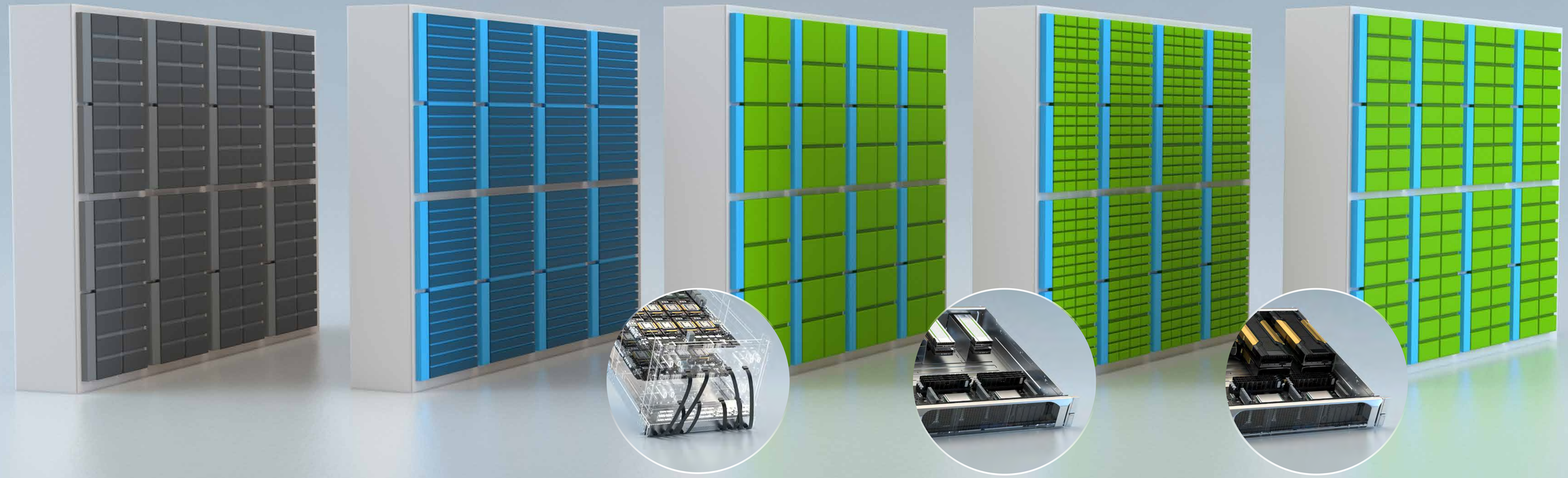
NVIDIA AI


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

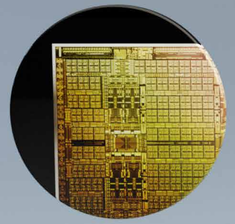

NVIDIA Merlin
NVIDIA Jarvis
APACHE Spark
TensorFlow
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton 推理服务器

现代云数据中心

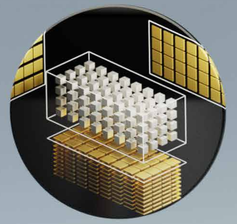
多元化应用 | 纵向及横向扩展工作负载 | 永无止境的需求



宣布推出 NVIDIA A100

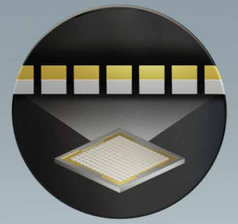


540 亿个 XTOR

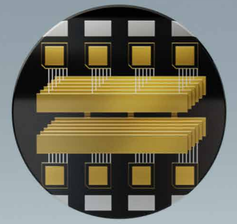


第 3 代 TENSOR CORE

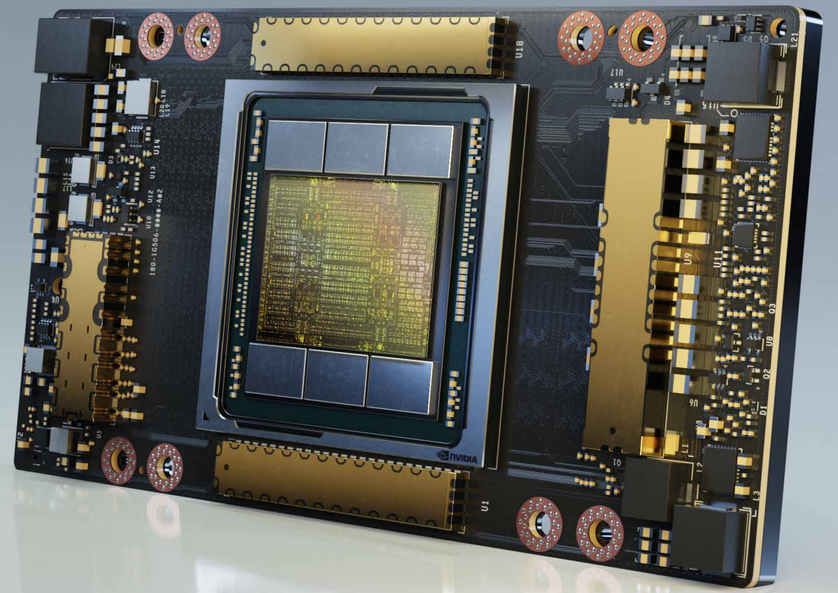
稀疏化加速



MIG

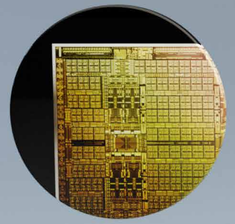


第 3 代 NVLINK 和 NVSWITCH

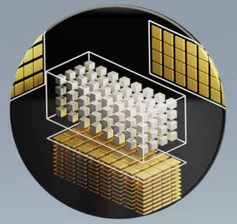


宣布推出 NVIDIA A100

TSMC 7nm | HBM2 – 1.6 TB/s | 3D 芯片封装

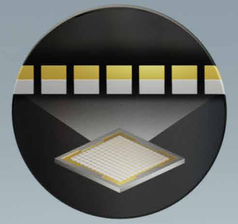


540 亿个 XTOR

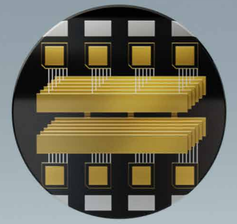


第 3 代 TENSOR CORE

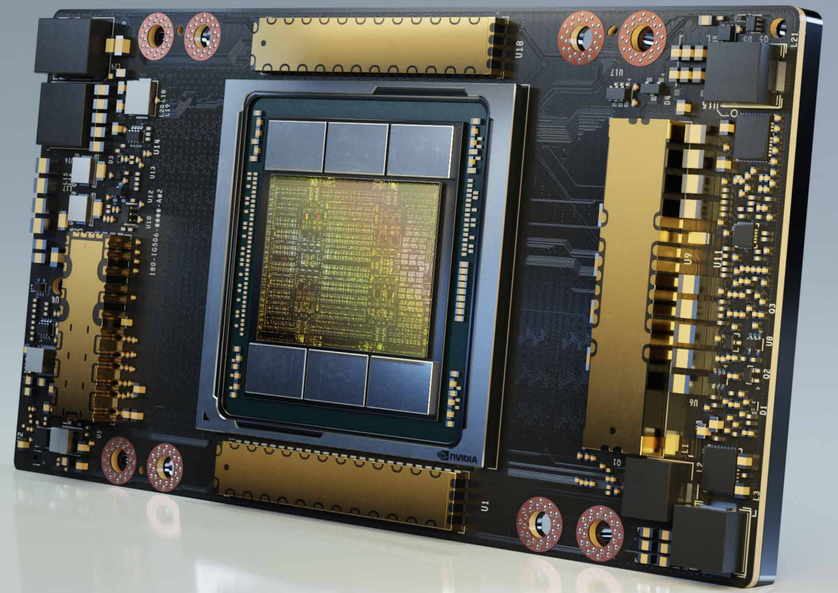
稀疏化加速



MIG

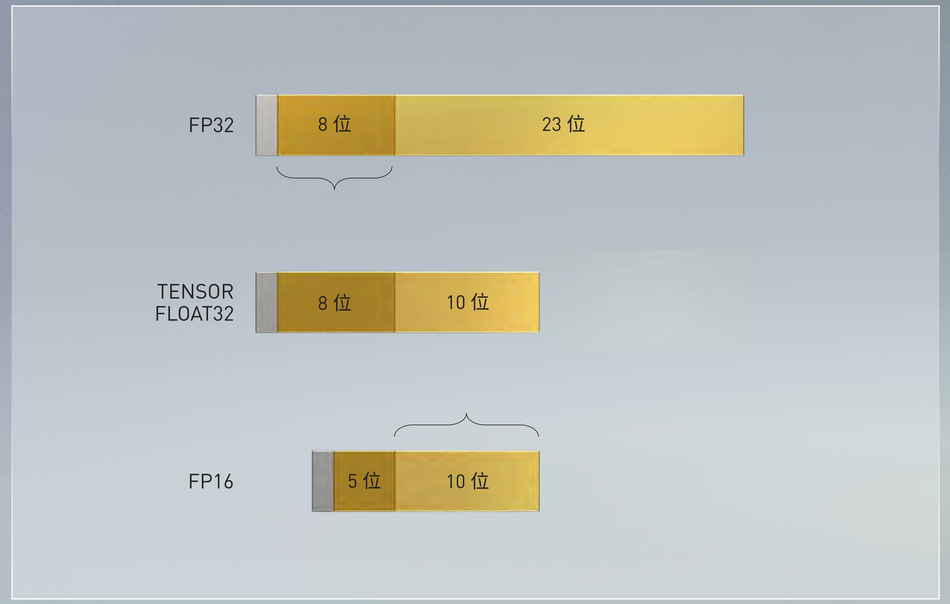


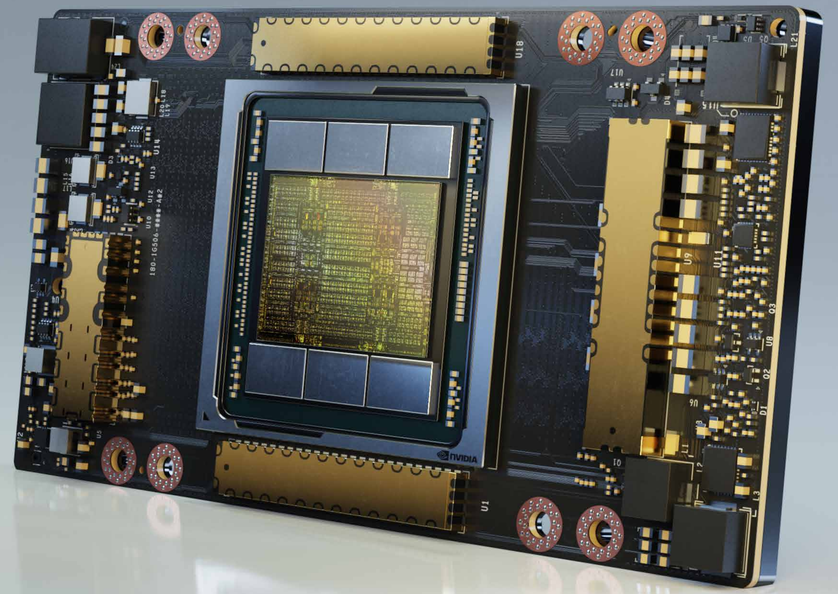
第 3 代 NVLINK 和 NVSWITCH



新的 TF32 TENSOR CORE

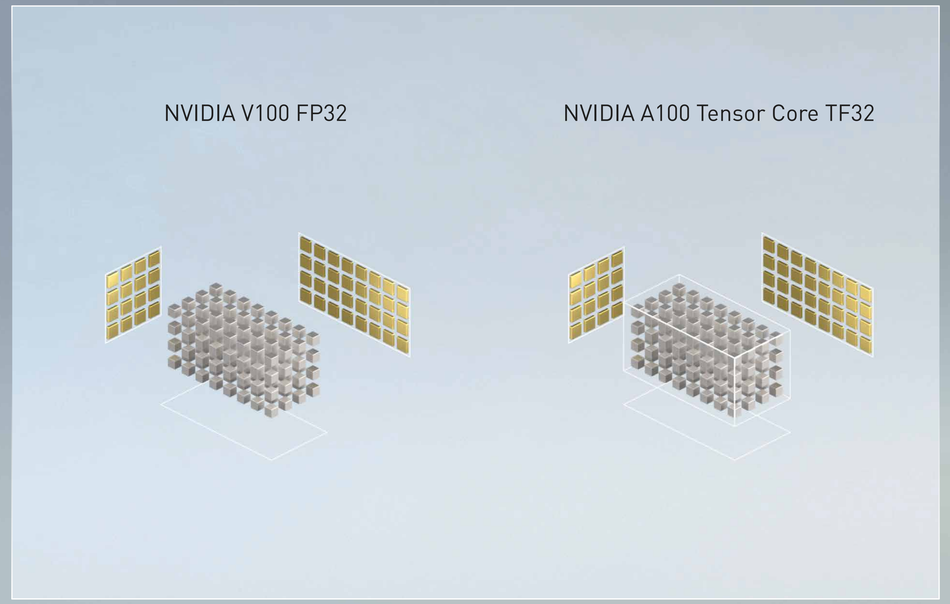
FP32 的指数范围和 FP16 的精度范围 | 以 FP32 输入，以 FP32 输出 | 无需更改代码即可加速训练

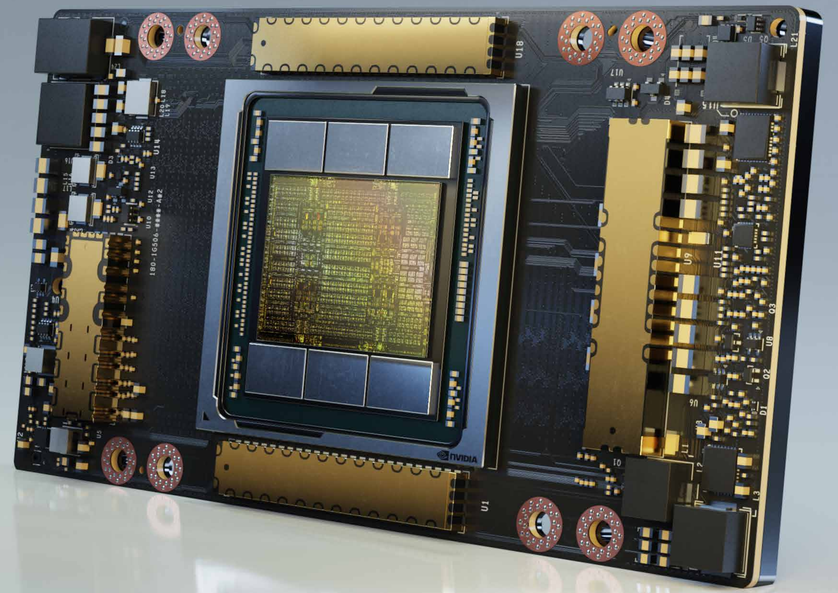




新的 TF32 TENSOR CORE

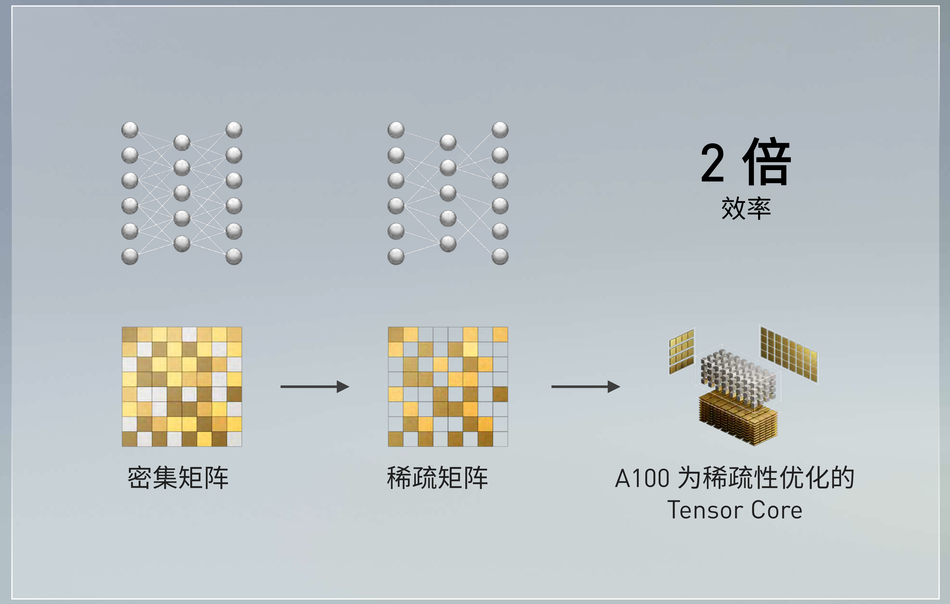
FP32 的指数范围和 FP16 的精度范围 | 以 FP32 输入，以 FP32 输出 | 无需更改代码即可加速训练

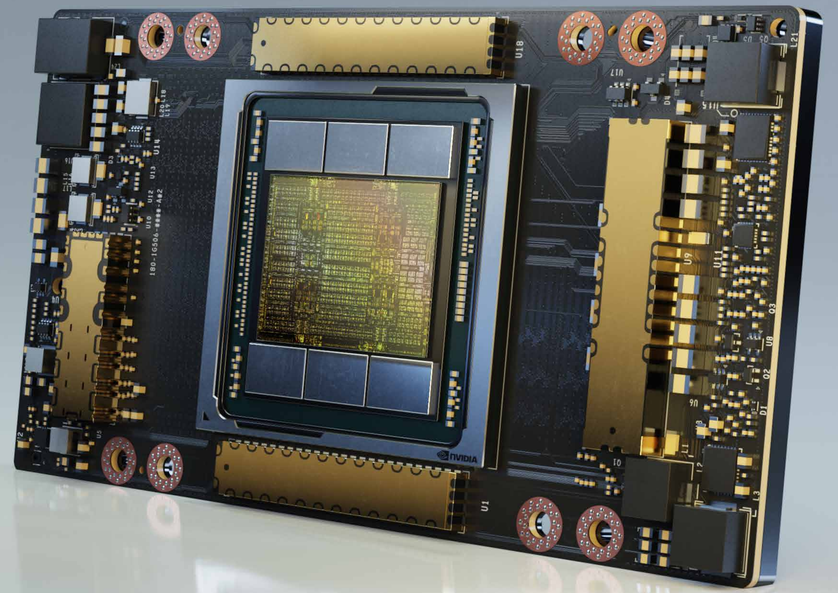




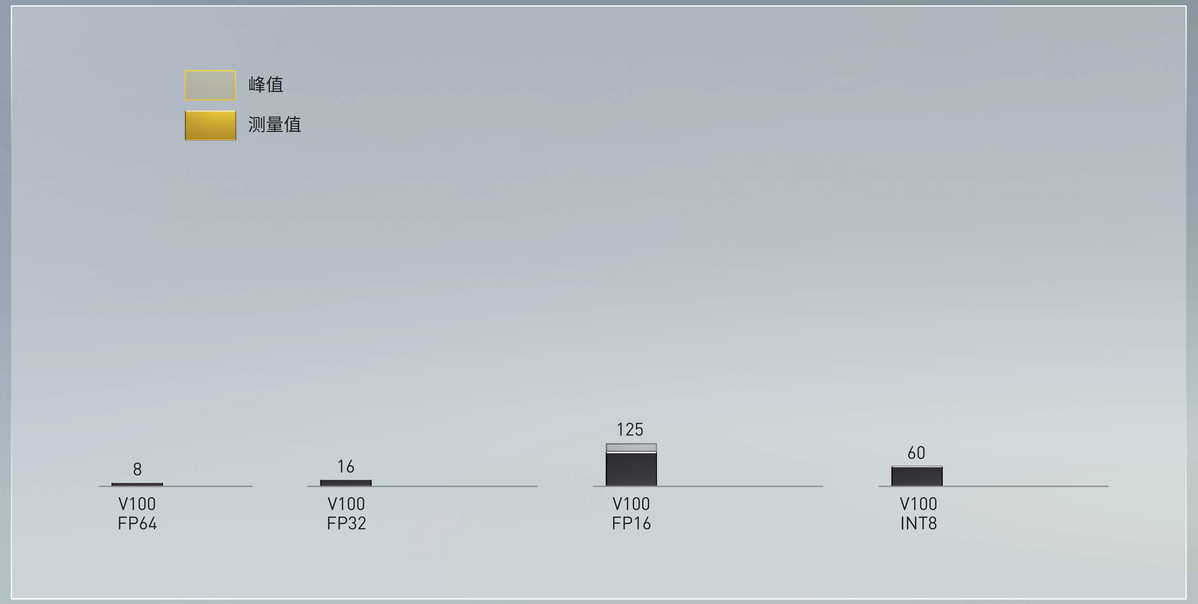
新的 TENSOR CORE 稀疏化加速

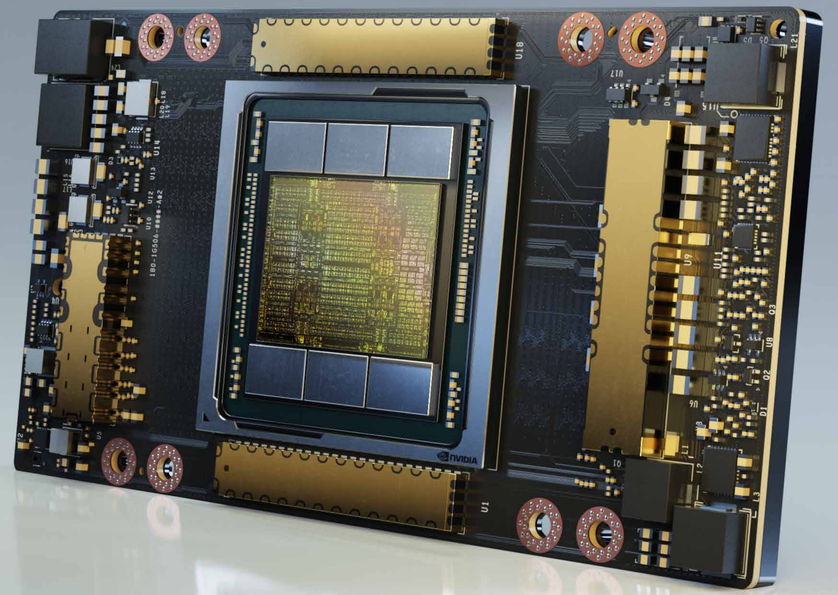
针对稀疏 AI 张量运算优化 | 执行速度提高 2 倍 | TF32、FP16、BFLOAT16、INT8 和 INT4 支持



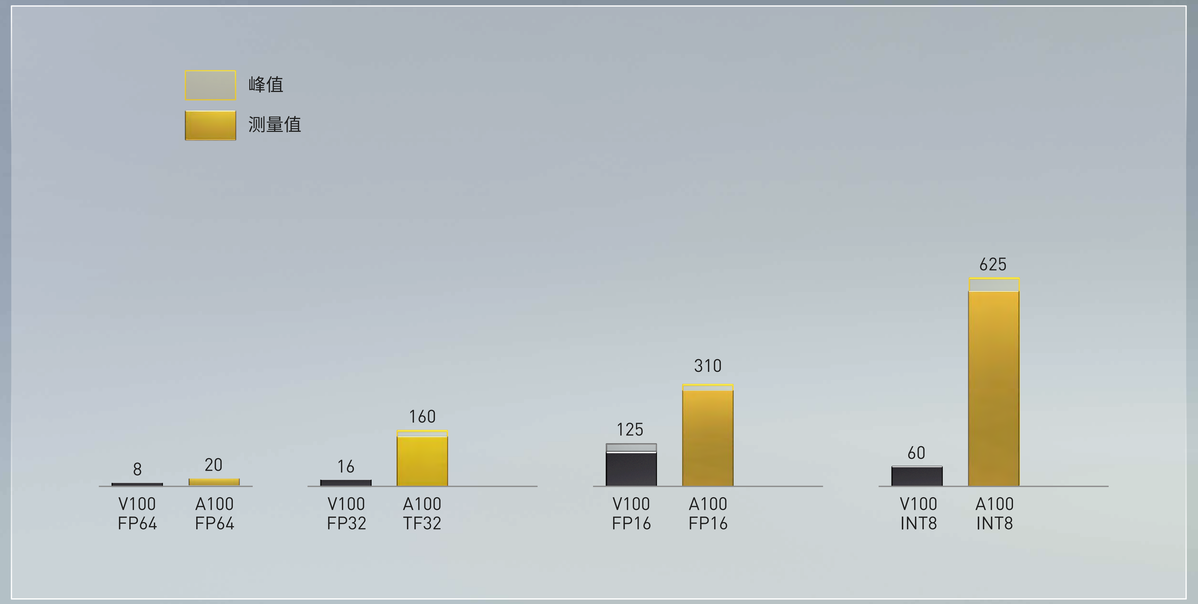


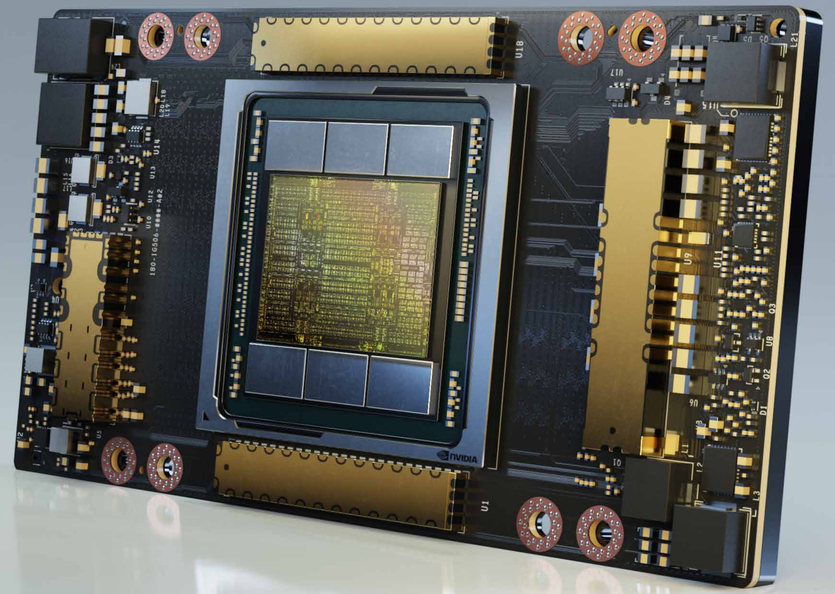
宣布推出 NVIDIA A100
巨大的代际飞跃



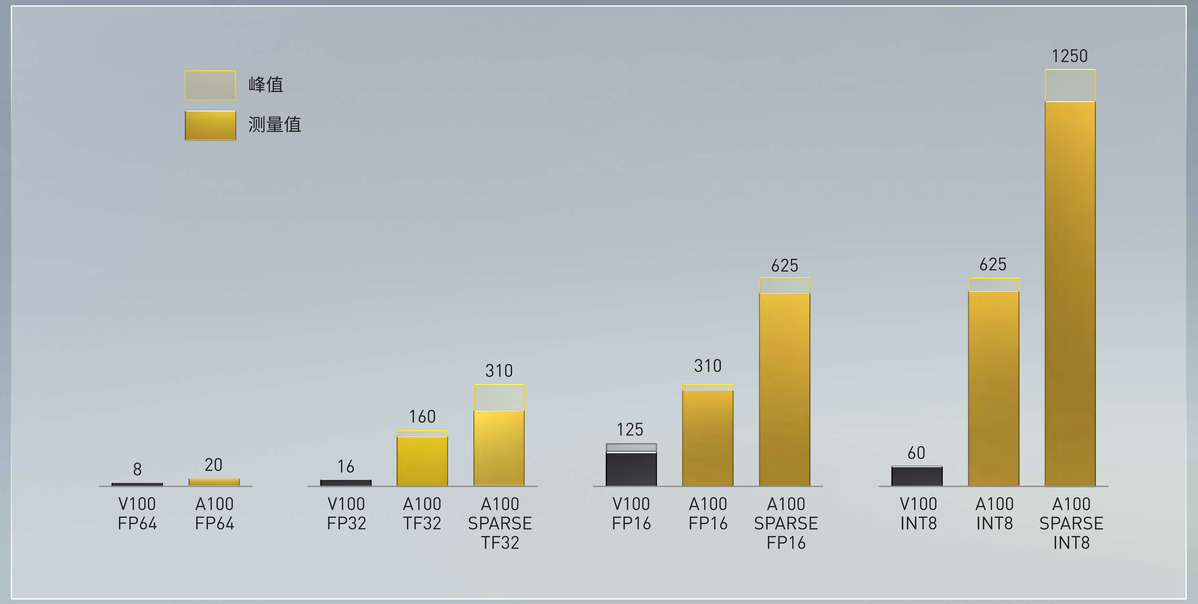


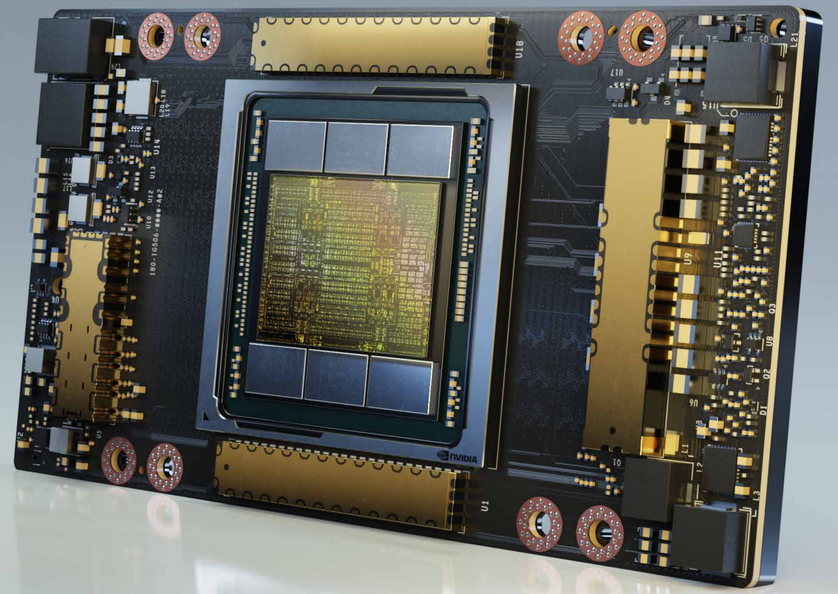
宣布推出 NVIDIA A100
巨大的代际飞跃



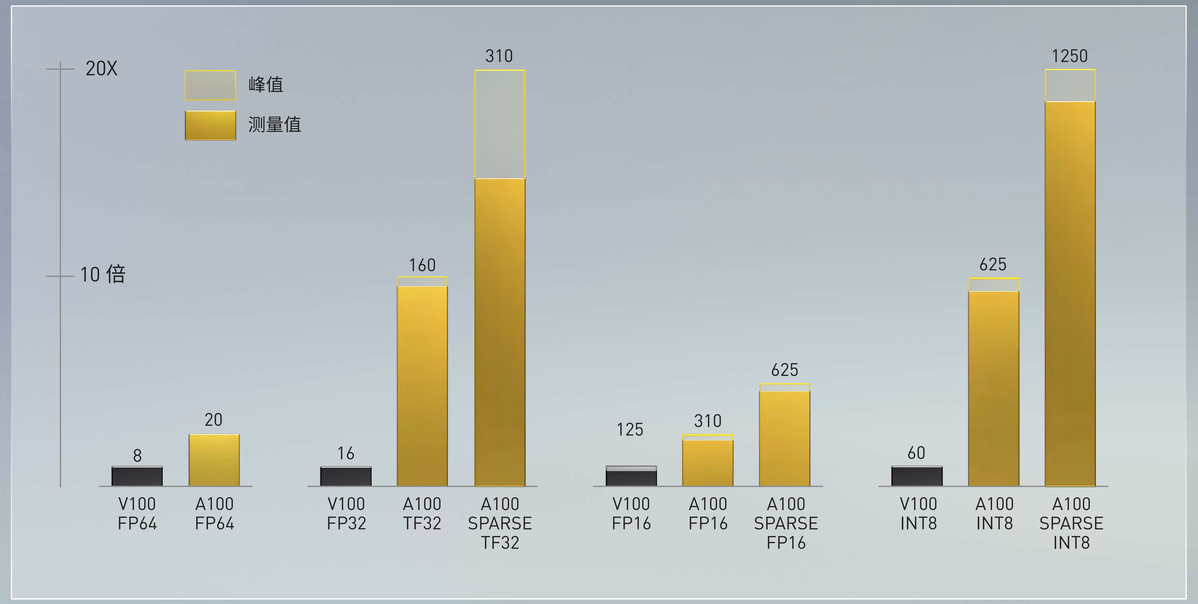


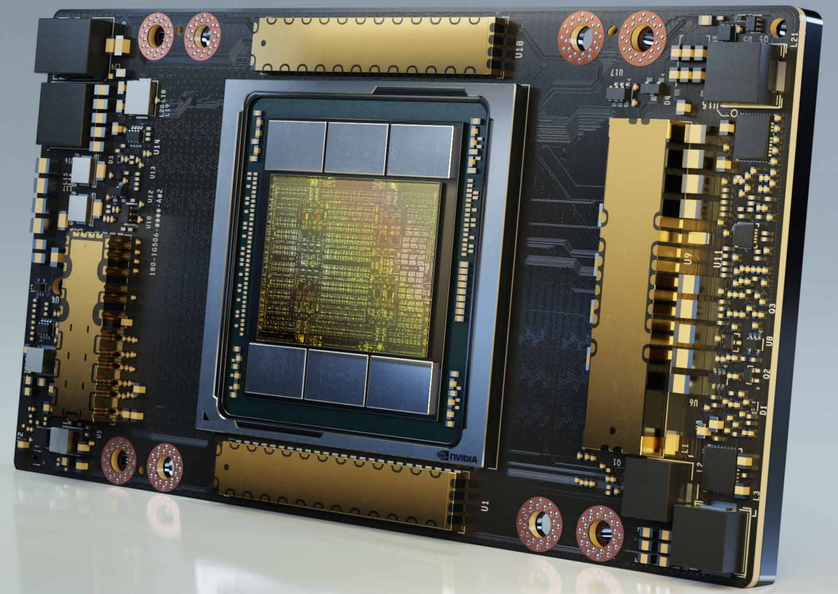
宣布推出 NVIDIA A100
巨大的代际飞跃





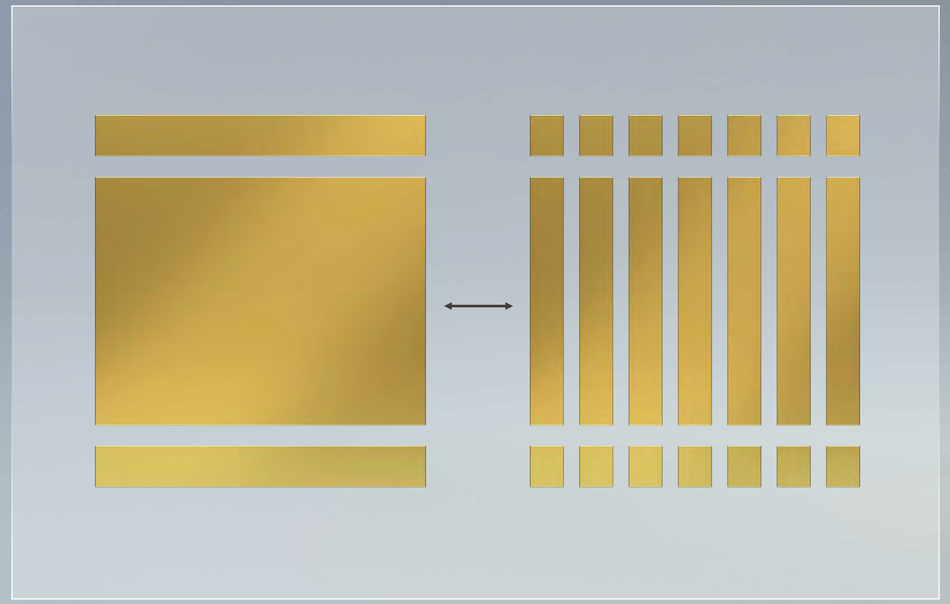
宣布推出 NVIDIA A100
巨大的代际飞跃

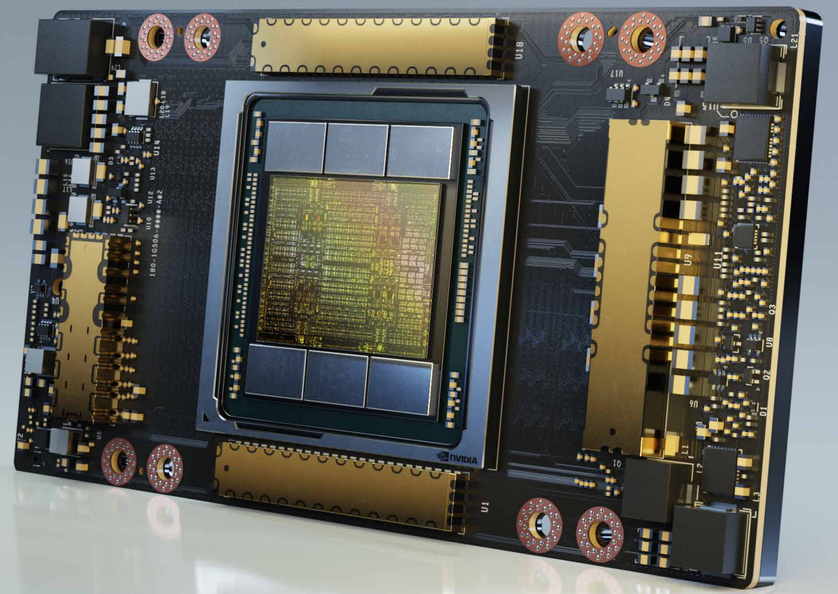




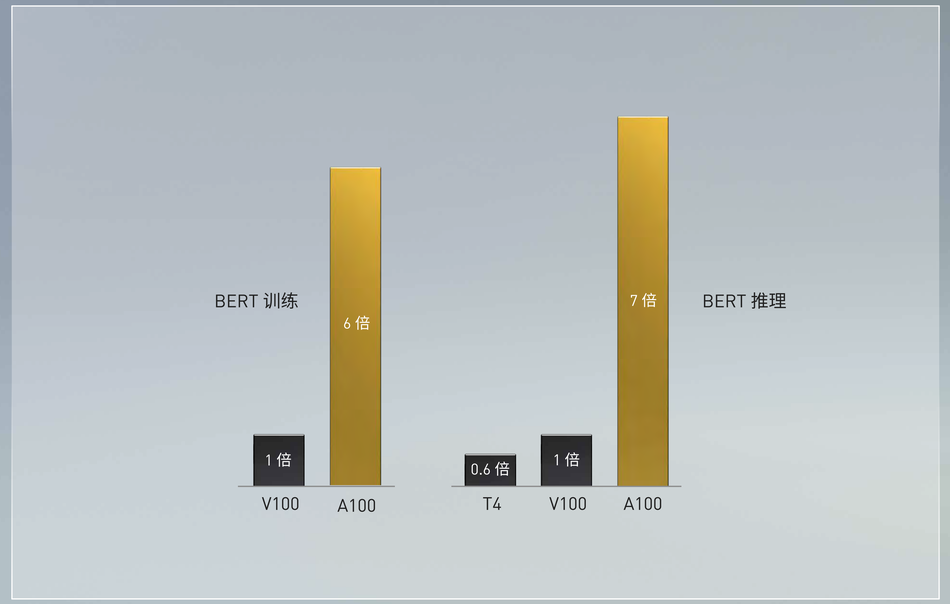
用于弹性 GPU 计算的新款多实例 GPU

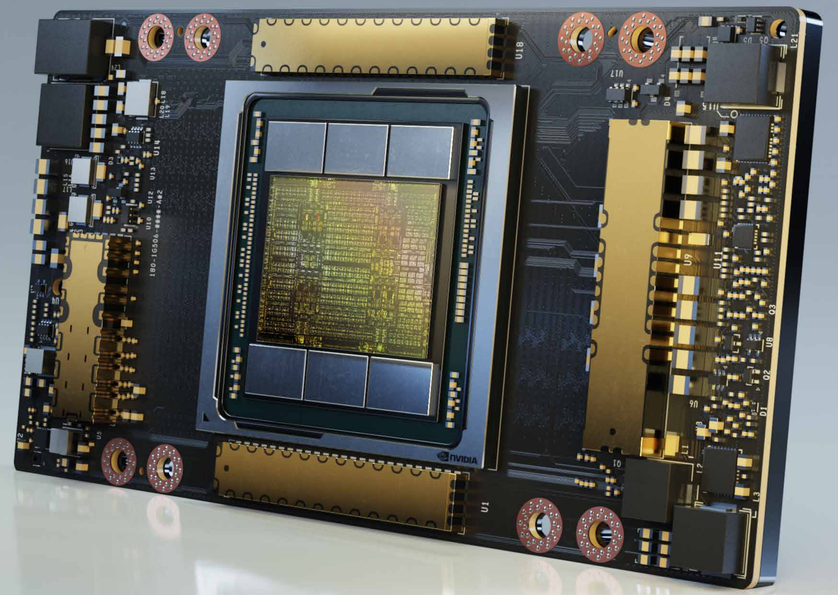
同时使用多个并发实例，吞吐量比 V100 高 7 倍



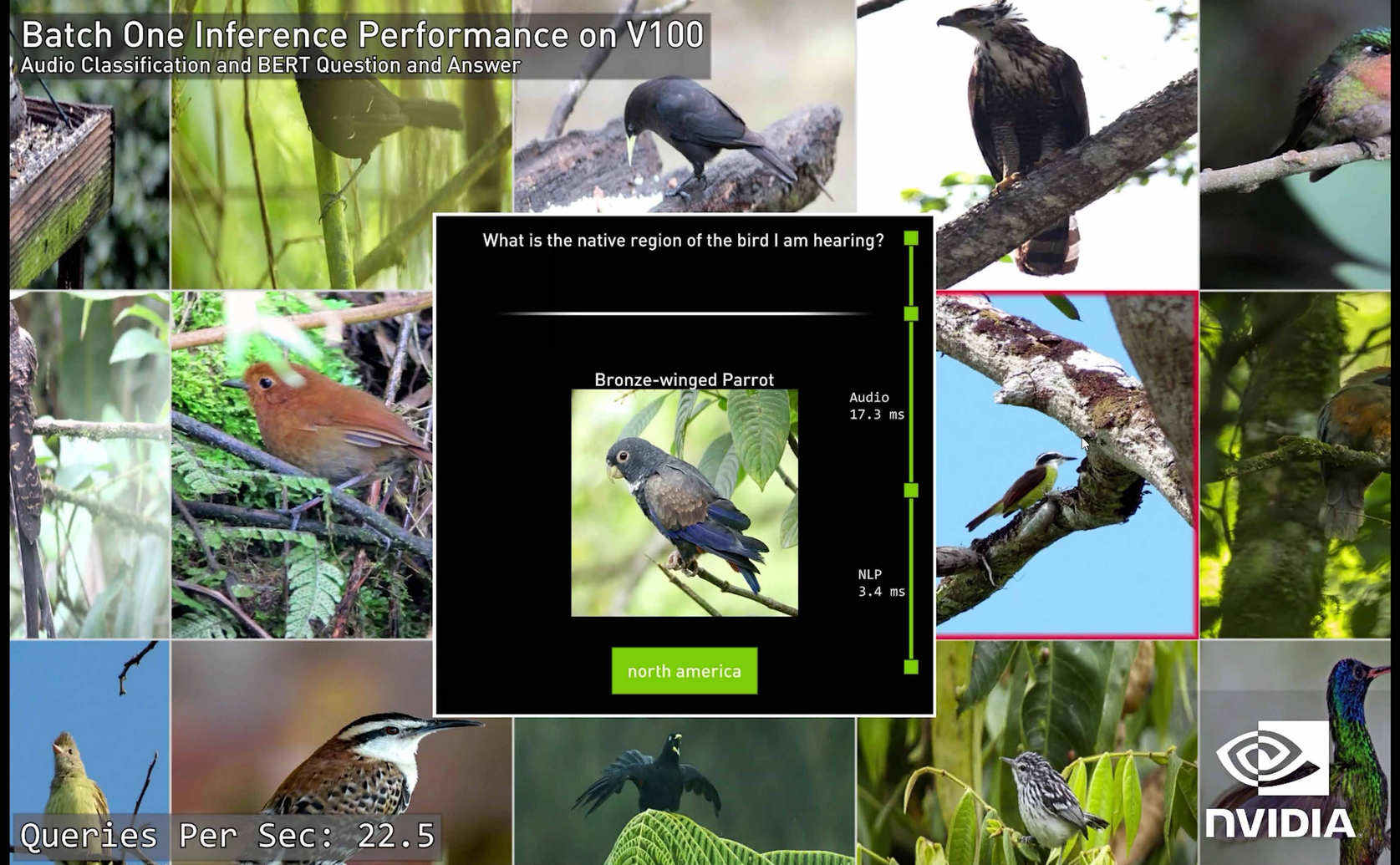


统一 AI 训练和推理加速

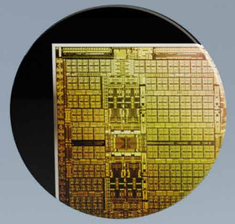




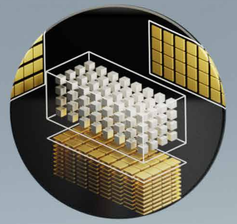
使用 Pytorch 的 BERT 预训练吞吐量，包括 (2/3) 阶段 1 和 (1/3) 阶段 2 | 阶段 1 序列长度 = 128，阶段 2 序列长度 = 512 V100：使用 8 个 V100 的 DGX-1 服务器（使用 FP32 精度）A100：使用 8 个 A100 的 DGX A100 服务器（使用 TF32 精度） | BERT-large 推理 | T4，V100：TRT 7.1，精度 = FP16，批尺寸 = 256 | A100 MIG：预先编译的 TRT，批尺寸 = 94，精度 = INT8，采用稀疏化优化


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

宣布推出 NVIDIA A100

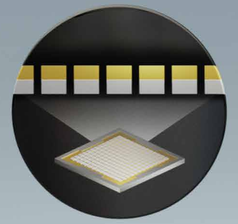


540 亿个 XTOR

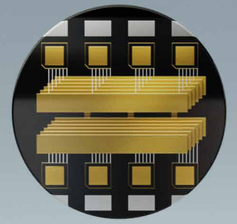


第 3 代 TENSOR CORE

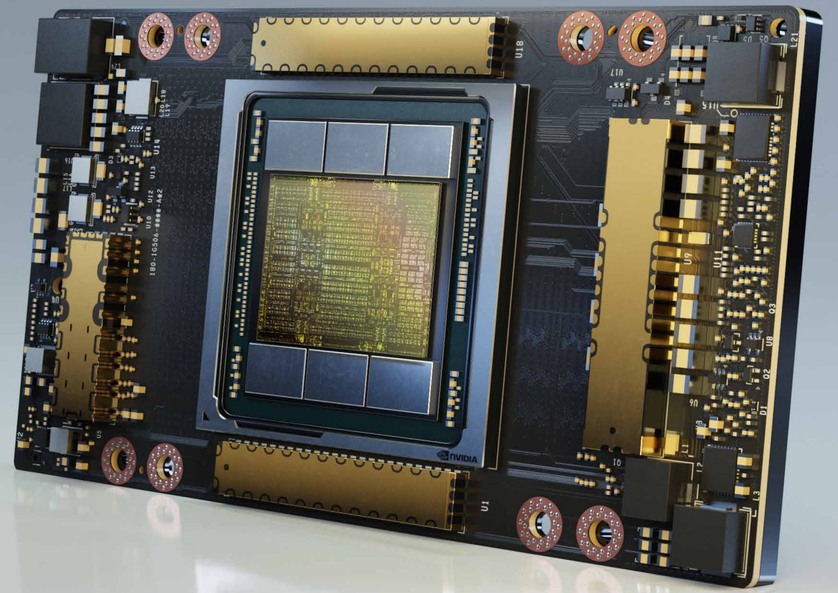
稀疏化加速



MIG



第 3 代 NVLINK 和 NVSWITCH



宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

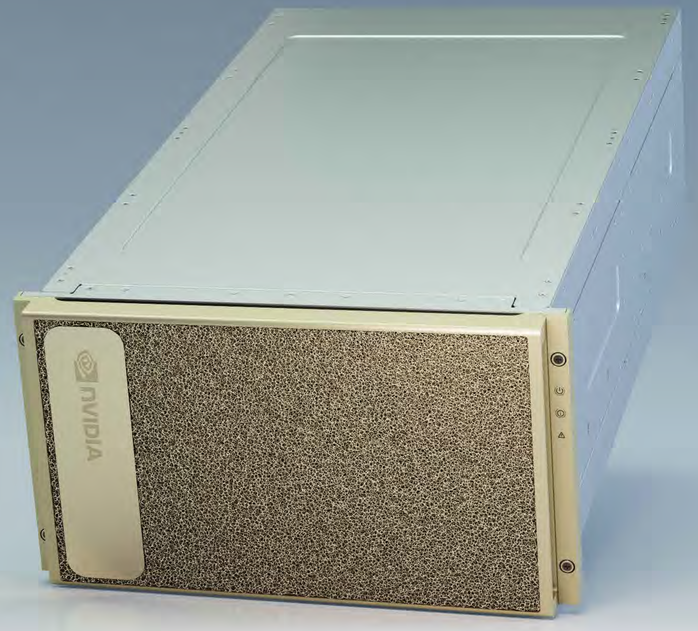
5 PetaFLOPS 的单节点性能

适用于端到端数据科学和 AI 的统一系统

完整加速技术栈 – Spark 3.0、RAPIDS、TensorFlow、PyTorch、Triton

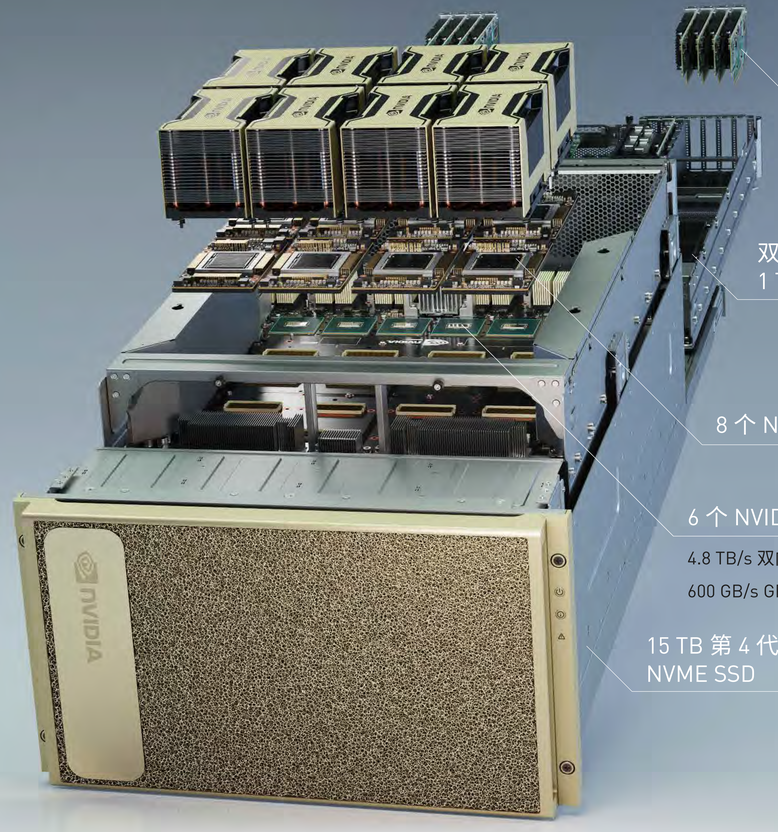
弹性纵向或横向扩展计算

使用 Mellanox 网络实现高可扩展性



宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

5 PetaFLOPS 的单节点性能



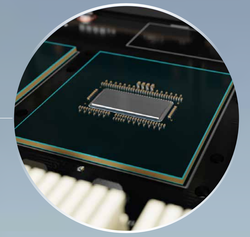
9 倍 Mellanox ConnectX-6 VPI
200 Gb/s 网络接口

双 64 核 AMD Rome CPU
1 TB RAM

8 个 NVIDIA A100 GPU

6 个 NVIDIA NVSwitch
4.8 TB/s 双向带宽
600 GB/s GPU 间带宽

15 TB 第 4 代
NVME SSD



宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

5 PetaFLOPS 的单节点性能

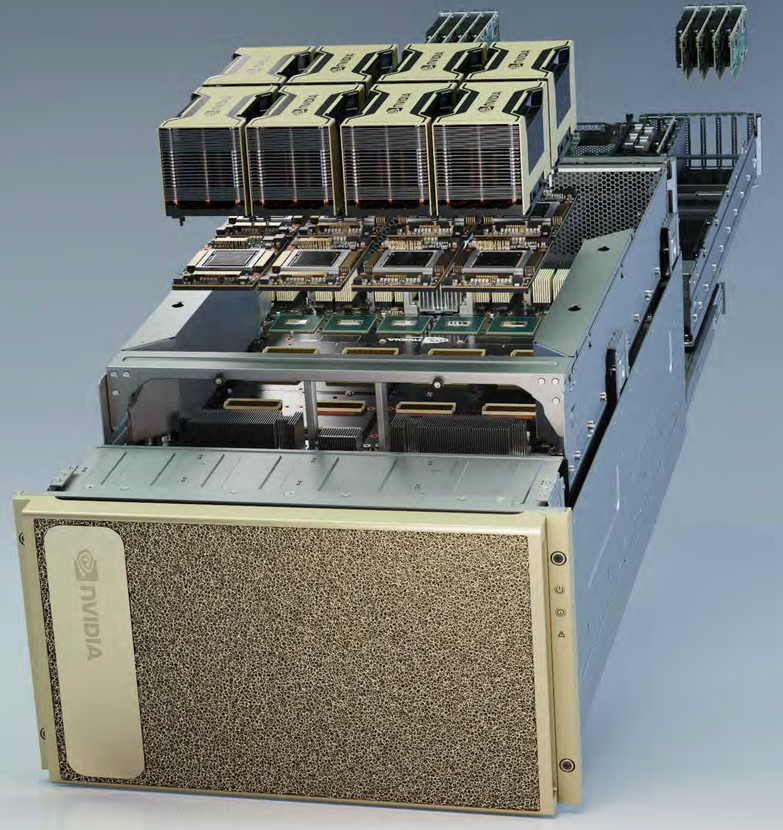
INT8 10 PetaOPS 峰值

FP16 5 PFLOPS 峰值

TF32 2.5 PFLOPS 峰值

FP64 156 TFLOPS 峰值

采用稀疏技术的 TensorCore



宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

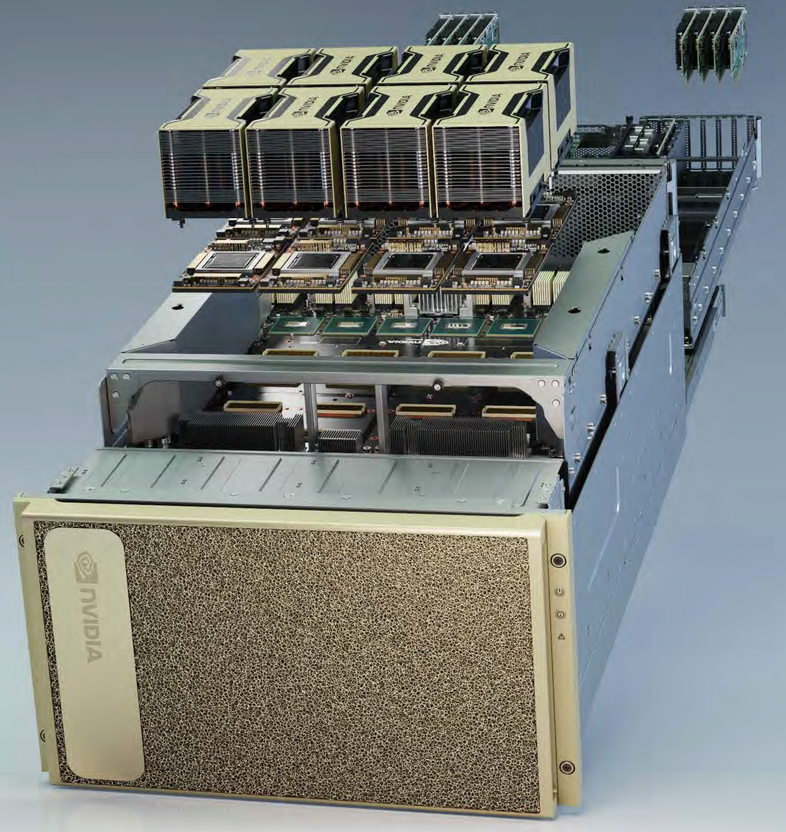
5 PetaFLOPS 的单节点性能

150X AI 计算

40X 显存带宽

40X IO 带宽

与高端 CPU 服务器相比较



宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

5 PetaFLOPS 的单节点性能

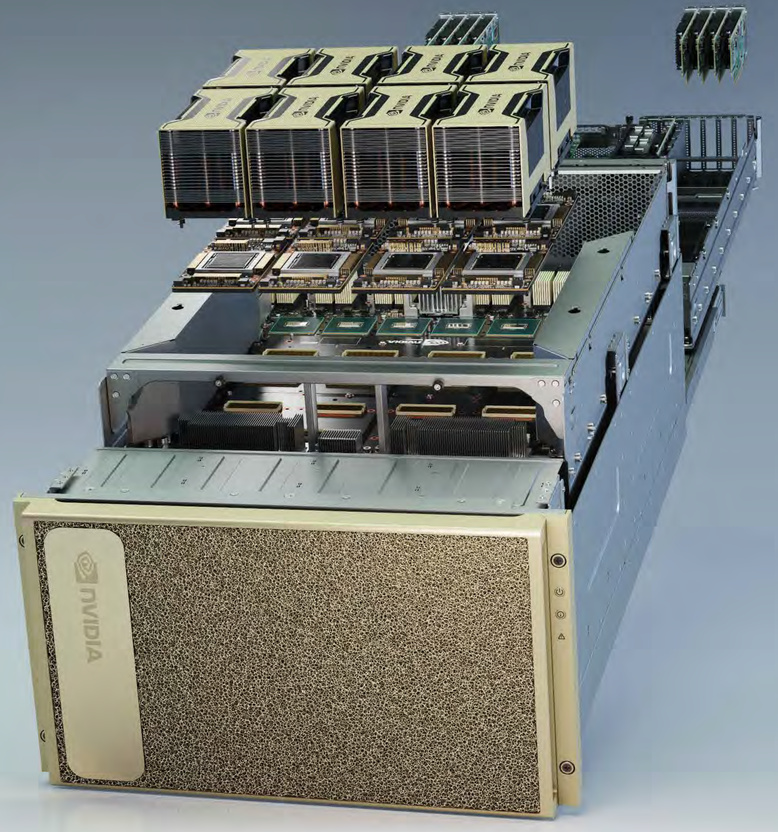
150X AI 计算

40X 显存带宽

40X IO 带宽

与高端 CPU 服务器相比较

目前售价为 19.9 万美元



介绍
NVIDIA A100 LIGHTHOUSE 客户
业界领军者优选的弹性数据中心加速器

云

Alibaba Cloud

aws

BAIDU AI CLOUD

Google Cloud

Microsoft Azure

ORACLE Cloud Infrastructure

Tencent Cloud

系统

Atos

CRAY a Hewlett Packard Enterprise company

DELL Technologies

FUJITSU

GIGABYTE

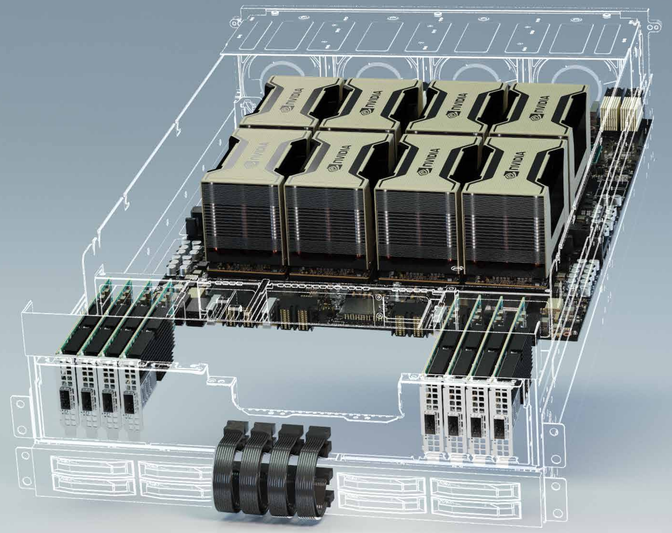
Hewlett Packard Enterprise

inspur

Lenovo

QCT

SUPERMICRO



当今的 AI 数据中心

50 个 DGX-1 服务器用于 AI 训练

600 个 CPU 服务器用于 AI 推理

1100 万美元

25 个机架

630 kW

DGX A100 AI

5 个 DGX A100 服务器用于 AI 训练和推理

100 万美元

1 个机架

28 kW

当今的 AI 数据中心

50 个 DGX-1 服务器用于 AI 训练

600 个 CPU 服务器用于 AI 推理

1100 万美元

25 个机架

630 kW

DGX A100 AI

5 个 DGX A100 服务器用于 AI 训练和推理

100 万美元

1 个机架

28 kW

PAGERANK 案例研究

通用爬网数据集
2.6 TB 图 – 128B 边缘

3000 个 CPU 服务器 – 105 个机架

PAGERANK 案例研究

通用爬网数据集
2.6 TB 图 – 128B 边缘

4 个 DGX A100 通过外部 NVLINK 连接

PAGERANK 案例研究

通用爬网数据集
2.6 TB 图 – 128B 边缘

3000 个 CPU 服务器 – 105 个机架

PAGERANK 案例研究

通用爬网数据集
2.6 TB 图 – 128B 边缘

4 个 DGX A100 通过外部 NVLINK 连接

宣布推出
NVIDIA DGX A100 SUPERPOD

140 个 DGX A100 服务器（1120 个 A100）
170 个 Mellanox Quantum 200G InfiniBand 交换机
280 Tb/s 的网络交换矩阵 – 15 公里长的光纤
4 PB 的全闪存网络存储
700 PFLOPS 的 AI 性能
不到 3 周时间完成构建

NVIDIA 扩展 SATURNV

扩展之前
1,800 个 DGX 服务器
1.8 ExaFLOPS

添加了 4 个 DGX SuperPOD
560 个 DGX A100 服务器 = 2.8 ExaFLOPS

4.6 ExaFLOPS 总算力

宣布推出
NVIDIA DGX A100
第 3 代集成 AI 系统

5 PetaFLOPS 的单节点性能

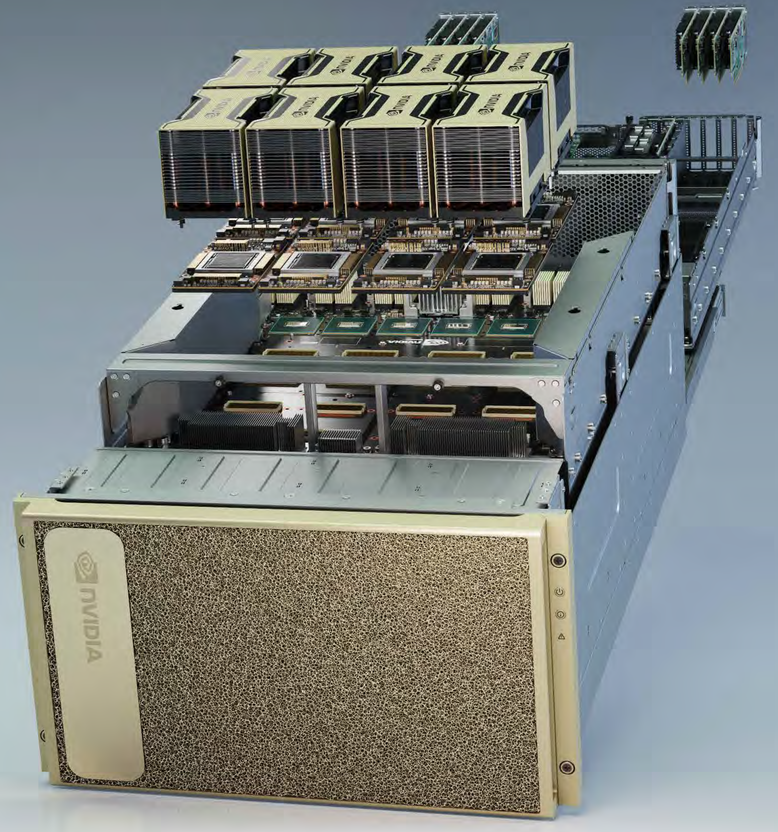
150X AI 计算

40X 显存带宽

40X IO 带宽

与高端 CPU 服务器相比较

目前售价为 19.9 万美元



万物智能革命

始终开启 | 即时感知、推理、行动 | 远程 | 万亿级

宣布推出
采用 MELLANOX CX6 DX 的 NVIDIA EGX A100

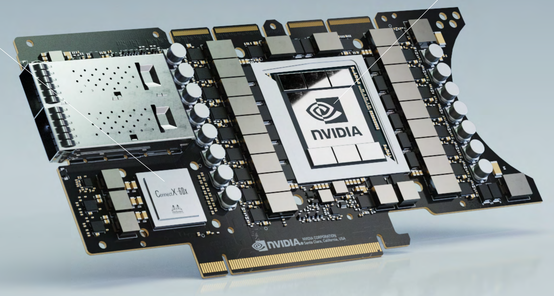
NVIDIA Mellanox ConnectX-6 DX

双 100 Gb/s 以太网或 InfiniBand

线速 TLS/IPSec 加密引擎

时间触发型电信级传输技术（5G 下为 5T）

ASAP² SR-IOV 和 VirtualIO 分载



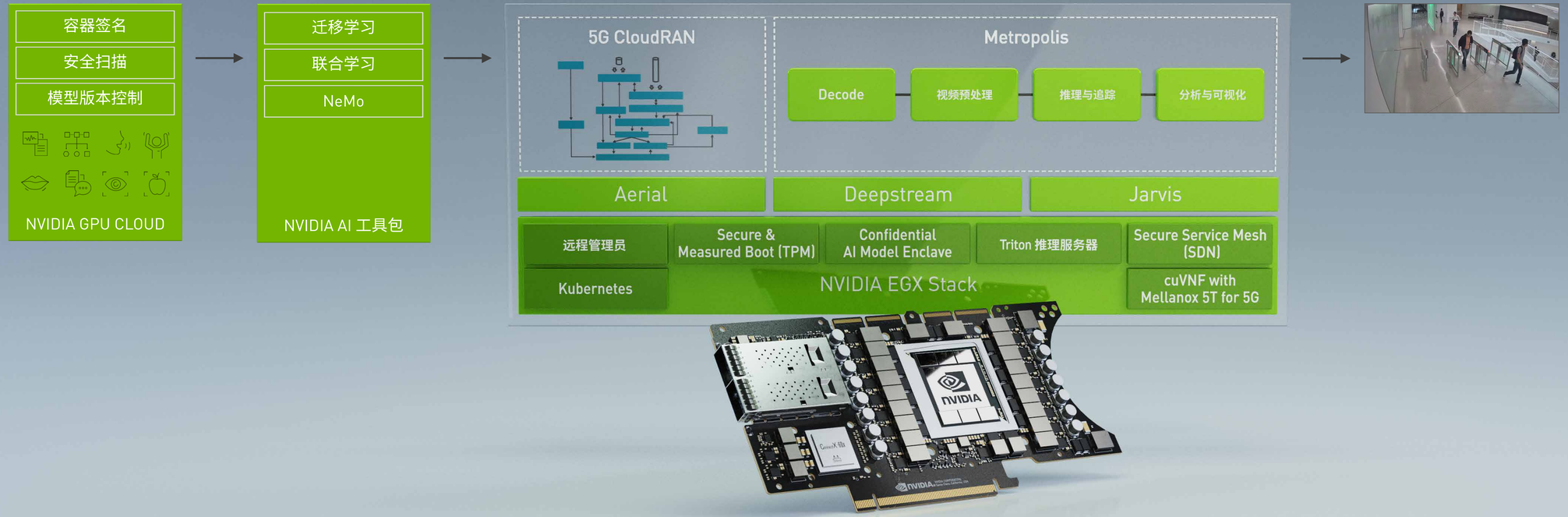
NVIDIA Ampere GPU

第 3 代 Tensor Core

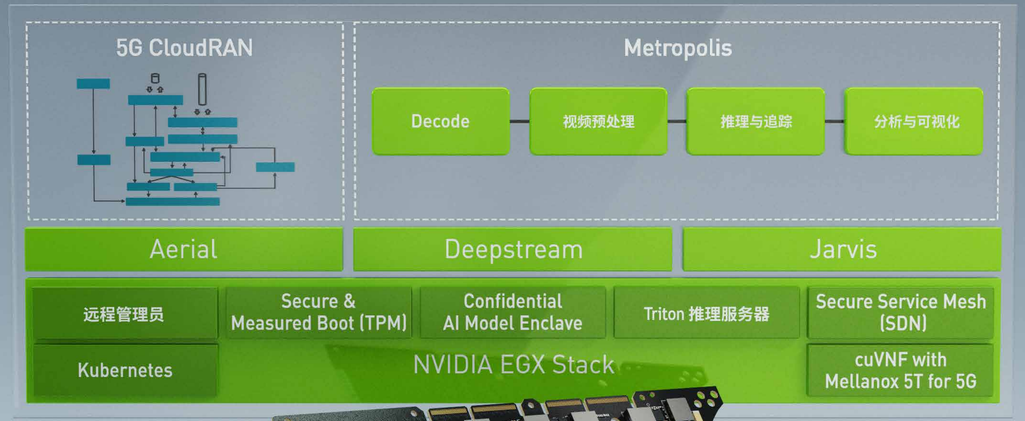
用于机密 AI 的新款安全引擎

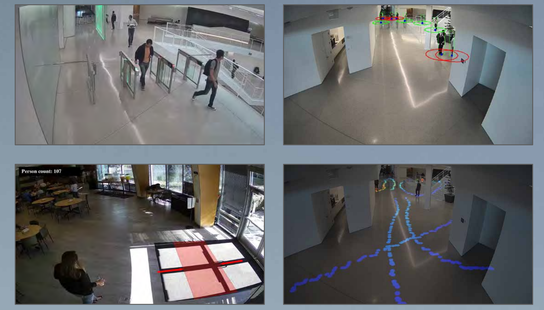
安全、经身份验证的引导

宣布推出
NVIDIA EGX 边缘 AI 平台



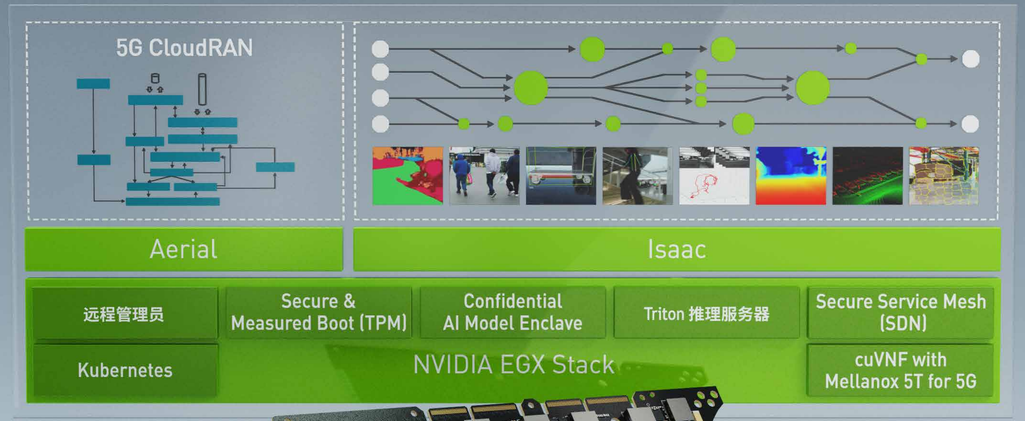
基于 NVIDIA EGX 的 METROPOLIS 视频 AI 和 AERIAL 5G





基于 NVIDIA EGX 的 METROPOLIS 机器人工厂和 AERIAL 5G

实体工厂



虚拟工厂
数字孪生


KEEP CLEAR

宣布
宝马选择 NVIDIA ISAAC 机器人

“选择的力量”

超过 40 款宝马车型，每款车 100 种选件

99% 的订单是定制/独特的

可能有 2^{100}种不同的配置

“原始部件进厂，部件托盘出厂”

每天收到 3000 万个部件

1800 家供应商，送货到 31 家工厂

23 万个部件编号

“准时化，顺序化”

每条生产线多达 10 款汽车

每 56 秒生产出一辆新车

分割机器人

采集机器人

放置机器人

智能运输机器人
(STR)

排序机器人

NVIDIA EGX 生态系统

5G 和 CloudRAN

ALTRAN · ERICSSON · MAVENIR · Viasat

安全和网络

CUMULUS · FORTINET · JUNIPER · nuagenetworks

基础设施

IBM · Red Hat · vmware · WIND

系统

Atos · DELL Technologies · Hewlett Packard Enterprise · Lenovo

云

Baidu 百度 · Microsoft

对话式 AI

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

智能视频分析

AIFI · DEEPVISION · MALONG · IRONYUN · SAFR

机器人

BMW · DENSO · FANUC · FOXCONN · KOMATSU

医疗

CHOOCH · DawnLight · GE Healthcare · UCSF · WHITEBOARD COORDINATION

工业

Micron · MUSASHI · SAMSUNG · SEAGATE · tsmc

行业领袖

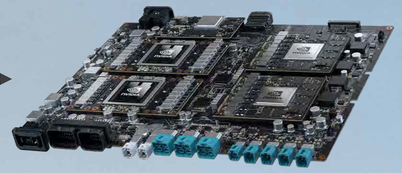
goli · P&G · SoftBank · UNITED STATES POSTAL SERVICE · Walmart

使用 ORIN 和 AMPERE 的 NVIDIA DRIVE
5W 到 2,000 TOPS – 同一种可编程架构

ADAS
挡风玻璃 NCAP
10 TOPS，5W

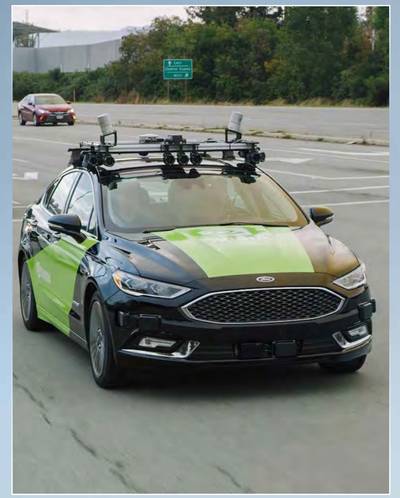
L2+ 级
自动驾驶
200 TOPS，45W

L5 级
Robotaxi
2,000 TOPS，800W

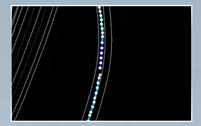


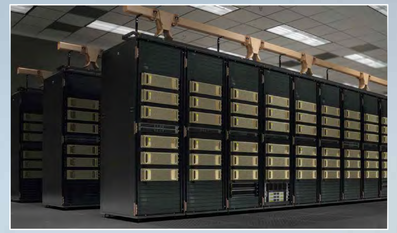
NVIDIA DRIVE – 软件定义的 AV 平台

收集数据



训练模型





模拟

DRIVE AV

DRIVE IX

DRIVE RC

NVIDIA DRIVE 全球生态系统

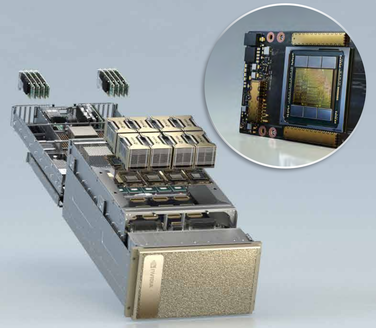
汽车	卡车	供应商	出行服务	初创公司	软件	地图	模拟
							
							
							
							

GTC 2020 公告

Mellanox TECHNOLOGIES

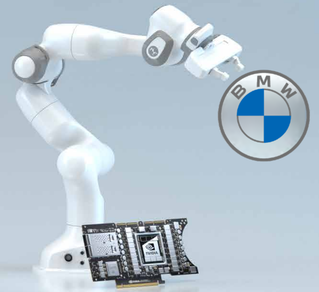
数据中心级计算

Omniverse RTX 服务器



A100 和 DGX A100

NVIDIA AI



EGX 和 ISAAC


NVIDIA