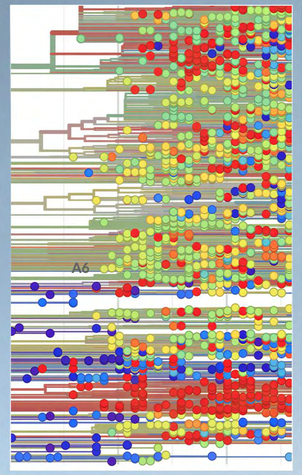
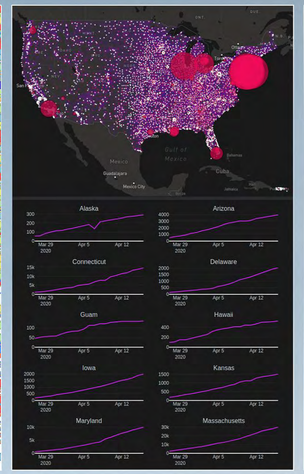

NVIDIA


THANK YOU
HEROES

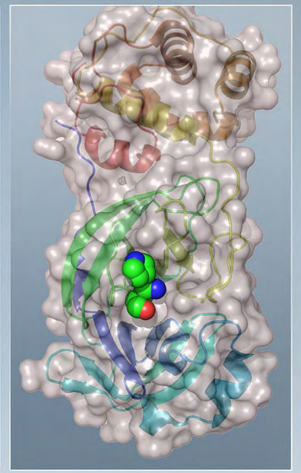
NVIDIA FIGHTS COVID-19



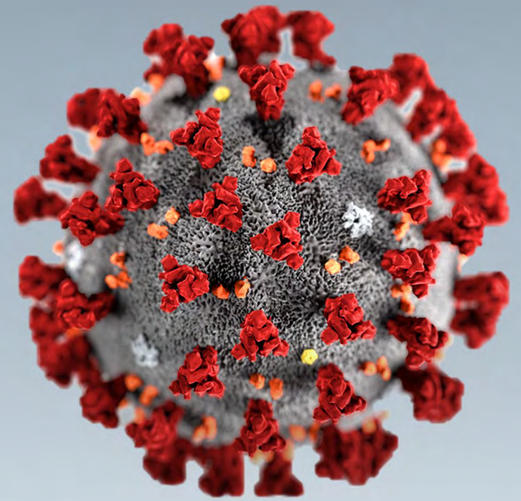
Oxford Nanopore
Sequence Virus Genome
in 7Hrs



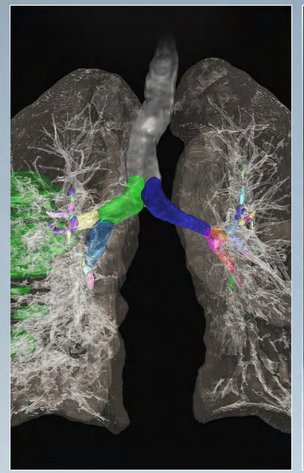
Plotly, NVIDIA
Real-Time
Infection Rate Analysis



ORNL, Scripps
Screen
1B Drug Compounds in
1 Day vs 1 Year

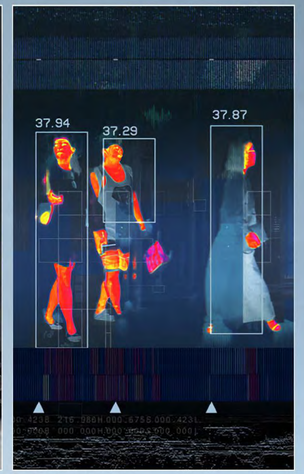


Structura, NIH, UT Austin
CryoSPARC
1st 3D Structure of Virus Spike Protein



NIH, NVIDIA
AI COVID-19
Classification

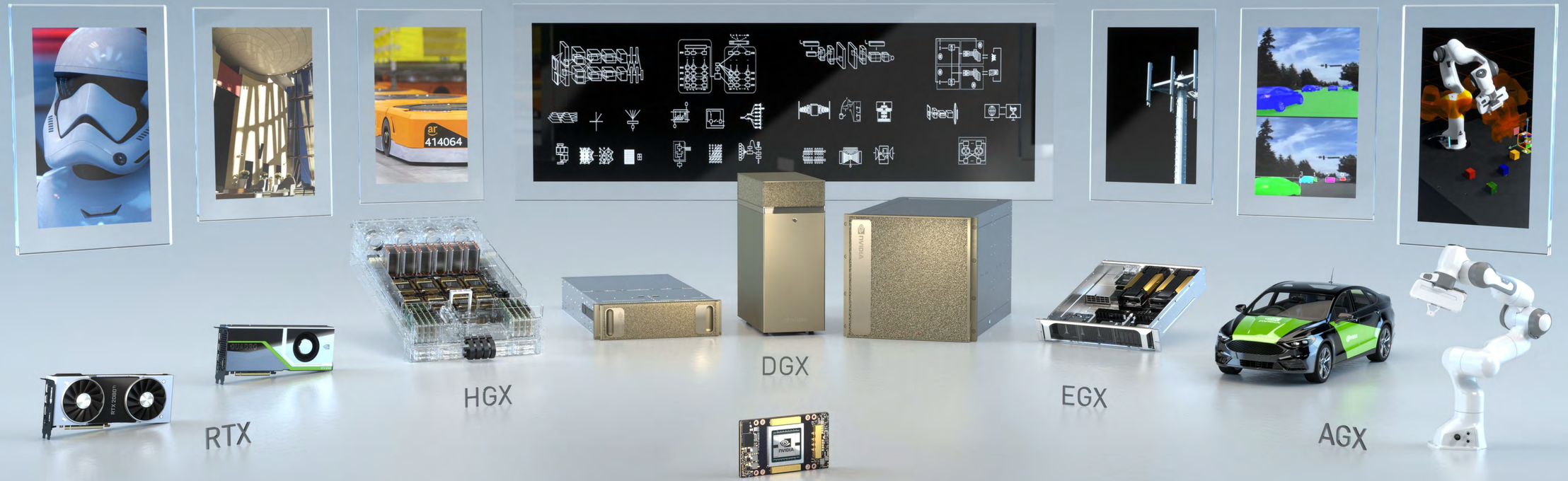
Kiwibot
Robot Medical Supply
Delivery



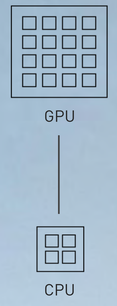
Whiteboard Coordinator
AI Elevated Body Temp
Screening System

Containment — Mitigation — Treatment — Tracking & Monitoring

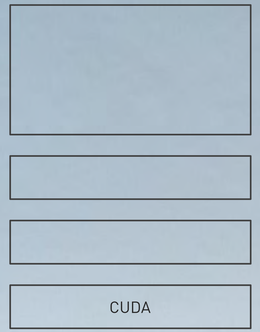
COMPUTING FOR THE DA VINCIS OF OUR TIME



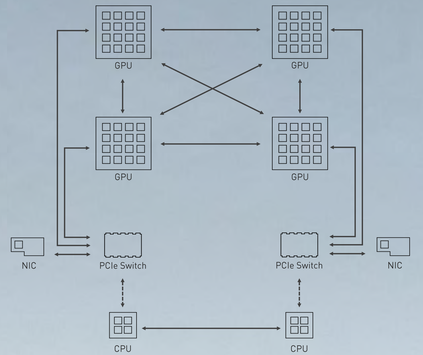
25 YEARS OF ACCELERATED COMPUTING



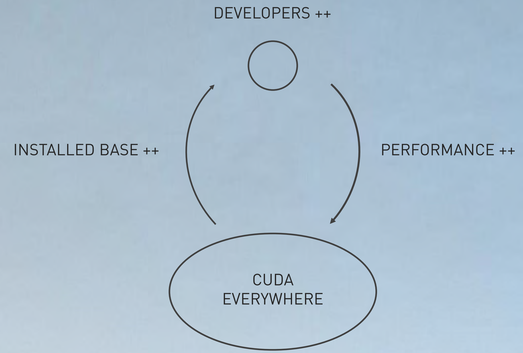
X-FACTOR SPEED-UP



FULL STACK

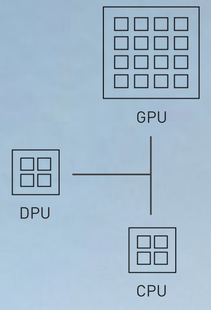


SYSTEMS

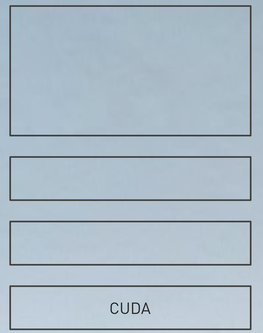


ONE ARCHITECTURE

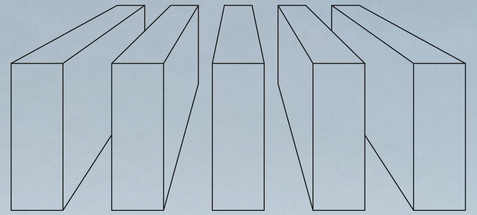
DATA-CENTER-SCALE ACCELERATED COMPUTING



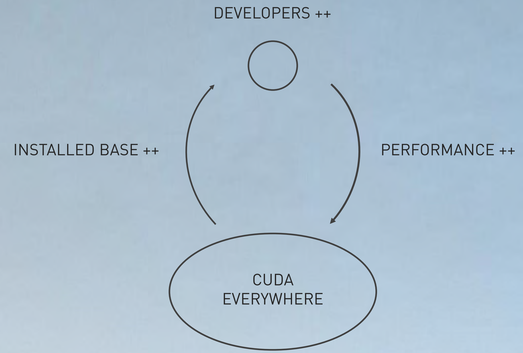
X-FACTOR SPEED-UP



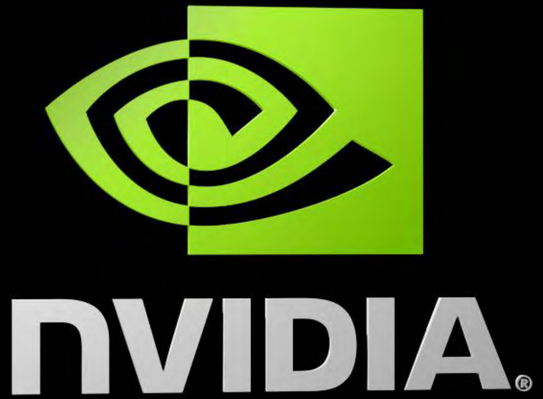
FULL STACK



DATA CENTER SCALE

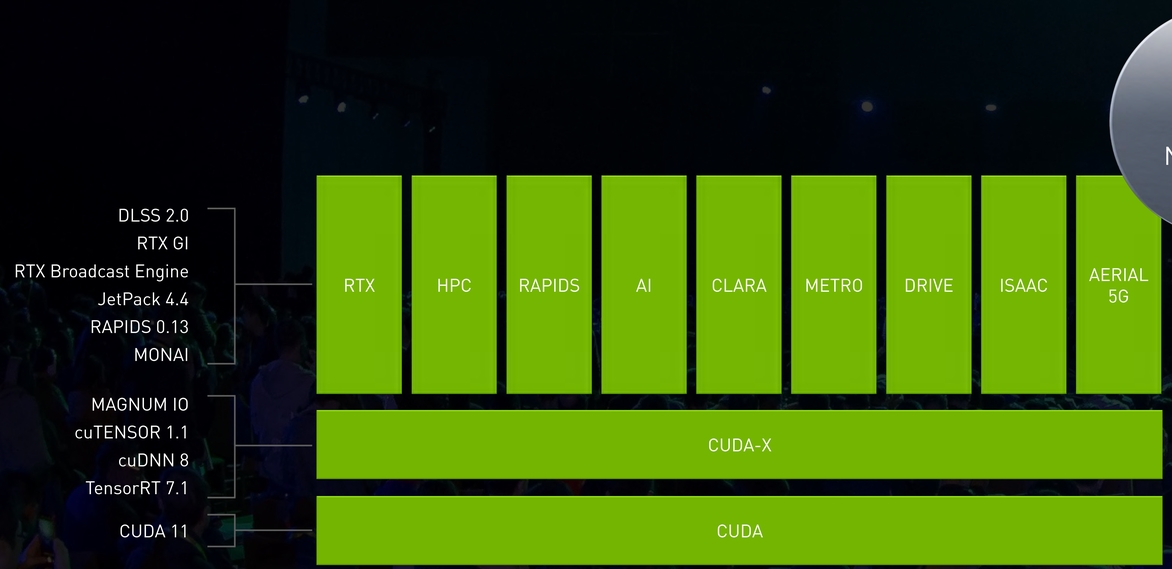


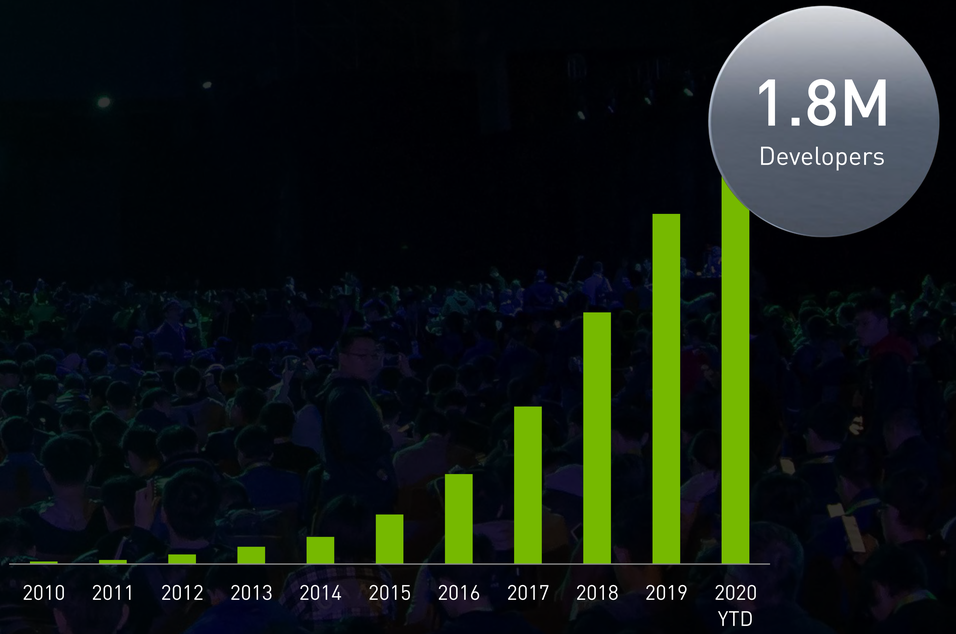
ONE ARCHITECTURE


NVIDIA


Mellanox
TECHNOLOGIES

FANTASTIC YEAR FOR OUR DEVELOPERS

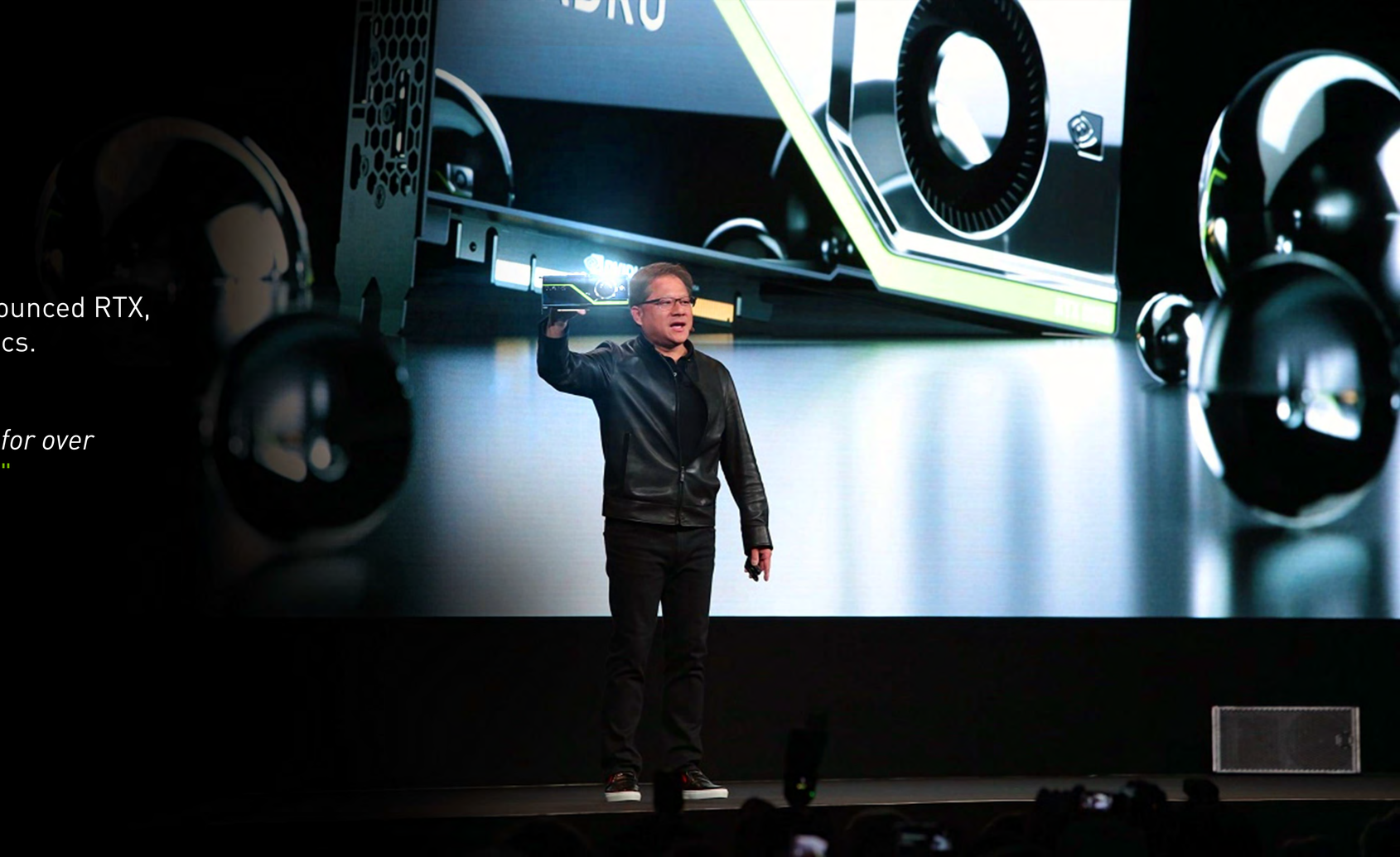




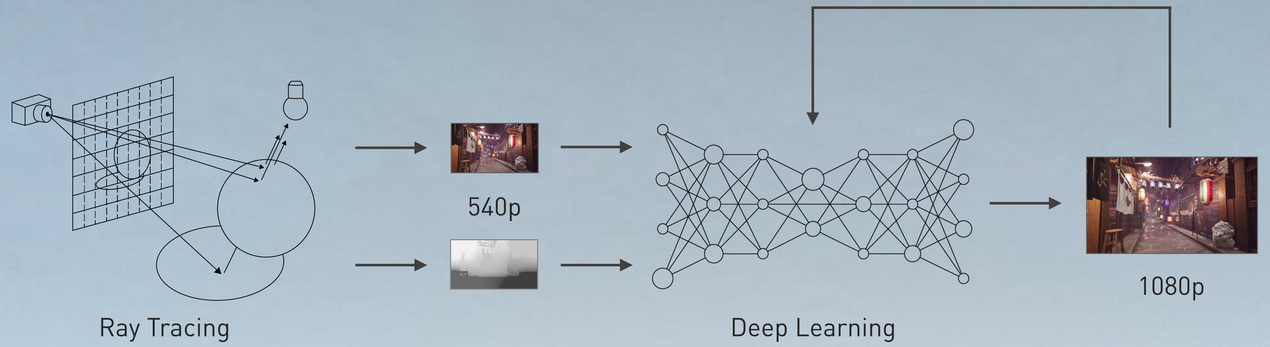
On August 13, 2018 at SIGGRAPH, NVIDIA announced RTX, the beginning of a new era in computer graphics.

"Real-time ray tracing, which had been promised for over 30 years, arrived a decade earlier than expected."

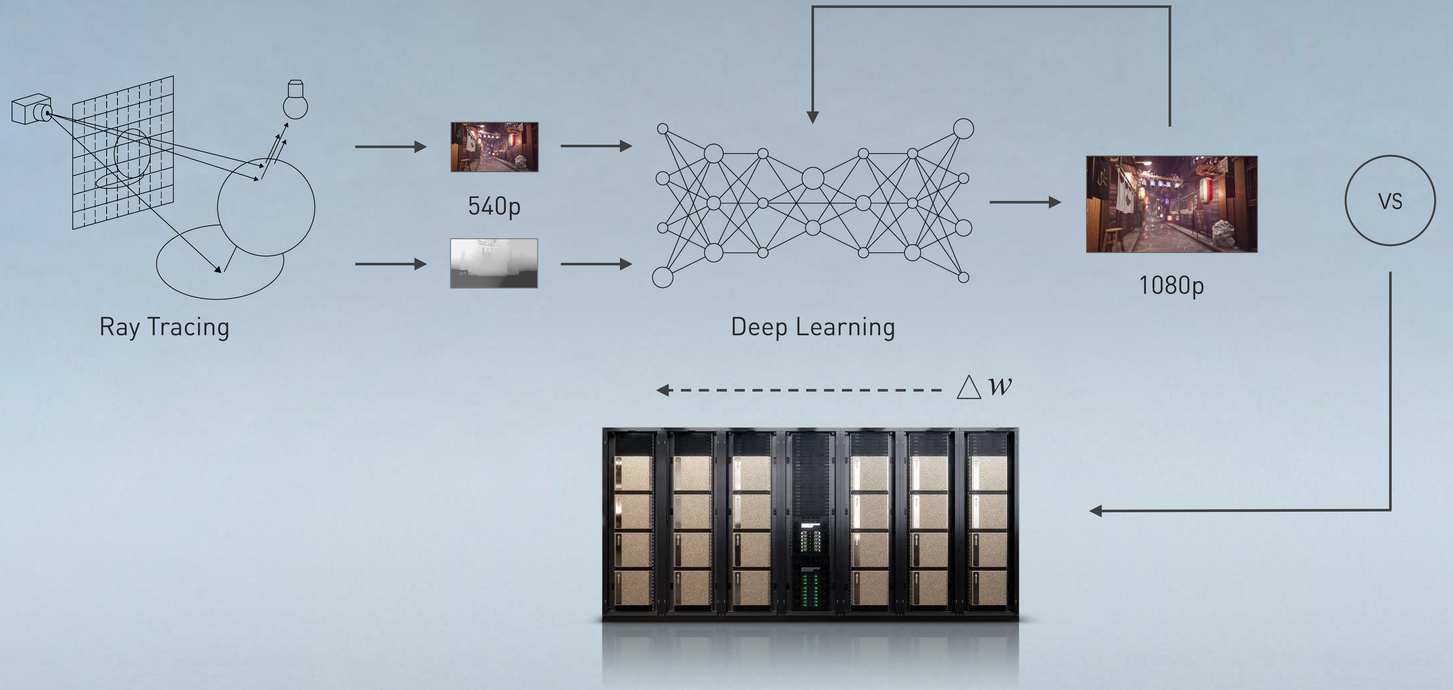
– Jon Peddie, JPR



NVIDIA RTX
A NEW ERA IN COMPUTER GRAPHICS — RAY TRACING AND AI

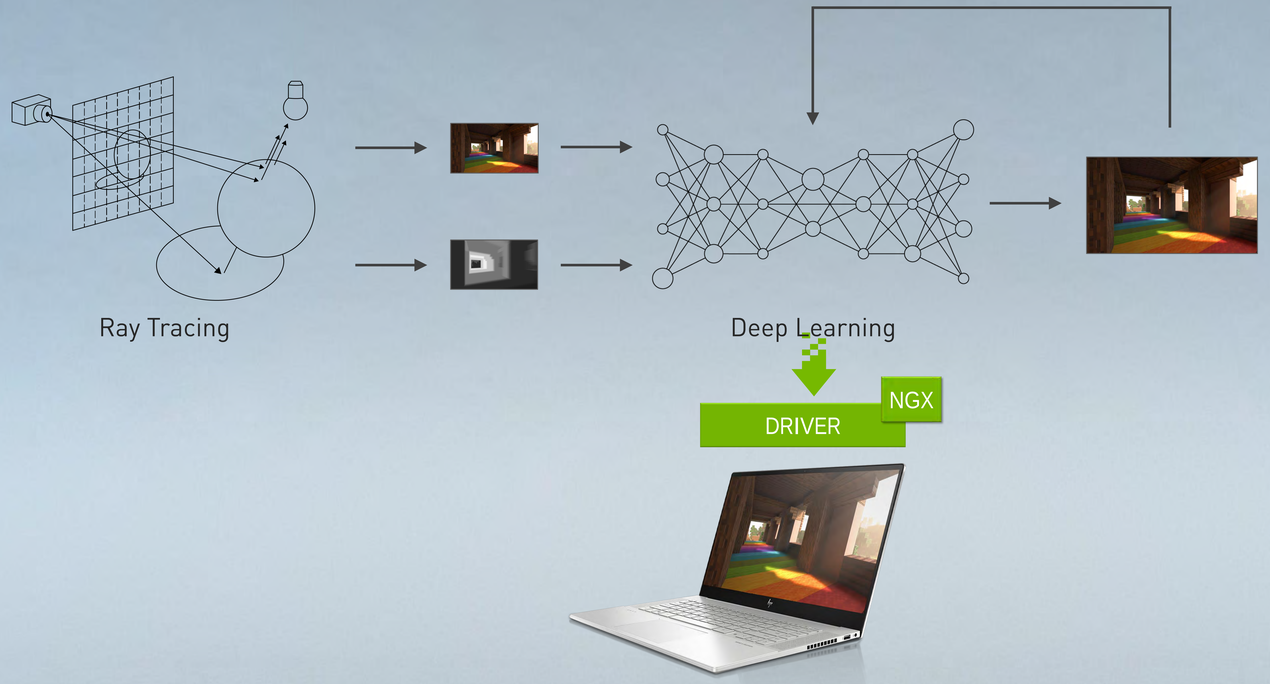


NVIDIA RTX
A NEW ERA IN COMPUTER GRAPHICS — RAY TRACING AND AI



Supercomputer Rendered - 16K Ground Truth

NVIDIA RTX
A NEW ERA IN COMPUTER GRAPHICS — RAY TRACING AND AI




Ground
Truth 16K


Native
720p


DLSS 1.0
720p > 1080p


DLSS 2.0
720p > 1080p

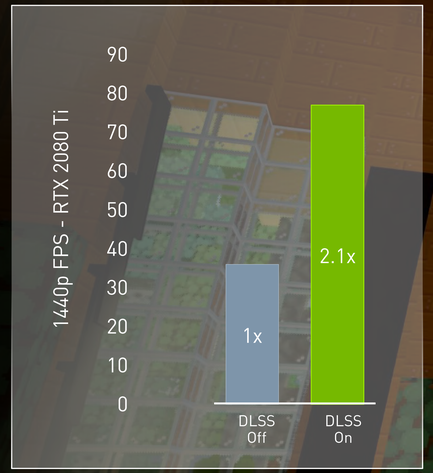

Native
1080p


Ground
Truth 16K


Native
540p


DLSS 2.0
540p > 1080p

MINECRAFT



MOJANG

“Gamechanger"

– Digital Foundry

“My God, it’s gorgeous"

– PCWorld

“Nothing short of awesome"

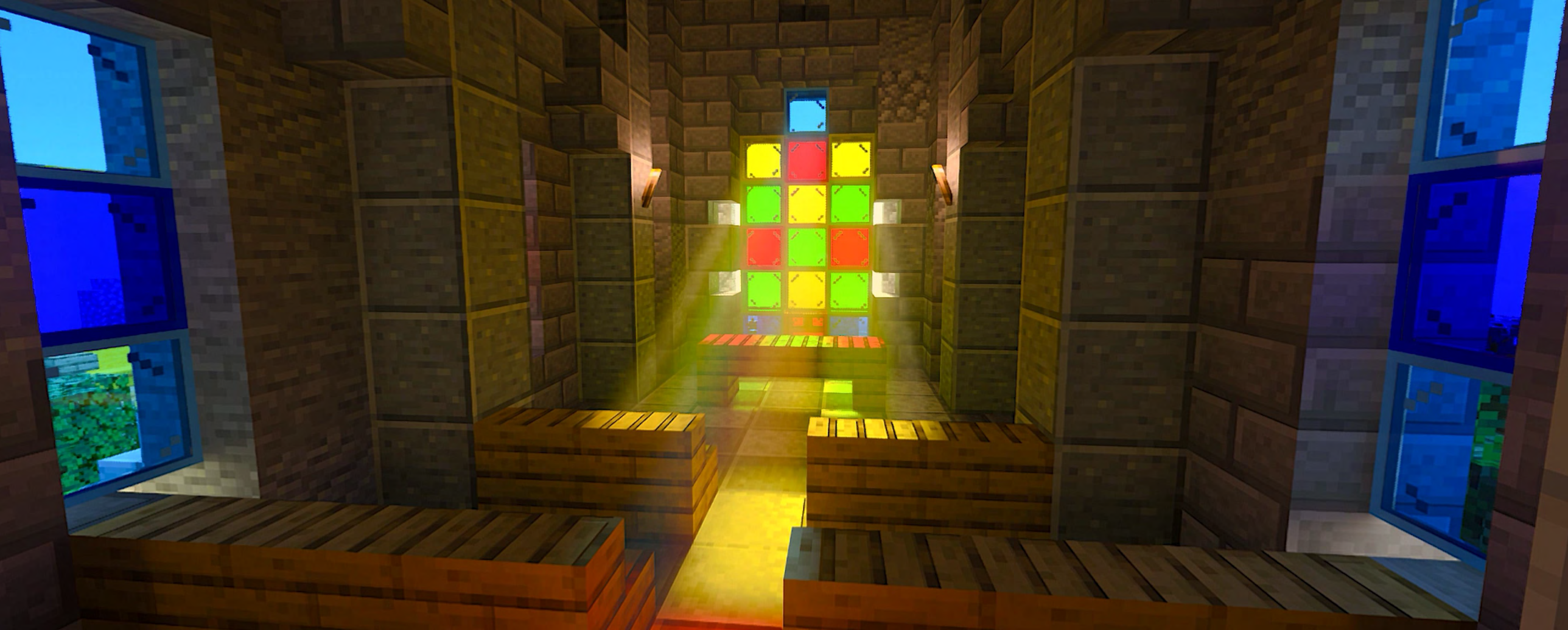
– IGN

“Stunning"

– PC Gamer

“Jaw-dropping"

– Trusted Reviews



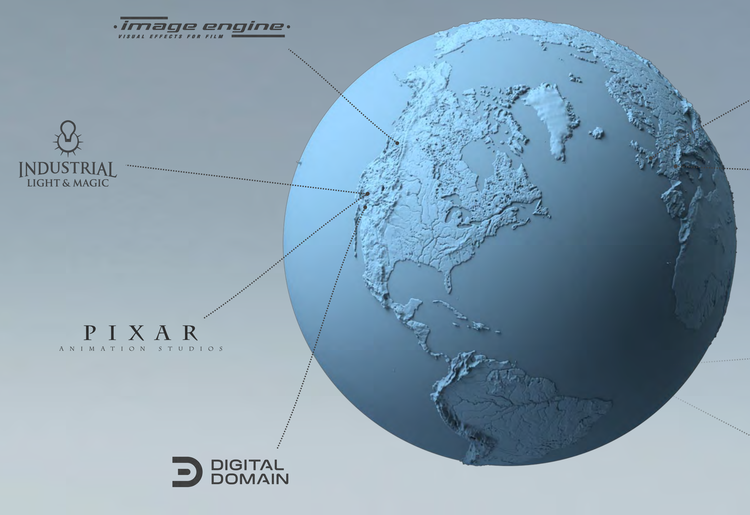
3D IMMENSELY COMPLEX

Different Tools and Giant Data Sets

Large Teams of Diverse Experts

Multiple Locations and Studios

Expensive



NVIDIA OMNIVERSE

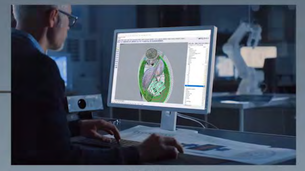
Design Workflow Collaboration Platform

Built on USD Universal Scene Description

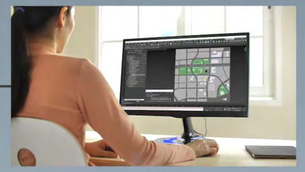
Built-In Interactive Renderer with Materials and Physics

Support PC and Linux; Streaming Clients for Macs, Android

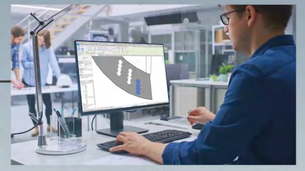
Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS



RHINO



MAX



REVIT

AR

Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS


HOORKA
OIL COLOUR

ANNOUNCING
NVIDIA RTX SERVER OPTIMIZED FOR REMOTE COLLABORATION

Design Workflow Collaboration with Omniverse
Interactive Production Rendering with Fully Ray-Traced Global Illumination
Quadro Virtual Workstations Validated with Design and Simulation

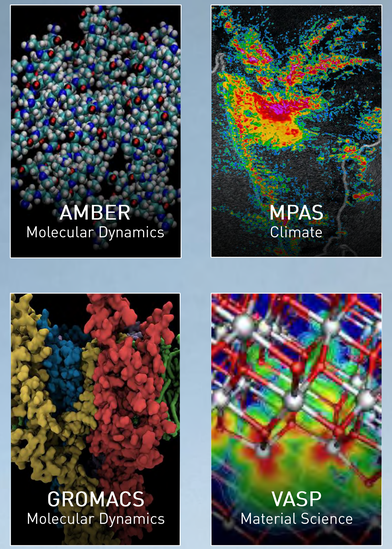
NVIDIA RTX Server Shipping Now

BOXX DELL Technologies

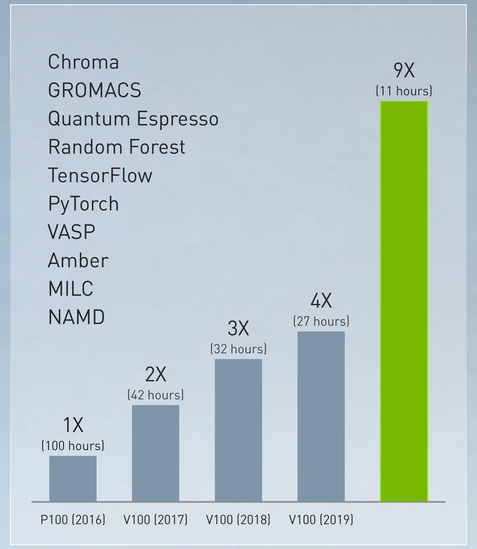
Hewlett Packard Enterprise SUPERMICRO

Leeza SOHO, Beijing
by ZAHA HADID ARCHITECTS

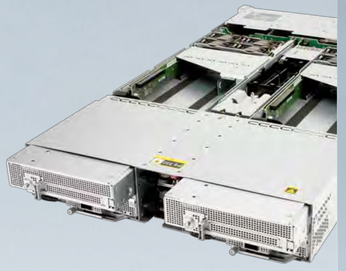
NVIDIA HPC



+700 CUDA ACCELERATED APPLICATIONS

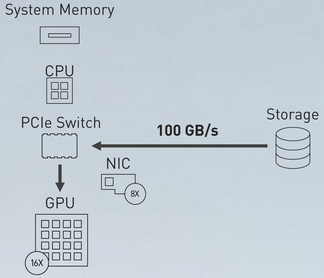


9X PERFORMANCE
IN 4 YEARS

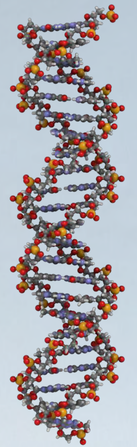


HPE Apollo70
Marvell ThunderX
ARM CPU

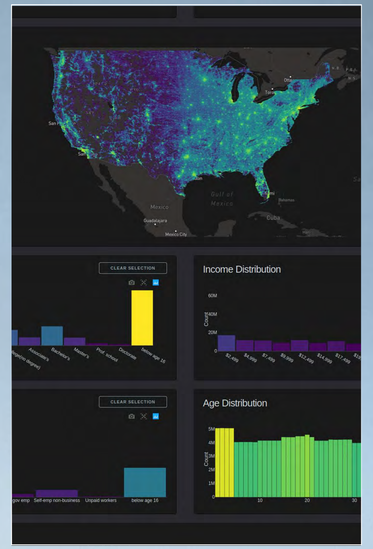
CUDA ON ARM



NVIDIA MAGNUM-IO
I/O Acceleration

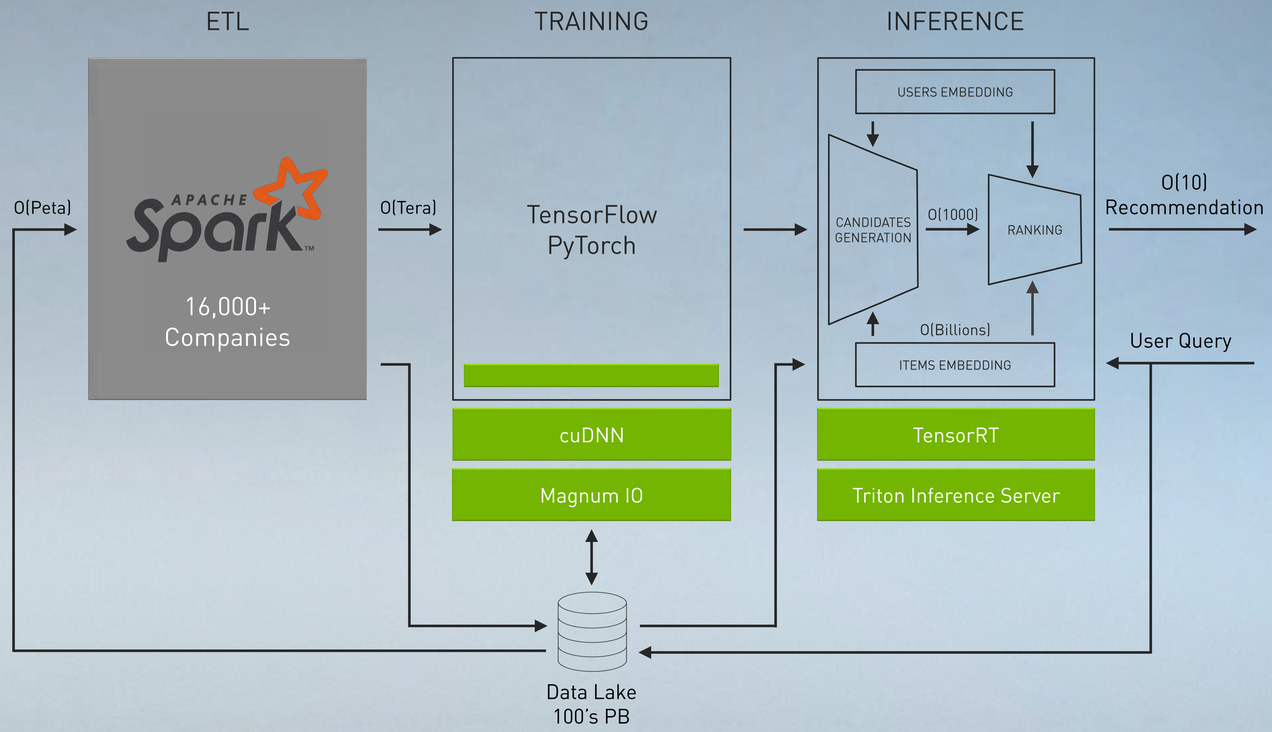


NVIDIA PARABRICKS
Genomics

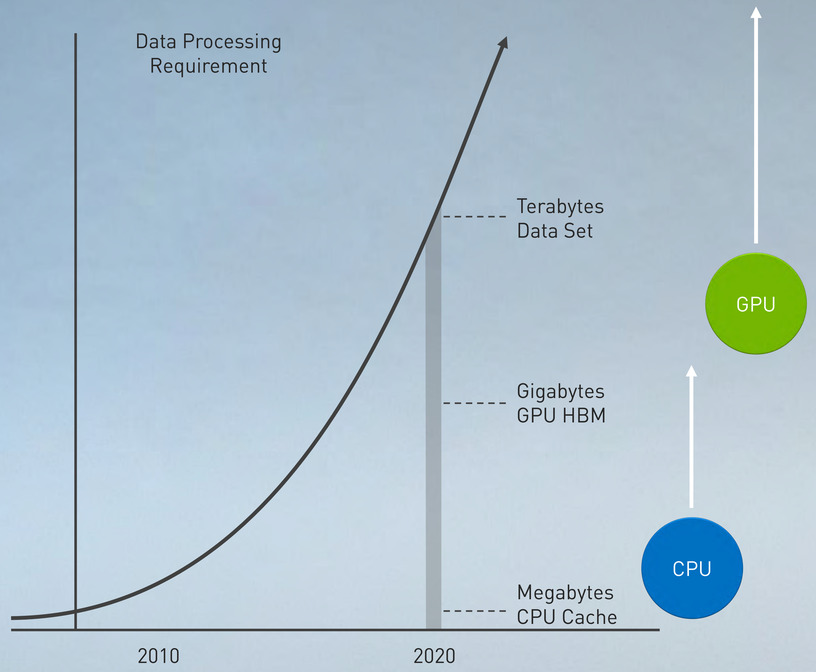


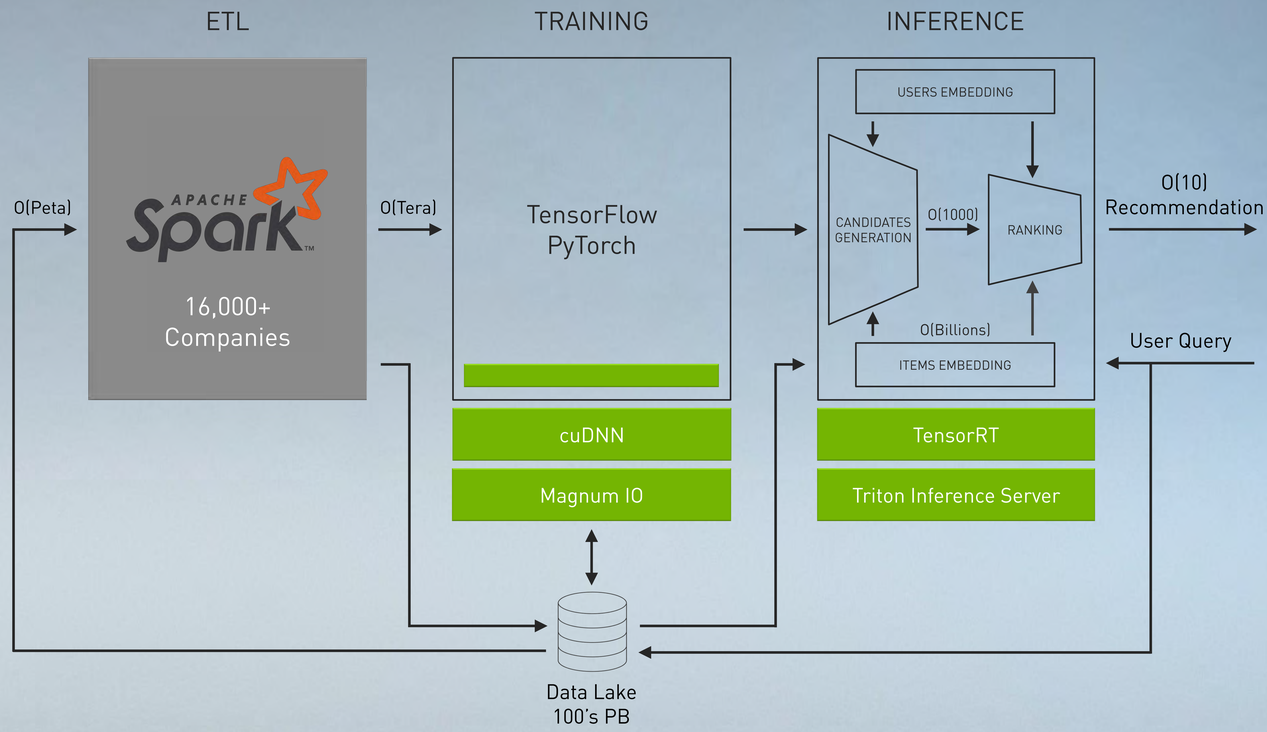
NVIDIA RAPIDS
Data Analytics

MACHINE LEARNING PIPELINE IS AN HPC CHALLENGE

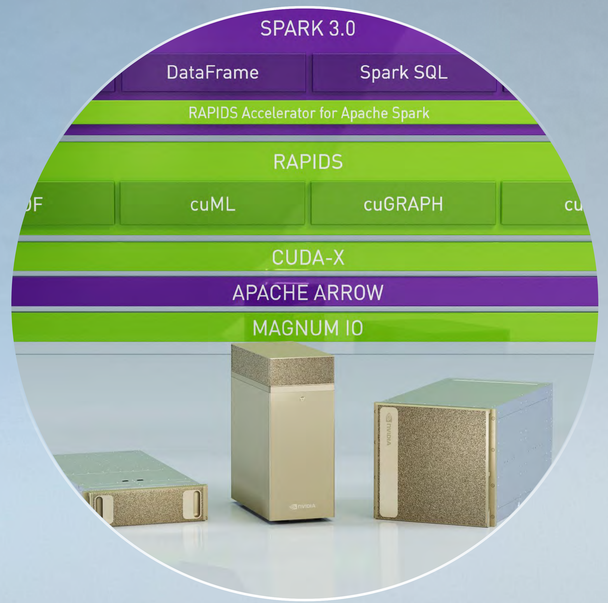


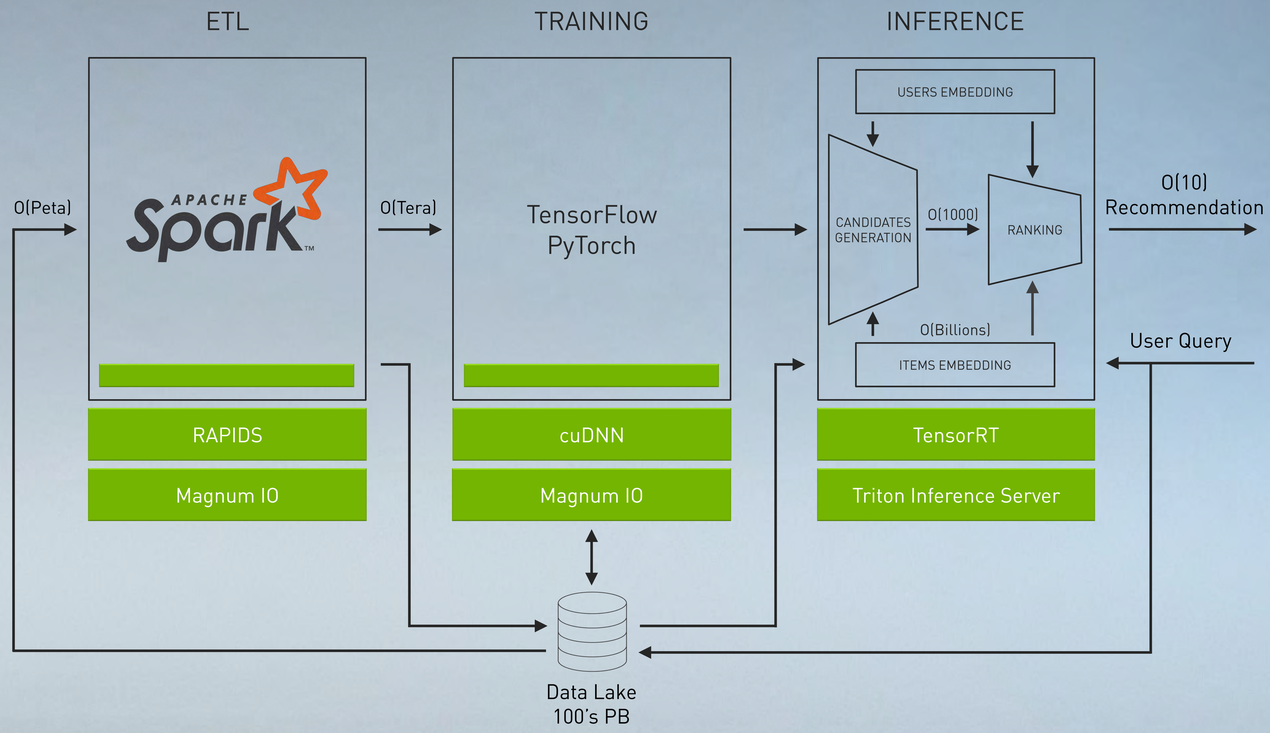
MACHINE LEARNING DRIVING EXPONENTIAL GROWTH IN DATA





ANNOUNCING
NVIDIA ACCELERATES SPARK 3.0

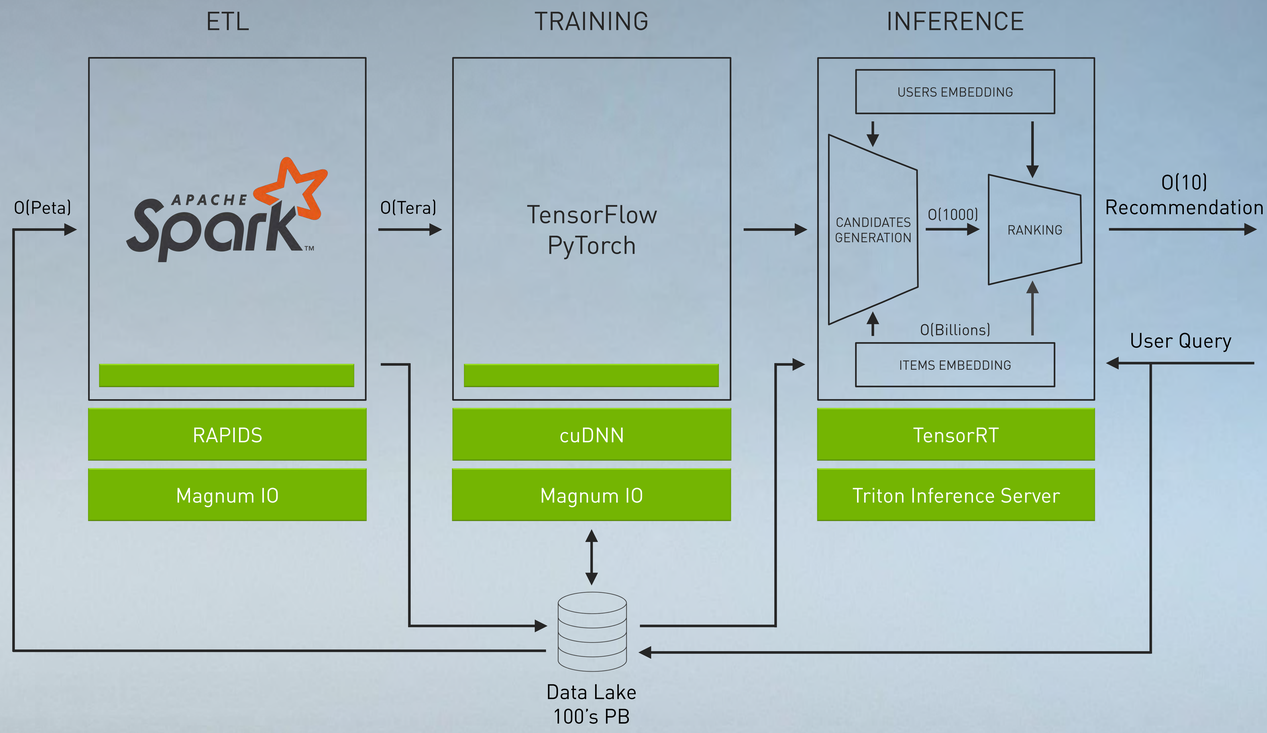




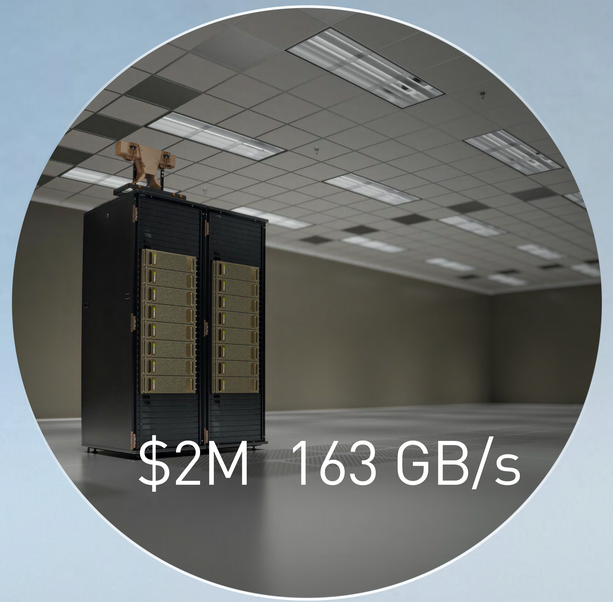
SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK

17 GB/s Throughput on TPCx-BB @ SF 10K

$1M | 18 2U CPU Systems | 2 Racks | 16 kW

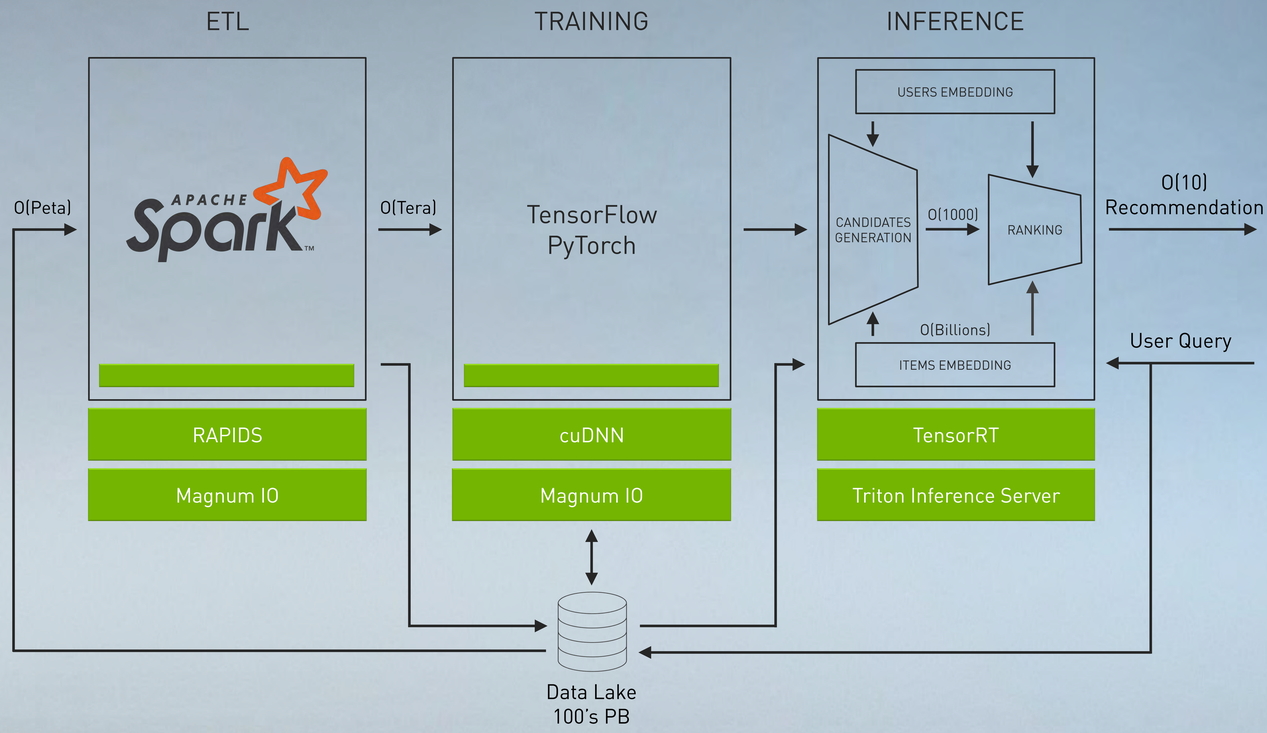


SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK



163 GB/s Throughput on RAPIDS Implementation of TPCx-BB @ SF 10K

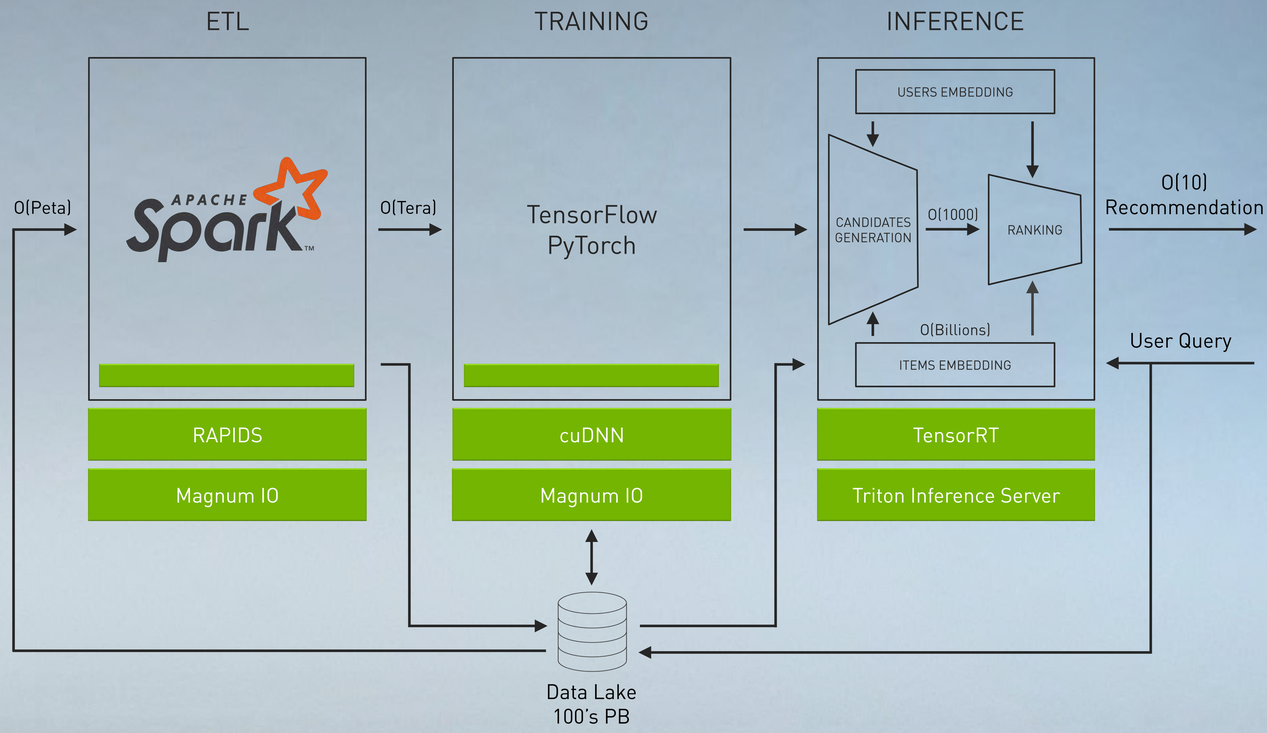
$2M | 16 DGX-1 | 2 Racks | 56 kW



SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK

163 GB/s Throughput on RAPIDS Implementation of TPCx-BB @ SF 10K

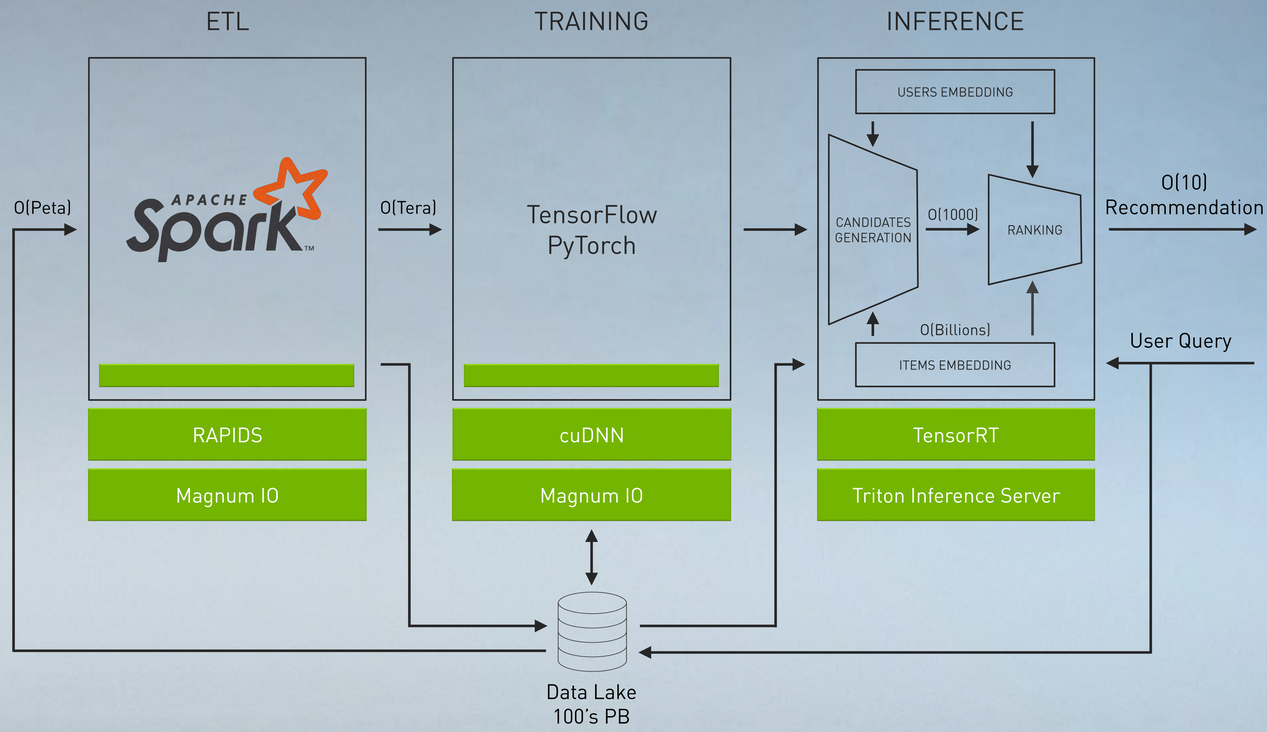
$2M | 16 DGX-1 | 2 Racks | 56 kW



SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK

Equivalent 163 GB/s Throughput on TPCx-BB @ SF 10K

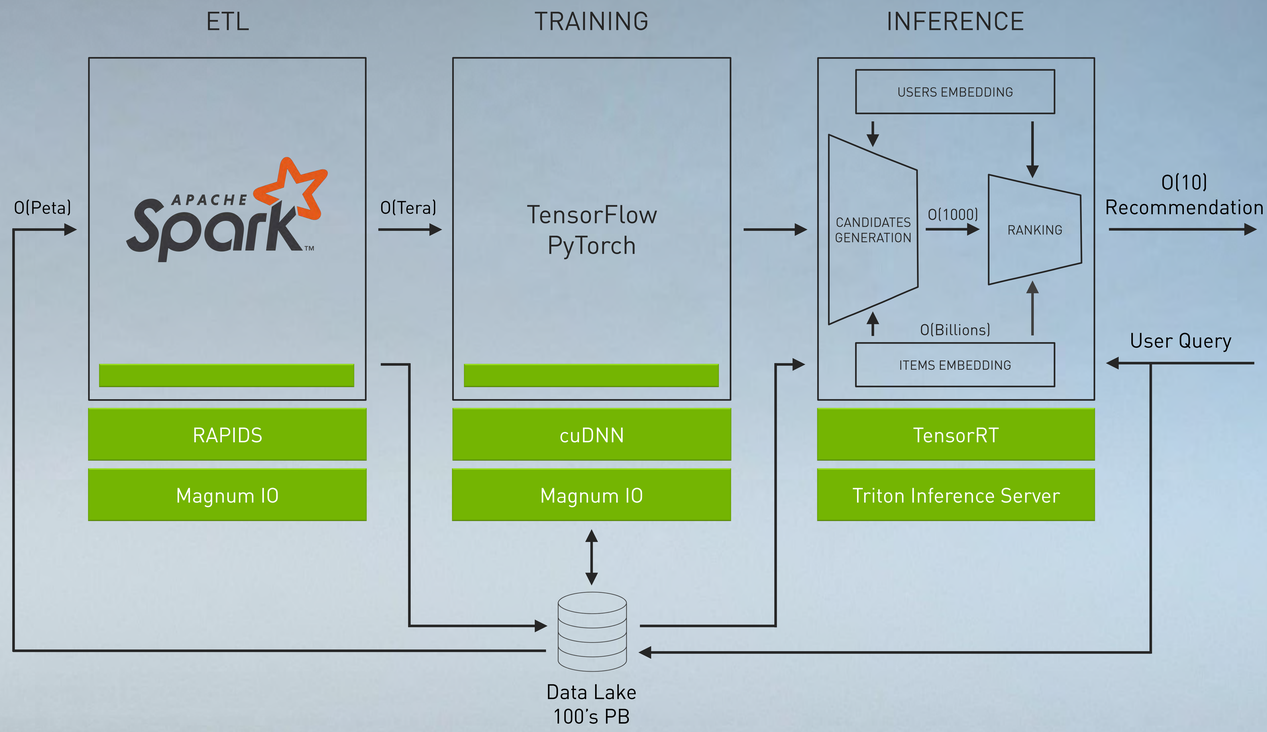
$10M | 167 2U CPU Systems | 11 Racks | 140 kW



SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK

163 GB/s Throughput on RAPIDS Implementation of TPCx-BB @ SF 10K

$2M | 16 DGX-1 | 2 Racks | 56 kW

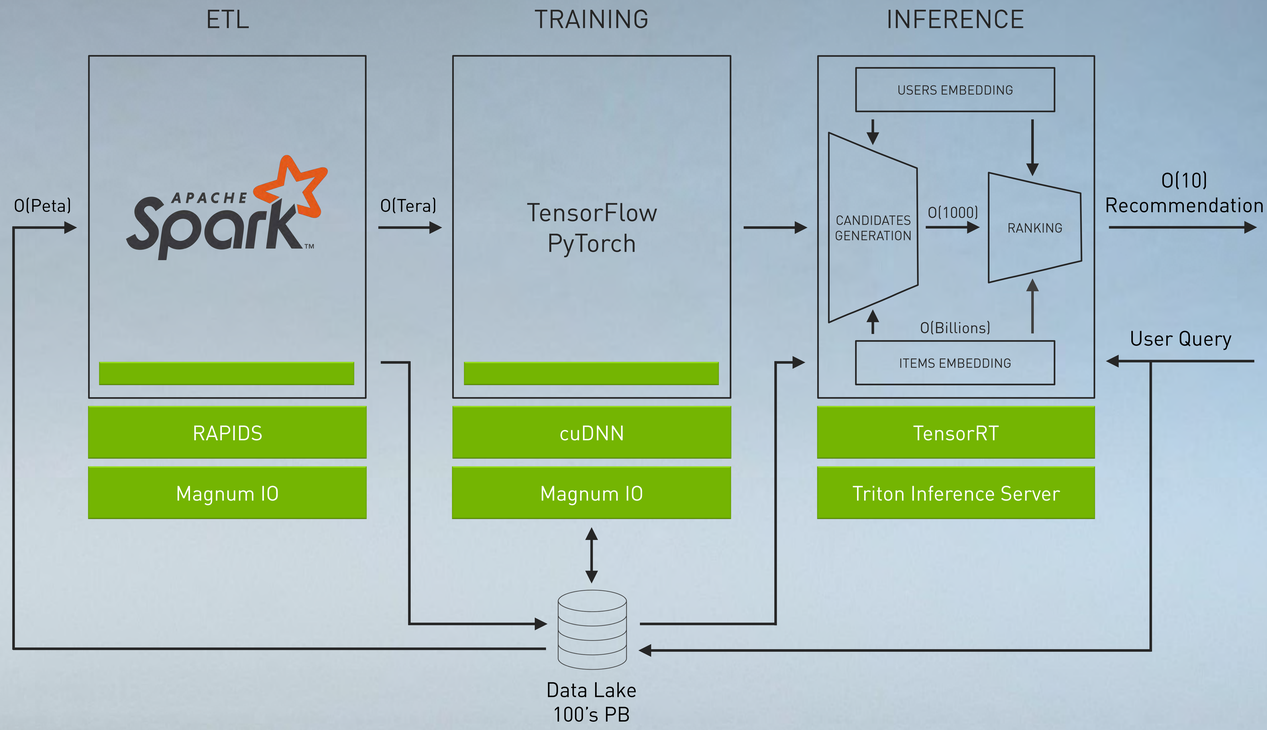


ANNOUNCING
DATABRICKS ACCELERATED WITH NVIDIA

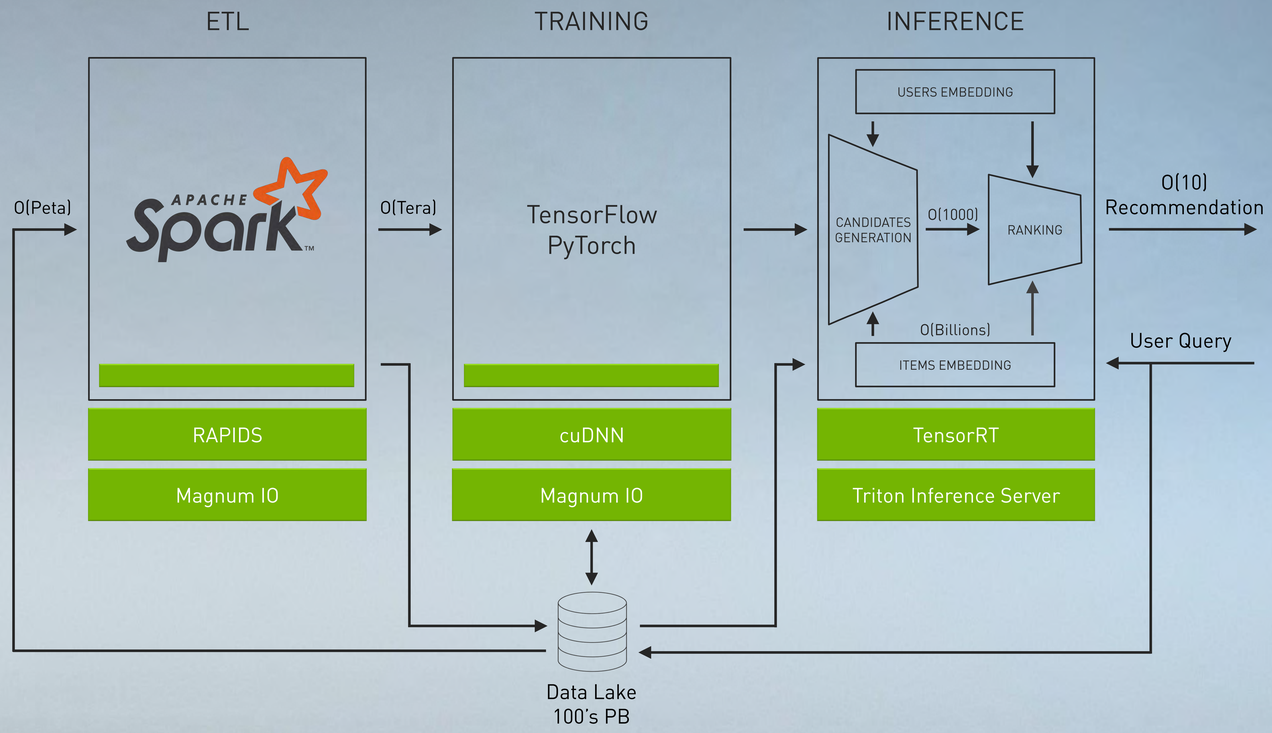
databricks

" These contributions lead to faster data pipelines, model training and scoring for more breakthroughs and insights with Apache Spark 3.0 and Databricks."

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

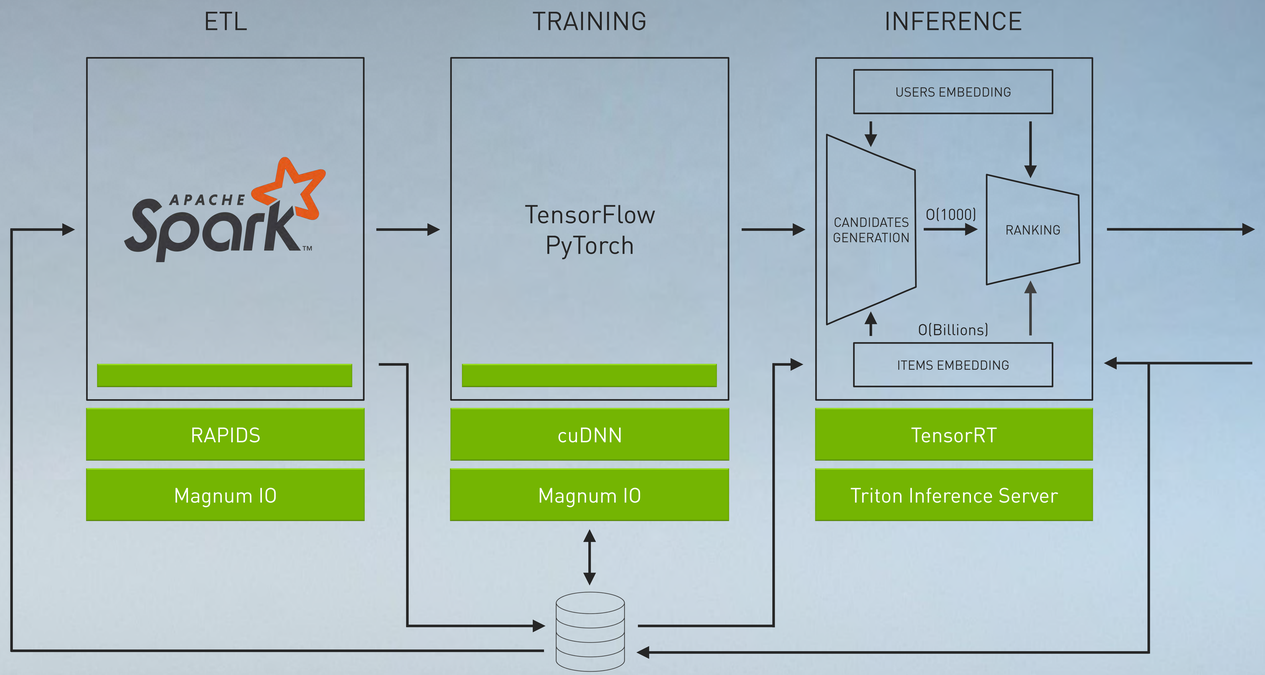


ANNOUNCING
CLOUD ANALYTICS PLATFORMS ACCELERATED WITH NVIDIA

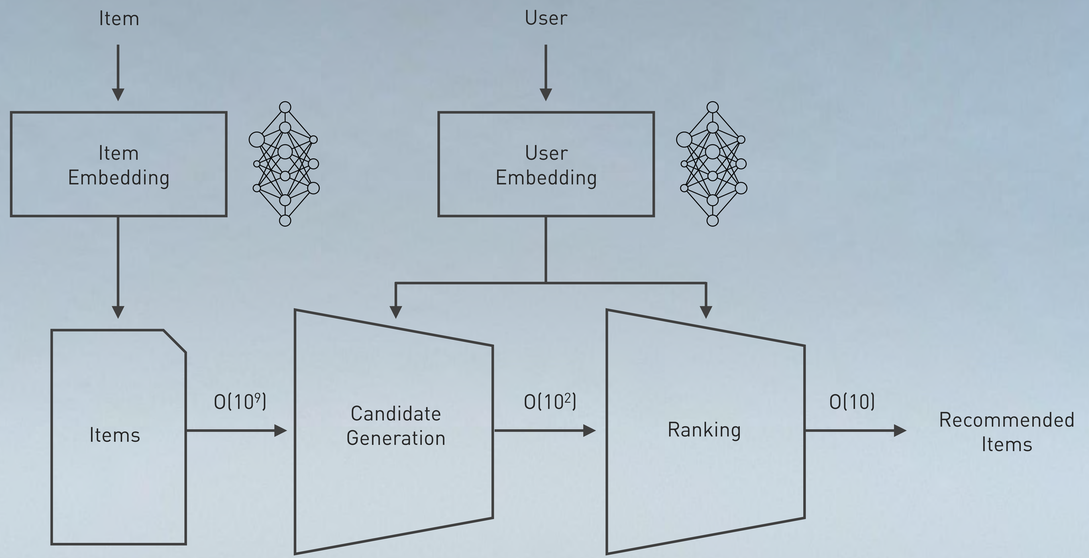


NVIDIA AI

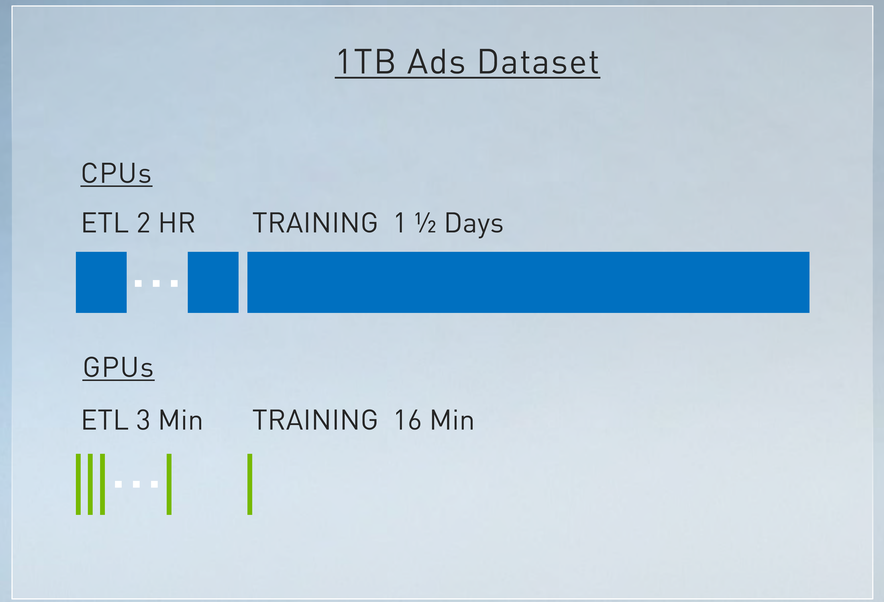

Amazon SageMaker
Azure Machine Learning
databricks
Google Cloud AI
Google Cloud Dataproc
SPARK
DASK
XGBOOST
TensorFlow
PyTorch
MXNE
RAPIDS
cuDF
cuML
cuGRAPH
cuDNN
CUDA-X
APACHE ARROW
MAGNUM IO

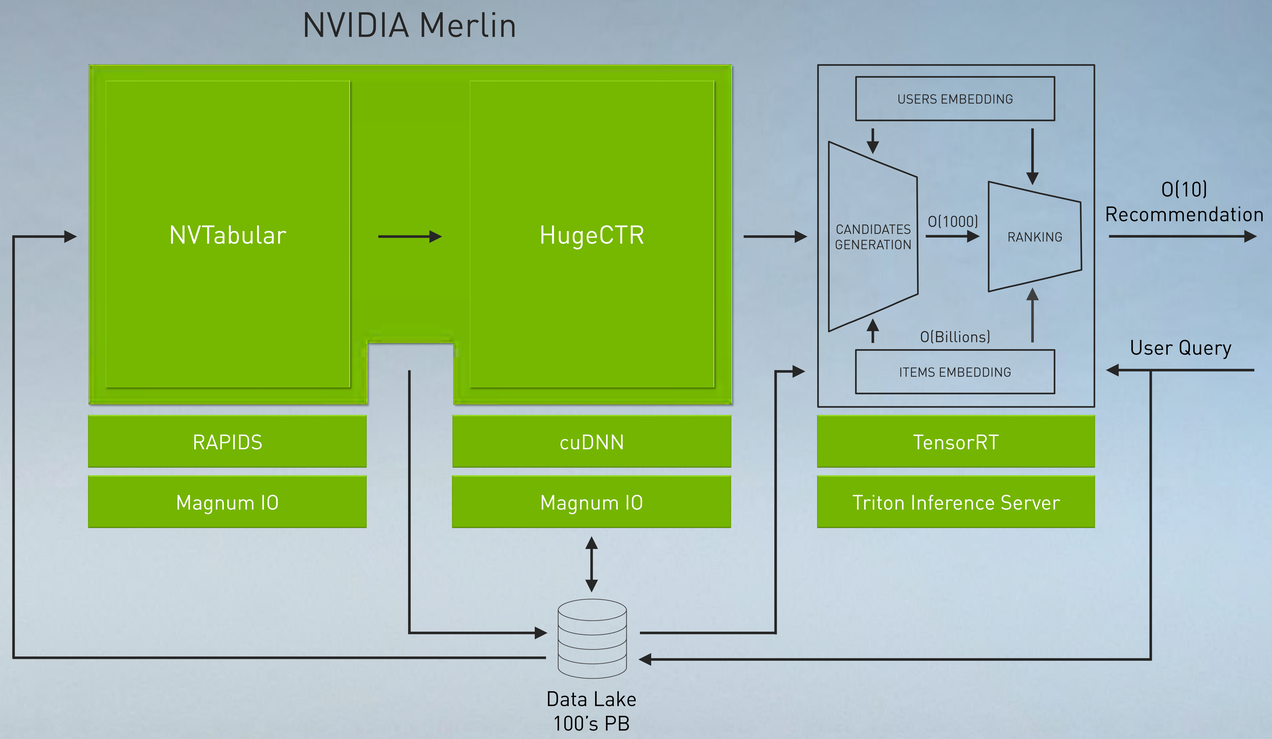

ETL
TRAINING
INFERENCE
APACHE Spark
TensorFlow
PyTorch
USERS EMBEDDING
CANDIDATES GENERATION
O(1000)
RANKING
O(Billions)
ITEMS EMBEDDING
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton Inference Server

RECOMMENDER SYSTEM IS THE ENGINE OF THE PERSONALIZED INTERNET



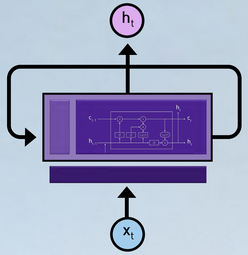
ANNOUNCING
NVIDIA MERLIN — DEEP RECOMMENDER APPLICATION FRAMEWORK



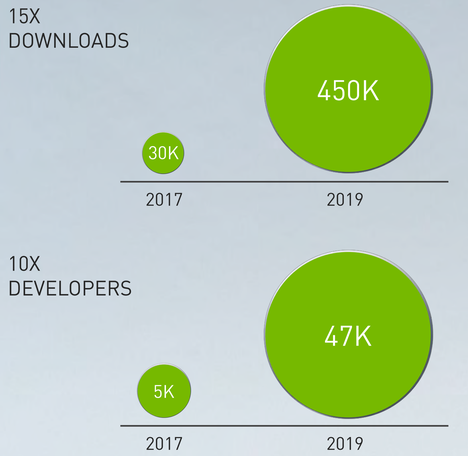


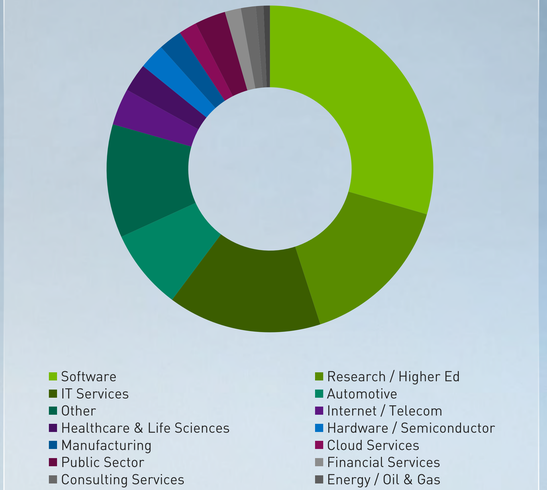
NVIDIA AI INFERENCE
15X TENSORRT DOWNLOADS IN 2 YEARS

TENSORRT 7

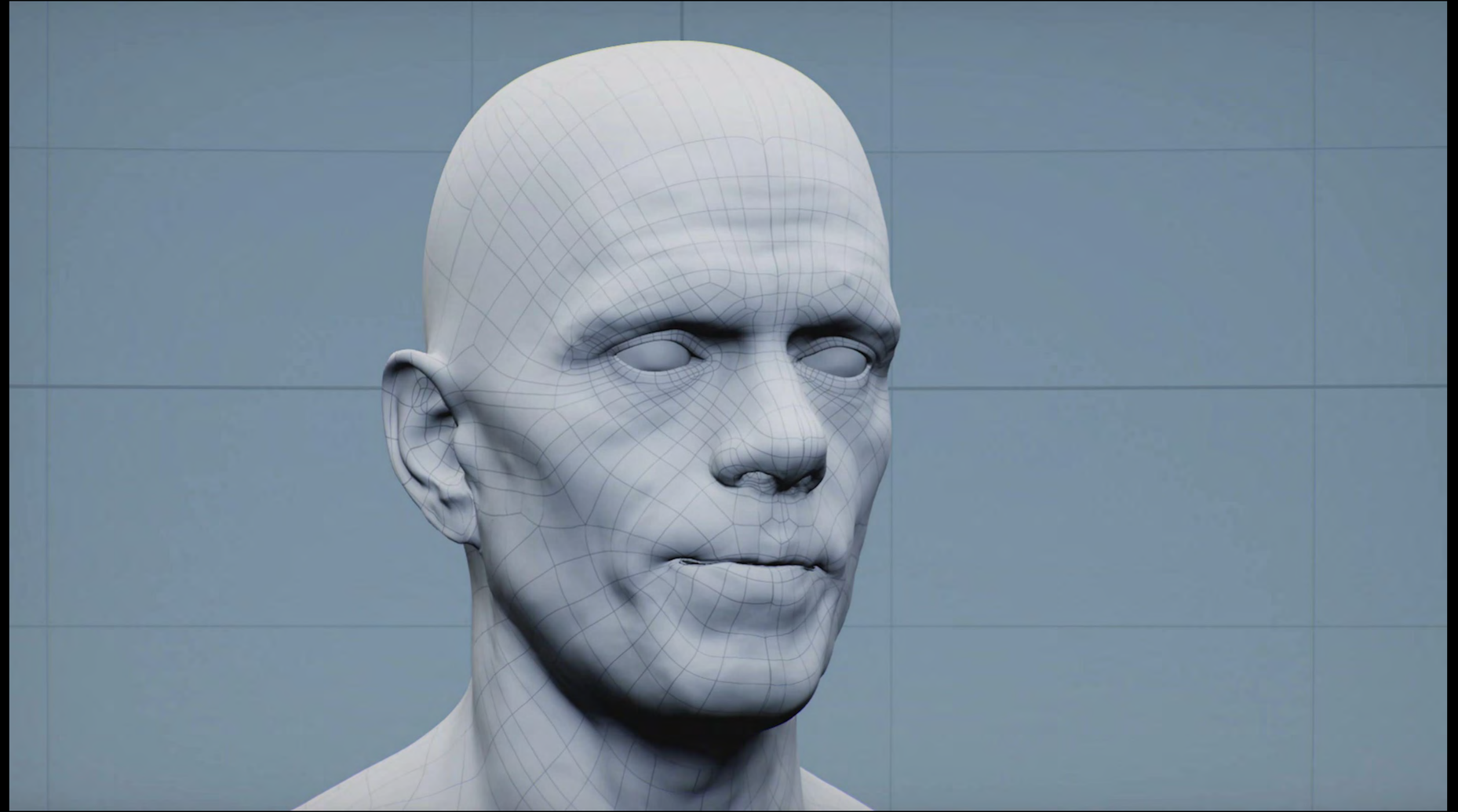


CNNs, Transformers, RNNs
1000+ Optimized Kernels
Automatic Mixed-Precision FP32, FP16 and INT8



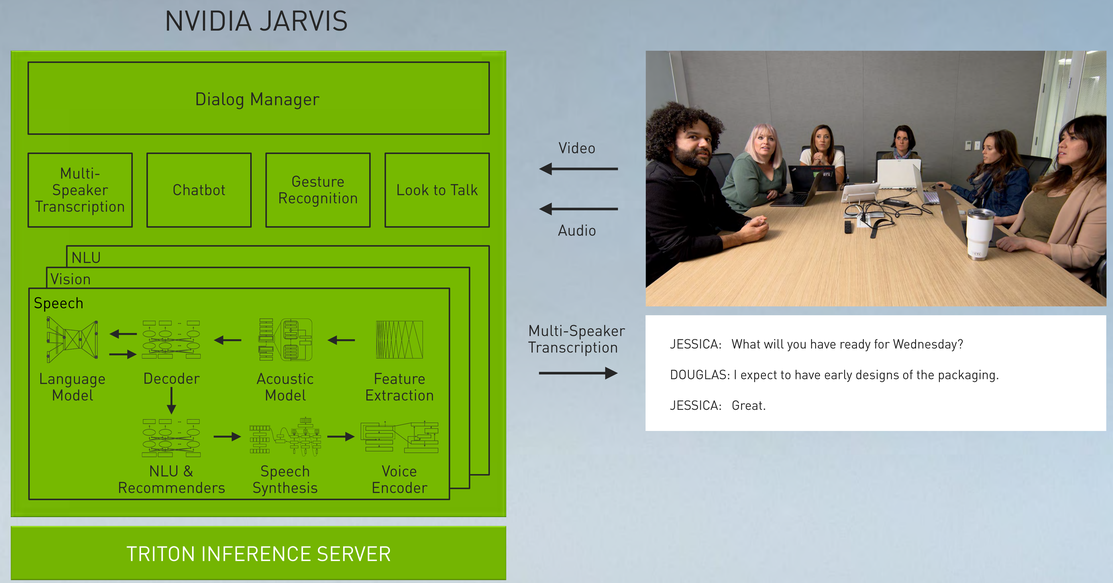


ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK

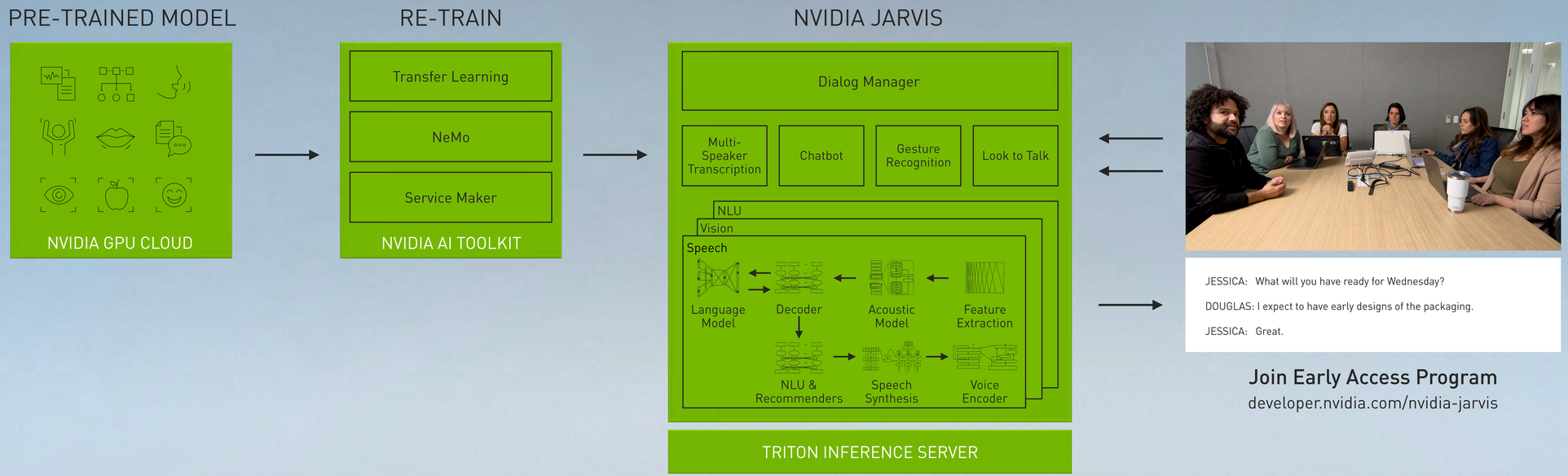


JARVIS DEMO
MISTY — INTERACTIVE 3D CHATBOT

ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK



ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK



Join Early Access Program

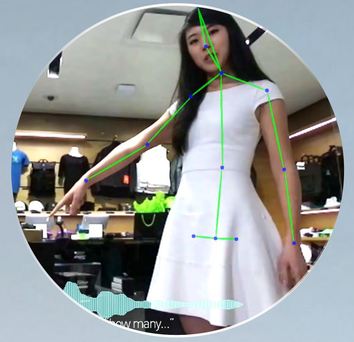
developer.nvidia.com/nvidia-jarvis

CONVERSATIONAL AI IS TRANSFORMING INDUSTRIES

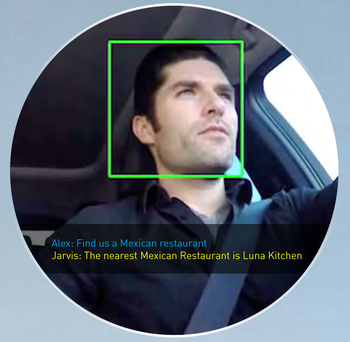
VIDEOCONFERENCE
CC, TRANSLATION, TRANSCRIPTION
200M Meetings per Day

CALL CENTER
500M Calls per Day

SMART SPEAKERS
150M Sold per Year



RETAIL ASSISTANTS
12M Retail Stores

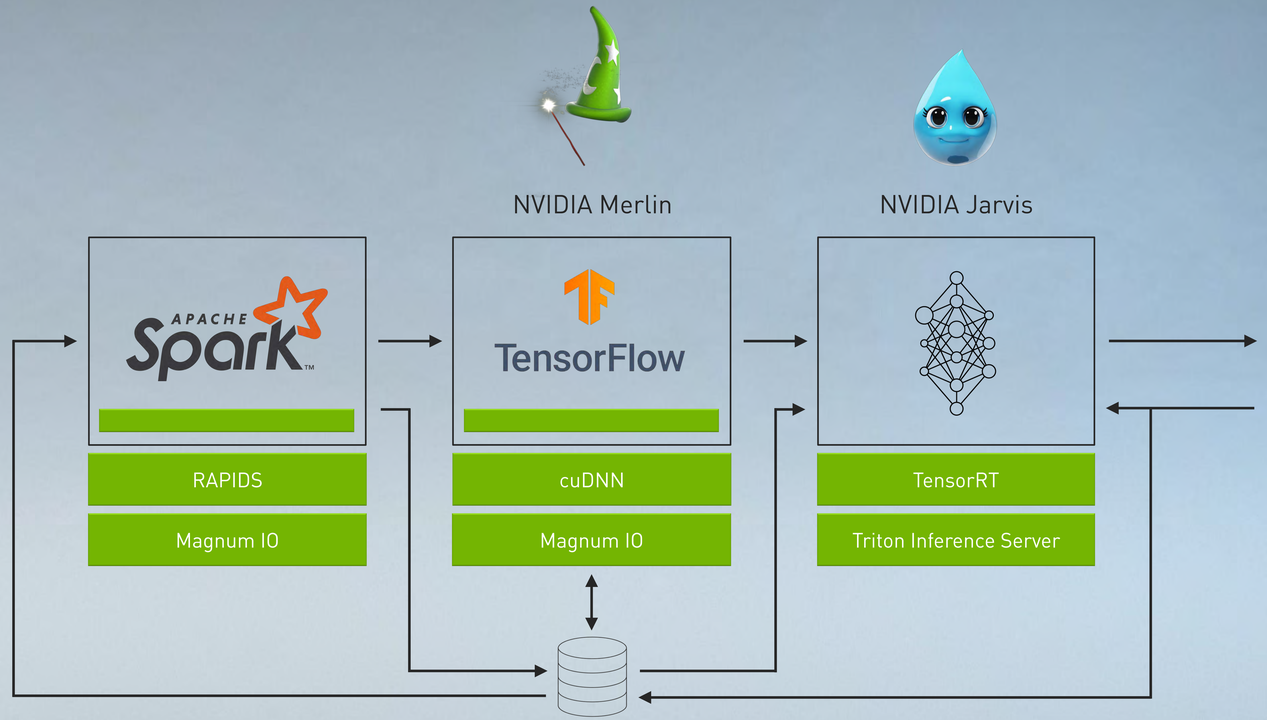


IN-CAR ASSISTANTS
75M New Cars per Year

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

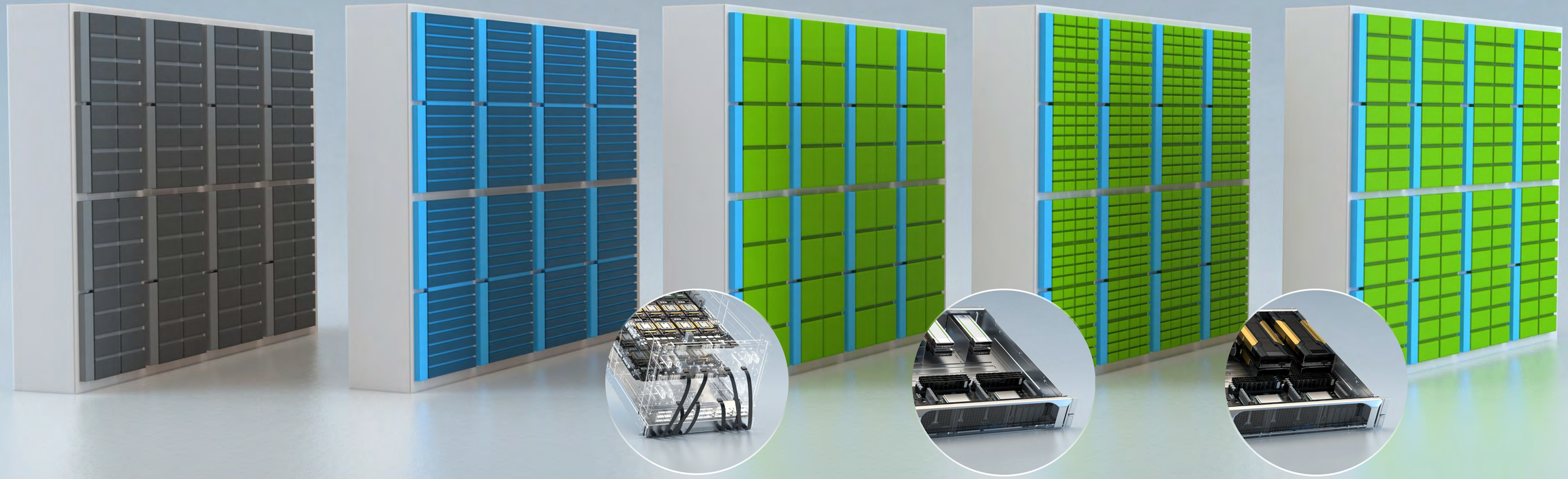
NVIDIA AI


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

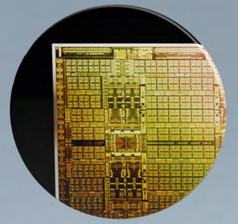

NVIDIA Merlin
NVIDIA Jarvis
APACHE Spark
TensorFlow
RAPIDS
Magnum IO
cuDNN
Magnum IO
TensorRT
Triton Inference Server

MODERN CLOUD DATA CENTER

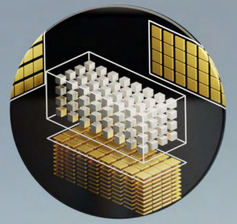
Diverse Applications | Scale-Up & Scale-Out Workloads | Insatiable Demand



ANNOUNCING NVIDIA A100

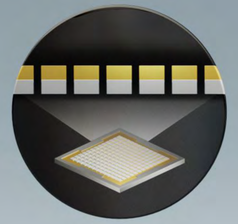


54 BILLION XTORS

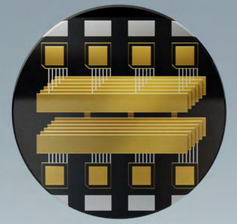


3RD GEN TENSOR CORES

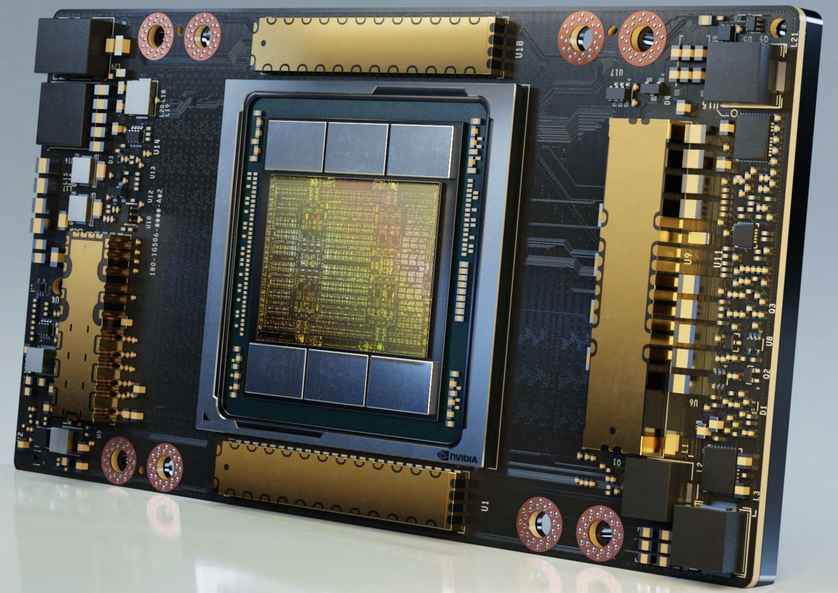
SPARSITY ACCELERATION



MIG

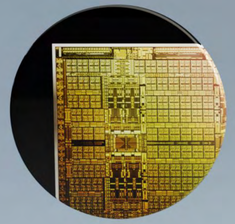


3RD GEN NVLINK & NVSWITCH

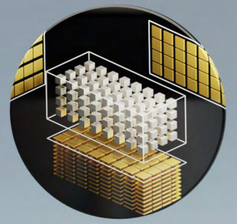


ANNOUNCING NVIDIA A100

TSMC 7nm | HBM2 — 1.6 Terabytes per Second | 3D Chip Stack

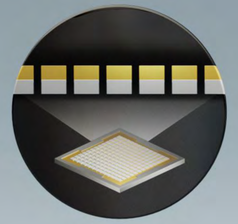


54 BILLION XTORS

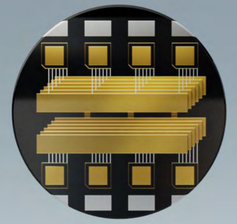


3RD GEN TENSOR CORES

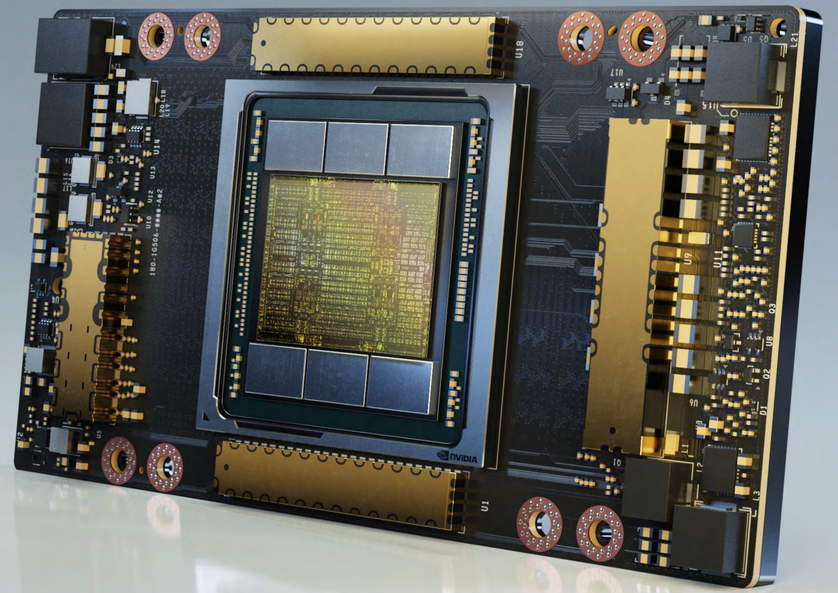
SPARSITY ACCELERATION



MIG

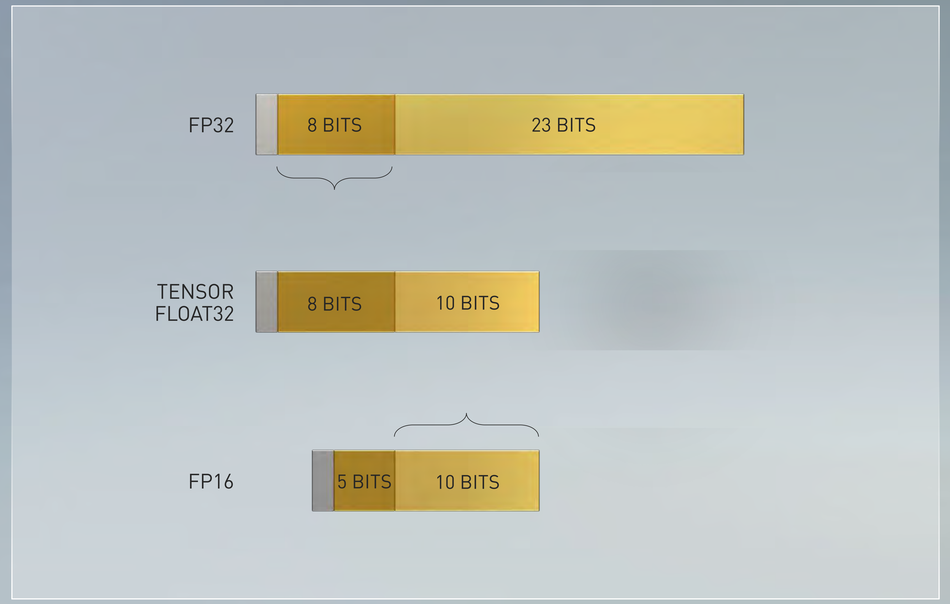


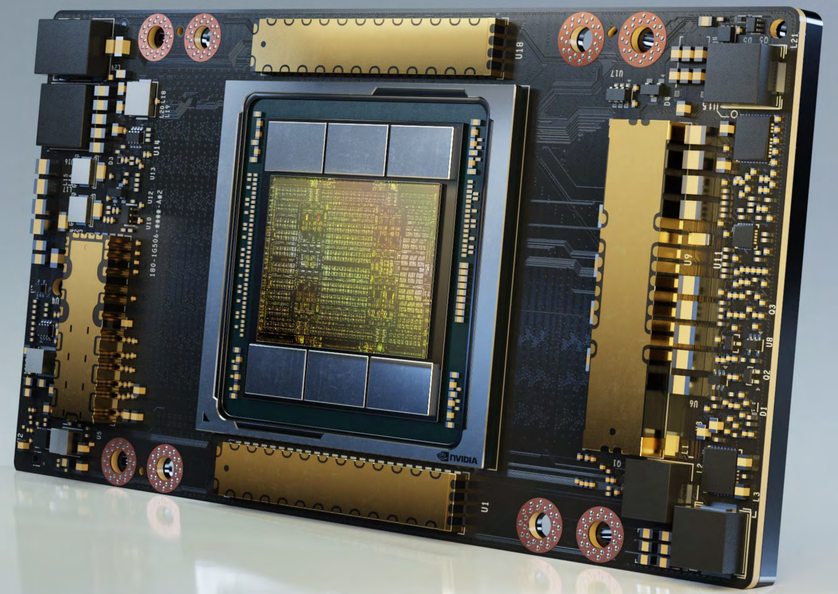
3RD GEN NVLINK & NVSWITCH



NEW TF32 TENSOR CORES

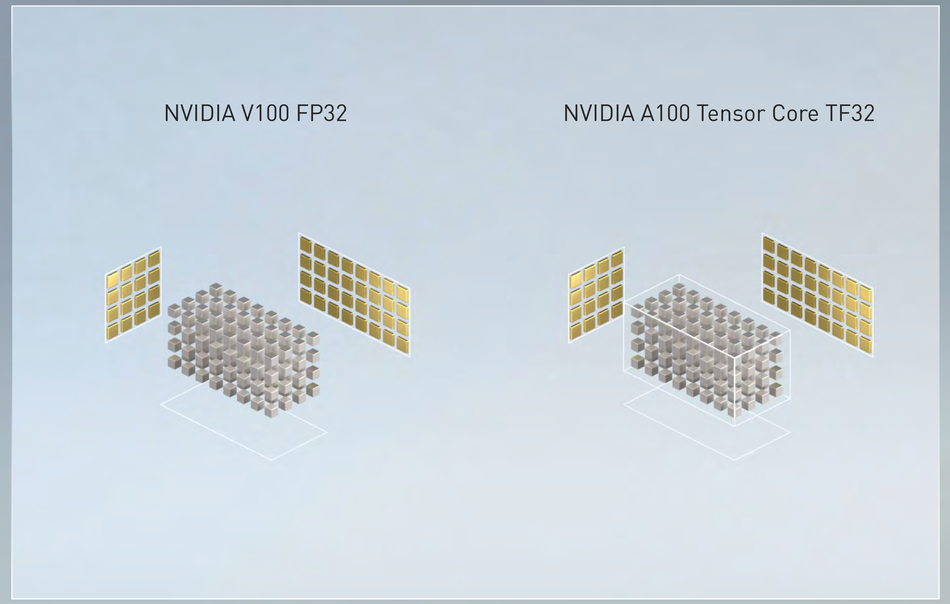
Range of FP32 and Precision of FP16 | Input in FP32 and Accumulation in FP32 | No Code Change Speed-up for Training

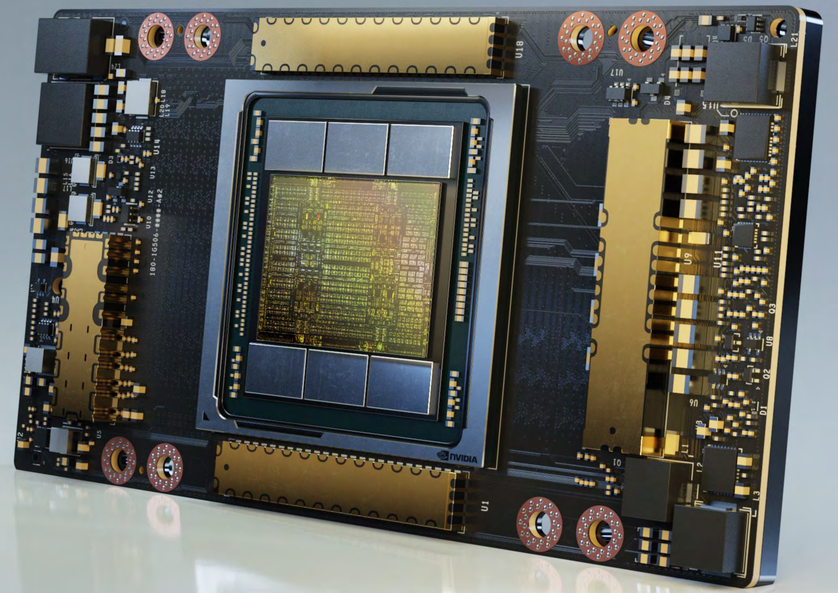




NEW TF32 TENSOR CORES

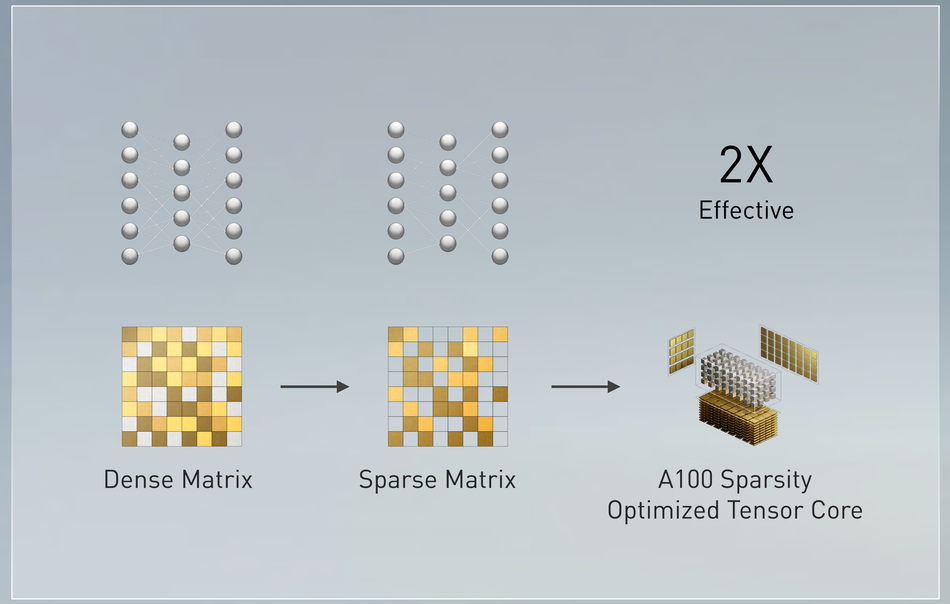
Range of FP32 and Precision of FP16 | Input in FP32 and Accumulation in FP32 | No Code Change Speed-up for Training

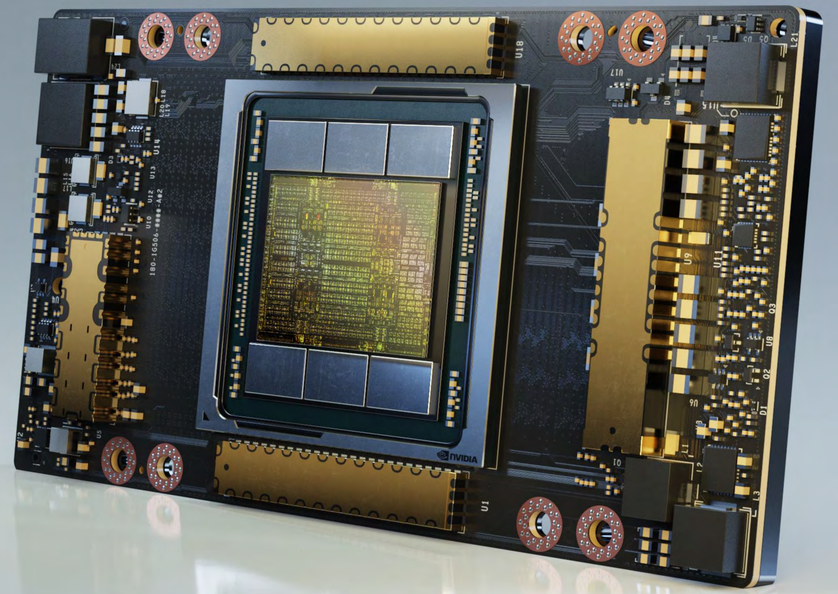




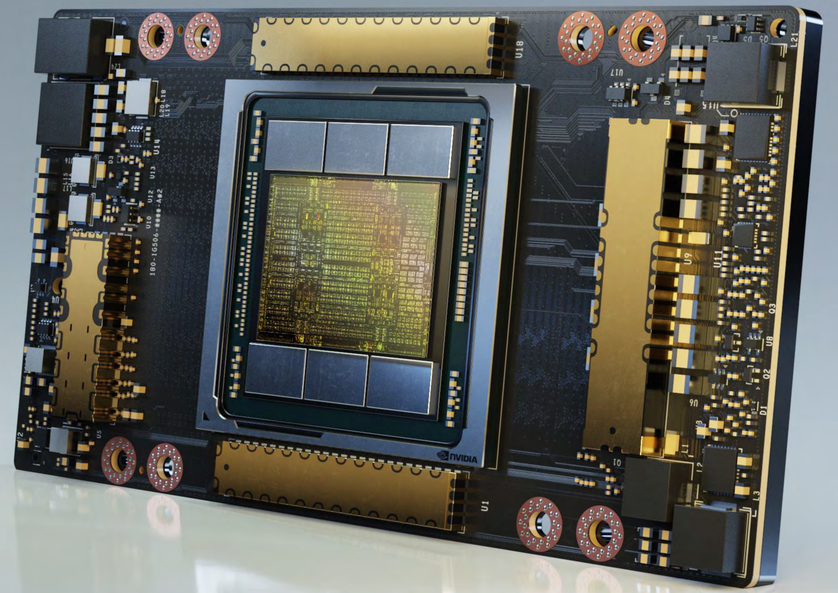
NEW TENSOR CORE ACCELERATION FOR SPARSITY

Optimized For Sparse AI Tensor Ops | 2X Faster Execution | Supported on TF32, FP16, BFLOAT16, INT8 and INT4

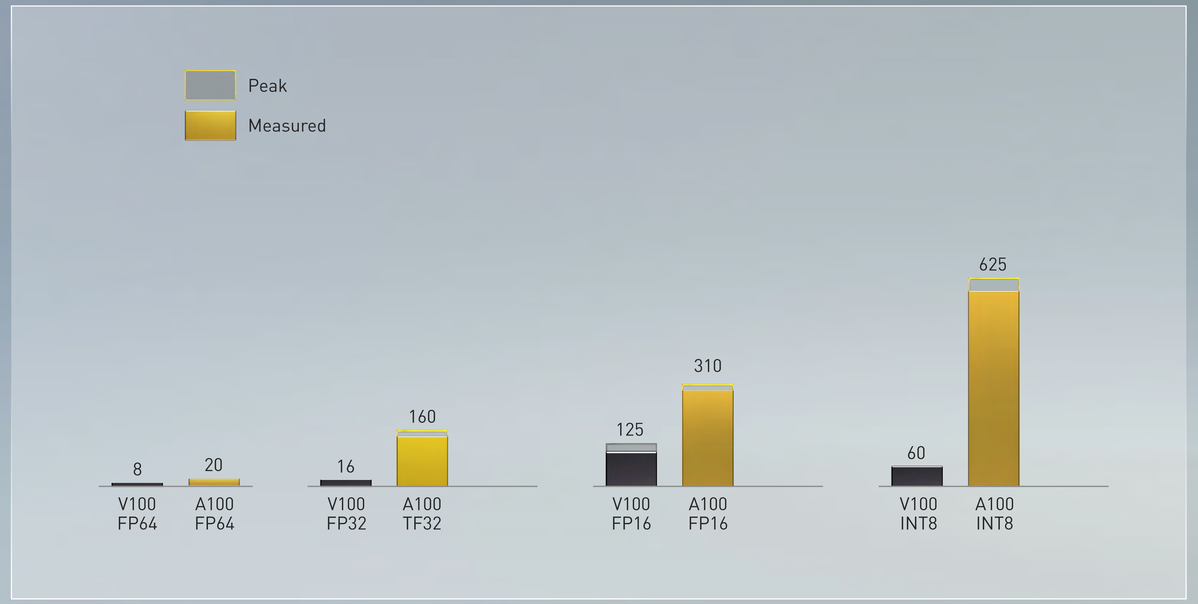


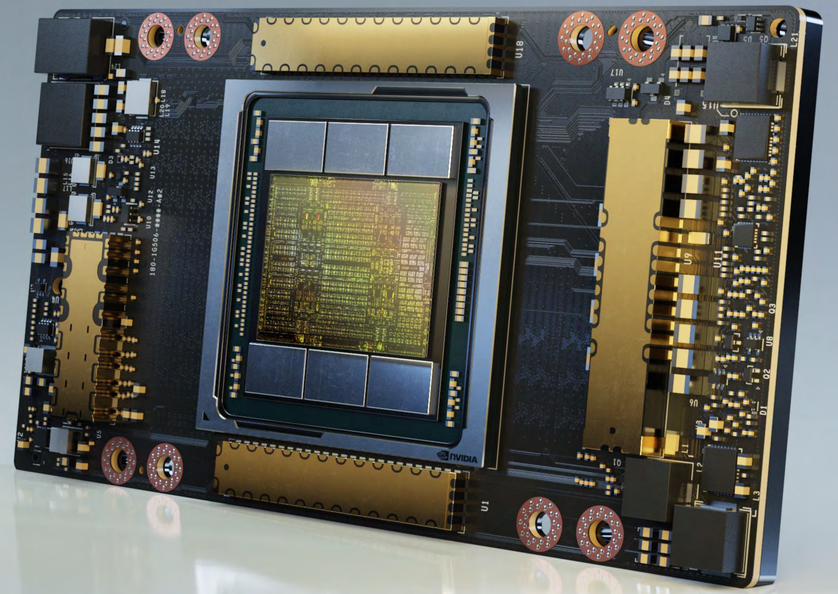


ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP

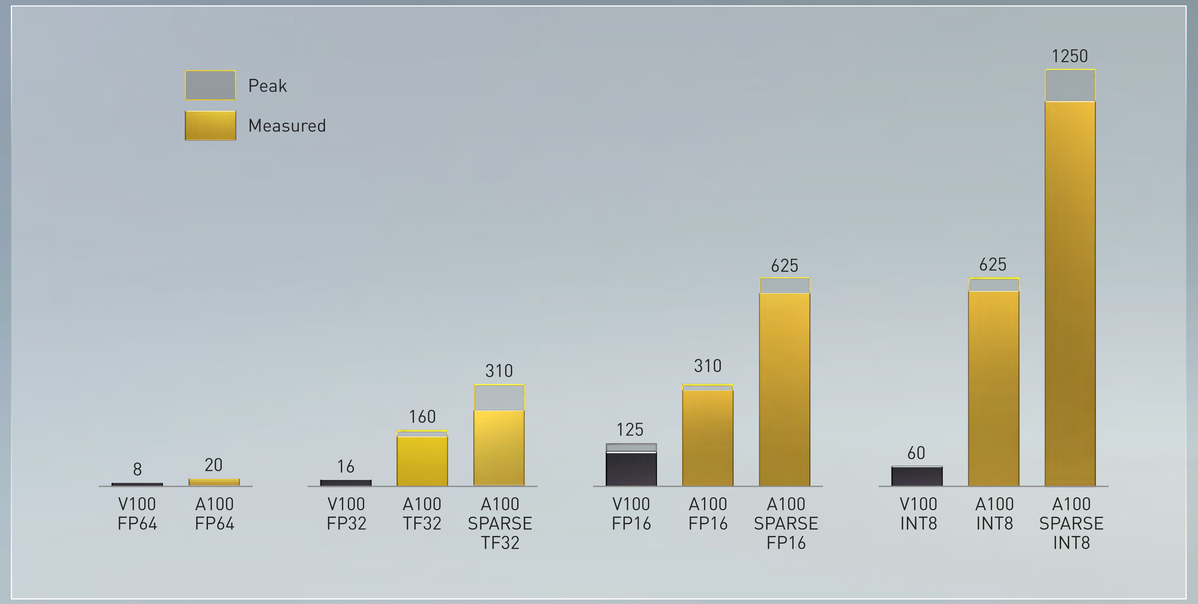


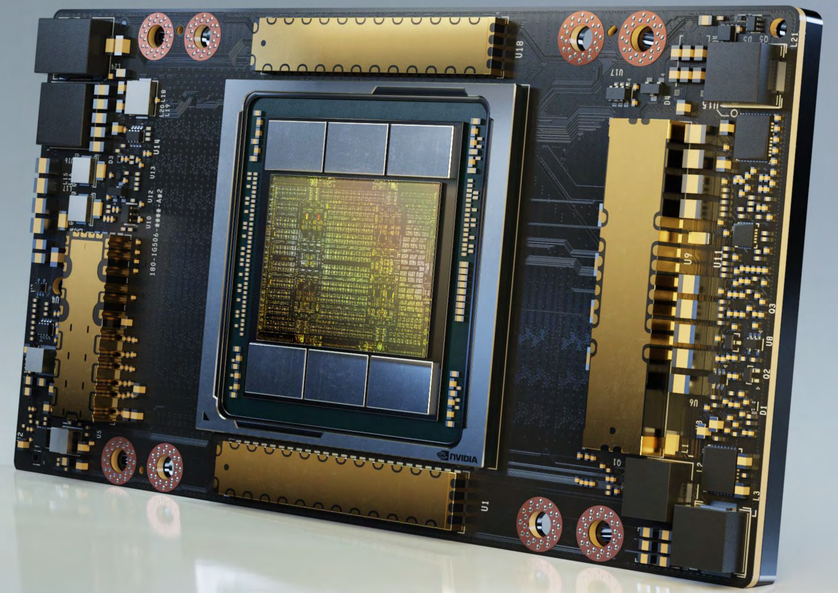
ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP



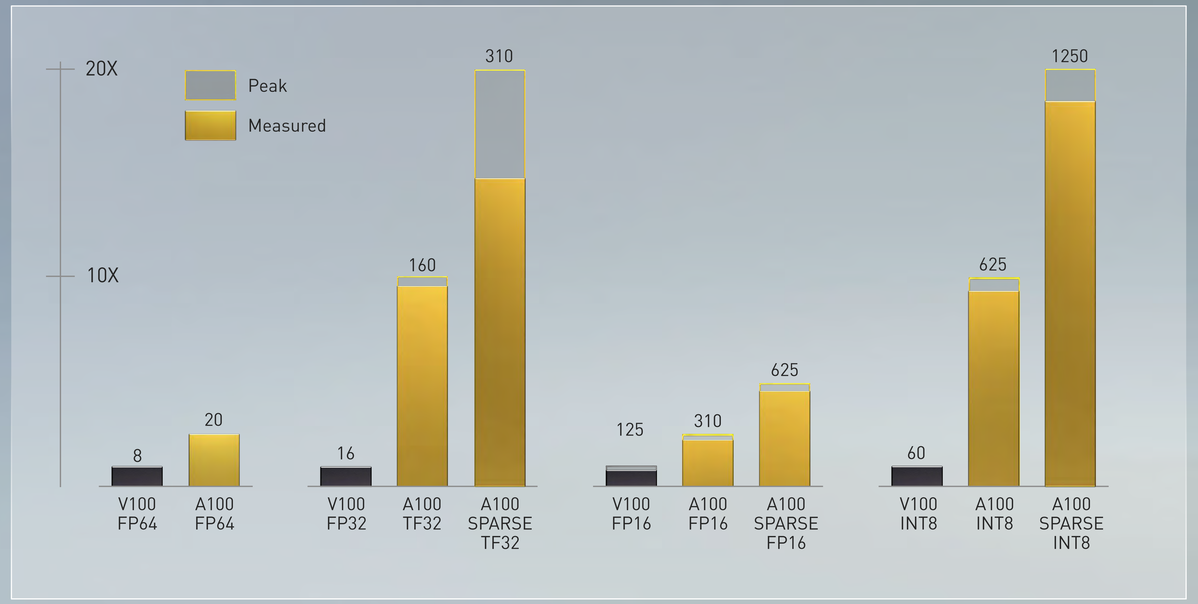


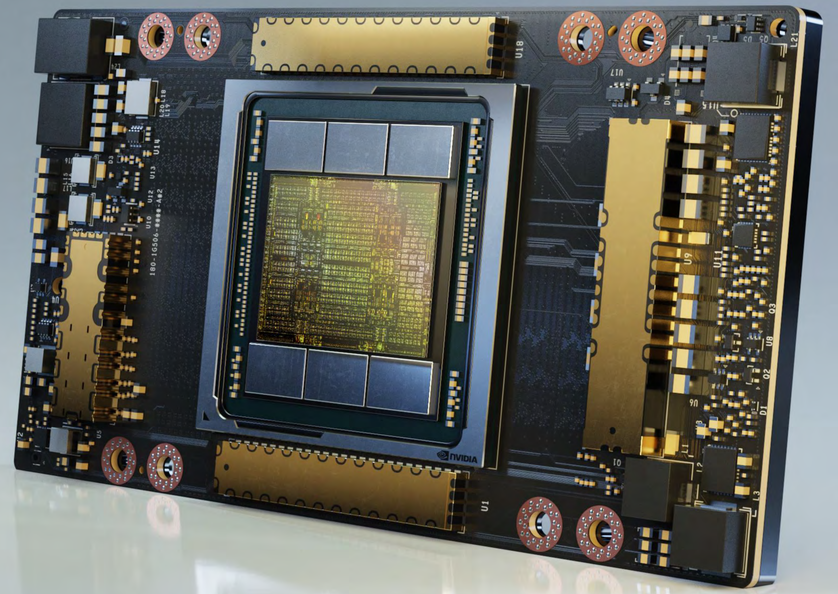
ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP





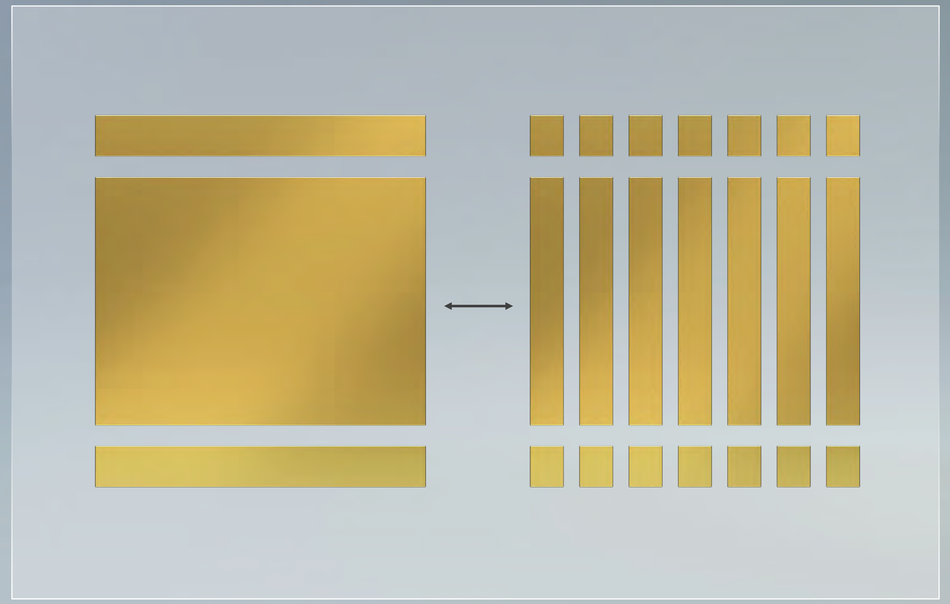
ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP

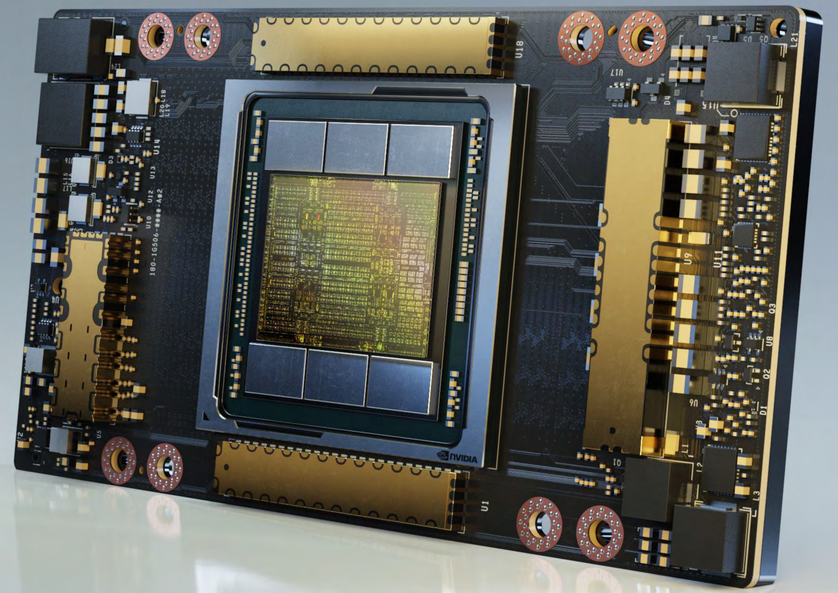




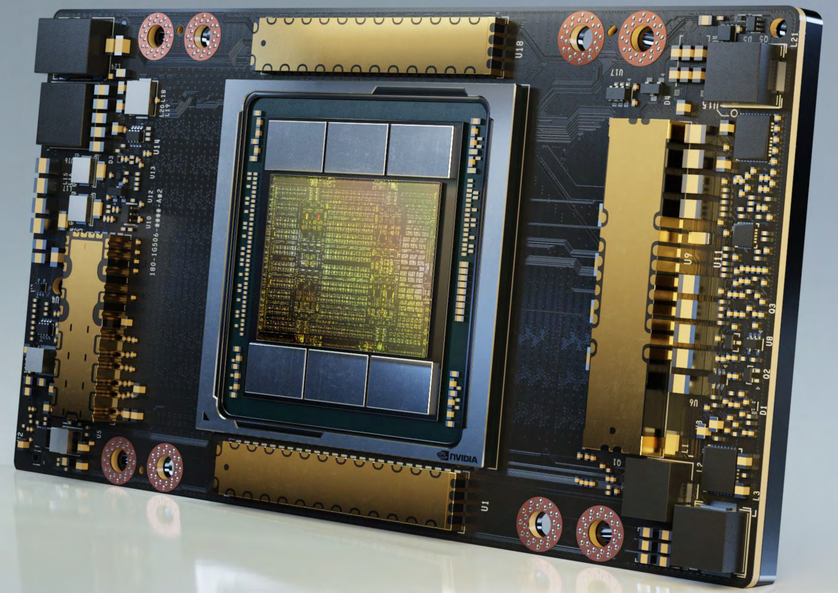
NEW MULTI-INSTANCE GPU FOR ELASTIC GPU COMPUTING

7x Higher Throughput of V100 with Simultaneous Instances per GPU

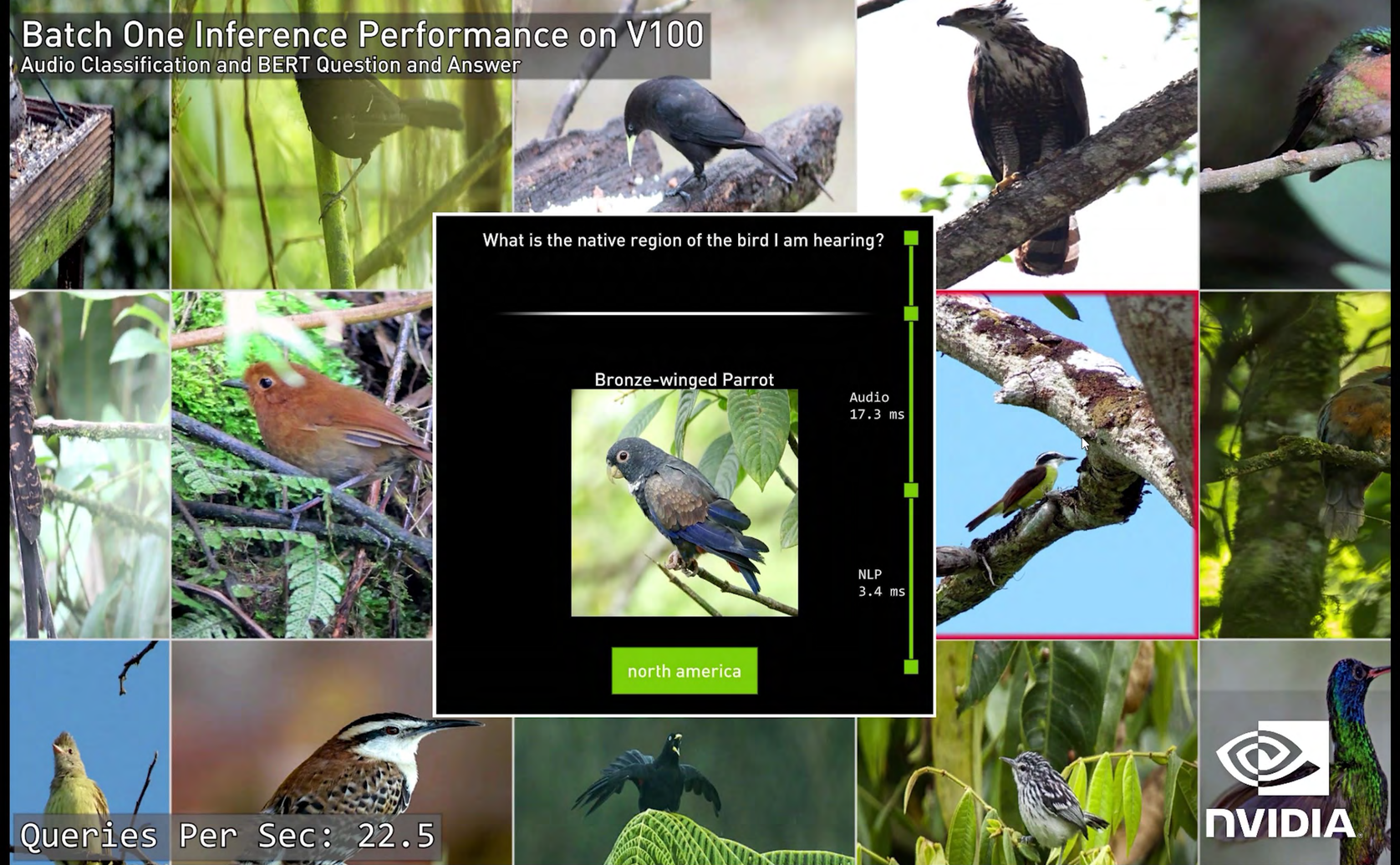




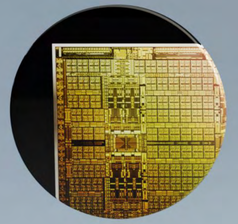
UNIFIED AI TRAINING AND INFERENCE ACCELERATION



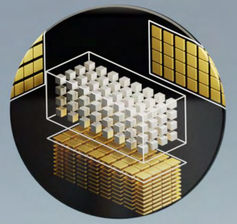
BERT Pre-Training Throughput using Pytorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 V100: DGX-1 Server with 8xV100 using FP32 precision A100: DGX A100 Server with 8xA100 using TF32 precision | BERT Large Inference | T4, V100: TRT 7.1, Precision = FP16, Batch Size =256 | A100 MIG: Pre-production TRT, Batch Size =94, Precision = INT8 with Sparsity


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

ANNOUNCING NVIDIA A100

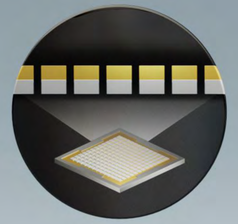


54 BILLION XTORS

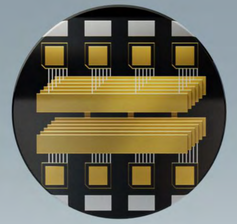


3RD GEN TENSOR CORES

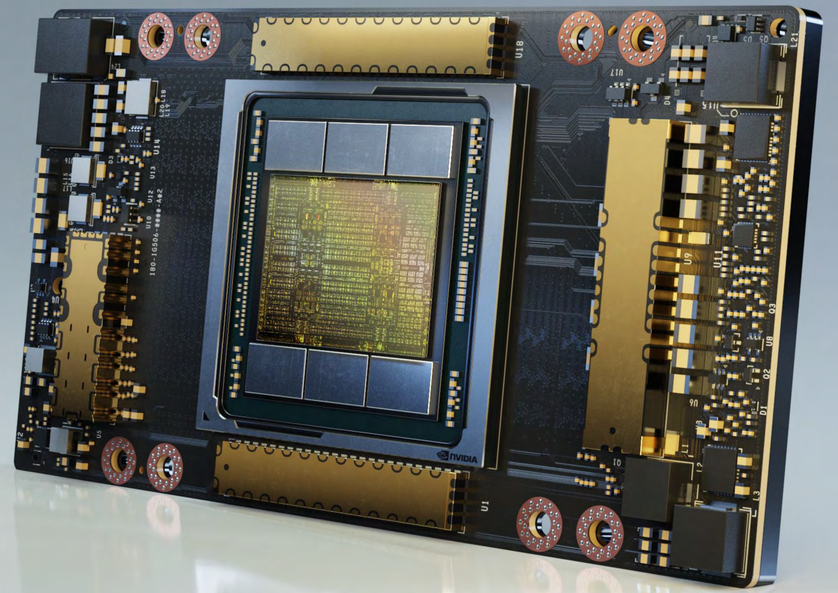
SPARSITY ACCELERATION



MIG



3RD GEN NVLINK & NVSWITCH



ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

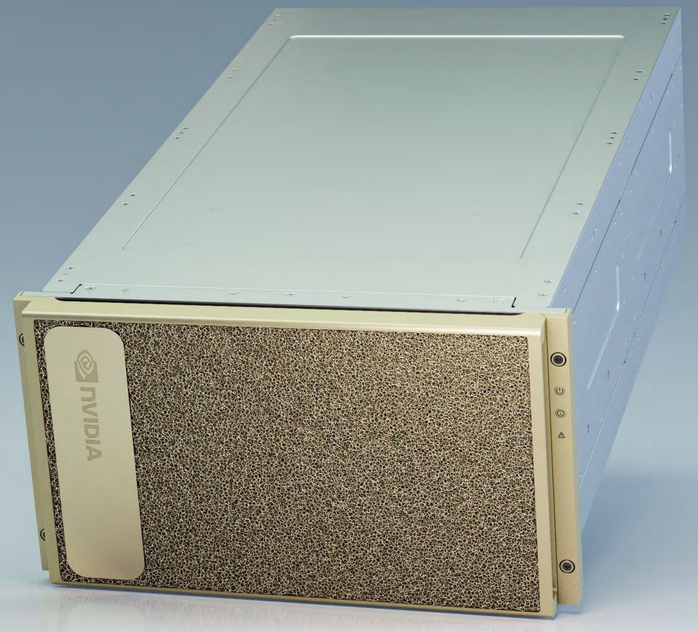
5 PetaFLOPS of Performance in a Single Node

Unified System for End-to-End Data Science and AI

Fully Accelerated Stacks — Spark 3.0, RAPIDS, TensorFlow, PyTorch, Triton

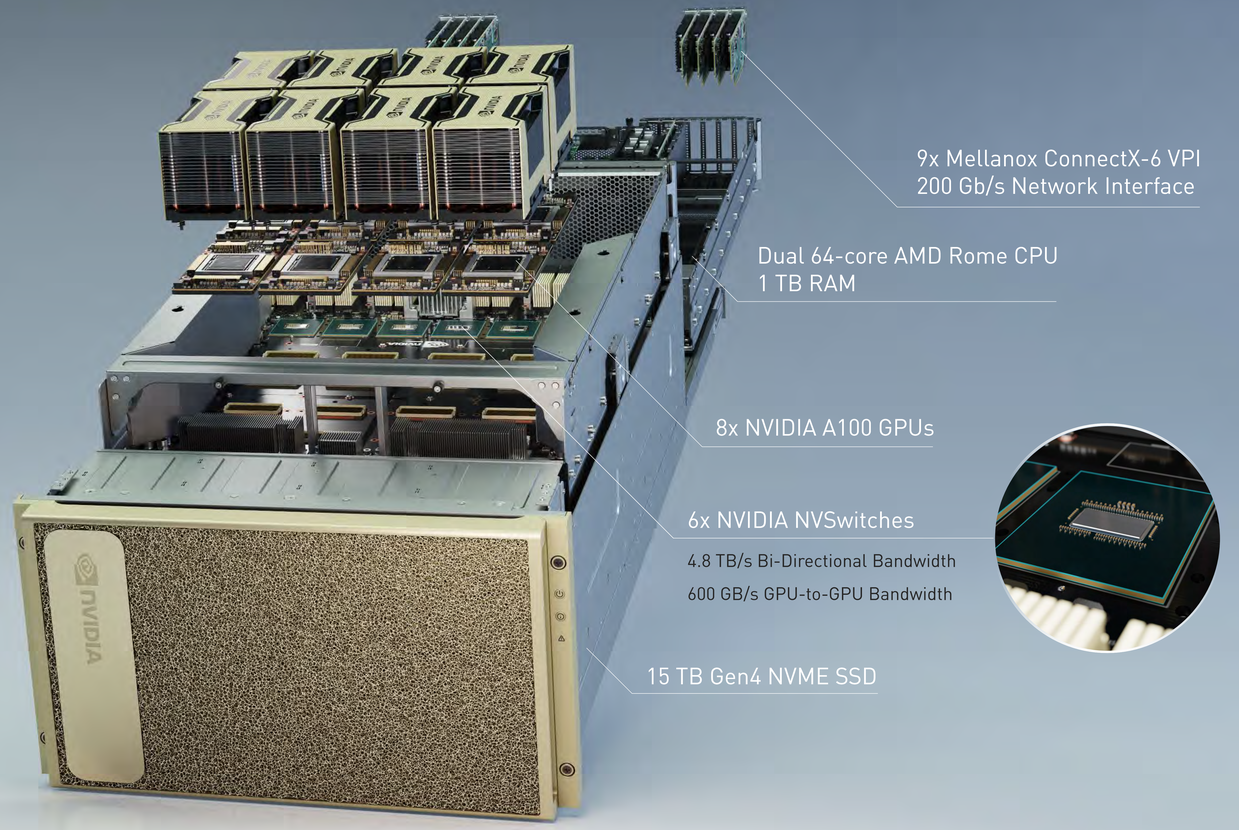
Elastic Scale-Up or Scale-Out Computing

High Scalability with Mellanox Networking



ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

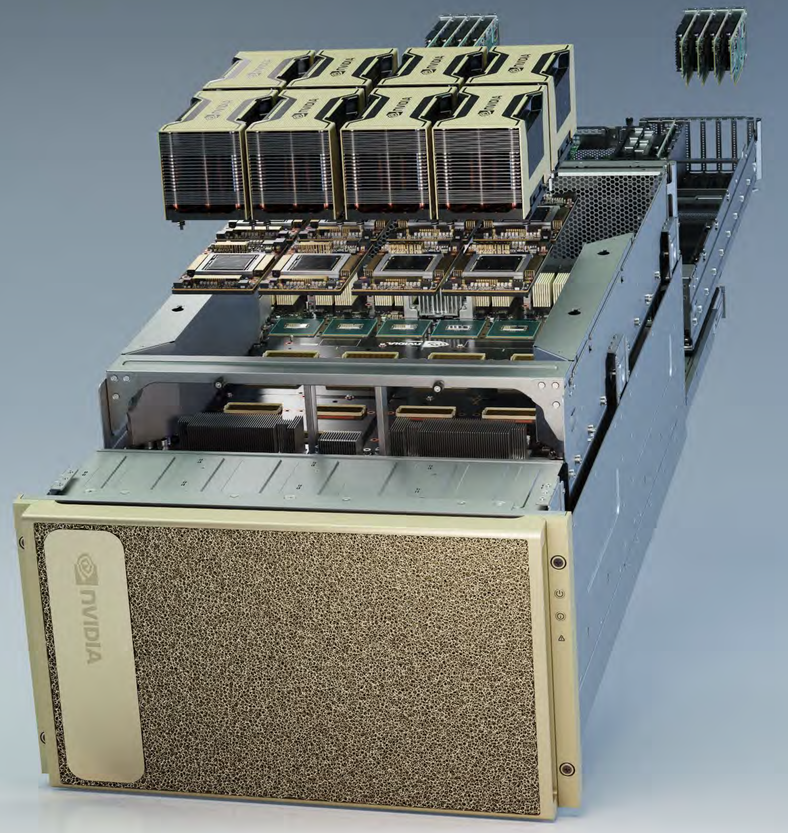


ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

INT8	10	PetaOPS Peak
FP16	5	PFLOPS Peak
TF32	2.5	PFLOPS Peak
FP64	156	TFLOPS Peak

TensorCore with Sparsity



ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

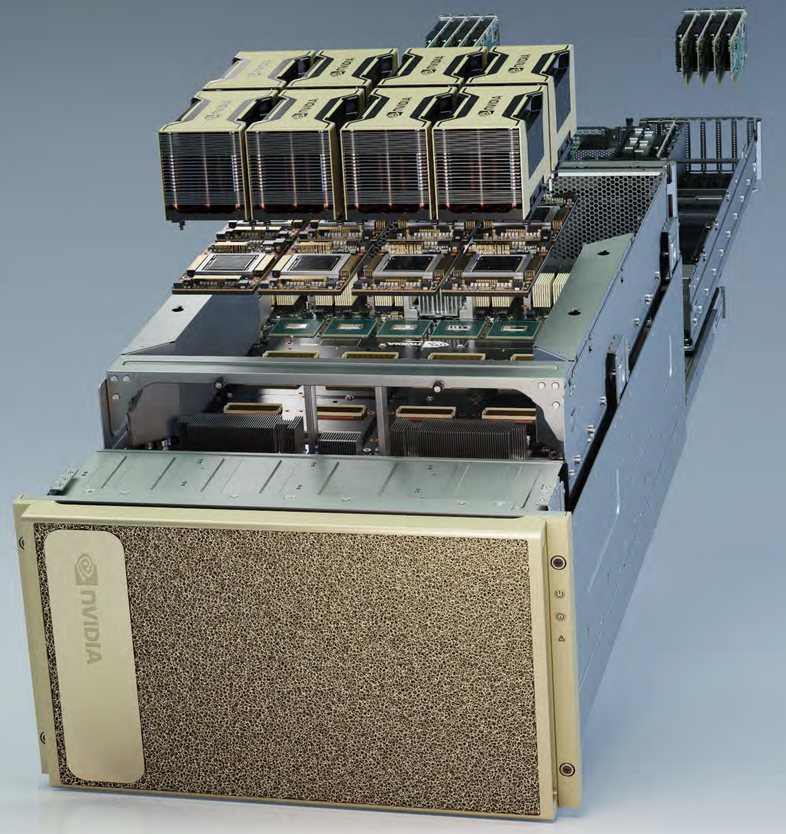
5 PetaFLOPS of Performance in a Single Node

150X AI Compute

40X Memory Bandwidth

40X IO Bandwidth

Compared to High-End CPU server



ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

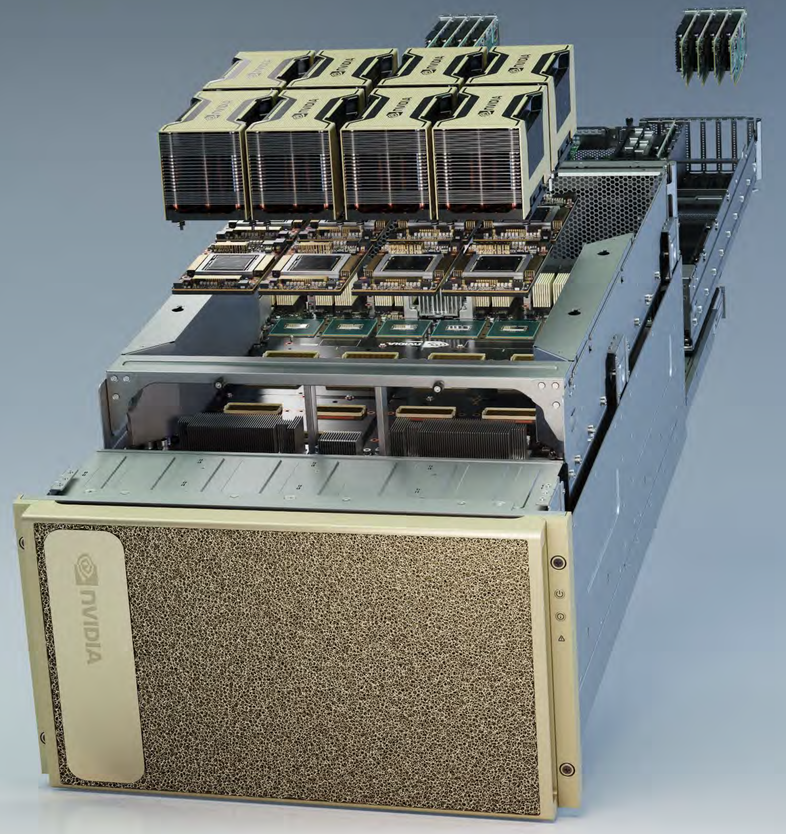
150X AI Compute

40X Memory Bandwidth

40X IO Bandwidth

Compared to High-End CPU server

Available Now at $199K



ANNOUNCING
NVIDIA A100 LIGHTHOUSE CUSTOMERS

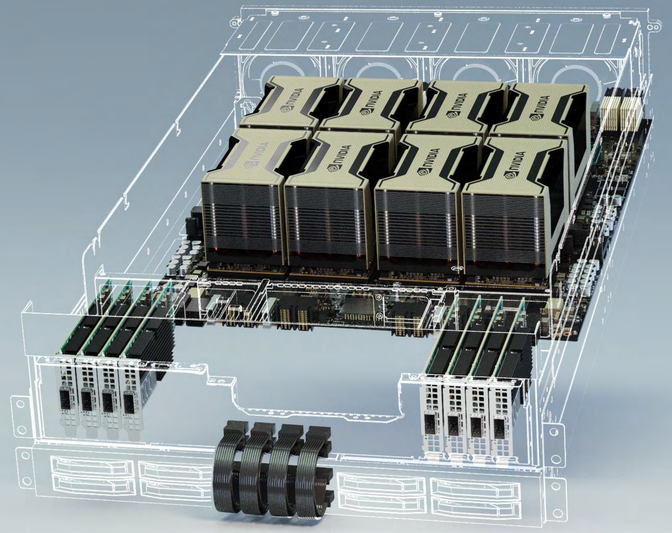
Elastic Data Center Accelerator Choice of Industry Leaders

CLOUD

Alibaba Cloud
aws
BAIDU AI CLOUD
Google Cloud
Microsoft Azure
ORACLE Cloud Infrastructure
Tencent Cloud

SYSTEMS

Atos
CRAY a Hewlett Packard Enterprise company
DELL Technologies
FUJITSU
GIGABYTE
Hewlett Packard Enterprise
inspur
Lenovo
QCT
SUPERMICRO



TODAY'S AI DATA CENTER

50 DGX-1 Systems for AI training

600 CPU Systems for AI Inference

$11M

25 Racks

630 kW

DGX A100 AI

5 DGX A100 Systems for AI Training and Inference

$1M

1 Rack

28 kW

TODAY'S AI DATA CENTER

50 DGX-1 Systems for AI training

600 CPU Systems for AI Inference

$11M

25 Racks

630 kW

DGX A100 AI

5 DGX A100 Systems for AI Training and Inference

$1M

1 Rack

28 kW

PAGERANK CASE STUDY

Common Crawl Data Set
2.6TB Graph – 128B Edge

3,000 CPU Servers – 105 Racks

PAGERANK CASE STUDY

Common Crawl Data Set
2.6TB Graph – 128B Edge

4 DGX A100 Connected via External NVLINK

PAGERANK CASE STUDY

Common Crawl Data Set
2.6TB Graph – 128B Edge

3,000 CPU Servers – 105 Racks

PAGERANK CASE STUDY

Common Crawl Data Set
2.6TB Graph – 128B Edge

4 DGX A100 Connected via External NVLINK

ANNOUNCING
NVIDIA DGX A100 SUPERPOD

140 DGX A100 Systems (1,120 A100)
170 Mellanox Quantum 200G InfiniBand Switches
280 Tb/s Network Fabric - 15km of Optical Cable
4 PB of All-Flash Networked Storage
700 PFLOPS of AI Performance
Built in under 3 Weeks

NVIDIA EXPANDS SATURNV

Before Expansion
1,800 DGX Systems
1.8 ExaFLOPS

Adding 4 DGX SuperPODs
560 DGX A100 = 2.8 ExaFLOPS

4.6 ExaFLOPS Total Capacity

ANNOUNCING
NVIDIA DGX A100
3RD GENERATION INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

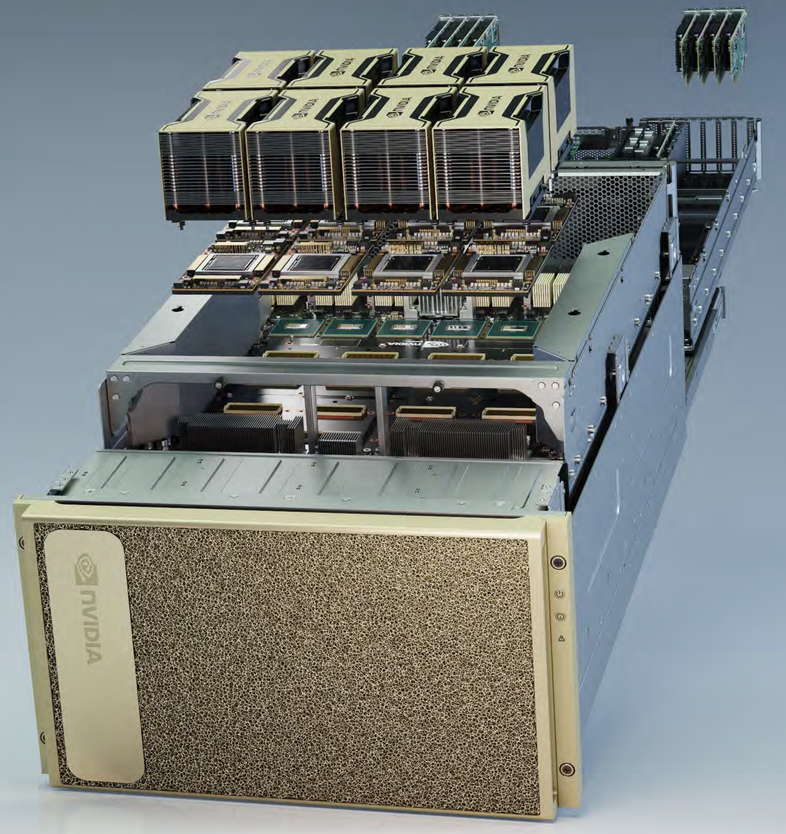
150X AI Compute

40X Memory Bandwidth

40X IO Bandwidth

Compared to High-End CPU server

Available Now at $199K



SMART EVERYTHING REVOLUTION

ALWAYS-ON | INSTANT SENSE-INFER-ACT | DISTANT | TRILLIONS

ANNOUNCING
NVIDIA EGX A100 WITH MELLANOX CX6 DX

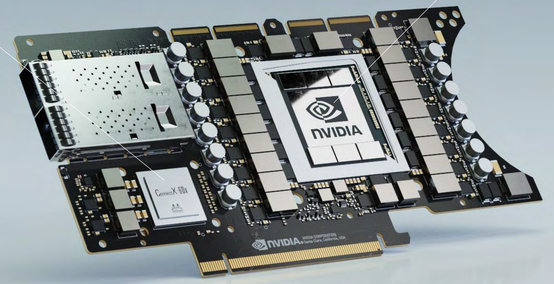
NVIDIA Mellanox ConnectX-6 DX

Dual 100 Gb/s Ethernet or InfiniBand

Line-speed TLS/IPSec Crypto Engine

Time Triggered Transmission Tech for Telco (5T for 5G)

ASAP^2 SR-IOV and VirtualIO Offload



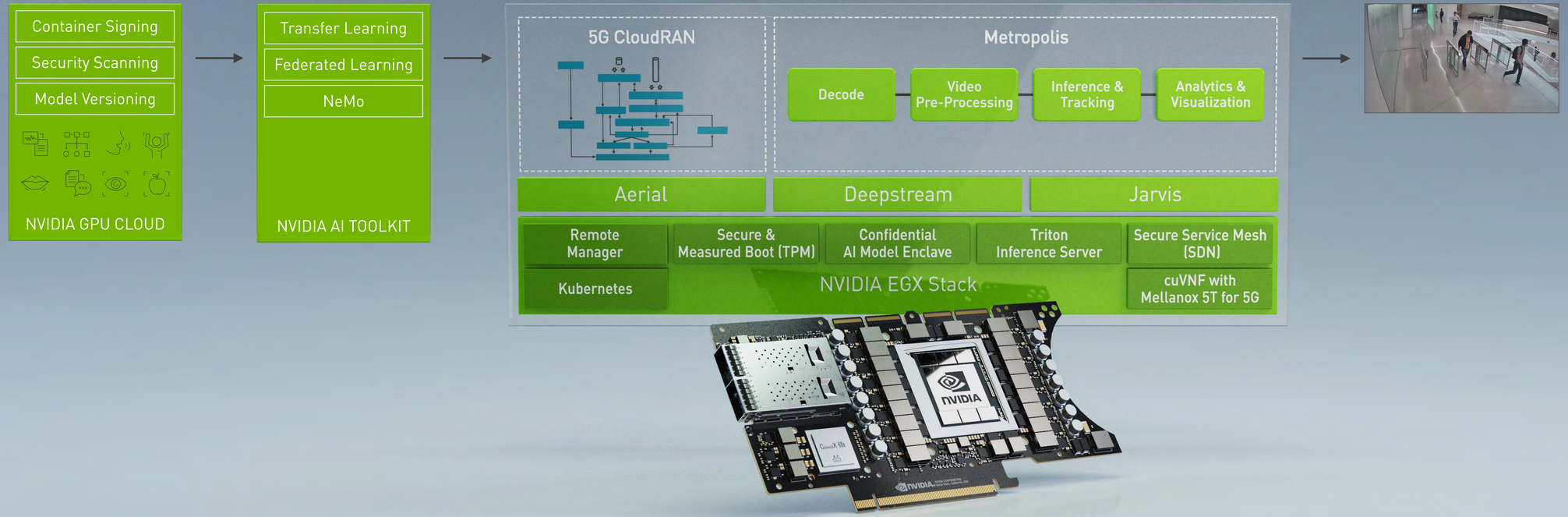
NVIDIA Ampere GPU

3rd generation Tensor Core

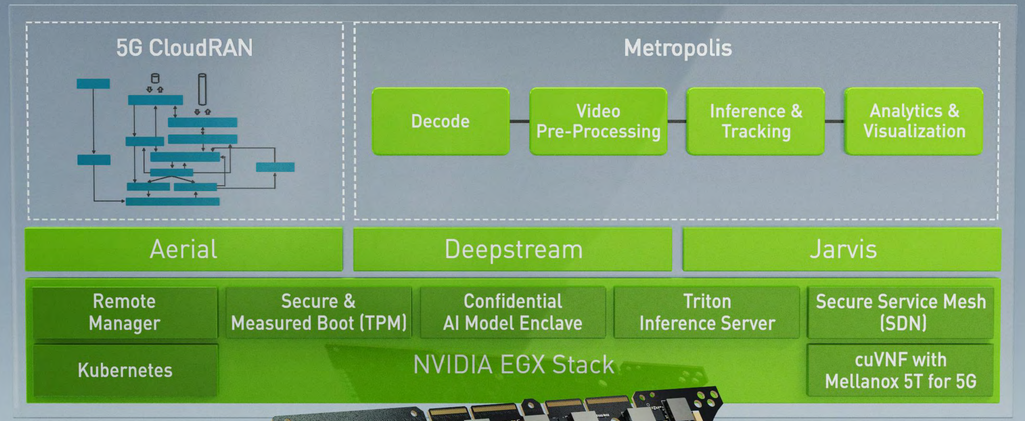
New Security Engine for Confidential AI

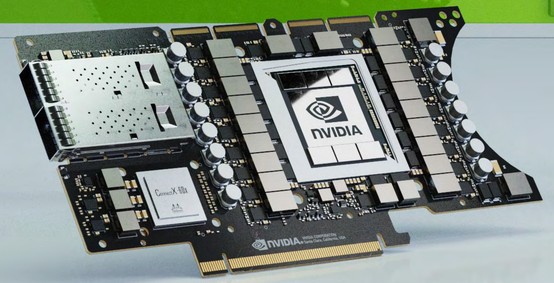
Secure, Authenticated Boot

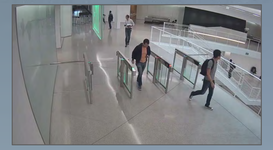
ANNOUNCING
NVIDIA EGX EDGE AI PLATFORM

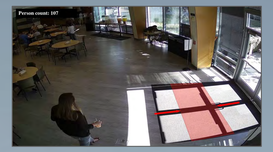


METROPOLIS VIDEO AI AND AERIAL 5G ON NVIDIA EGX



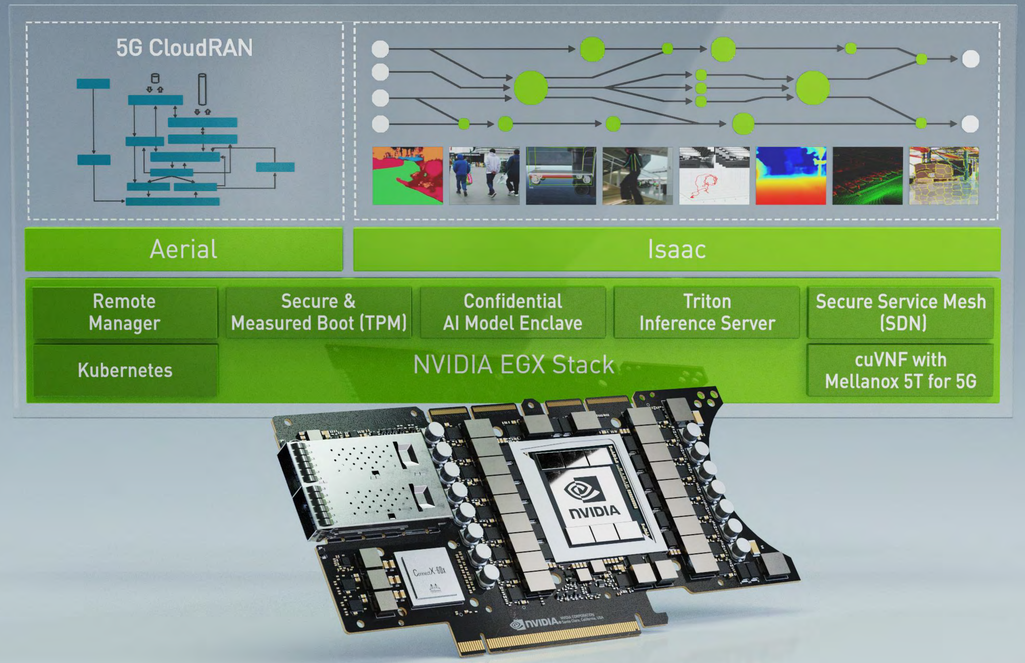






ISAAC ROBOTIC FACTORY AND AERIAL 5G ON NVIDIA EGX

Actual Factory



Virtual Factory
Digital Twin


KEEP CLEAR

ANNOUNCING
BMW SELECTS NVIDIA ISAAC ROBOTICS

“The Power of Choice”

Over 40 BMW models, 100 options/car

99% of orders are custom/unique

2^{100} different possible configurations

“Raw Parts In, Parts Trays Out”

30M parts come in every day

1.800 suppliers, to 31 factories

230k part numbers

“Just in Time, Just in Sequence”

Up to 10 cars per line

New car every 56 seconds

SplitBot

PickBot

PlaceBot

Smart Transport Robot (STR)

SortBot

NVIDIA EGX ECOSYSTEM

5G & CloudRAN

altran · ERICSSON · MAVENIR · Viasat

SECURITY & NETWORKING

CUMULUS · FORTINET · JUNIPER · nuagenetworks From Nokia

INFRASTRUCTURE

IBM · Red Hat · vmware · WIND

SYSTEMS

Atos · DELL Technologies · Hewlett Packard Enterprise · Lenovo

CLOUD

Baidu 百度 · Microsoft

CONVERSATIONAL AI

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

INTELLIGENT VIDEO ANALYTICS

AIFI · DEEPVISION · MALONG · IRONYUN · SAFR from RYAN

ROBOTICS

DENSO · FANUC · FOXCONN · KOMATSU

MEDICAL

CHOOCH · DawnLight · GE Healthcare · UCSF · WHITEBOARD COORDINATOR

INDUSTRIAL

Micron · MUSASHI · SAMSUNG · SEAGATE · TSMC

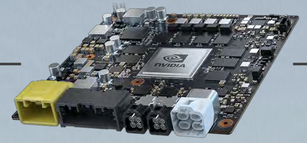
INDUSTRY LEADERS

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

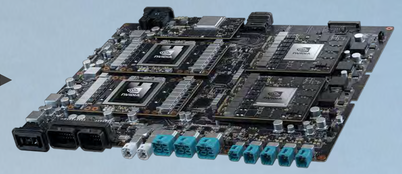
NVIDIA DRIVE WITH ORIN AND AMPERE
5W TO 2,000 TOPS — ONE PROGRAMMABLE ARCHITECTURE

ADAS
Windshield NCAP
10 TOPS, 5W

L2+
Autopilot
200 TOPS, 45W

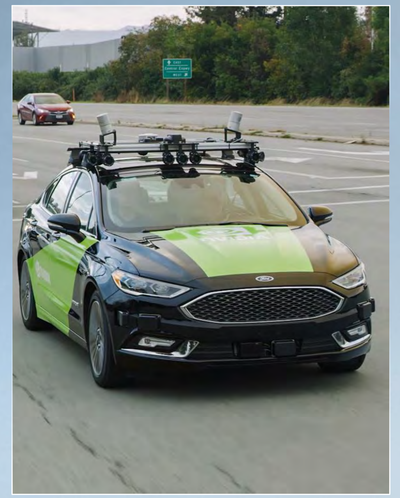


L5
Robotaxi
2,000 TOPS, 800W

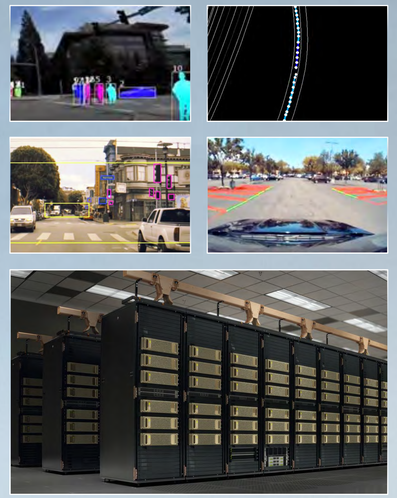


NVIDIA DRIVE — SOFTWARE-DEFINED AV PLATFORM

COLLECT DATA

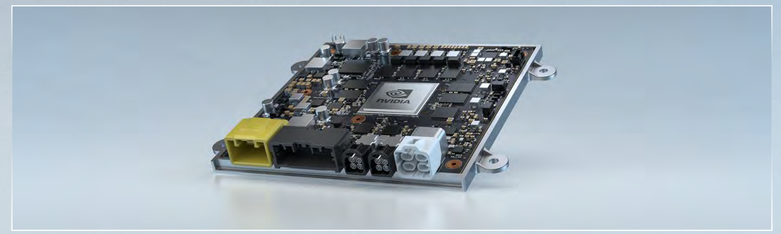


TRAIN MODELS



SIMULATE

DRIVE AV



DRIVE IX

DRIVE RC

NVIDIA DRIVE GLOBAL ECOSYSTEM

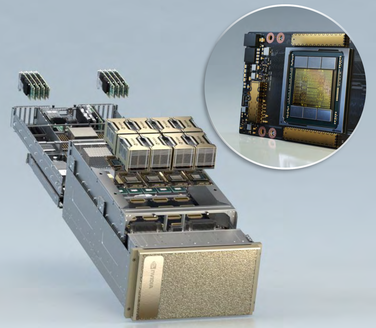
CARS	TRUCKS	SUPPLIERS	MOBILITY SERVICES	STARTUPS	SOFTWARE	MAPPING	SIMULATION
							
							
							
							

GTC 2020 ANNOUNCEMENTS

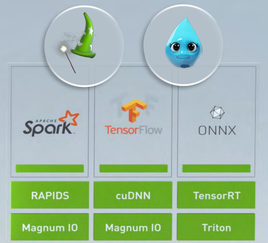
Mellanox TECHNOLOGIES

Data-Center-Scale Computing

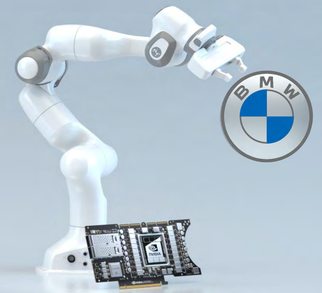
Omniverse RTX Server



A100 and DGX A100



NVIDIA AI



EGX and ISAAC


NVIDIA