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
25 YEARS OF ACCELERATED COMPUTING

X-FACTOR SPEED-UP

FULL STACK

SYSTEMS

ONE ARCHITECTURE

DEVELOPERS ++

PERFORMANCE ++

CUDA EVERYWHERE
DATA-CENTER-SCALE ACCELERATED COMPUTING

- X-FACTOR SPEED-UP
- FULL STACK
- DATA CENTER SCALE
- ONE ARCHITECTURE

DEVELOPERS ++
INSTALLED BASE ++
CUDA EVERYWHERE
PERFORMANCE ++
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

Ray Tracing

Deep Learning

540p

1080p

\[ \Delta w \]

Supercomputer Rendered - 16K Ground Truth
NVIDIA RTX
A NEW ERA IN COMPUTER GRAPHICS — RAY TRACING AND AI

Ray Tracing

Deep Learning

DRIVER NGX
“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
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
MACHINE LEARNING PIPELINE IS AN HPC CHALLENGE
MACHINE LEARNING DRIVING EXPONENTIAL GROWTH IN DATA

Data Processing Requirement

- Terabytes Data Set
- Gigabytes GPU HBM
- Megabytes CPU Cache

GPU

CPU

ETL

- TensorFlow
- PyTorch

TRAINING

- cuDNN
- Magnum I/O

INFEERENCE

- TensorRT
- Triton Inference Server

Users Embeddings

Candidates Generation

Q(1000)

RANKING

Recommendation

User Query

Data Lake 100's PB

0(10)
SPARK 3.0 BUILT ON STATE-OF-THE-ART FOUNDATION
RAPIDS SHATTERS ETL BENCHMARK

$1M  17 GB/s

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

$2M 163 GB/s

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

5X
PRICE-PERF

163 GB/s Throughput on RAPIDS Implementation of TPCx-BB @ SF 10K
$2M | 16 DGX-1 | 2 Racks | 56 kW

ETL

TRAINING

INFERENCE

RAPIDS
Magnum IO

TensorFlow
PyTorch

cuDNN
Magnum IO

TensorRT
Triton Inference Server

Data Lake
100's PB

Q(Terab)

Q(Petab)

User Query

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

$10M 140 kW

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

1/5th COST 1/3rd POWER

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
RECOMMENDER SYSTEM IS THE ENGINE OF THE PERSONALIZED INTERNET
ANNOUNCING
NVIDIA MERLIN — DEEP RECOMMENDER APPLICATION FRAMEWORK

1TB Ads Dataset

<table>
<thead>
<tr>
<th>CPUs</th>
<th>ETL 2 HR</th>
<th>TRAINING 1 1/2 Days</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>GPUs</th>
<th>ETL 3 Min</th>
<th>TRAINING 16 Min</th>
</tr>
</thead>
</table>

NVIDIA Merlin

- NVTabular
- HugeCTR
- RAPIDS
- cuDNN
- Magnum IO

TensorRT

Triton Inference Server

DATA LAKE 100's PB

INPUT

USER QUERY

OUTPUT

RANKING

OBJECTIVE

USER EMBEDDINGS

CANDIDATE GENERATION

O(1000)

Recommendation

O(10)
NVIDIA AI INFERENC E
15X TENSORRT DOWN LOADS IN 2 YEARS

TENSORRT 7

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

15X DOW NDLOADS

<table>
<thead>
<tr>
<th>Year</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>50K</td>
</tr>
<tr>
<td>2019</td>
<td>450K</td>
</tr>
</tbody>
</table>

10X DEVELOPERS

<table>
<thead>
<tr>
<th>Year</th>
<th>Developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>5K</td>
</tr>
<tr>
<td>2019</td>
<td>47K</td>
</tr>
</tbody>
</table>

Research / Higher Ed
Automotive
Internet / Telecom
Hardware / Semiconductor
Financial Services
Energy / Oil & Gas

Software
IT Services
Other
Healthcare & Life Sciences
Manufacturing
Public Sector
Consulting Services
ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK

Jarvis

ASR  NLU  TTS  A2F

Omniverse
JARVIS DEMO
MISTY — INTERACTIVE 3D CHATBOT

ASR  NLU  TTS  A2F

Jarvis

Omniverse
ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK
ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK

PRE-TRAINED MODEL
- Transfer Learning
- NaMo
- Service Maker

NVIDIA GPU CLOUD

RE-TRAIN
- NVIDIA AI TOOLKIT

NVIDIA JARVIS
- Dialog Manager
- Multi-Speaker Transcription
- Chatbot
- Gesture Recognition
- Look to Talk

TRITON INFERENCER SERVER
- Speech
  - Language Model
  - Decoder
  - Acoustic Model
  - Feature Extraction
- NLU & Recommenders
- Speech Synthesis
- Voice Encoder

JESSICA: What will you have ready for Wednesday?
TODD: We have only early designs of the packaging.
JESSICA: Great.

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
MODERN CLOUD DATA CENTER
Diverse Applications | Scale-Up & Scale-Out Workloads | Insatiable Demand
ANNOUNCING NVIDIA A100
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

**FP32**
- 8 BITS
- 23 BITS

**TENSOR FLOAT32**
- 8 BITS
- 10 BITS

**FP16**
- 5 BITS
- 10 BITS
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

NVIDIA V100 FP32

NVIDIA A100 Tensor Core TF32
NEW TENSOR CORE ACCELERATION FOR SPARSITY

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

Dense Matrix → Sparse Matrix → A100 Sparsity Optimized Tensor Core

2X Effective
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
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
What is the native region of the bird I am hearing?

Bronze-winged Parrot

Audio
17.3 ms

NLP
3.4 ms

north america
ANNOUNCING NVIDIA A100
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

- Dual 64-core AMD Rome CPU
- 1 TB RAM
- 8x NVIDIA A100 GPUs
- 6x NVIDIA NVSwitches
- 4.8 TB/s Bi-Directional Bandwidth
- 600 GB/s GPU-to-GPU Bandwidth
- 15 TB Gen4 NVMe SSD
- 9x Mellanox ConnectX-6 VPI 200 Gb/s Network Interface
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
TODAY’S AI DATA CENTER

50 DGX-1 Systems for AI training
600 CPU Systems for AI Inference
$11M
25 Racks
630 kW

$11M  630 kW
DGX A100 AI

5 DGX A100 Systems for AI Training and Inference
$1M
1 Rack
28 kW

$1M  28 kW
1/10th COST
1/20th POWER
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

52 Billion Edges / Sec
PAGERANK CASE STUDY

Common Crawl Data Set
2.6TB Graph – 128B Edge
4 DGX A100 Connected via External NVLINK

688 Billion Edges / Sec
13X PERFORMANCE
1/75th COST
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
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 (ST for 5G)
- ASAP² 5R-I0V and Virtual I/O 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
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
NVIDIA DRIVE WITH ORIN AND AMPERE
5W TO 2,000 TOPS — ONE PROGRAMMABLE ARCHITECTURE

- ADAS Windshield NCAP
  10 TOPS, 5W

- L3+
  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
GTC 2020 ANNOUNCEMENTS

- Data-Center-Scale Computing
- Omniverse RTX Server
- A100 and DGX A100
- NVIDIA AI
- EGX and ISAAC