ACCELERATED COMPUTING
1000X EVERY 10 YEARS
ONE ARCHITECTURE

ACCELERATION STACKS

GAMING | HPC | TRANSPORTATION | HEALTHCARE

PRO VIZ | AI | ROBOTICS | AI IOT
COMPUTERS WRITING SOFTWARE
RISE OF GPU COMPUTING

GTC Attendees — 7X in 5 Yrs
- 2013: 25K
- 2018: 200K

CUDA Downloads — 5X in 5 Yrs
- 2013: 20M
- 2018: 100M

Applications:
- MEDICAL IMAGING
- BIOINFORMATICS
- COMPUTATIONAL FLUID DYNAMICS
- NUMERICAL ANALYTICS
- DEEP LEARNING
- MATERIALS
- RAY TRACING
- WEATHER AND CLIMATE
THE HOLY GRAIL OF COMPUTER GRAPHICS

Turner Whitted | 1979
“Multi-bounce Recursive Ray Tracing”
1.2 Hours for 512x512 on VAX 11/780
## NEW QUADRO RTX
**WORLD’S FIRST RAY TRACING GPU**

<table>
<thead>
<tr>
<th>QUADRO RTX 5000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16 GB / 32 GB</td>
<td>6 GIGA RAYS</td>
</tr>
<tr>
<td>$2,300</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QUADRO RTX 6000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24 GB / 48 GB</td>
<td>10 GIGA RAYS</td>
</tr>
<tr>
<td>$6,300</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QUADRO RTX 8000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>48 GB / 96 GB</td>
<td>10 GIGA RAYS</td>
</tr>
<tr>
<td>$10,000</td>
<td></td>
</tr>
</tbody>
</table>
RTX OPENS $250B VISUAL EFFECTS INDUSTRY

RTX SERVER — 60X FASTER THAN CPU NODE

RTX ACCELERATED RENDERERS

PHOTOREAL VFX

AEC
DESIGN
VISUALIZATION
FILM & TELEVISION
NEW GEFORCE RTX

GRAPHICS REINVENTED

GEFORCE RTX 2070
8 TFLOPS | 8 TIOPS | 63 Tensor TFLOPS | 6 GIGA RAYS
FROM $499

GEFORCE RTX 2080
11 TFLOPS | 11 TIOPS | 85 Tensor TFLOPS | 8 GIGA RAYS
FROM $699

GEFORCE RTX 2080 Ti
14 TFLOPS | 14 TIOPS | 114 Tensor TFLOPS | 10 GIGA RAYS
FROM $999
RTX RESETS $100B GAMING INDUSTRY

FASTEST GAMING GPU | PLAYS AT 4K 60FPS

DEEP LEARNING IMAGING WITH 114 TFLOPS TENSOR CORE

RAY TRACING: “THE HOLY GRAIL OF GRAPHICS”
NEW NVIDIA DGX-2
THE LARGEST GPU EVER CREATED

2 PFLOPS
512GB HBM2
16 TB/sec Memory Bandwidth
10 kW | 160 kg
WORLD’S LEADING AI PLATFORM

**IMAGES**

- **FASTEST SINGLE CHIP**
  - 24 hours

- **FASTEST SINGLE NODE**
  - 108 minutes

- **FASTEST AT SCALE**
  - 6.6 minutes

**TRANSLATION**

- **FASTEST SINGLE NODE**
  - 5 hours

- **FASTEST AT SCALE**
  - 32 minutes

Images: Single Chip: Resnet-50 Training on Tesla V100 with NVIDIA NGC MXNet Container 18.09 Pre-release by NVIDIA. Single Node: Resnet-50 V1 Training on NVIDIA DGX-2 with NVIDIA NGC MXNet Container 18.09 by NVIDIA. At scale: Resnet-50 Training with 2048 P40 GPUs by Tencent.

Translation: Single Node Strong NMT (Transformer) Training on WMT ‘14 English-German Translation with DGX-1 by Facebook Research. At Scale Strong NMT (Transformer) Training on WMT ‘14 English-German Translation with 16 DGX-1 Systems by Facebook Research.
NVIDIA PARTNERS WITH JAPAN LEADERS IN AI & HPC

AIST - ABCI
Japan’s Fastest AI Supercomputer

FujiFilm
AI Supercomputer, Accelerating Medical Imaging

Weathernews
Breakthrough Virtual Radar with dAIagnosis

NTT
Accelerating COREVO AI

PFN
AI Supercomputer
NVIDIA PARTNERS WITH JAPAN LEADERS IN AI & HPC

- **AIST - ABCI**: Japan's Fastest AI Supercomputer
- **FujiFilm**: AI Supercomputer, Accelerating Medical Imaging
- **Weathernews**: Breakthrough Virtual Radar with dAIgnosis
- **NTT**: Accelerating COREVO AI
- **PFN**: AI Supercomputer
ANNOUNCING TESLA T4

Tesla T4
Universal Inference Acceleration

Multi-Precision Tensor Core

Giant Leap
12X Pascal FP Inference

TensorRT 5.0
Support for Tensor Core

21X ASR
Deep Speech 2

36X NLP
GNMT

20X RECOM
Deep Recognition

8X TTS
WaveNet

18X TTS
Deep Neural WaveNet

36X NLP
Abstractive MT

27X TTS
Recurrent and Fast

12X Pascal FP Inference

5.5

65 TF FP16
130 TOPS INT8
260 TOPS INT4

22

260

5.5

130

260

75W

FP16

INT8

INT4

FLOAT

INT8

INT4
ANNOUNCING TESLA T4

Tesla P4 and TensorRT Adoption

World's Leading Systems Makers

World's First 1 PetaFLOPS Inference Machine
TRADITIONAL
HYPERSCALE
DATACENTER

200 CPU Servers
Speech | NLP | Video
60 kW
GPU-ACCELERATED HYPERSCALE INFERENCE SERVER

1 Server with 16 Tesla T4 GPUs
Speech | NLP | Video
2 kW

5 Racks in a Box
ANNOUNCING NVIDIA TENSORRT HYPERSCALE

Kubernetes and Docker on NVIDIA GPUs
New Inference Serving Engine
Multiple Model Types and Frameworks Concurrently
Maximize Datacenter Throughput and Utilization
ANNOUNCING NVIDIA TENSORRT HYPERSONELE

DNN Models

NV DL SDK
NV Docker

TensorRT Inference Server

Kubernetes
WORLD’S LEADING AI PLATFORM

 IMAGES

 FASTEST INFERENCE
 1 millisecond

 HIGHEST INFERENCE THROUGHPUT
 6,250 images/second

 HIGHEST INFERENCE EFFICIENCY
 56 images/second/watt

 TRANSLATION

 HIGHEST INFERENCE THROUGHPUT
 13,160 words/second

NVIDIA TENSORRT HYPERSCALE

Tesla T4

TensorRT 5

TensorRT Inference Server

- 21X ASR (Deep Speech 2)
- 36X NLP (GNMT)
- 20X RECOM (Deep Recom)
- 8X TTS (WaveNet)
- 27X VIDEO/IMAGE (ResNet-50)

Deep Speech 2

Deep Recom

WaveNet

ResNet-50
XAVIER
WORLD'S FIRST AUTONOMOUS MACHINE PROCESSOR

Most Complex SOC Ever Made
9 Billion Transistors, 350mm², 12nFFN
-8,000 Engineering Years
ANNOUNCING NVIDIA AGX EMBEDDED AI HPC

High-speed SerDes — 109 Gbps + 320 Gbps I/O
Up to 320 TOPS Tensor Ops
Up to 25 TFLOPS FP32
Up to 16 GiGA Rays
Starting from 15W
NVIDIA DRIVE

SENSOR PROCESSING
MAPPING & LOCALIZATION
PATH & TASK PLANNING
PERCEPTION
SITUATION UNDERSTANDING
DIVERSITY & REDUNDANCY
NVIDIA PARTNERS WITH JAPAN LEADERS IN AV

TOYOTA
Production Cars

TIER IV
Last Mile Delivery

ZMP
Robotaxis

ISUZU TRUCKS
Autonomous Trucks
NVIDIA DRIVE PLATFORM ADOPTION ACROSS TRANSPORTATION

CARS
- TOYOTA
- MERCEDES-BENZ
- HONDA

MOBILITY SERVICES
- UBER
- TIER 4
- ZENITY

TRUCKS
- ISUZU
- DAF
- BOSCH

TIER ONES
- DENSO
- ZENUN
- ORION

MAPPING
- HONDA
- NUVIS
- NAVTEQ

SENSORS
- SONY
- Intel
- Pioneer
ANNOUNCING
DRIVE AGX XAVIER DEVKIT
SCALABLE AV COMPUTING PLATFORM

Architected for Safety: 30 TOPS to 320 TOPS
Runs DRIVE Software 1.0
Full Support for CUDA and TensorRT
OTA Ready

Available October 1, 2018
NVIDIA ISAAC GEMS

- Global Localization
- LQR Path Planner
- Depth Estimation
- Human Pose Estimation
- Object / People Detection
- Map Editor
- Visual Odometry
- Physical Simulation
- Gesture Recognition
- ASR
ANNOUNCING
YAMAHA MOTOR ADOPTS
JETSON AGX FOR
AUTONOMOUS MACHINES
NVIDIA PARTNERS WITH JAPAN LEADERS IN ROBOTICS & AI IOT
$100B IMAGING INDUSTRY GOING TO AI HPC

- High-speed I/O
- Sensor Processing
- 10-10,000 GB/sec
- Image Recon
- 5-50 TFLOPS
- 1-12 GPUs per machine
- 50-1,800W
- Visualization

- Imaging modalities:
  - High-speed I/O
  - Sensor Processing
  - 10-10,000 GB/sec
  - Image Recon
  - 5-50 TFLOPS
  - 1-12 GPUs per machine
  - 50-1,800W
  - Visualization

- Applications:
  - Sequencers
  - Digital Pathology
  - Cryo-EM
  - Liquid Biopsy
  - Robotic Surgery
  - Ultrasound
  - Endoscopy
  - Mammography
  - 3D: CT, MRI, PET
  - Radiation Therapy
  - Ultrasound
  - Endoscopy
  - Mammography
  - 3D: CT, MRI, PET
  - Radiation Therapy
$100B IMAGING INDUSTRY GOING TO AI HPC

High-speed I/O
Sensor Processing

Image Recon

Visualization

CPU

FPGA

GPU

ULTRASOUND
ENDOSCOPY
MAMMOGRAPHY
3D: CT, MRI, PET
RADIATION THERAPY

SEQUENCERS
DIGITAL PATHOLOGY
CRYO-EM
LIQUID BIOPSY
ROBOTIC SURGERY

ULTRASOUND
ENDOSCOPY
MAMMOGRAPHY
3D: CT, MRI, PET
RADIATION THERAPY

SEQUENCERS
DIGITAL PATHOLOGY
CRYO-EM
LIQUID BIOPSY
ROBOTIC SURGERY

$100B IMAGING INDUSTRY GOING TO AI HPC
ANNOUNCING CLARA AGX

High-speed I/O Sensor Processing → Image Recon → AI → Visualization

ULTRASOUND | ENDOSCOPY | MAMMOGRAPHY | 3D: CT, MRI, PET | RADIATION THERAPY

SEQUENCERS | DIGITAL PATHOLOGY | CRYO-EA | LIQUID BIOPSY | ROBOTIC SURGERY

Single Chip Medical Instrument
SerDes, Imaging, CV, AI, Visualization on One Chip
30 TOPS DL Processing
ANNOUNCING CLARA AGX

HIGH-SPEED I/O SENSOR PROCESSING → IMAGE RECON → AI → VISUALIZATION

Scale up to 200 TOPS DL Processing
8 GIGA Rays
200W
ANNOUNCING JETSON AGX XAVIER DEVKIT
WORLD’S FIRST EDGE AI COMPUTER

Jetpack Acceleration Lib SDK
Isaac Robotics SDK
Order Today at developer.nvidia.com/buy-jetson
ANNOUNCING NEW NVIDIA PLATFORMS

- DRIVE AGX | DRIVE SDK
- JETSON AGX | ISAAC SDK
- TESLA T4 | TENSORRT HYERSCALE
- QUADRO RTX
- GeForce RTX
- NVIDIA AGX
- QUADRO RTX
- GeForce RTX
- Tesla T4
- NVIDIA AGX
- Jetson AGX
- Isaac SDK