ACCELERATED COMPUTING
1000X EVERY 10 YEARS
COMPUTERS
WRITING SOFTWARE
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

QUADRO RTX 5000
16 GB / 32 GB | 6 GIGA RAYS
$2,300

QUADRO RTX 6000
24 GB / 48 GB | 10 GIGA RAYS
$6,300

QUADRO RTX 8000
48 GB / 96 GB | 10 GIGA RAYS
$10,000
RTX OPENS $250B VISUAL EFFECTS INDUSTRY

RTX SERVER — 60X FASTER THAN CPU NODE

RTX ACCELERATED RENDERERS

PHOTOREAL VFX

DESIGN

AEC

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 V1 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. AtScale 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 dAIgnosis

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

Support for Tensor Core

- FP16
- INT8
- INT4

PASCAL
- TF
- FP16
- 130 TOPS INT8
- 260 TOPS INT4
- 65 TF FP16
- 75W

TURING
- FP16
- INT8
- INT4
- 22 TOPS FP16
- 45 TOPS INT8
- 130 TOPS INT4
- 260 TOPS INT4

WaveNet

Deep Recom

Deep Speech 2

ASR

NLP

RECOM

TTS

RECOM

Turing

Turing

Deep Recom

ResNet-50

12X Pascal FP Inference

65 TF FP16

130 TOPS INT8

260 TOPS INT4

0
50
100
150
200
250
300

TOPS INT4

TOPS INT4

TOPS INT4

TOPS INT4

75W

22

45

130

260
ANNOUNCING TESLA T4

World's First 1 PetaFLOPS Inference Machine

World's Leading Systems Makers

Tesla P4 and TensorRT Adoption

QuantaGRID 16-T4 Server
TRADITIONAL HYPERSONE DATA CENTER

200 CPU Servers
Speech | NLP | Video
60 kW
GPU-ACCELERATED HYPERSONE SCALE INFEERENCE SERVER

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

5 Racks in a Box
ANNOUNCING NVIDIA TENSORRT HYPERSONE

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

- NV DL SDK
- NV Docker
- DNN Models
- TensorRT Inference Server
- Kubernetes
WORLD’S LEADING AI PLATFORM


- **Translation**: GMNT Inference on Newstest2015 test dataset on Tesla V100.
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

TTS WaveNet

Deep Speech 2

ASR

RECOM Deep Recom

GNMT

VIDEO/IMAGE

ResNet-50

NLP
XAVIER
WORLD'S FIRST AUTONOMOUS MACHINE PROCESSOR

Most Complex SOC Ever Made
9 Billion Transistors, 350mm², 12nFFN
~8,000 Engineering Years
ANNNOUNCING
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
- Perception
- Diversity & Redundancy
- Path & Task Planning
- Situation Understanding
NVIDIA DRIVE

TRAINING

SIMULATING

DRIVING

- Cars
- Pedestrians
- Lanes
- Path
- Signs
- Lights
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
- UBER
- @Tier IV
- ZENITY
- SUBARU
- NVIDIA

MOBILITY SERVICES
- ISUZU

TRUCKS
- DENSO
- ZENNI

TIER ONES
- @Galactica S

MAPPING
- SONY
- ORION
- Pioneer

SENSORS
- NVIDIA
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

SENSOR PROCESSING

MAPPING & LOCALIZATION

PATH & TASK PLANNING

PERCEPTION

SITUATION UNDERSTANDING

DIVERSITY & REDUNDANCY
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

KOMATSU  
#1 Construction

DENSO  
#1 Auto Parts

YAMAHA MOTORS  
#1 Mobility Machines

FANUC  
#1 FA Robotics

KAWADA TECHNOLOGIES  
Collaborative Robots

MUSASHI  
Factory Automation

CANON  
Factory Automation

PANASONIC  
Smart City
$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

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

High-speed I/O
Sensor Processing
Image Recon
Visualization

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

SEQUENCERS  DIGITAL PATHOLOGY  CRYO-EM  LIQUID BIOPSY  ROBOTIC SURGERY

CPU
FPGA
GPU
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-EM | LIQUID BIOPSY | ROBOTIC SURGERY

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

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

SEQUENCERS  DIGITAL PATHOLOGY  CRYO-EA  LIQUID BIOPSY  ROBOTIC SURGERY

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 HYPERSCALE
- NVIDIA AGX
- QUADRO RTX
- GEFORCE RTX