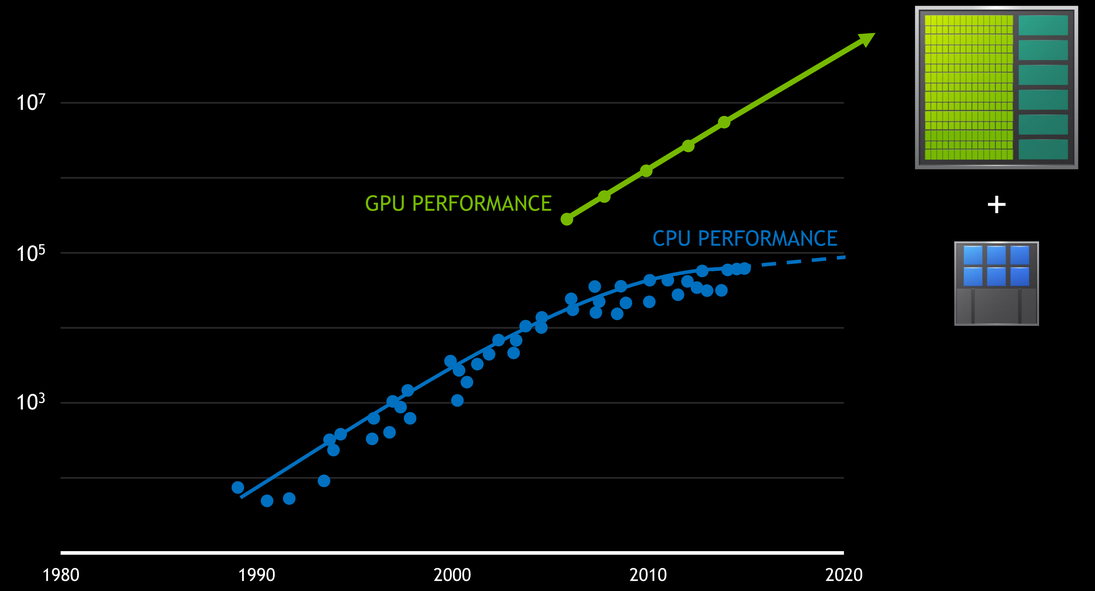
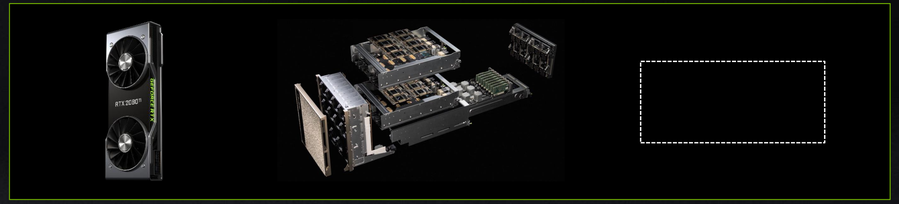
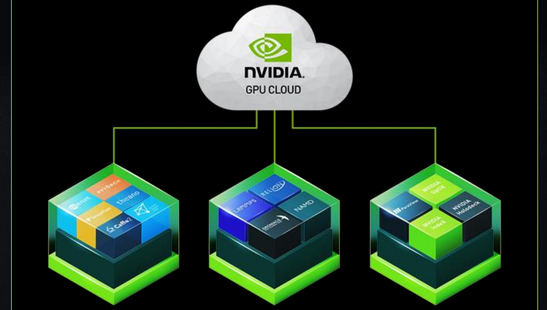
NVIDIA

ACCELERATED COMPUTING
1000X EVERY 10 YEARS



ONE ARCHITECTURE

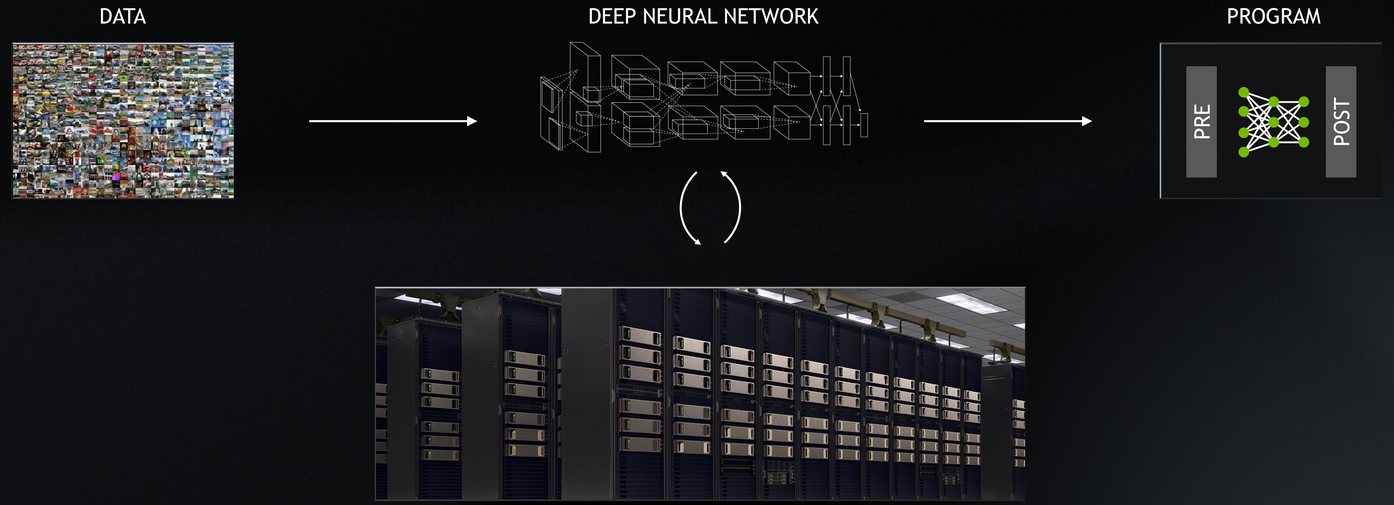




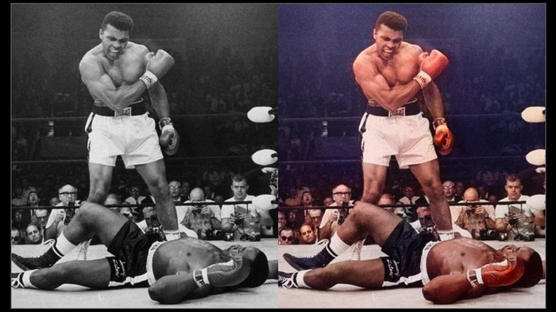
ACCELERATION STACKS

GAMING	HPC	TRANSPORTATION	HEALTHCARE
PRO VIZ	AI	ROBOTICS	AI IOT

COMPUTERS WRITING SOFTWARE



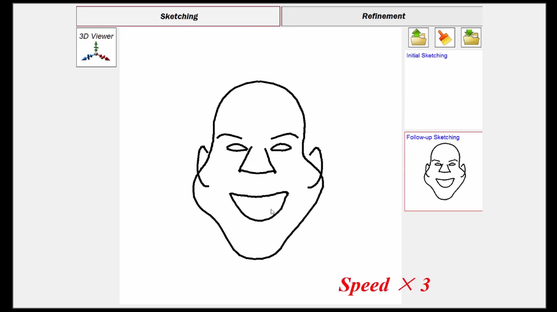
AMAZING SOFTWARE



COLORIZING IMAGES
UC Berkeley

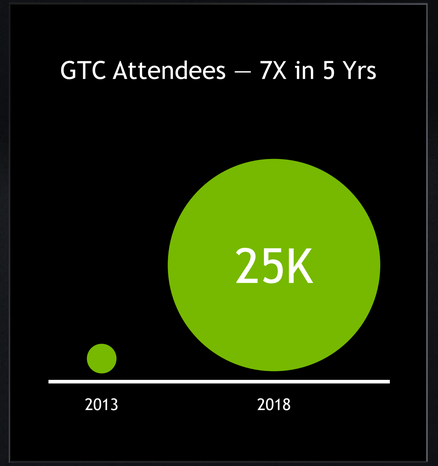
SEGMENTATION
NVIDIA

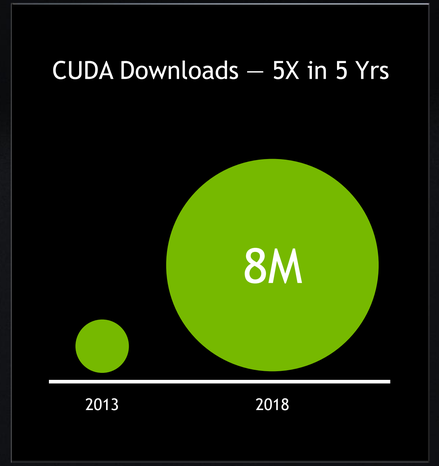
COLORIZING HAIR
L'Oréal

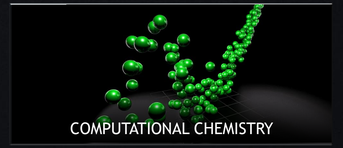


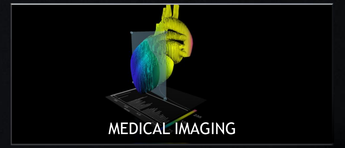
SKETCH TO FACE
The University of Hong Kong

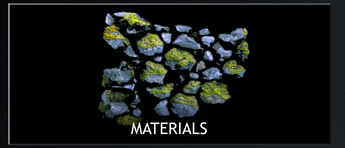
RISE OF GPU COMPUTING

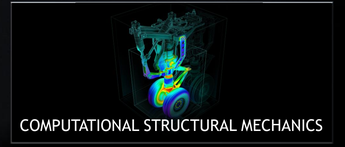


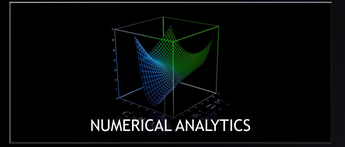


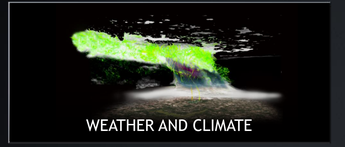




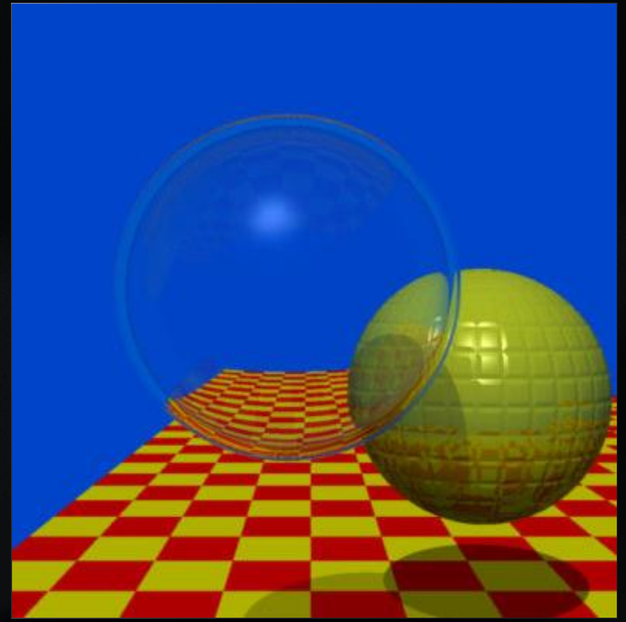








THE HOLY GRAIL OF COMPUTER GRAPHICS

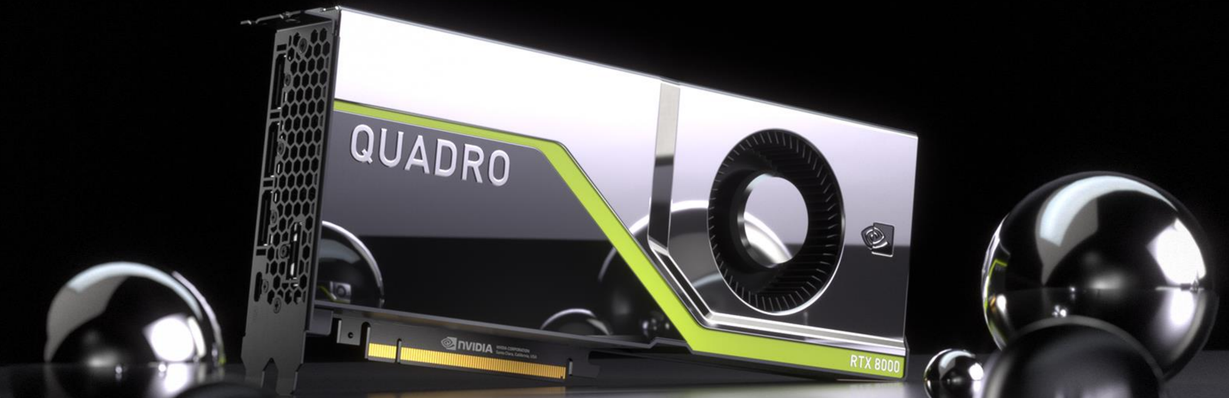


Turner Whitted | 1979
“Multi-bounce Recursive Ray Tracing”
1.2 Hours for 512x512 on VAX 11/780

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

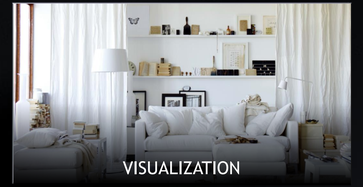


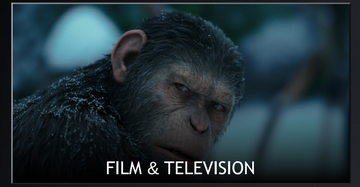
RTX OPENS $250B VISUAL EFFECTS INDUSTRY

RTX SERVER — 60X FASTER THAN CPU NODE

Adobe Dimension CC
AUTODESK ARNOLD
clarisse
DaVinci Resolve
3DEXPERIENCE CATIA
SOLIDWORKS
IC.IDO
octanerender
ParaView
REDSHIFT
RENDERMAN
SEED
SIEMENS NX
SPEOS
SUBSTANCE DESIGNER
unity
UNREAL ENGINE
VMD Visual Molecular Dynamics
V-ray
weta DIGITAL

RTX ACCELERATED RENDERERS





PHOTOREAL VFX

NEW GEFORCE RTX

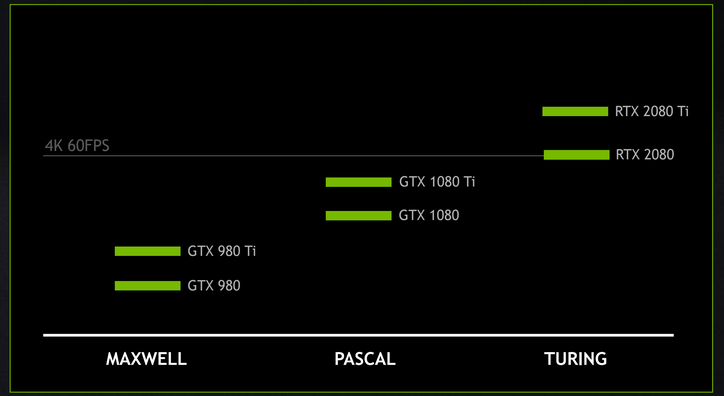
GRAPHICS REINVENTED

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

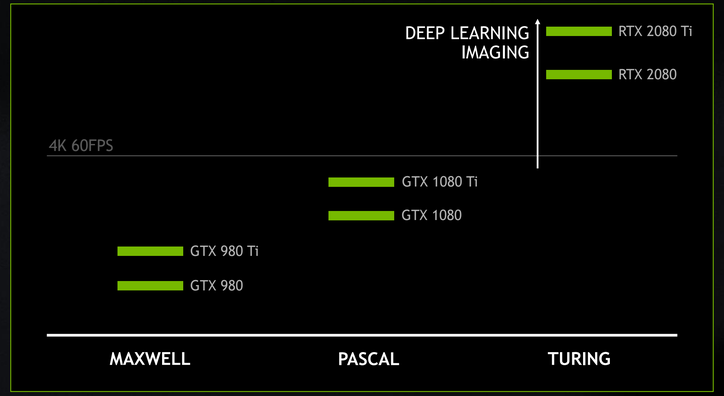
GEFORCE RTX 2080 — FROM $699
11 TFLOPS | 11 TIOPS | 85 Tensor TFLOPS | 8 GIGA RAYS

GEFORCE RTX 2080 Ti — FROM $999
14 TFLOPS | 14 TIOPS | 114 Tensor TFLOPS | 10 GIGA RAYS

RTX RESETS $100B GAMING INDUSTRY



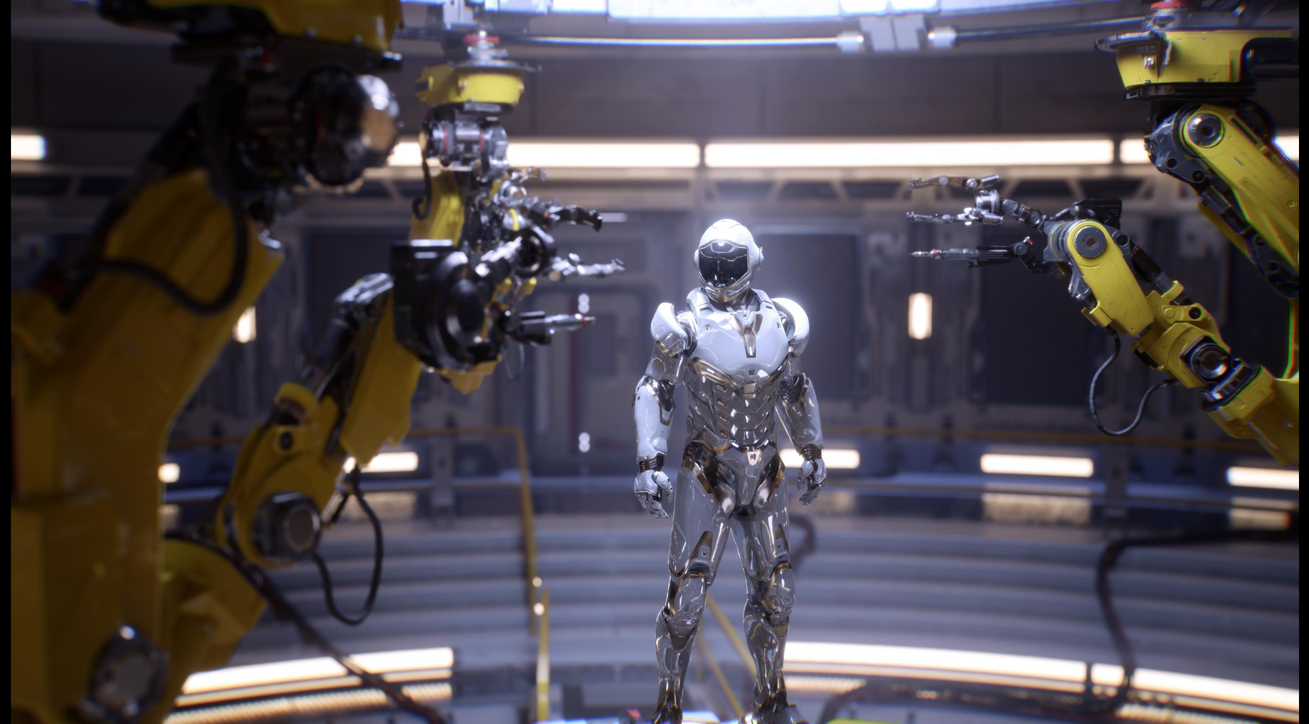
FASTEST GAMING GPU | PLAYS AT 4K 60FPS



DEEP LEARNING IMAGING WITH 114 TFLOPS TENSOR CORE

RAY TRACING: “THE HOLY GRAIL OF GRAPHICS”

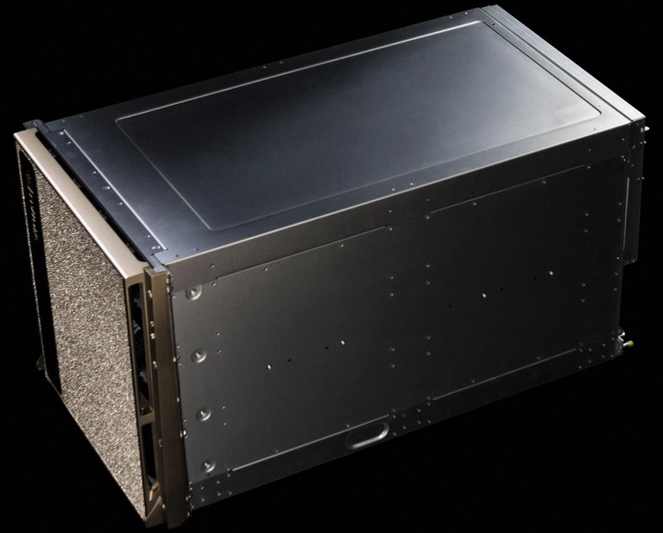
RTX
ON



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

24
hours

108
minutes

6.6
minutes

TRANSLATION

5
hours

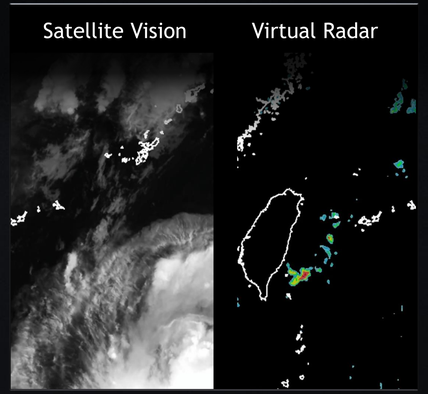
32
minutes

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. 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 Supercompute

FujiFilm
AI Supercomputer, Accelerating Medical Imaging



Weathernews
Breakthrough Virtual Radar with dAIgnosis

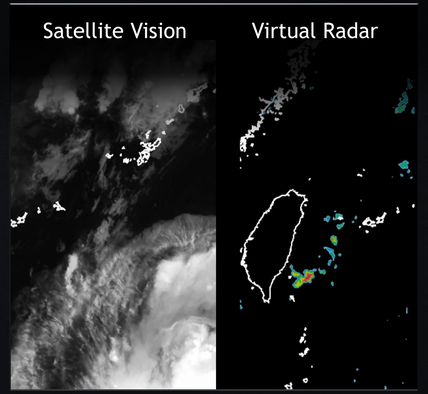
NTT
Accelerating COREVO AI

PFN
AI Supercomputer

NVIDIA PARTNERS WITH JAPAN LEADERS IN AI & HPC

AIST - ABCI
Japan's Fastest AI Supercomputer

FujiFilm
AI Supercomputer, Accelerating Medical Imaging



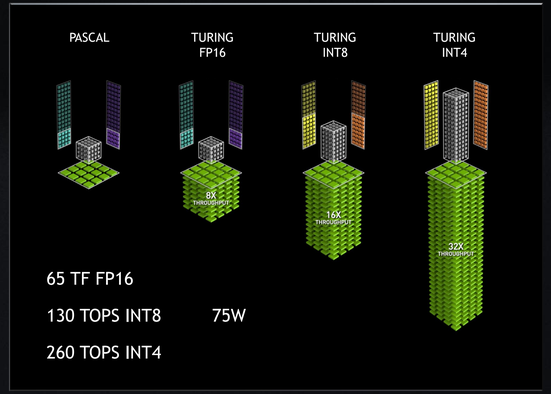
Weathernews
Breakthrough Virtual Radar with dAIgnosis

NTT
Accelerating COREVO AI

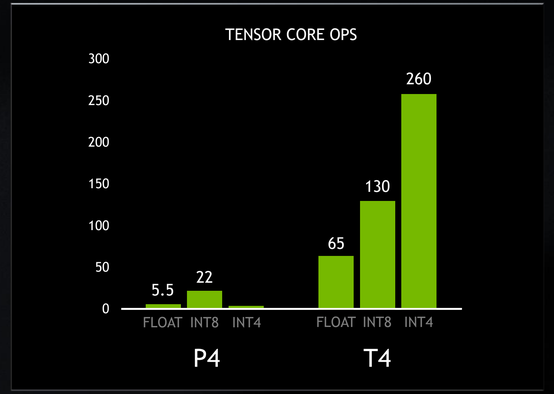
PFN
AI Supercomputer

ANNOUNCING TESLA T4

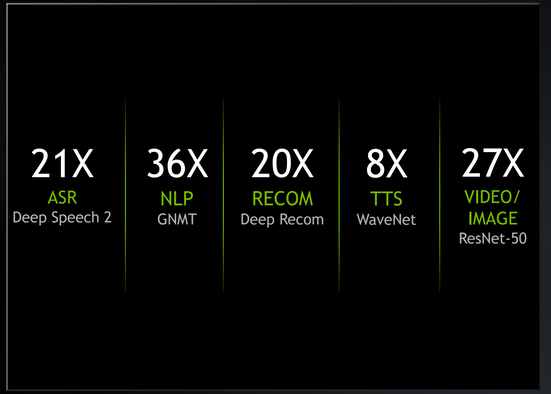
Tesla T4
Universal Inference Acceleration



Multi-Precision Tensor Core

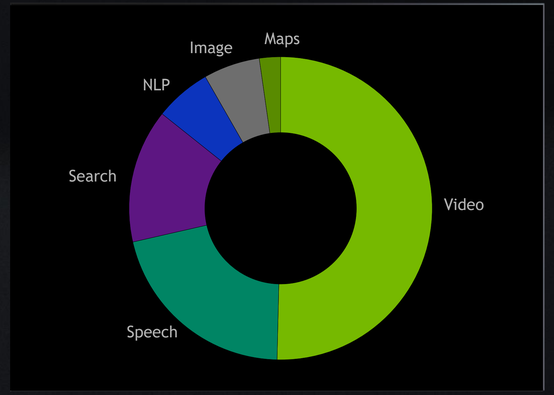


Giant Leap
12X Pascal FP Inference

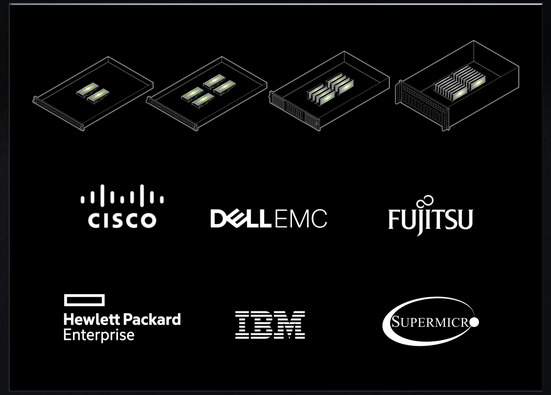


TensorRT 5.0
Support for Tensor Core

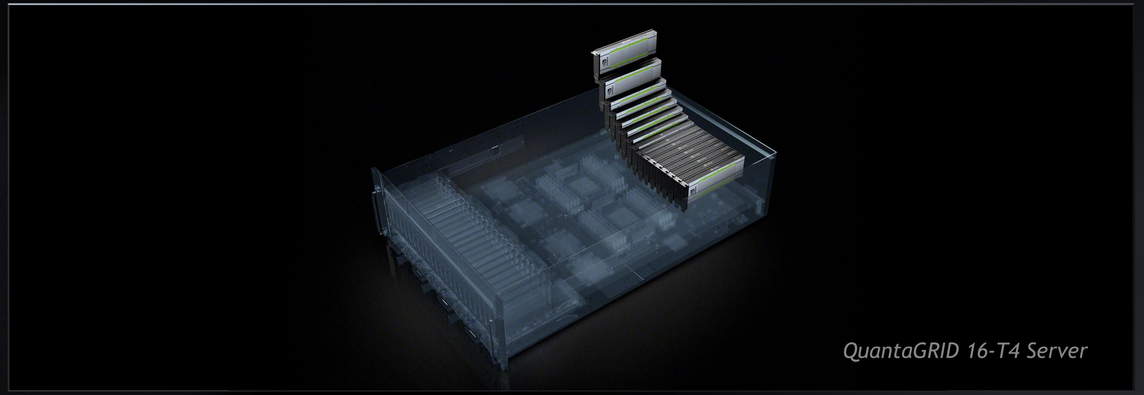
ANNOUNCING TESLA T4



Tesla P4 and TensorRT Adoption



World's Leading Systems Makers



World's First 1 PetaFLOPS Inference Machine

TRADITIONAL HYPERSCALE DATACENTER

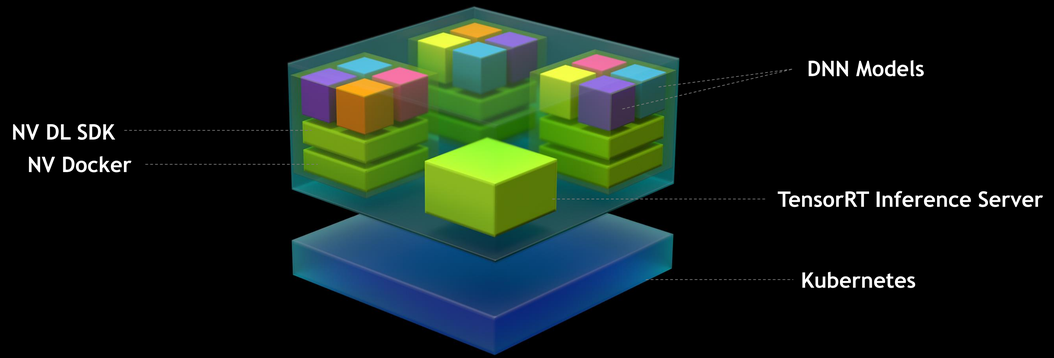
200 CPU Servers
Speech | NLP | Video
60 kW

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



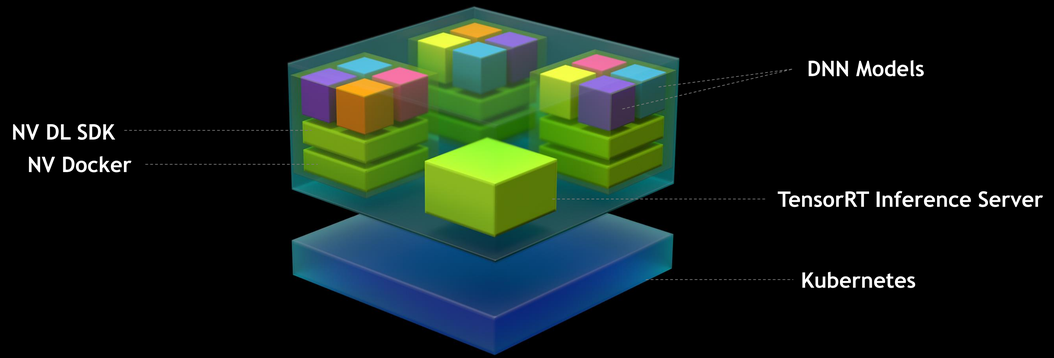
Kubernetes and Docker on NVIDIA GPUs

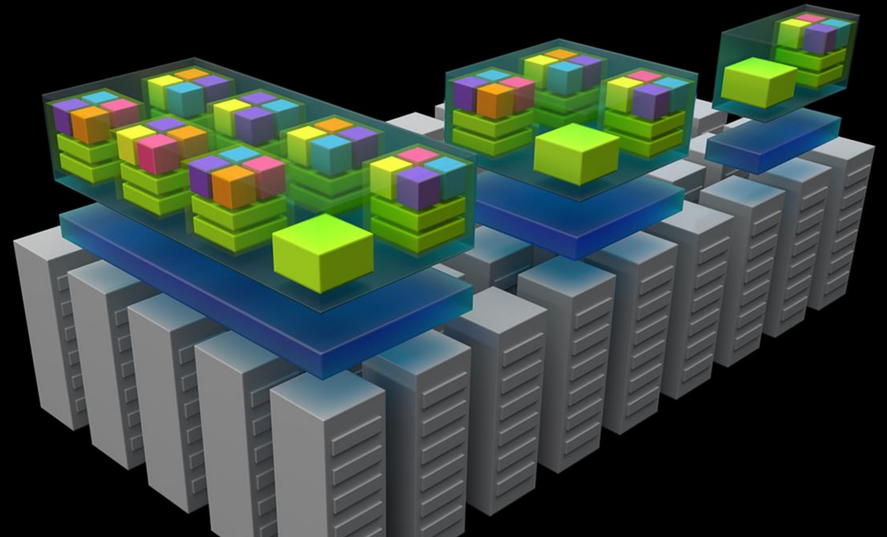
New Inference Serving Engine

Multiple Model Types and Frameworks Concurrently

Maximize Datacenter Throughput and Utilization

ANNOUNCING NVIDIA TENSORRT HYPERSCALE





WORLD'S LEADING AI PLATFORM

IMAGES

1
millisecond

6,250
images/second

56
images/second/watt

TRANSLATION

13,160
words/second

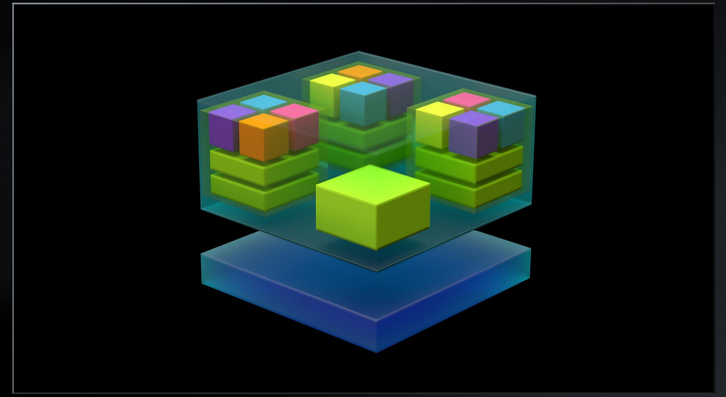
Images: Fastest Inference: Resnet-50 Inference on Tesla V100 with TensorRT, Batch size: 1, Int8 optimized. Inference Throughput: Resnet-50 Inference on Tesla V100 with TensorRT, Batch size: 128, Int8 Optimized. Inference Efficiency: Resnet-50 Inference on Tesla T4, Int8 optimized, batch size = 32
Translation: GMNT Inference on Newstest2015 test dataset on Tesla V100.

NVIDIA TENSORRT HYPERSCALE

Tesla T4

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

TensorRT 5

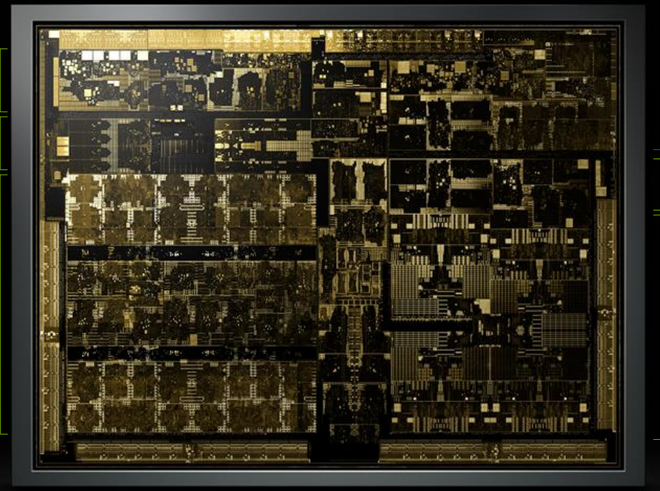


TensorRT Inference Server

XAVIER

WORLD'S FIRST AUTONOMOUS MACHINE PROCESSOR

Most Complex SOC Ever Made
9 Billion Transistors, 350mm^2, 12nFFN
~8,000 Engineering Years



16 Lane CSI
109 Gbps CPHY 1.1
1Gb Ethernet

DLA
5.7 TFLOPS FP16
11.4 TOPS INT8

Industry Standard High-Speed IO
PCIe Gen4 Root and Endpoint
USB 3.1 Gen2 Host and Device
UFS 2.1 Embedded Storage

Multimedia Engines
1.2 GPIX/s Encode
1.8 GPIX/s Decode
4 GPIX/s Video Image Compositor

Vision Accelerator
1.7 TOPS

Stereo & Optical Flow Engine
2x 3.1 TOPS

ISP
2.4 GPIX/s
Native Full-range HDR
Tile-based Processing

Volta Tensor Core GPU
FP32 / FP16 / INT8 Multi-Precision
512 CUDA Tensor Cores
2.8 CUDA TFLOPS (FP16)
22.6 Tensor Core DL TOPS

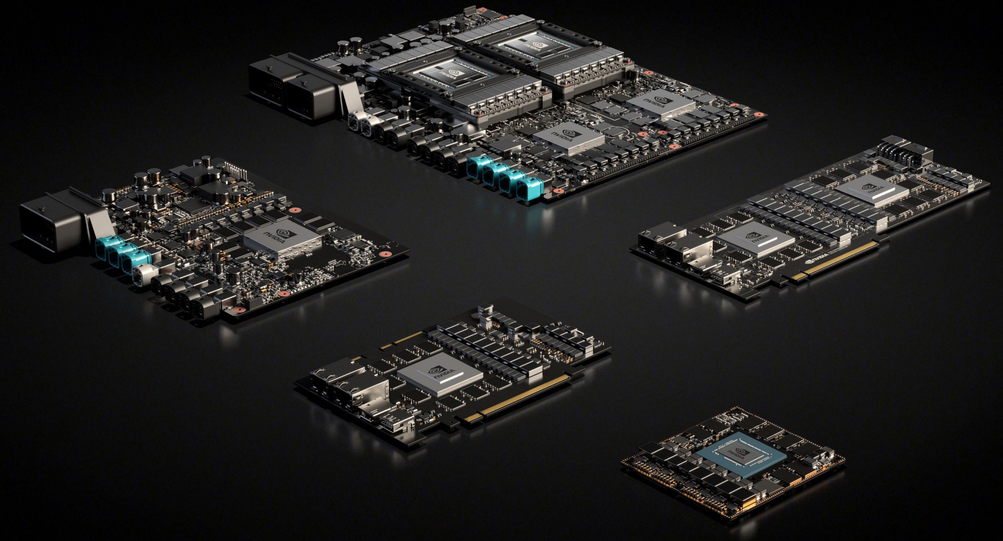
Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6

256-Bit LPDDR4X
137 GB/s

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



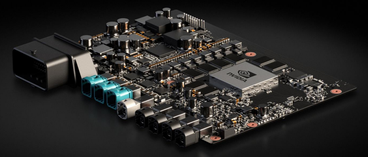
NVIDIA DRIVE

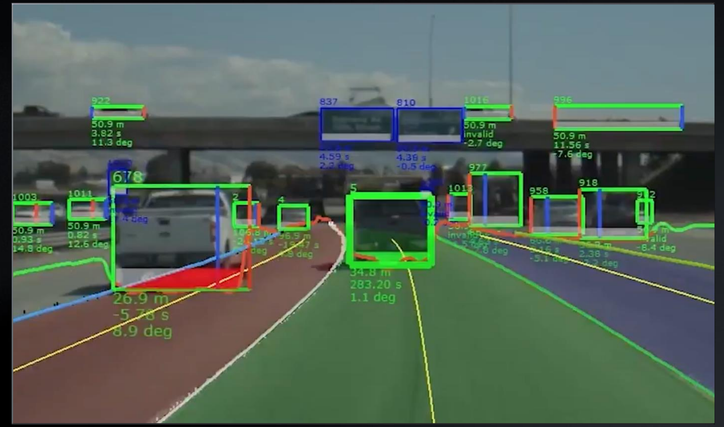
SENSOR PROCESSING

MAPPING & LOCALIZATION | PATH & TASK PLANNING

PERCEPTION | SITUATION UNDERSTANDING

DIVERSITY & REDUNDANCY



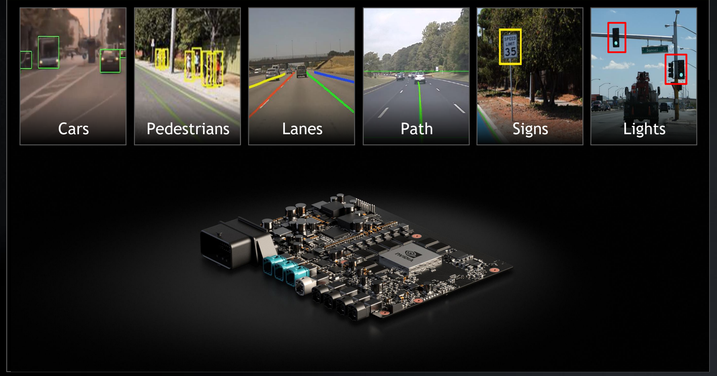


NVIDIA DRIVE

TRAINING

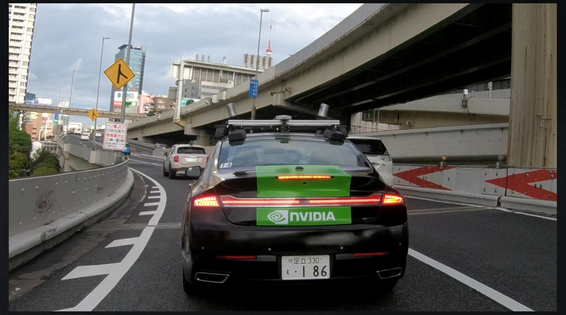
SIMULATING

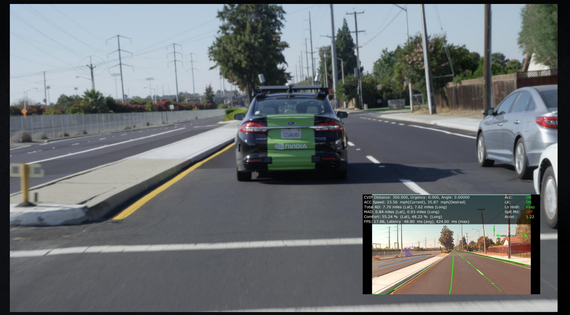
DRIVING



0
20
40
60
80
100
120
140
MPH
km/h
23 mph
BRAKE
ACC

DRIVE AV





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	MOBILITY SERVICES		TRUCKS	TIER ONES	MAPPING	SENSORS

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

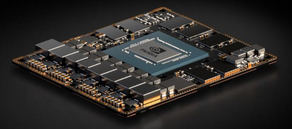
NVIDIA ISAAC

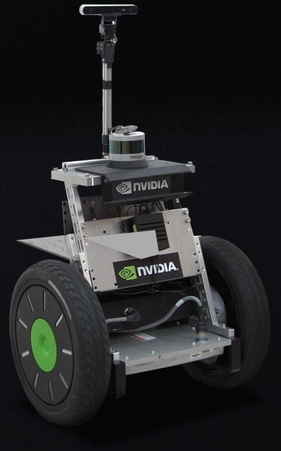
SENSOR PROCESSING

MAPPING & LOCALIZATION | PATH & TASK PLANNING

PERCEPTION | SITUATION UNDERSTANDING

DIVERSITY & REDUNDANCY



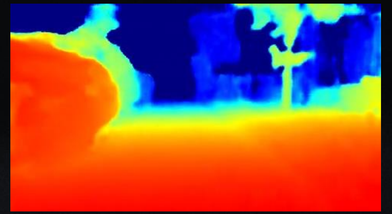


Reality
Simulation

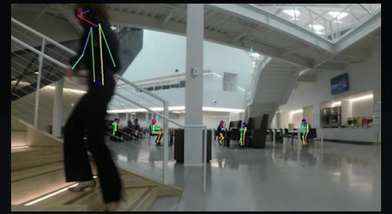
NVIDIA ISAAC GEMS

Global Localization

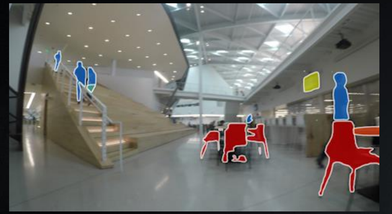
LQR Path Planner



Depth Estimation

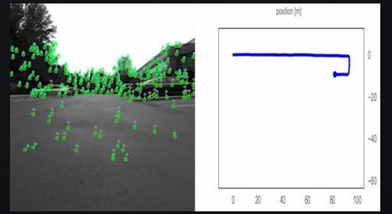


Human Pose Estimation



Object / People Detection

Map Editor



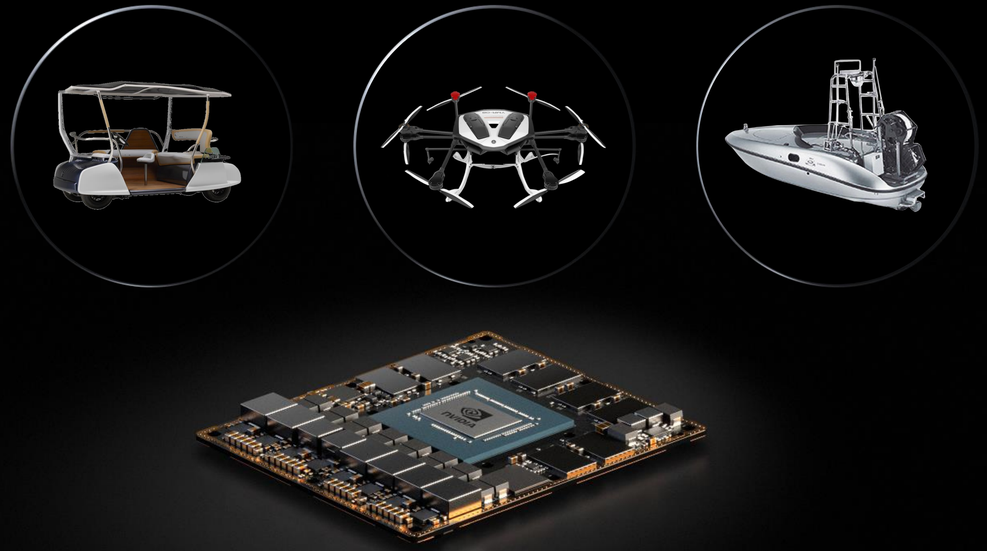
Visual Odometry

Physical Simulation

Gesture Recognition

ASR

ANNOUNCING YAMAHA MOTOR ADOPTS JETSON AGX FOR AUTONOMOUS MACHINES



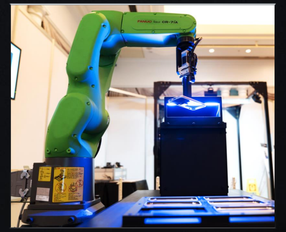
NVIDIA PARTNERS WITH JAPAN LEADERS IN ROBOTICS & AI IOT

KOMATSU
#1 Construction

DENSO
#1 Auto Parts

YAMAHA MOTORS
#1 Mobility Machines

CANON
Factory Automation



FANUC
#1 FA Robotics

KAWADA TECHNOLOGIES
Collaborative Robots

MUSASHI
Factory Automation

PANASONIC
Smart City

$100B IMAGING INDUSTRY GOING TO AI HPC

ULTRASOUND

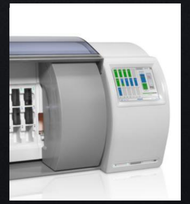
ENDOSCOPY

MAMMOGRAPHY

3D: CT, MRI, PET

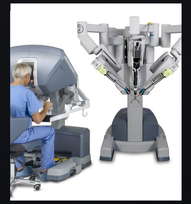
RADIATION THERAPY

SEQUENCERS


DIGITAL PATHOLOGY

CRYO-EM

LIQUID BIOPSY


ROBOTIC SURGERY

10-10,000 GB/sec

5-50 TFLOPS
1-12 GPUs per machine
50-1,800W

$100B IMAGING INDUSTRY GOING TO AI HPC

ULTRASOUND

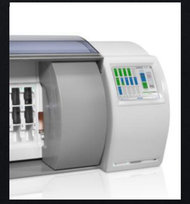
ENDOSCOPY

MAMMOGRAPHY

3D: CT, MRI, PET

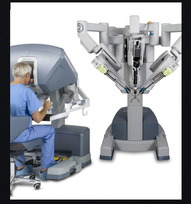
RADIATION THERAPY

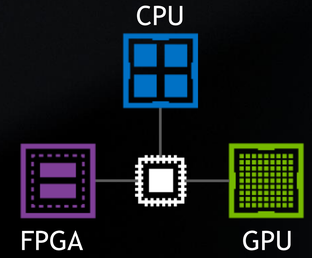
SEQUENCERS


DIGITAL PATHOLOGY

CRYO-EM

LIQUID BIOPSY


ROBOTIC SURGERY



ANNOUNCING CLARA AGX

ULTRASOUND

ENDOSCOPY

MAMMOGRAPHY

3D: CT, MRI, PET

RADIATION THERAPY

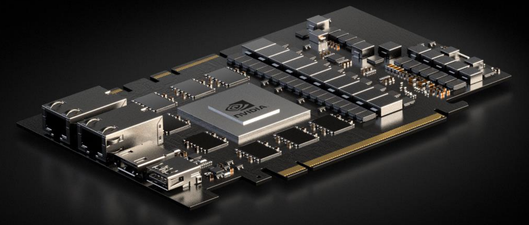
SEQUENCERS

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

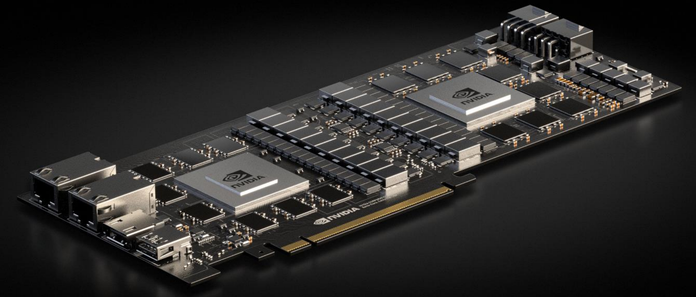
SEQUENCERS

DIGITAL PATHOLOGY

CRYO-EM

LIQUID BIOPSY

ROBOTIC SURGERY



Scale Up to 200 TOPS DL Processing

8 GIGA Rays

200W

ANNOUNCING
JETSON AGX XAVIER DEVKIT

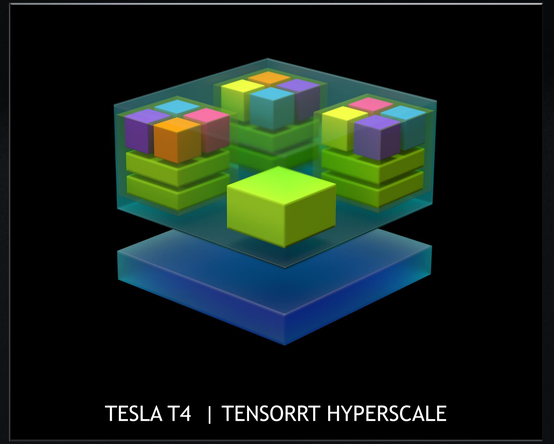
WORLD'S FIRST EDGE AI COMPUTER

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

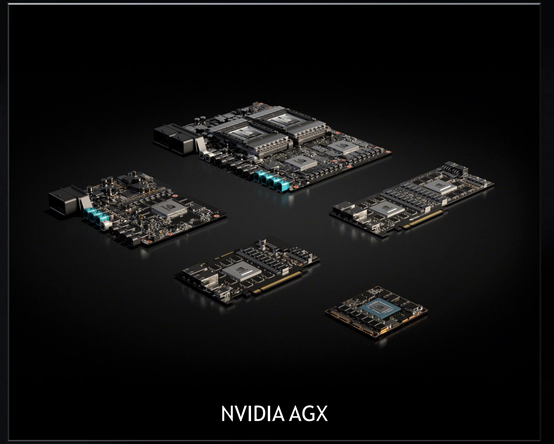
ANNOUNCING NEW NVIDIA PLATFORMS

QUADRO RTX

GEFORCE RTX



TESLA T4 | TENSORRT HYPERSCALE



NVIDIA AGX

DRIVE AGX | DRIVE SDK

JETSON AGX | ISAAC SDK

NVIDIA