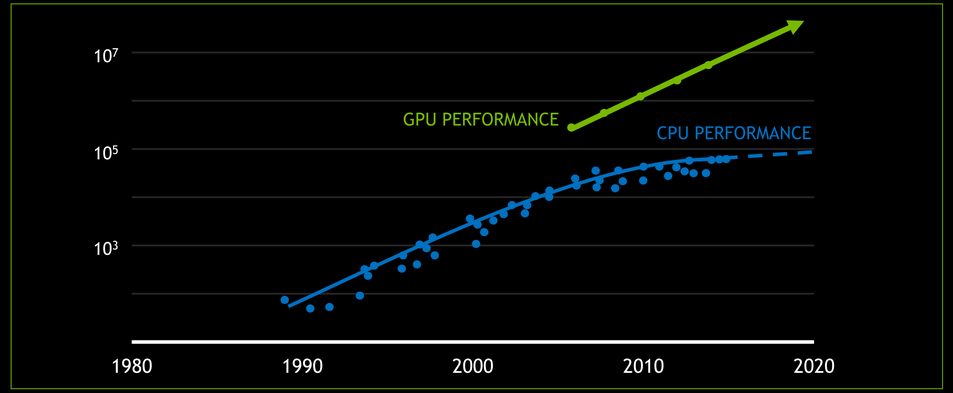
Accelerated Platforms: The Future of Computing

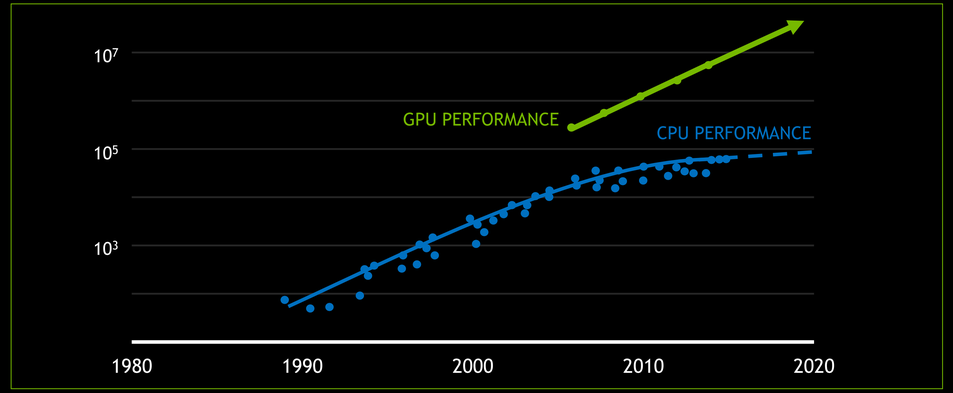
Marc Hamilton, VP Solutions Architecture & Engineering, NVIDIA | Korea AI Conference 2018

Forces Shaping Computing

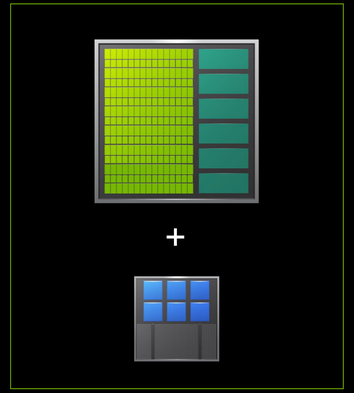


Beyond Moore’s Law

Forces Shaping Computing

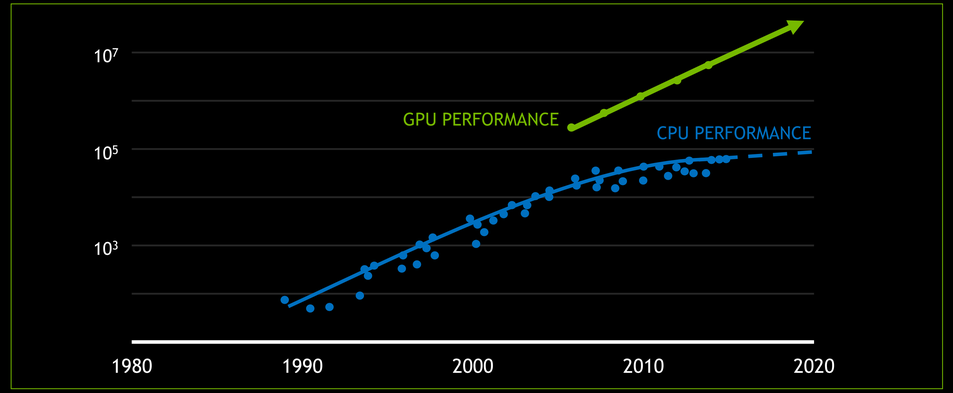


Beyond Moore's Law — 1000x Every 10 Years

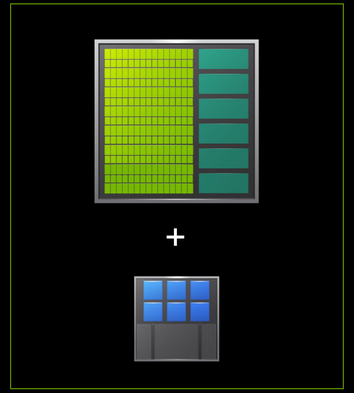


Accelerated Computing

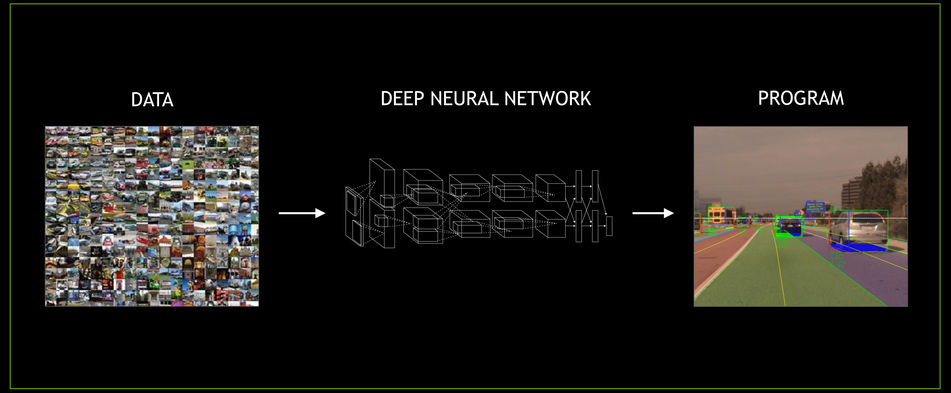
Forces Shaping Computing



Beyond Moore's Law — 1000x Every 10 Years



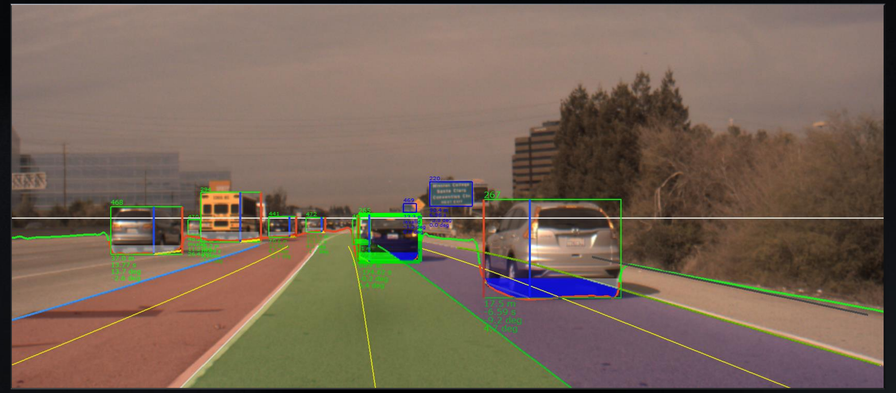
Accelerated Computing



The Deep Learning Revolution

NVIDIA is

Accelerators for Demanding Applications

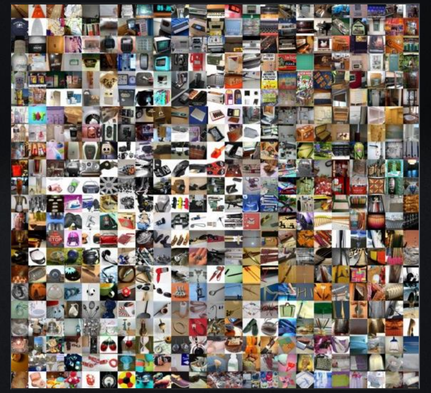


Platforms That Provide Complete Solutions

Graphics | AI | Healthcare | Autonomous Vehicles | Data Science | Robotics

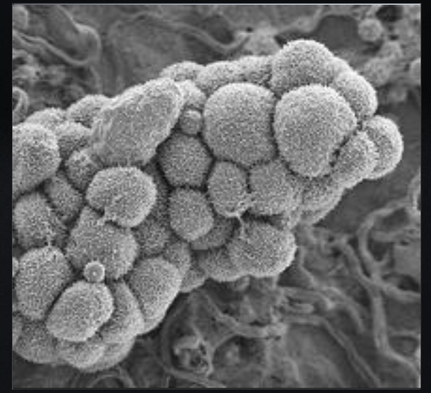
Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

Deep Learning Everywhere



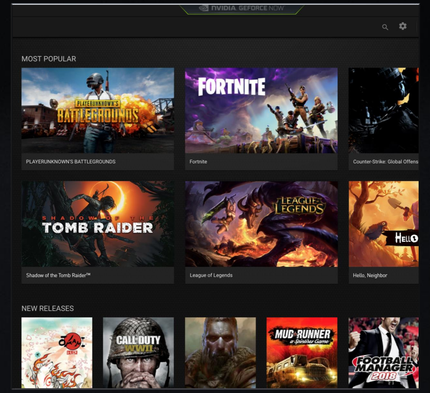
Internet & Cloud

Image Classification | Speech Recognition
Language Translation | Language Processing
Sentiment Analysis | Recommendations



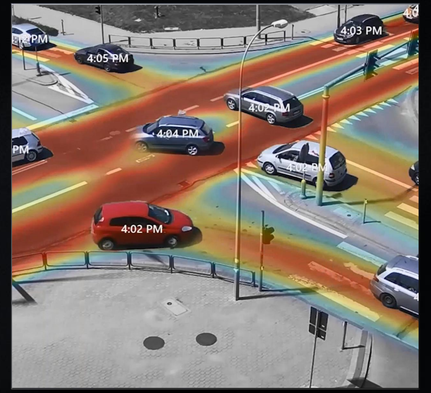
Medicine & Biology

Cancer Cell Detection | Diabetic Grading
Drug Discovery



Media & Entertainment

Video Captioning | Video Search
Real Time Translation



Intelligent Video Analytics

Traffic Analysis | Retail Analytics
Access Control

Transportation

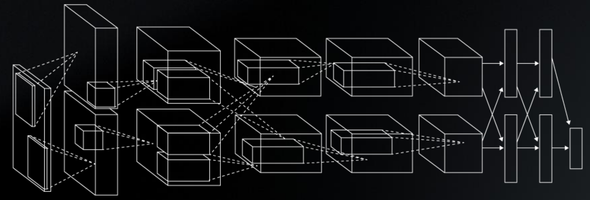
Pedestrian Detection | Lane Tracking
Traffic Sign Recognition

AI IMPROVES SEMICON INSPECTION ACCURACY

In the semiconductor industry, inaccurate or false positive fault inspections can lead to huge product losses. SK Hynix deployed an AI fault detection solution with NVIDIA Tesla GPUs, DGX Station, Jetson TX2, CUDA and TensorRT. With its new deep learning-based tool, SK Hynix achieved over 90% inspection accuracy.

NVIDIA SK hynix

Deep Learning Was Enabled by Hardware



Deep Learning is Gated by Hardware

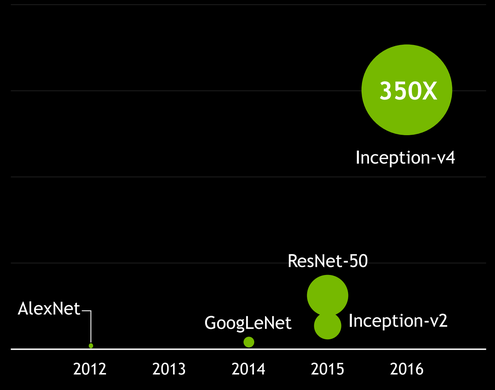
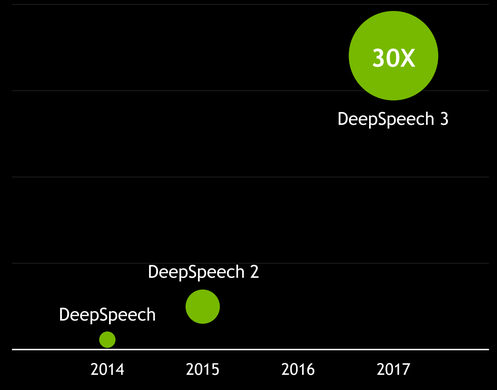
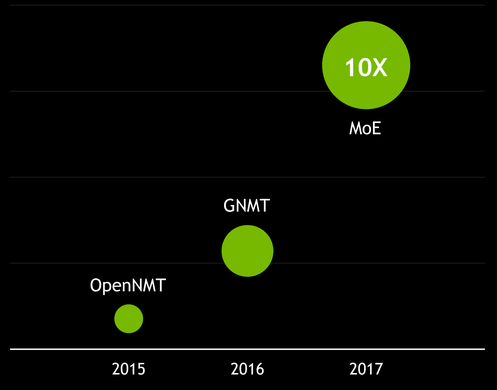


Image Network Complexity
GOPS * Bandwidth



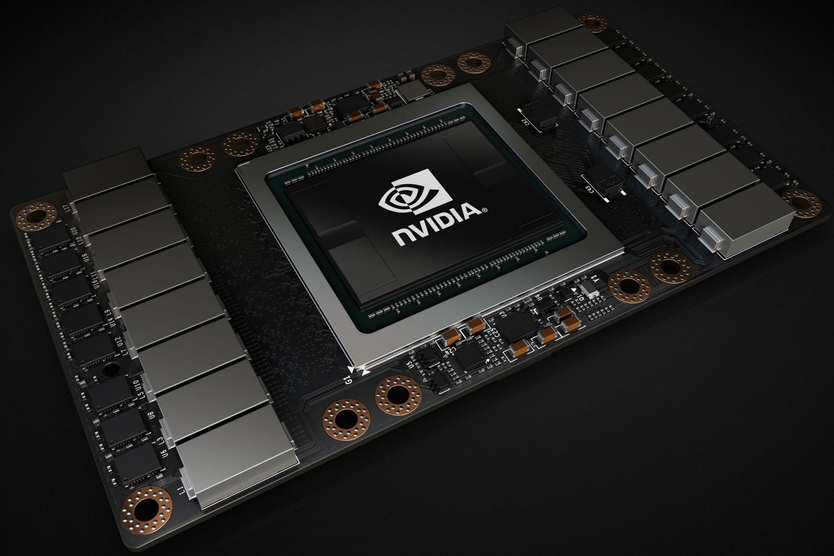
Speech Network Complexity
GOPS * Bandwidth



Translation Network Complexity
GOPS * Bandwidth

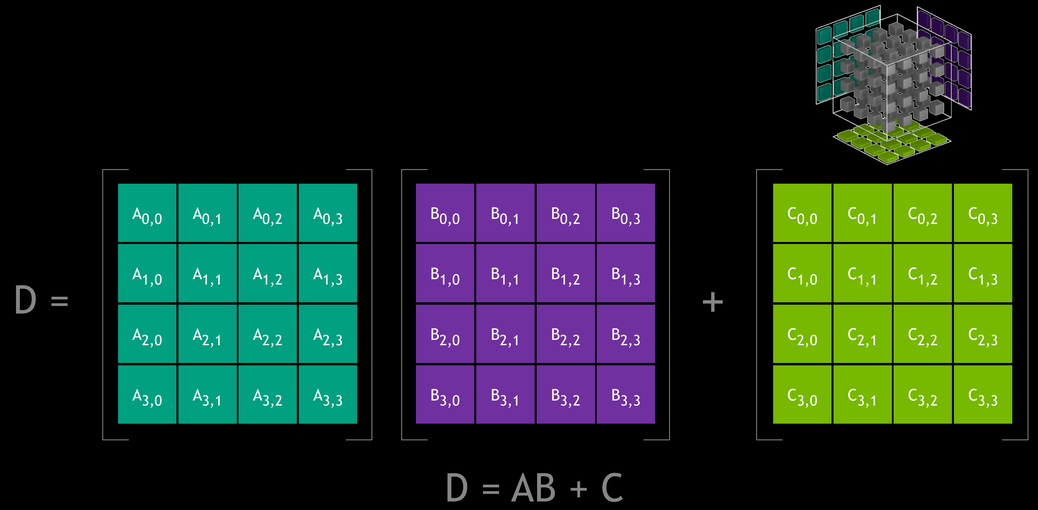
Tesla V100 Tensor Core GPU

21B Transistors | TSMC 12nm FFN | 815mm^2
5,120 CUDA Cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
125 Tensor TFLOPS
20 MB SM RF | 16 MB Cache
32 GB HBM2 @ 900 GB/s
300 GB/s NVLink

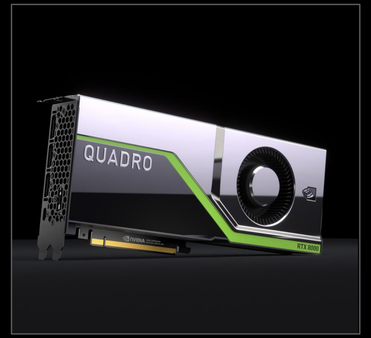


Tensor Core

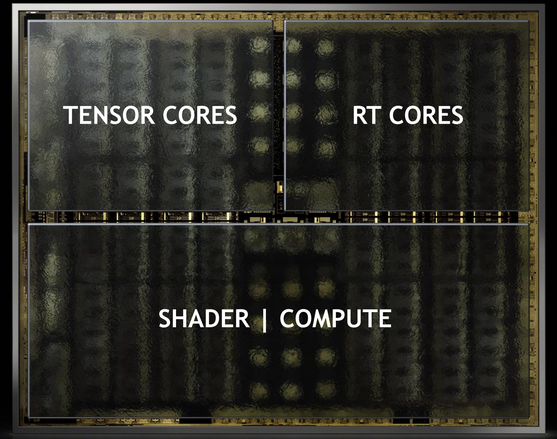
Mixed Precision Matrix Math
4x4 Matrices



Turing Accelerates Inference



Quadro RTX 8000
4,608 CUDA Cores | 576 Tensor Cores
48 GB GDDR6 Memory
130.5 TFLOPS FP16
261 TOPS INT8 | 522 TOPS INT4
672 GB/s DRAM BW
250 GB/s (Bidirectional) NVLink Channels
295 W

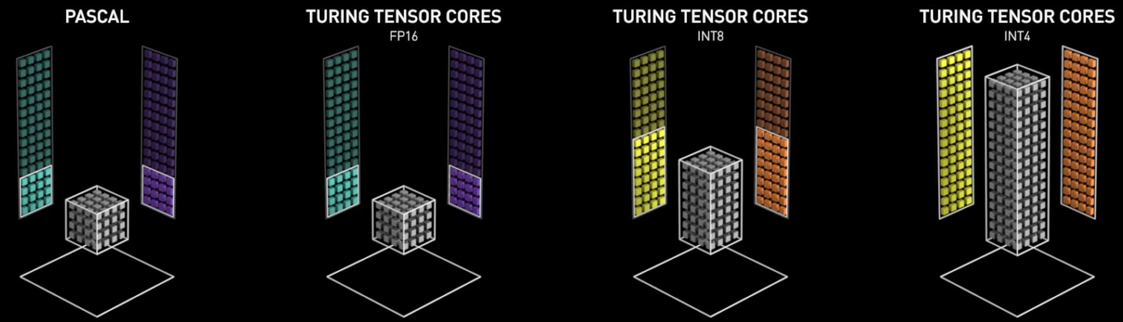


Tesla T4
2,560 CUDA Cores | 320 Tensor Cores
16 GB GDDR6 Memory
65 TFLOPS FP16
130 TOPS INT8 | 260 TOPS INT4
320 GB/s DRAM BW
70 W

Turing Tensor Core - Optimized for Inference

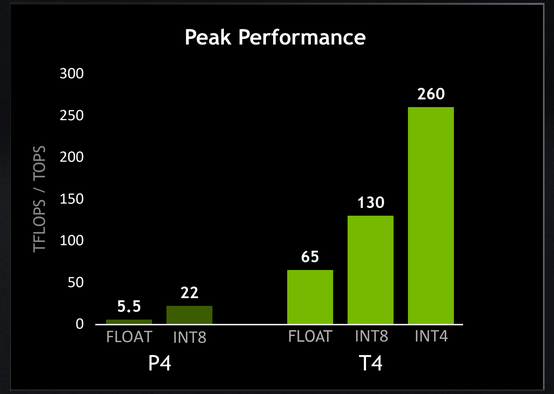
Multi-Precision for AI Inference

T4: 65 TFLOPS FP16 | 130 TOPS INT8 | 260 TOPS INT4
RTX 8000: 130 TFLOPS FP16 | 260 TOPS INT8 | 520 TOPS INT4



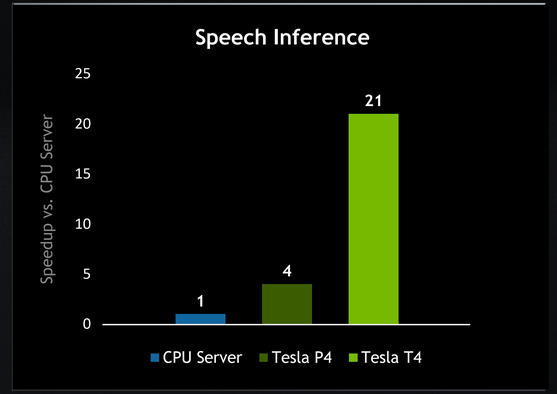
World's Most Performant Inference Platform

Up to 36X Faster Than CPUs | Accelerates All AI Workloads



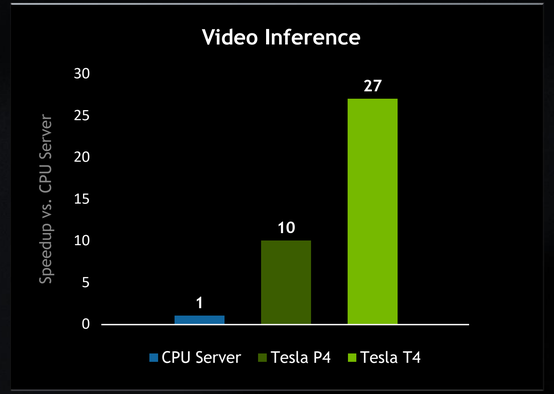
Speedup: 6X Faster

Int8 Ops vs P4



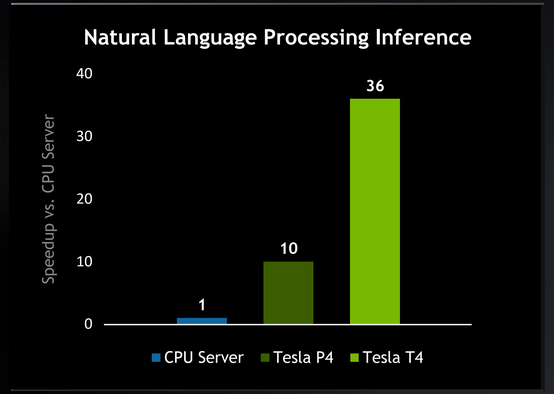
Speedup: 21X Faster

DeepSpeech 2



Speedup: 27X Faster

ResNet-50 (7ms Latency Limit)

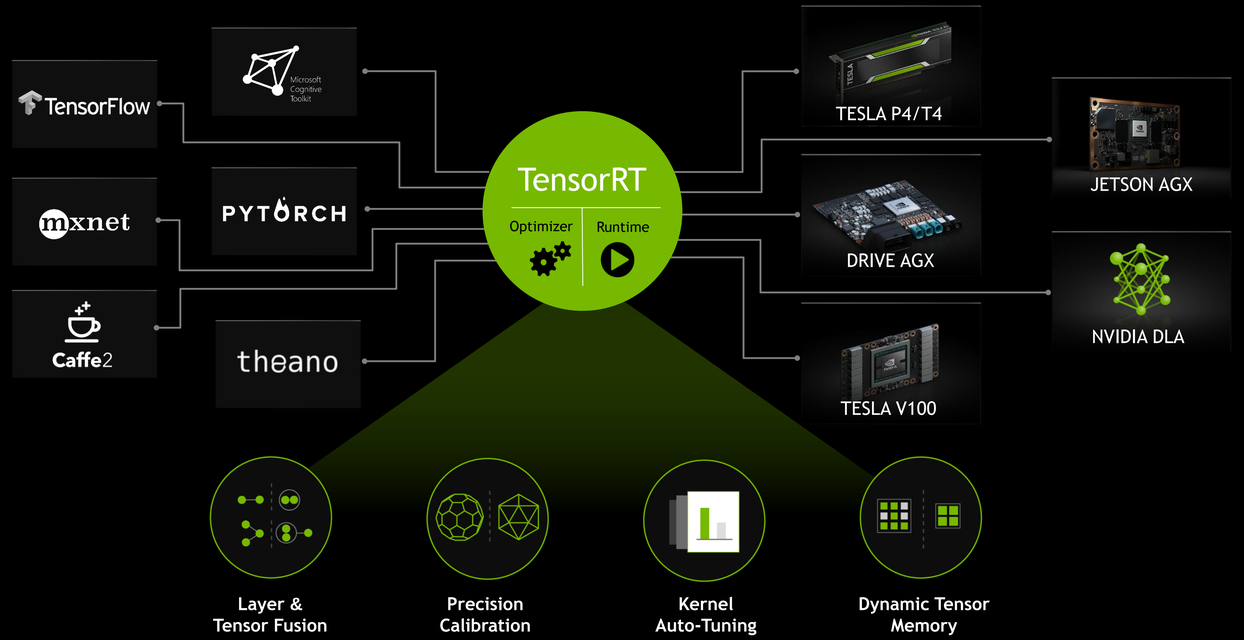


Speedup: 36X Faster

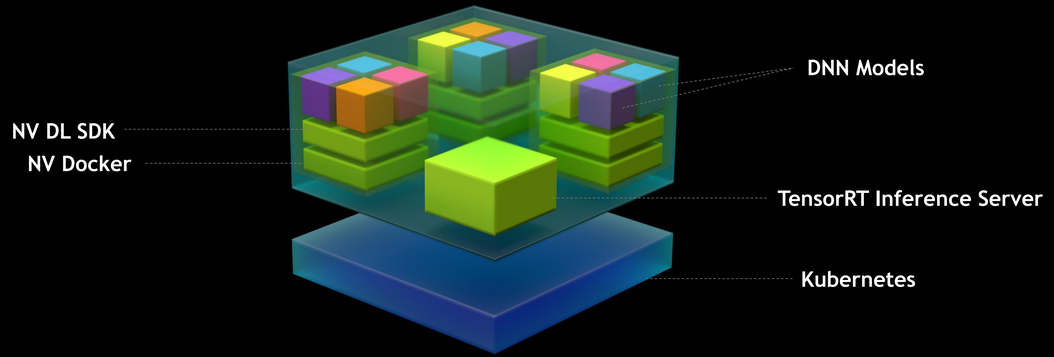
GNMT

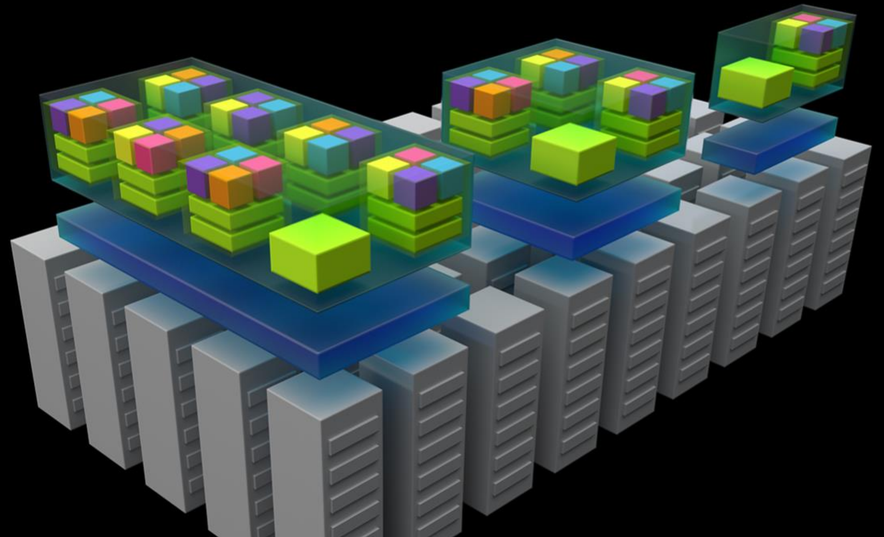
NVIDIA TensorRT 5

Multi-Precision Acceleration of All Frameworks
Containerized Inference Serving Engine
Docker and Kubernetes Integration



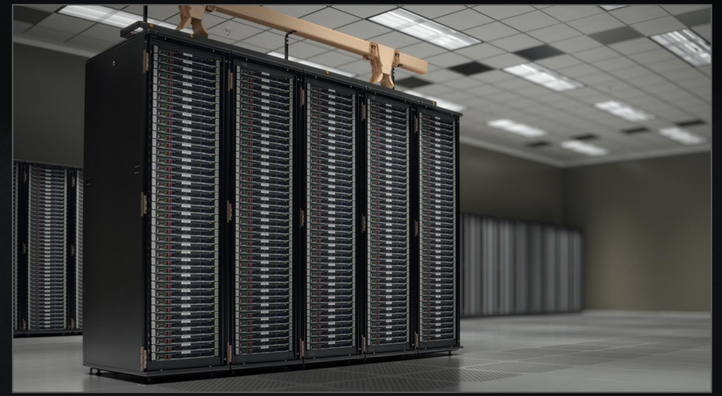
TensorRT Inference Server





Space and Power Reduction

Game-Changing Inference Performance



Inference Workload

200 CPU Servers | 60 KWatts | Speech, NLP and Video

=

Inference Workload

1 T4 Accelerated Server | 2 KWatts | Speech, NLP and Video

Accelerating Deep Learning

Accelerating Graphics

Accelerating Science

Accelerating Data Science

Scaling

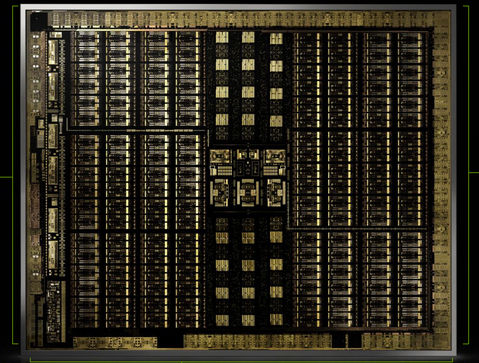
Accelerating Autonomous Vehicles

Accelerating Robotics

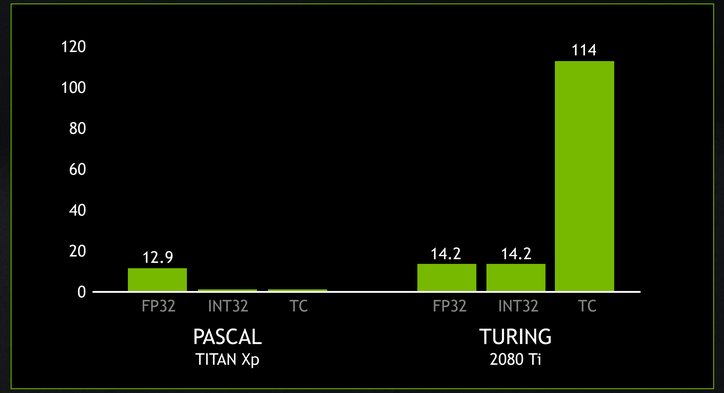
Conclusion

Turing Revolutionizes Graphics

Quadro RTX 8000
4,608 CUDA Cores | 576 Tensor Cores
72 RT Cores
48 GB GDDR6 Memory
130 TFLOPS FP16 | 260 TOPS INT8 | 520 TOPS INT4
336 GB/s DRAM BW
250 GB/s NVLINK Channels
295W



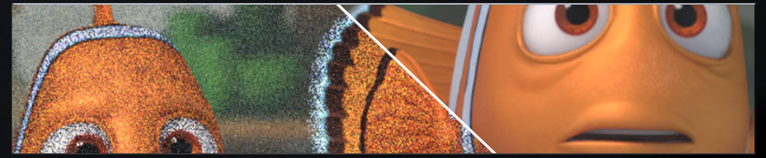
Deep Learning for Imaging



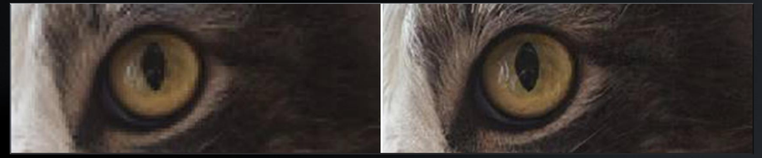
Turing – 9X Peak FLOPS

Colorizing | UC Berkeley

In-Painting | NVIDIA



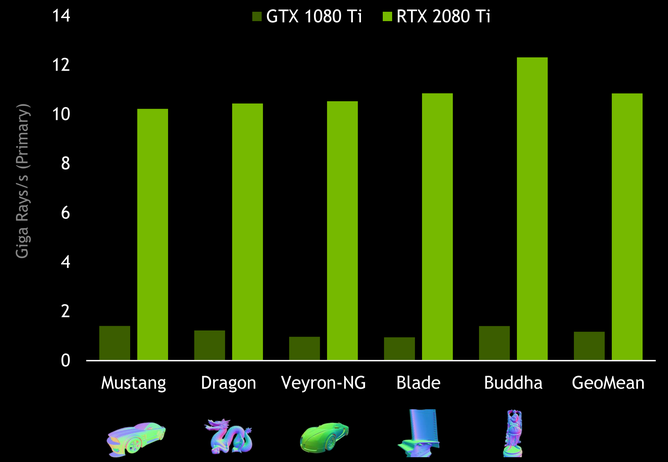
Denoising | Disney Research, Pixar, UCSB



SuperRez | NVIDIA

Turing Ray Tracing Performance

>10 Giga Rays



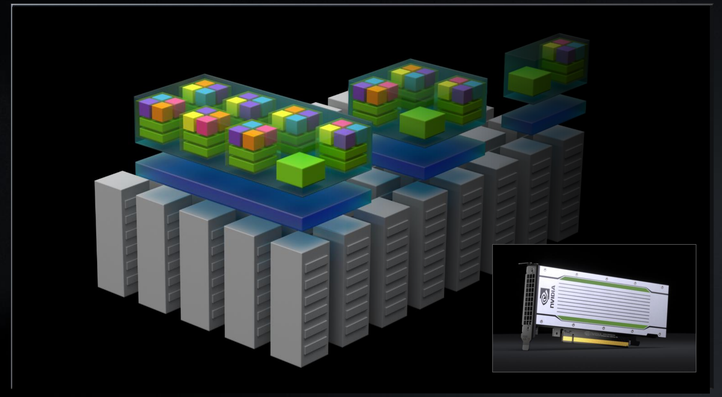
GTX 1080 Ti	RTX 2080 Ti
11.3 TFLOPS	68 RT Cores
1.1 Giga Rays	10+ Giga Rays
10 TFLOPS / Giga Ray	~10X faster than 1080 Ti

70
PORSCHE

Turing — A Giant Leap

Gaming Reinvented

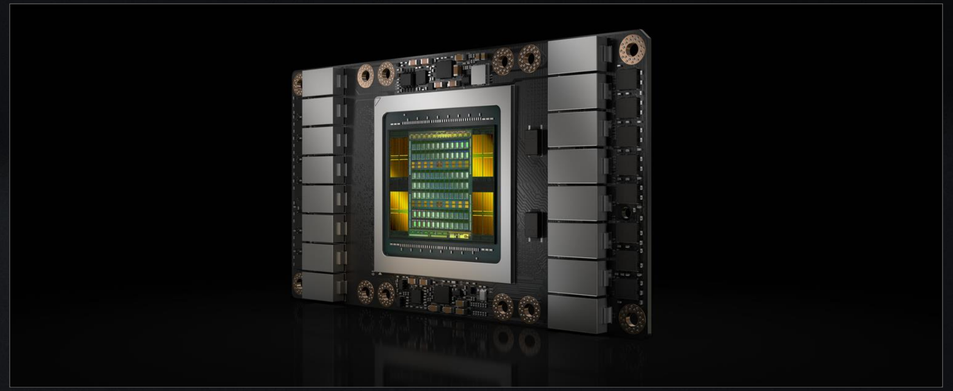
World’s First Ray Tracing GPU



Universal Deep Learning Accelerator

Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

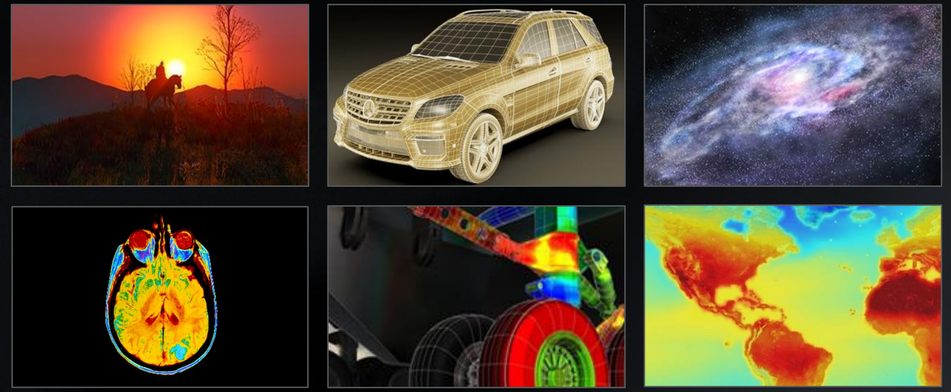
Accelerating Science



VASP
AMBER
NAMD
GROMACS
Gaussian
Simulia Abaqus
WRF
OpenFOAM
ANSYS
LS-DYNA
BLAST
LAMMPS
ANSYS Fluent
Quantum Espresso
GAMESS

Top 15 HPC Applications

Intersect360 Research, Nov 2017
"HPC Application Support for GPU Computing"



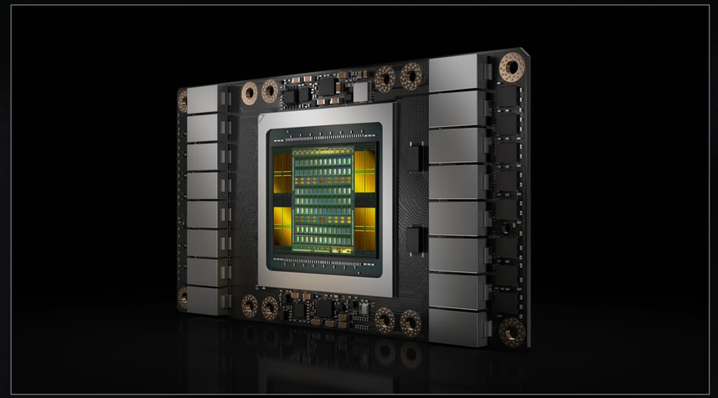
600 Accelerated Applications

NVIDIA Powers World's Fastest Supercomputer

Summit Becomes First System to Scale the 100 PetaFLOPS Milestone

122 PF HPC | 3 EF AI

=



27,648 Volta V100 Tensor Core GPUs

NVIDIA Powers Fastest Supercomputers in US, Europe, Japan, Industry

17 of World's 20 Most Energy-Efficient Supercomputers

ORNL Summit
World's Fastest
27,648 GPUs | 122 PF

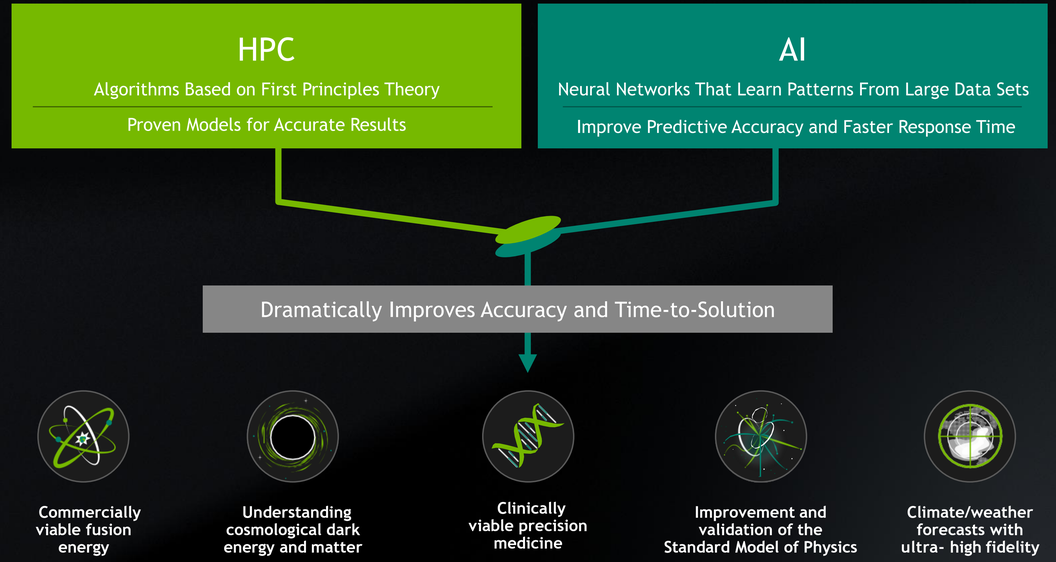
LLNL Sierra
US 2nd Fastest
17,280 GPUs | 72 PF

ABCI
Japan's Fastest
4,352 GPUs | 20 PF

Piz Daint
Europe's Fastest
5,320 GPUs | 20 PF

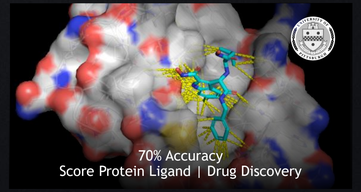
ENI HPC4
Fastest Industrial
3,200 GPUs | 12 PF

AI — A New Instrument for Science



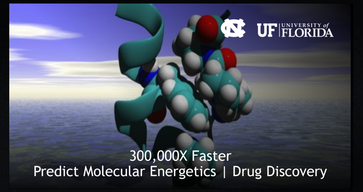
AI for Science

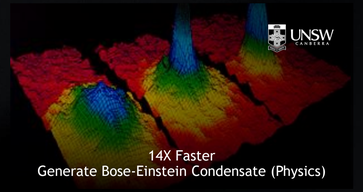
Transformative Tool to Accelerate the Pace of Scientific Innovation



Improves Accuracy

Enabling Realization of Full Scientific Potential





Accelerates Time to Solution

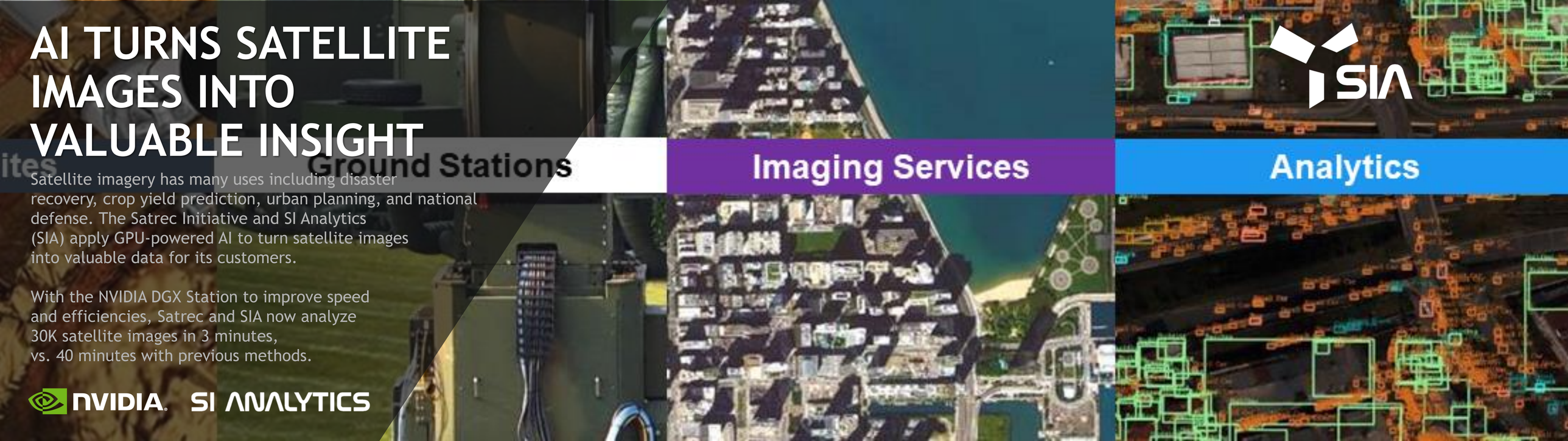
Unlocking Science in Exciting New Ways

AI TURNS SATELLITE IMAGES INTO VALUABLE INSIGHT

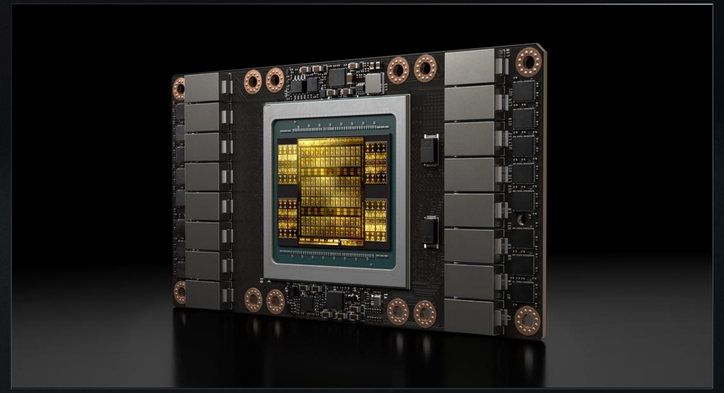
Satellite imagery has many uses including disaster recovery, crop yield prediction, urban planning, and national defense. The Satrec Initiative and SI Analytics (SIA) apply GPU-powered AI to turn satellite images into valuable data for its customers.

With the NVIDIA DGX Station to improve speed and efficiencies, Satrec and SIA now analyze 30K satellite images in 3 minutes, vs. 40 minutes with previous methods.

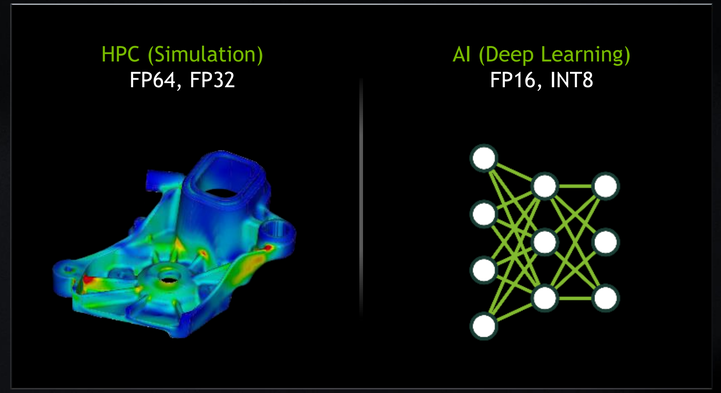
NVIDIA SI ANALYTICS



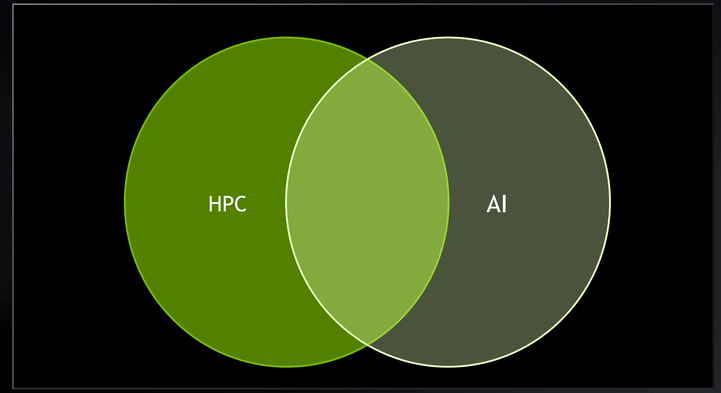
Tensor Core GPU Fuses HPC & AI Computing



Volta Tensor Core GPU



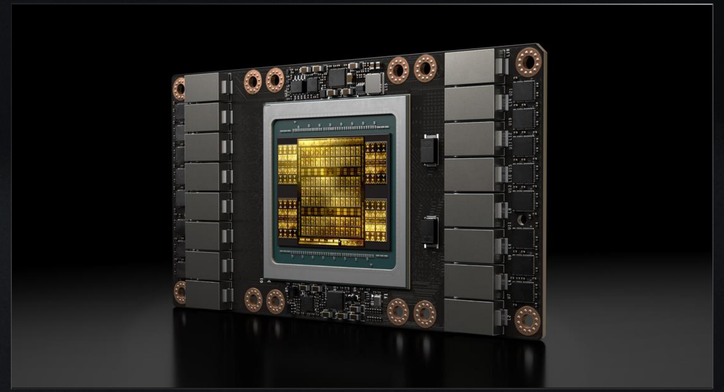
Multi-Precision Computing



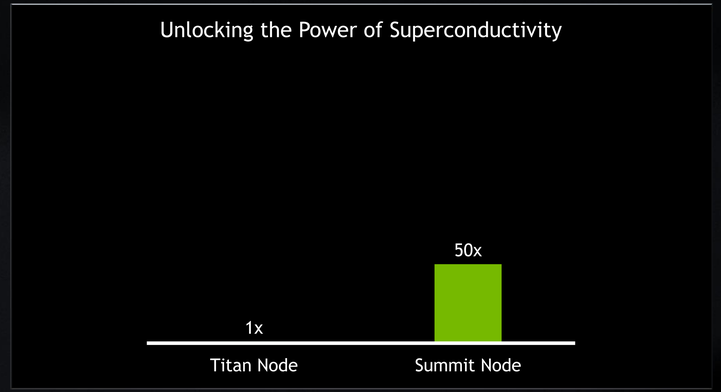
Fusion of HPC & AI

Tensor Core GPU Delivering Breakthrough Performance

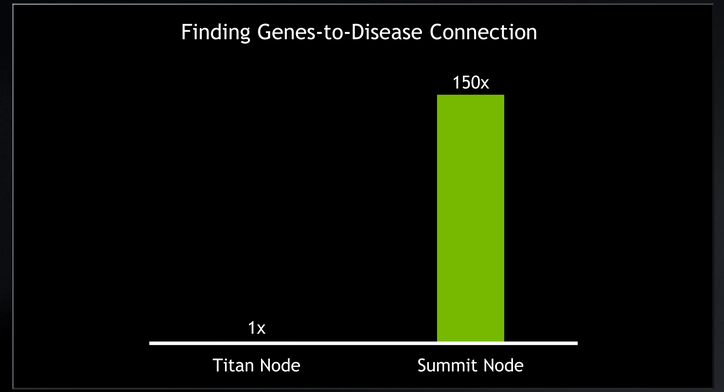
Multi-Precision Computing for Advancing Science



Volta Tensor Core GPU



Materials APP QMCPack (FP64, FP32)



Genomics APP CoMet (FP16)

AI IMPROVES QUALITY AND PRODUCTION YIELD

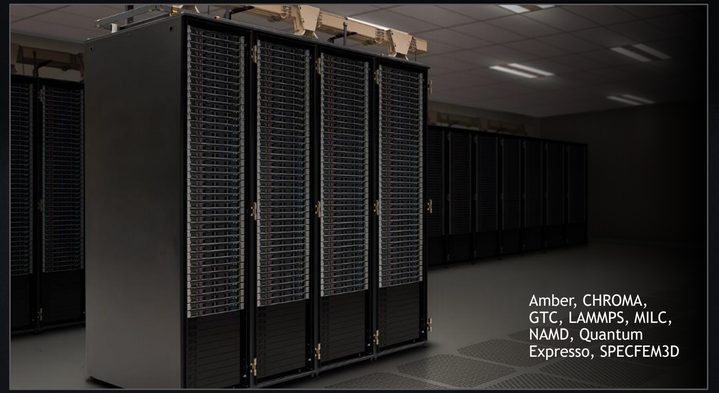
Delivering products of impeccable quality is a great opportunity for manufacturers to differentiate, but it raises the bar for detecting the smallest product defects. LG Consulting and Solutions (LG CNS) is using AI to identify product defects for the entire LG Electronics and LG Display production lines.

With NVIDIA Tesla P4, Jetson TX2, DGX-1 and TensorRT to speed training and inference, LG CNS achieved 1.5% higher yield, 65% quality improvement, and reduced inspection-induced employee fatigue.

NVIDIA LG CNS

Reduced Cost, Space, Power

5X Better HPC TCO for Same Throughput



Mixed HPC Workload

160 Self-hosted Skylake CPU Servers | 96 KWatts

=

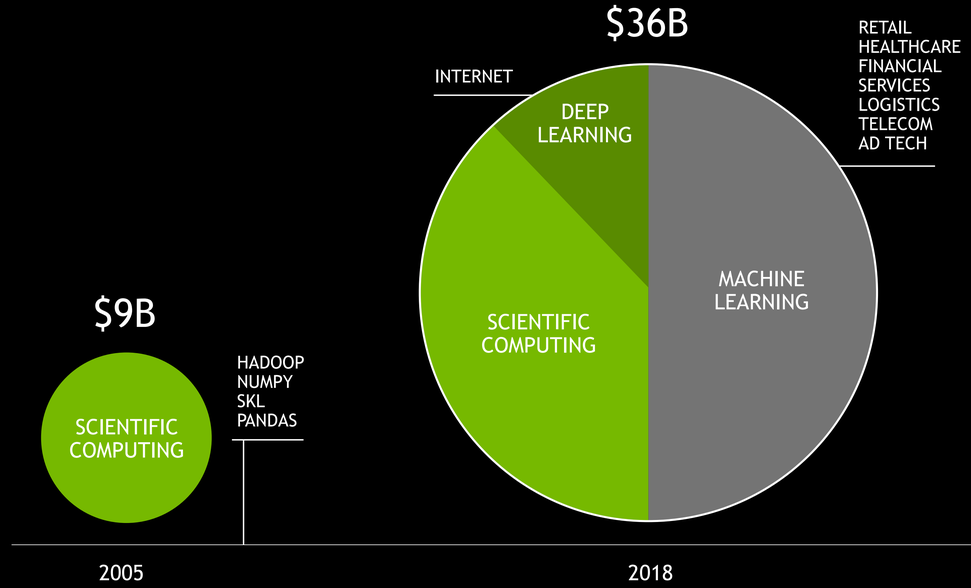
Mixed HPC Workload

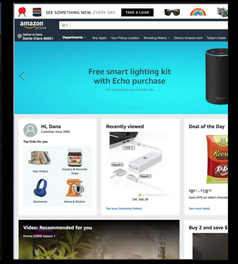
8 Accelerated Servers with 4 V100 GPUs | 13 KWatts

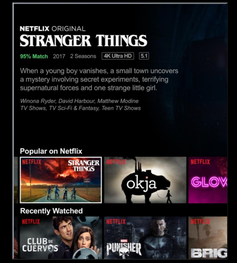
1/5 the Cost 1/7 the Space 1/7 the Power

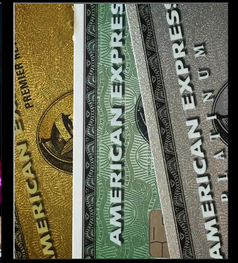
Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

The New HPC Market

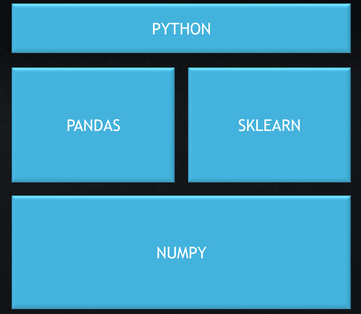








The Defacto Data Science Platform



python™

1991 | Guido van Rossum
Interpreted language emphasizing readability

NumPy

2006 | Travis Oliphant
Multi-dimensional arrays, math functions

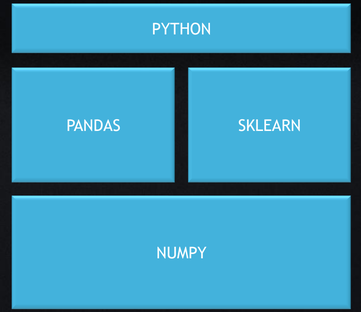
pandas

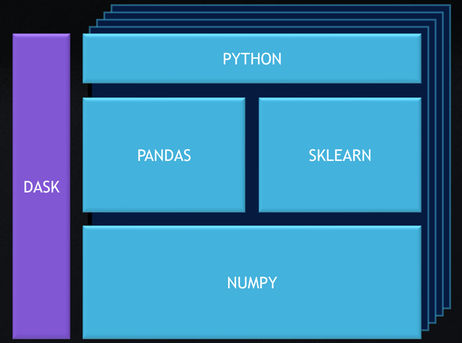
2008 | Wes McKinney
Data manipulation and analysis

scikit learn

2010 | Inria
Machine learning library

The Defacto Data Science Platform



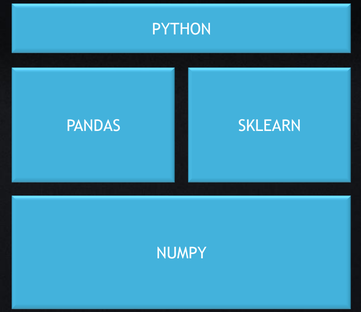


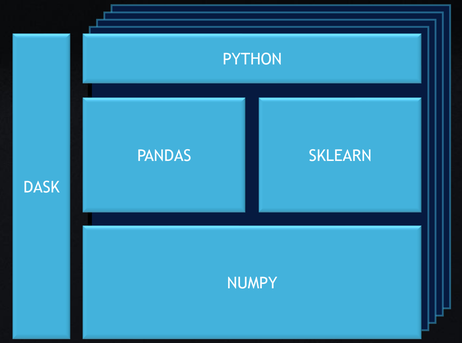
DASK

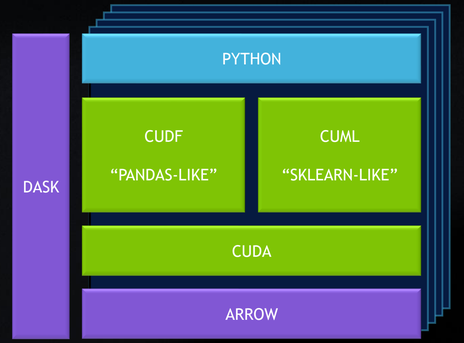
Matthew Rocklin

Parallel processing in Python data analytics
Dynamic task scheduling
Collection of parallel arrays, data frames, lists

RAPIDS — Accelerated Data Science





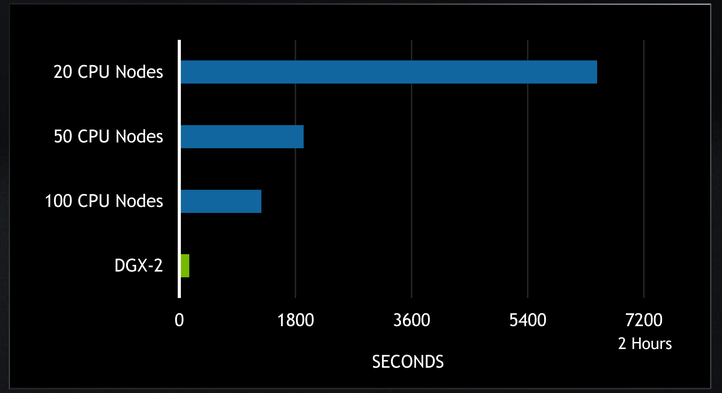


2016 | Wes McKinney

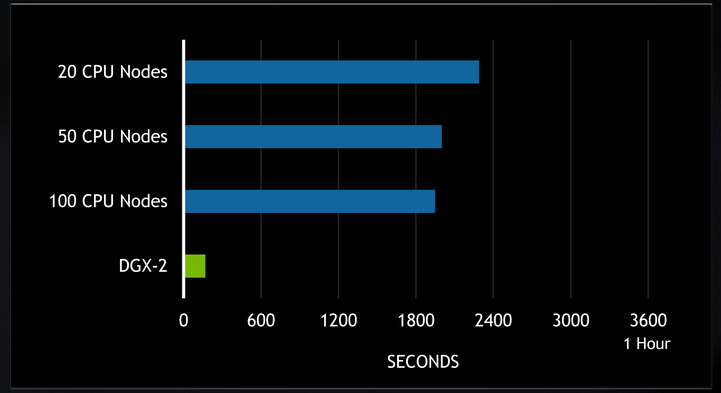
Cross-language platform for in-memory data
Columnar memory format
Vectorized execution engine
Zero-copy IPC
Designed with GPU in mind

RAPIDS — Accelerated Data Science

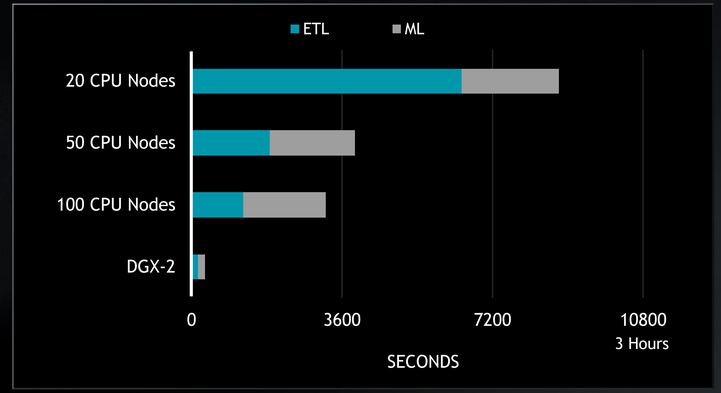
RAPIDS: Dramatic ML Acceleration



ETL

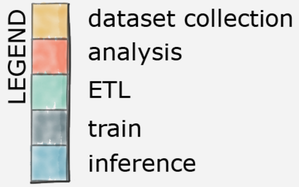


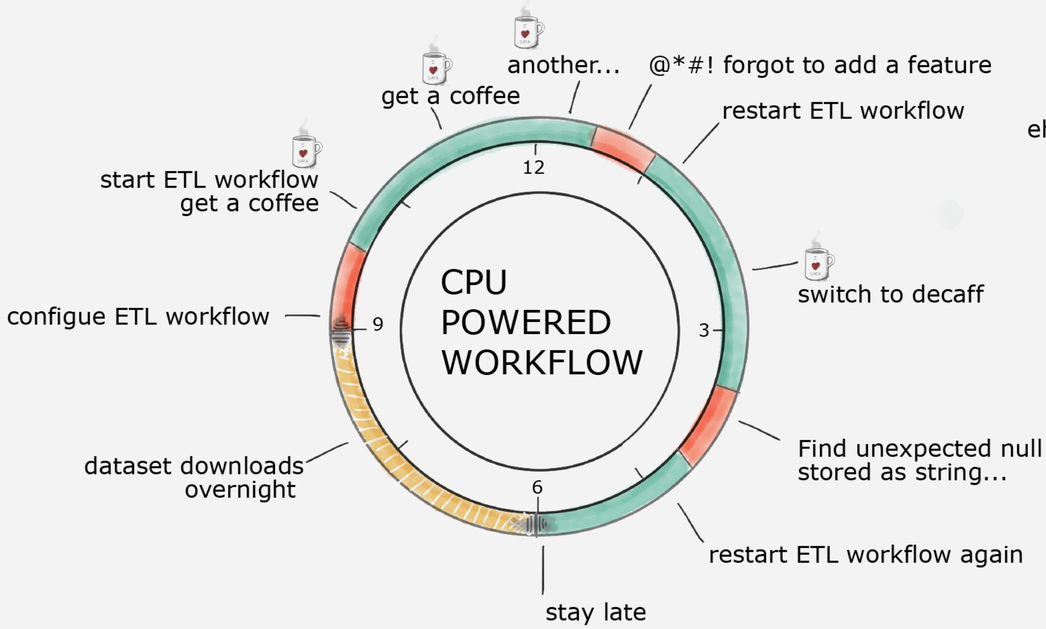
ML

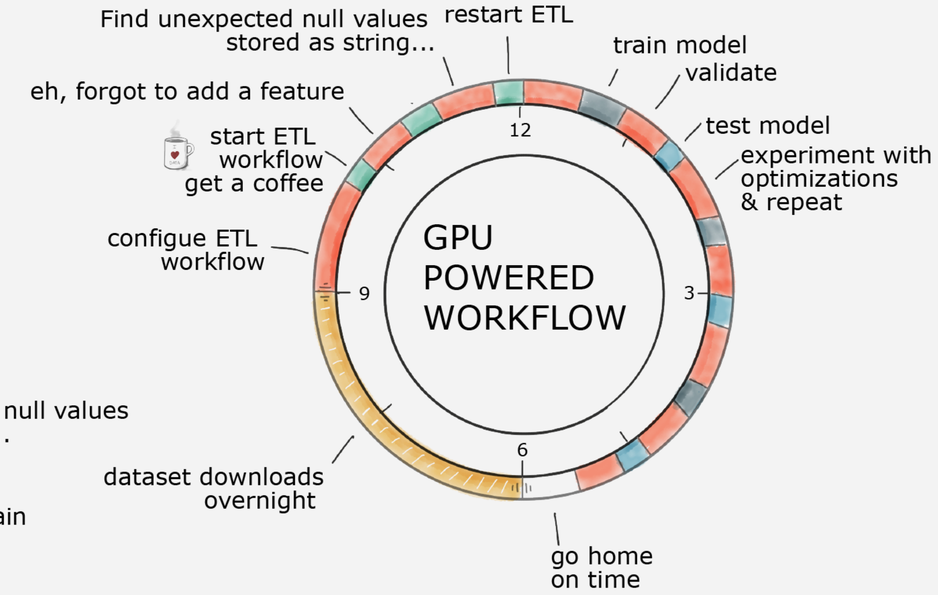


End-to-End

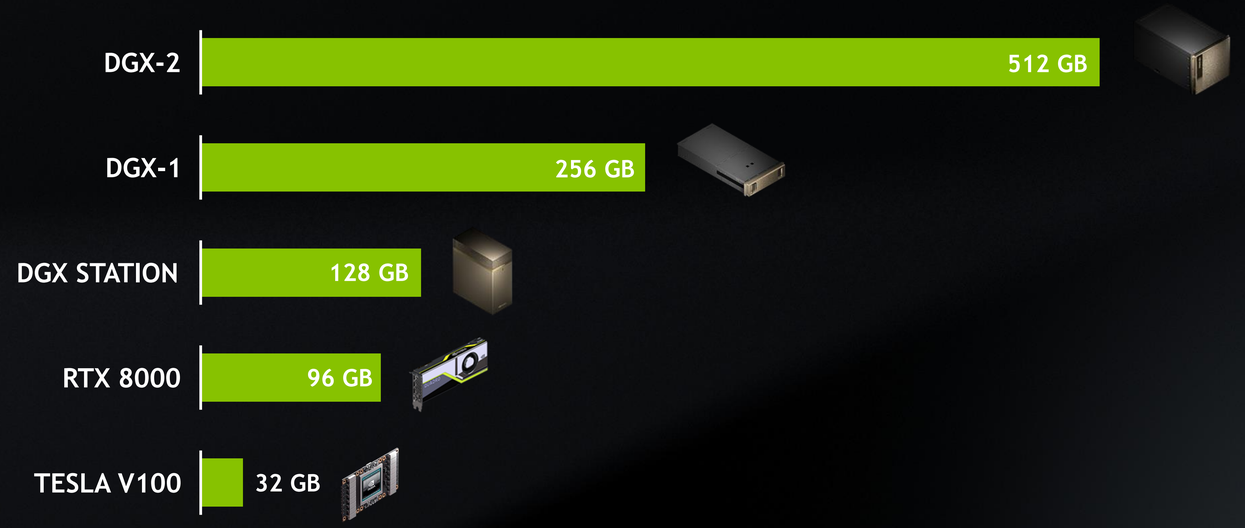
DAY IN THE LIFE OF A DATA SCIENTIST







Enterprise-Scale Data Science

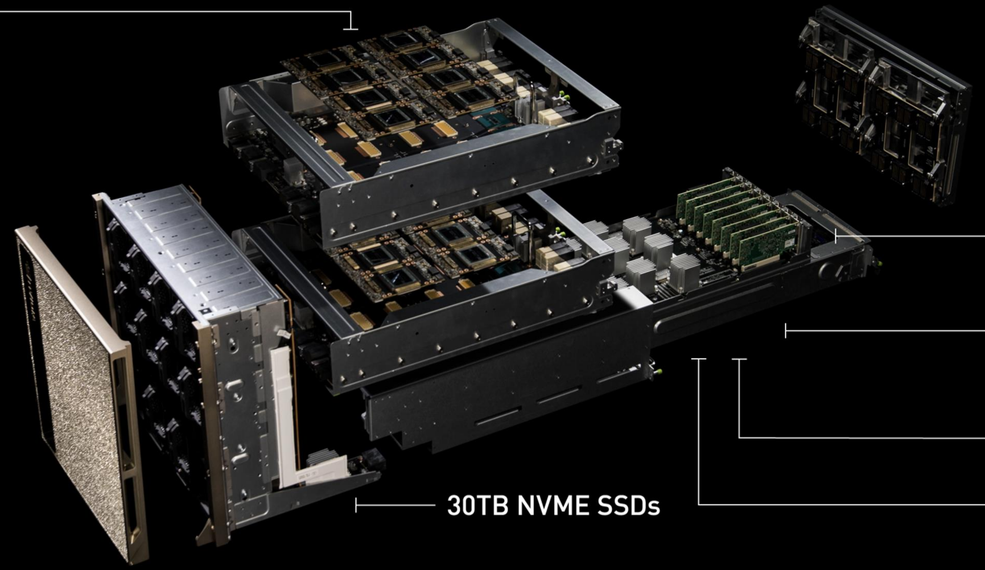


Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

New
NVIDIA DGX-2

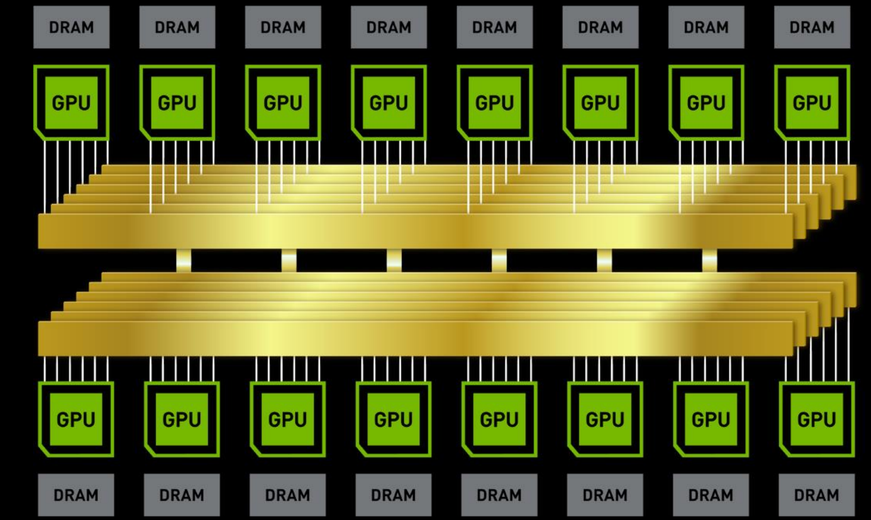
The Largest GPU Ever Created

2 PFLOPS
512 GB HBM2
16 TB/sec Memory Bandwidth
10 kW | 160 kg



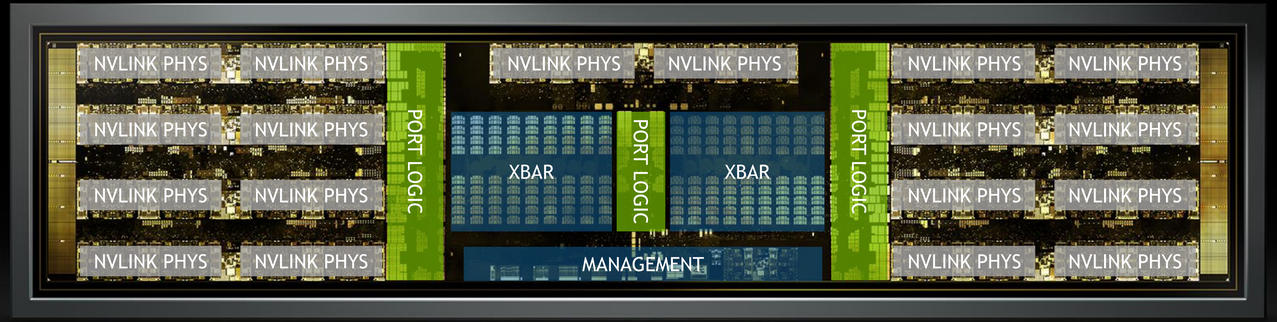
“The World’s Largest GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-Chip Memory Fabric Semantic Extended Across All GPUs
512 GB HBM2 and 14.4 TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



NVSwitch

Parameter	Spec
Bidirectional Bandwidth per NVLink	51.5 GB/s
NRZ Lane Rate (x8 per NVLink)	25.78125 Gbps
Transistors	2 Billion
Process	TSMC 12FFN
Die Size	106 mm^2
Bidirectional Aggregate Bandwidth	928 GB/s
NVLink Ports	18
Mgmt Port (Config, Maintenance, Errors)	PCIe
LD/ST BW Efficiency (128B pkts)	80.0%
Copy Engine BW Efficiency (256B pkts)	88.9%



Traditional Machine Learning Cluster

300 Servers
$3M
180 kW

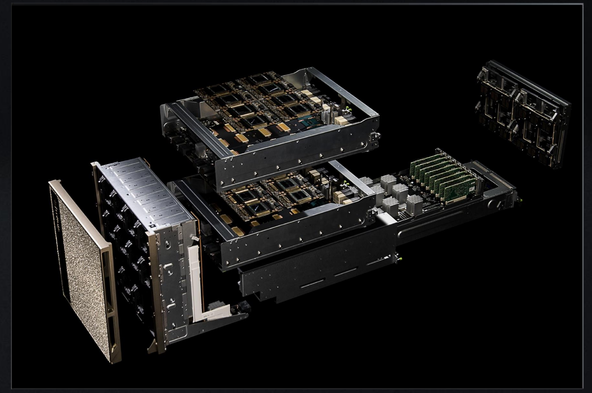
GPU-Accelerated Machine Learning Cluster

DGX-2 and Rapids for Predictive Analytics

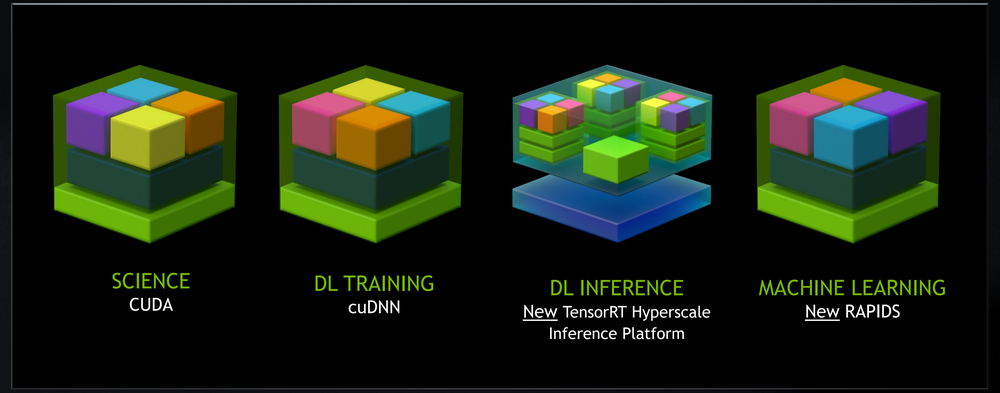
1 DGX-2
10 kW

1/8 the Cost 1/15 the Space 1/18 the Power

NVIDIA Accelerated HPC Platform



Dense HPC



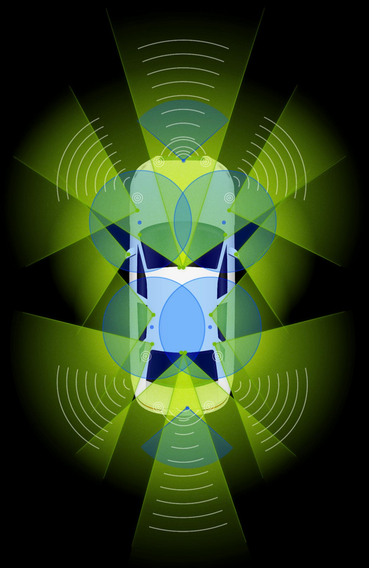
NVIDIA HPC Acceleration Stacks

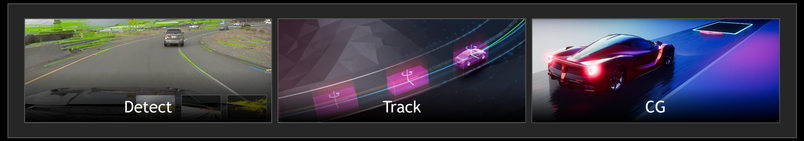
Hyperscale HPC

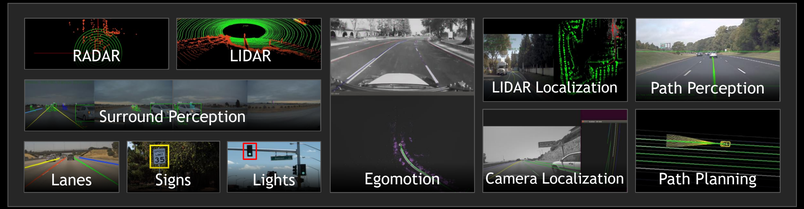
Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

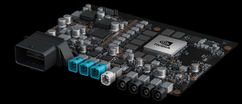
NVIDIA DRIVE
Software-Defined Car

Powerful and Efficient for AI, CV, AR, HPC
Rich Software Development Platform
370+ Partners Developing on DRIVE





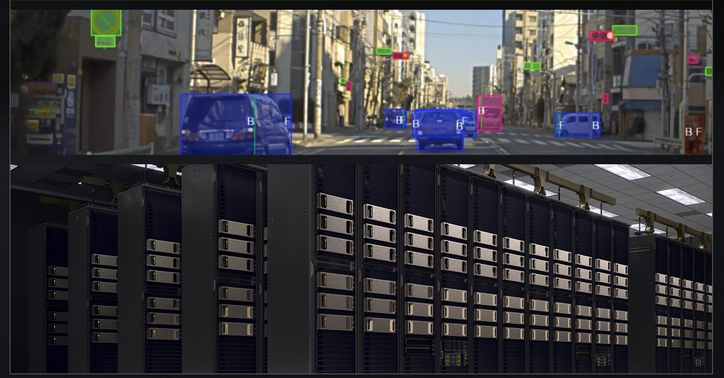




DRIVE AGX Xavier
DRIVE AGX Pegasus

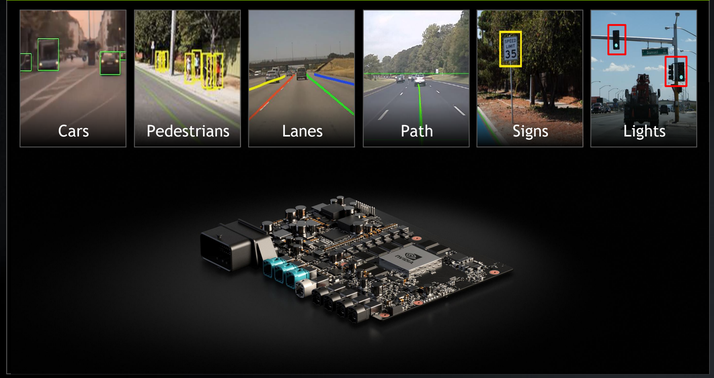
NVIDIA DRIVE

TRAINING



SIMULATING

DRIVING



NVIDIA AGX PEGASUS TEST DRIVE

OCTOBER 2, 2018

80 KILOMETERS

4 HIGHWAY INTERCHANGES

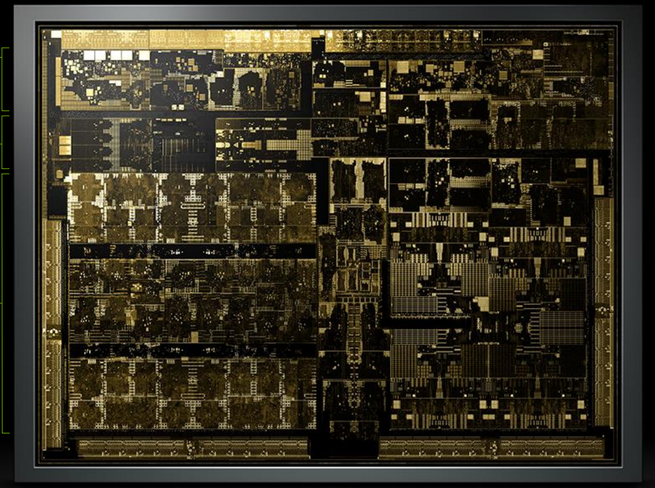
10 LANE CHANGES

0 DISENGAGEMENTS

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

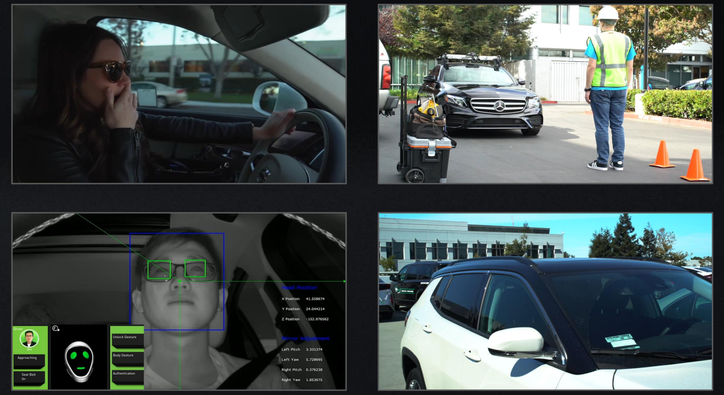
ISP
2.4 GPIX/s
Native Full-Range HDR
Tile-Based Processing

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

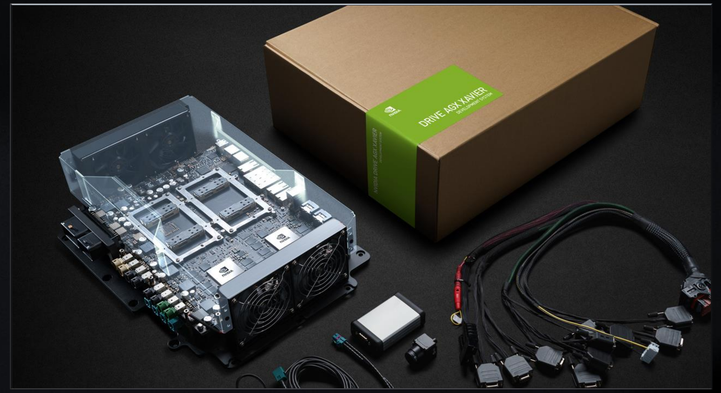
Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6

256-Bit LPDDR4X
137 GB/s

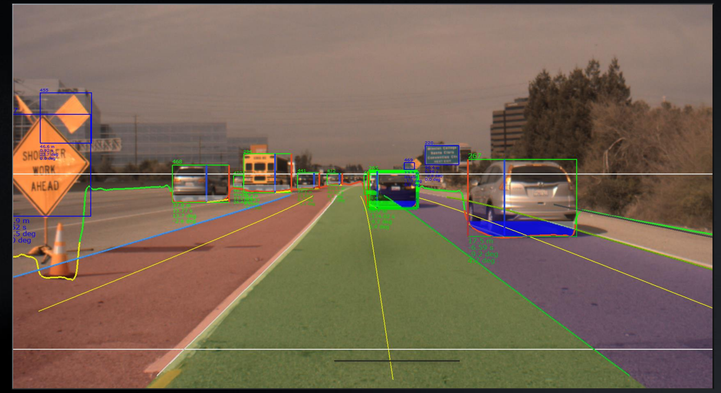
NVIDIA DRIVE — World's First Autonomous Vehicle Platform



DRIVE IX
Available Now



DRIVE AGX Xavier Developer Kit
Available Now



DRIVE AV
Available Now

Accelerating Deep Learning
Accelerating Graphics
Accelerating Science
Accelerating Data Science
Scaling
Accelerating Autonomous Vehicles
Accelerating Robotics
Conclusion

NVIDIA Isaac

SENSOR PROCESSING

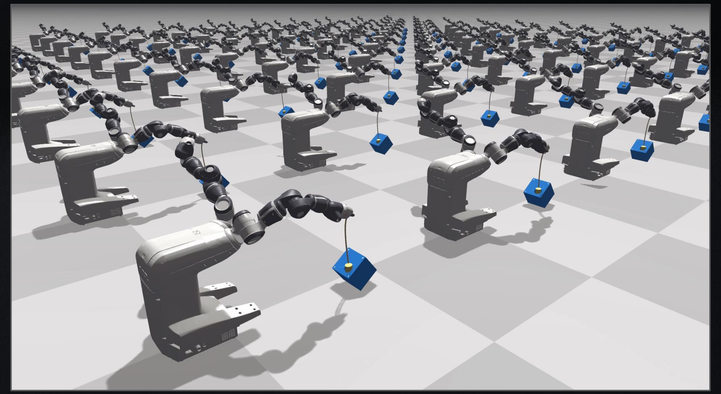
MAPPING & LOCALIZATION | PATH & TASK PLANNING

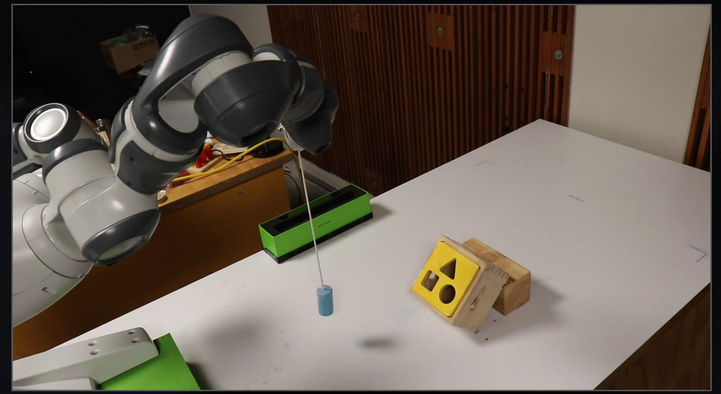
PERCEPTION | SITUATION UNDERSTANDING

DIVERSITY & REDUNDANCY

Efficient Learning of Robust Policies

Randomize Physical Parameters to Match Real World Rollouts





[Chebotar-Handa-Makoviychuk-Macklin-Ratliff-Fox: 18]

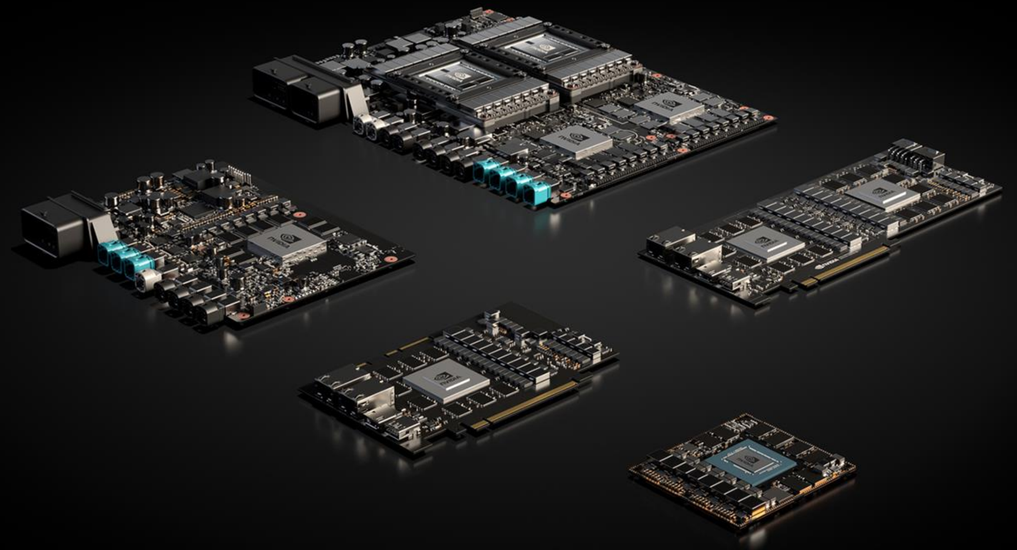
BETTER HUMAN-ROBOT IN TERACTION, BETTER SERVICE ROBOTS

From assisting the elderly to helping with every day chores, service robots hold the promise to make lives easier. But complex crowded environments could mean collisions between robots, humans and objects.
Naver Labs designs autonomous robots with better human interaction. With its obstacle avoidance, AROUND G, built on NVIDIA Jetson Xavier for real-time preprocessing and neural network inference—critical in navigating real world environments—Naver Labs aims to popularize service robots in the coming future.

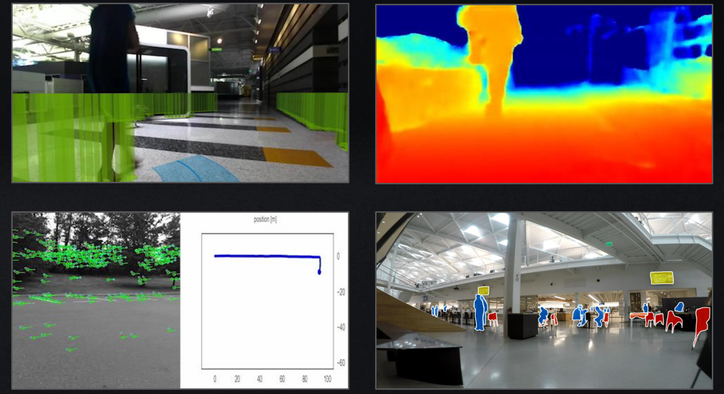
NVIDIA. LABS

New
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 Jetson AGX — World's First Edge AI Computer



Isaac Gems

Jetson AGX Xavier Developer Kit
Available Now

Isaac Sim

AI DELIVERS BUSINESS VALUE

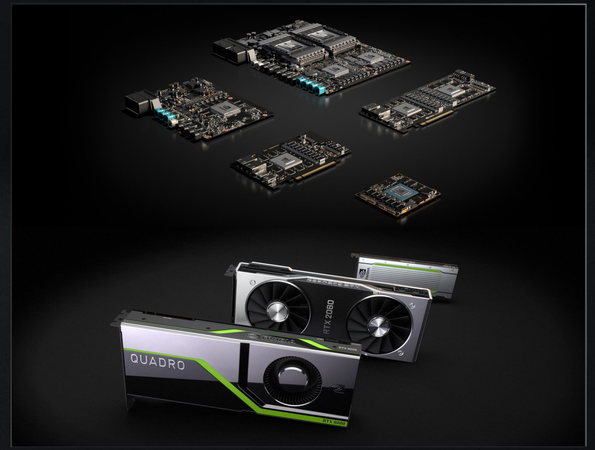
SI ANALYTICS

SK hynix

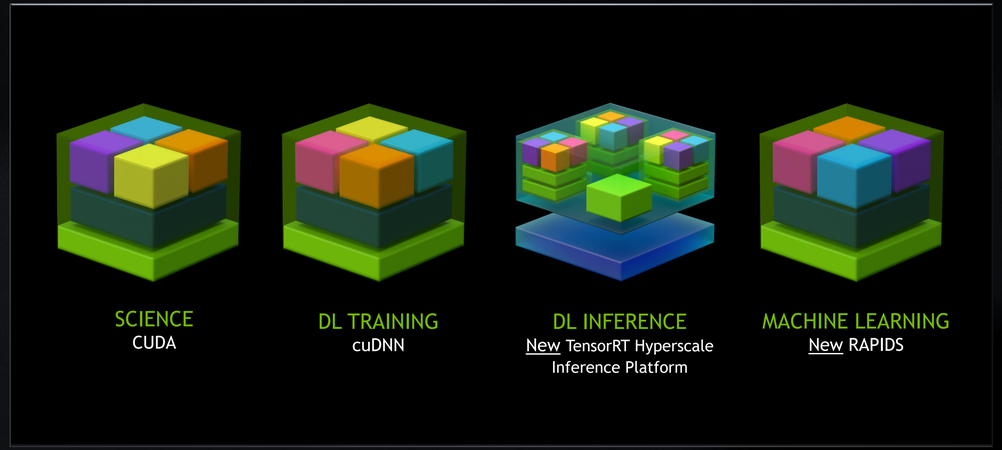
LG CNS

LABS

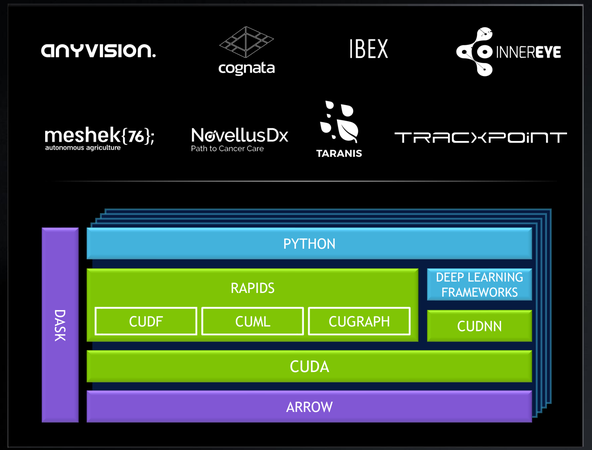
New NVIDIA Platforms



CUDA GPUs



NVIDIA HPC Acceleration Stacks



Ecosystem

NVIDIA