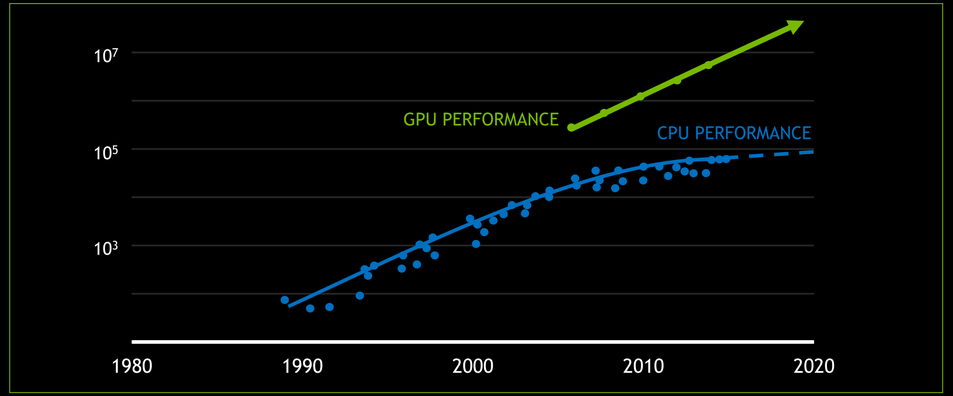
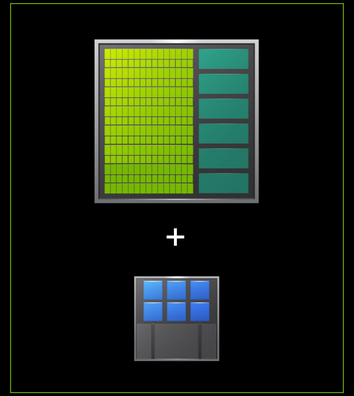
NVIDIA

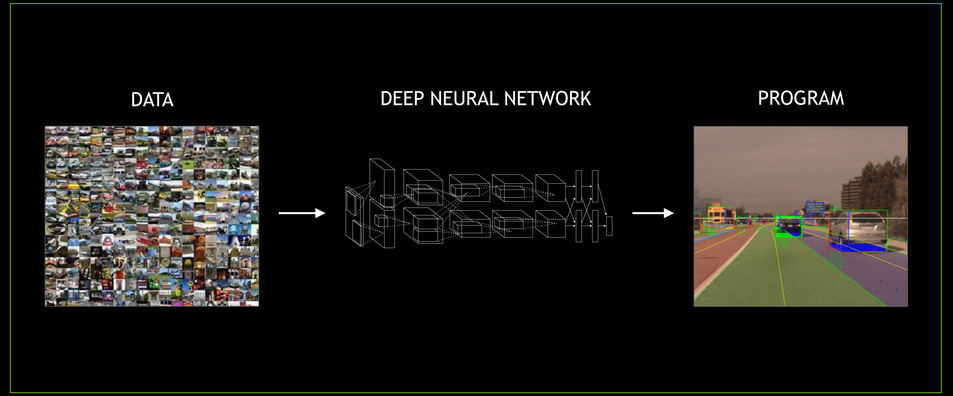
FORCES SHAPING COMPUTING



BEYOND MOORE'S LAW — 1000X EVERY 10 YEARS

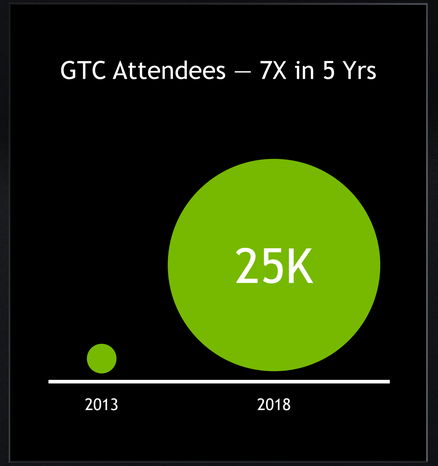


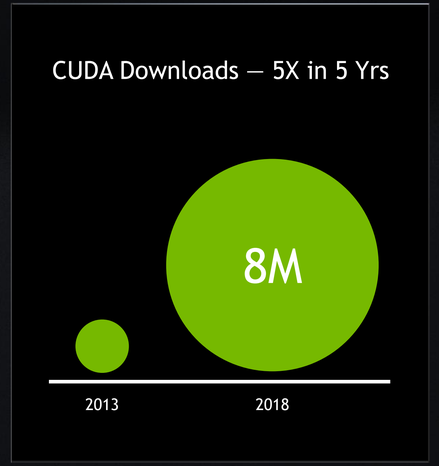
ACCELERATED COMPUTING

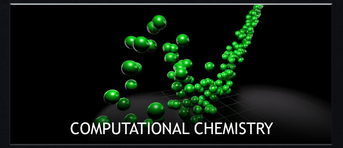


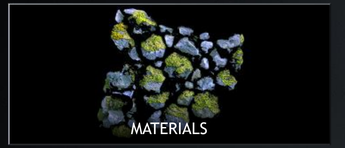
COMPUTERS WRITING SOFTWARE

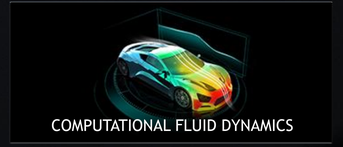
RISE OF GPU COMPUTING

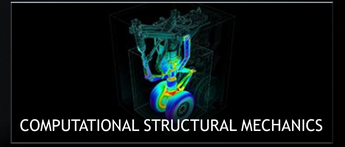


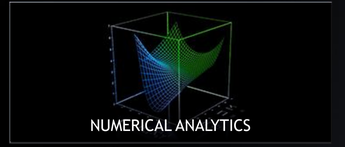


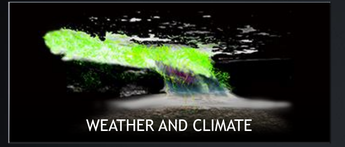




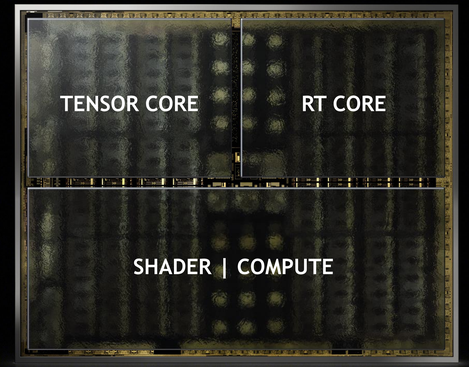


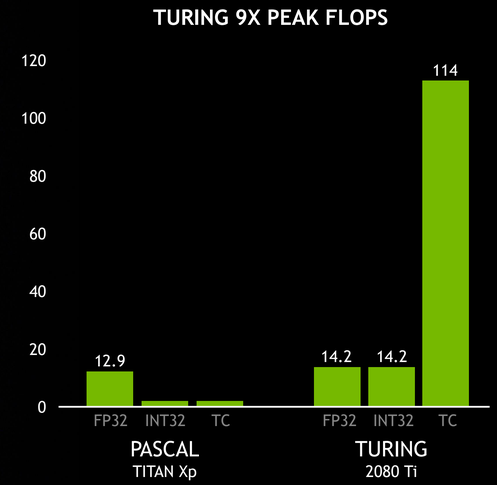






TURING — A GIANT LEAP

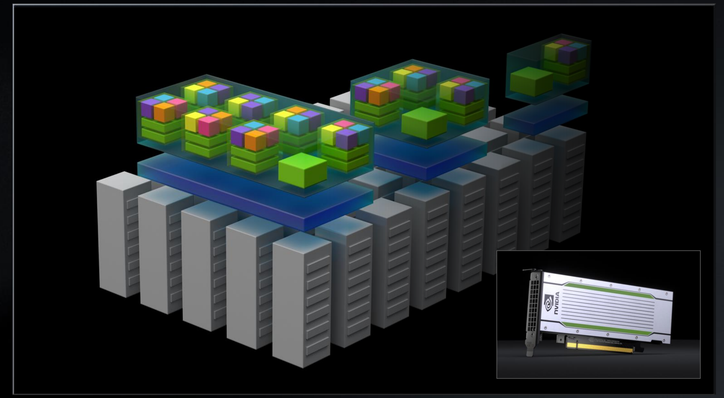




TURING — A GIANT LEAP

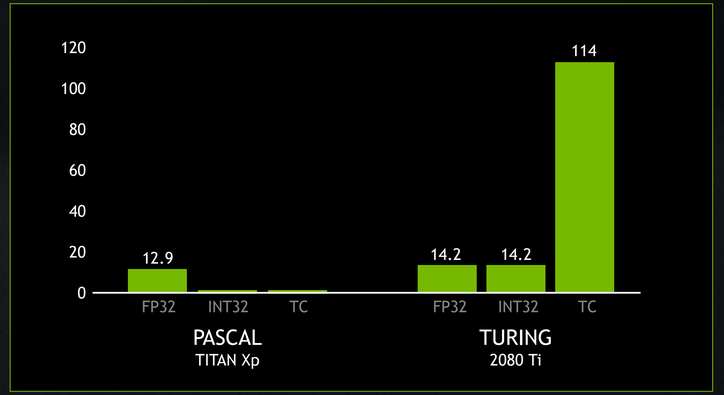
GAMING REINVENTED

WORLD'S FIRST RAY TRACING GPU



UNIVERSAL DEEP LEARNING ACCELERATOR

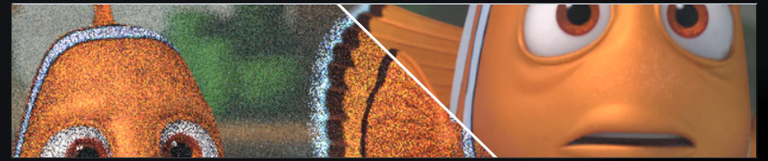
GRAPHICS REINVENTED



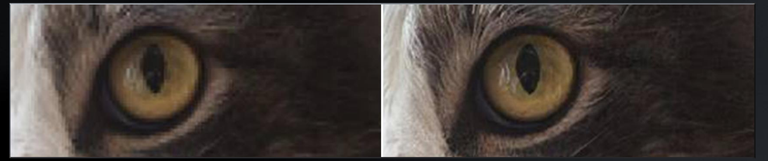
Turing — 9X Peak FLOPS

COLORIZING | UC Berkeley

IN-PAINTING | NVIDIA

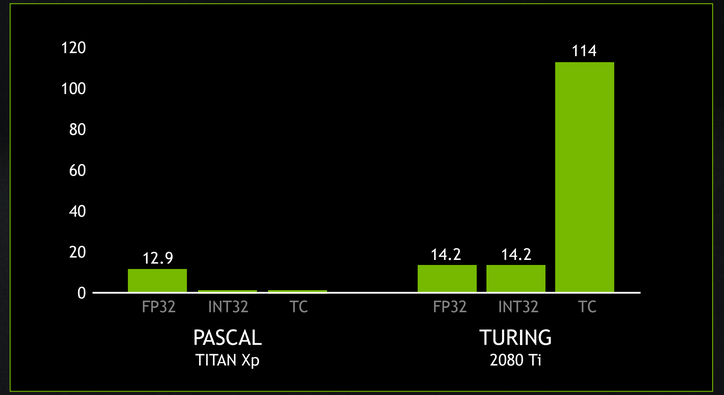


DENOISING | Disney Research, Pixar, UCSB

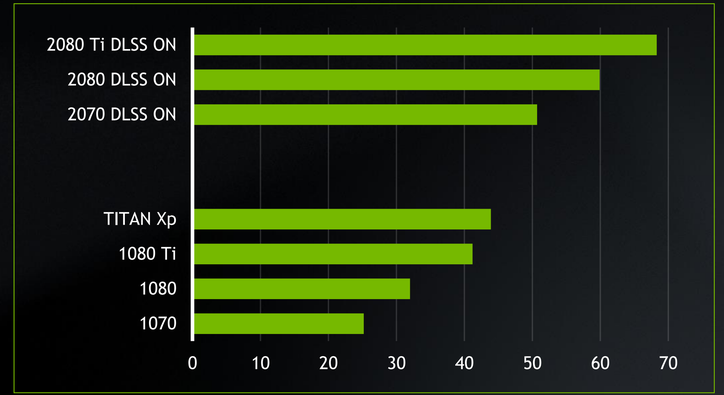


SUPERREZ | NVIDIA

GRAPHICS REINVENTED



Turing – 9X Peak FLOPS

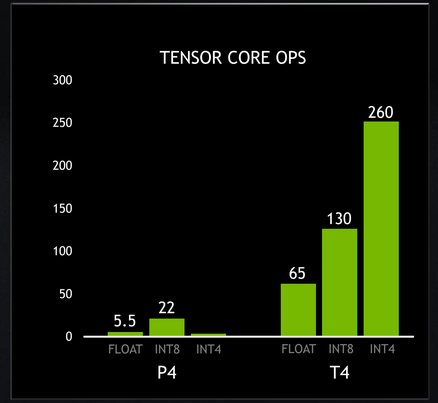


Hybrid Rendering with Tensor Core

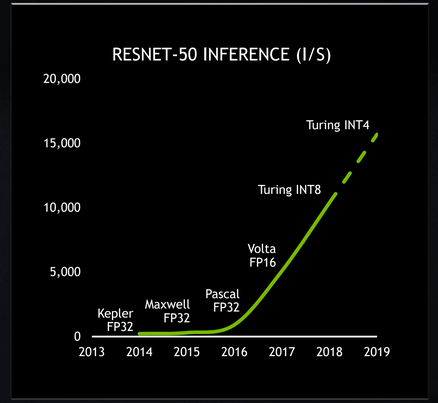
70
PORSCHE

UNIVERSAL DEEP LEARNING ACCELERATOR

Tesla T4

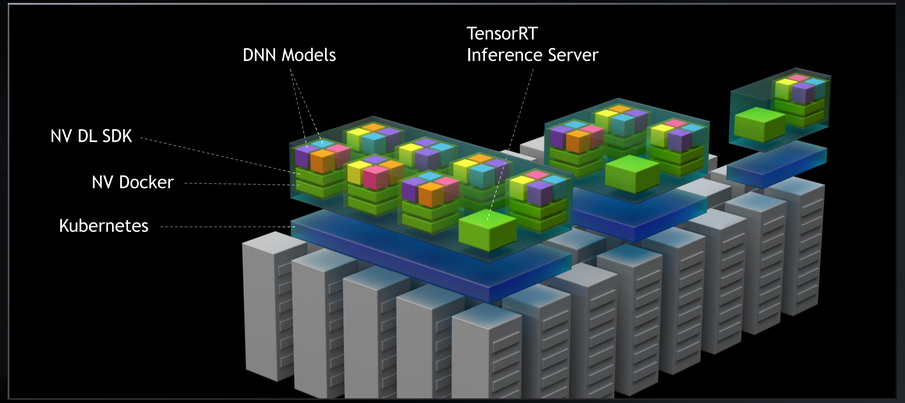


Multi-precision Tensor Core

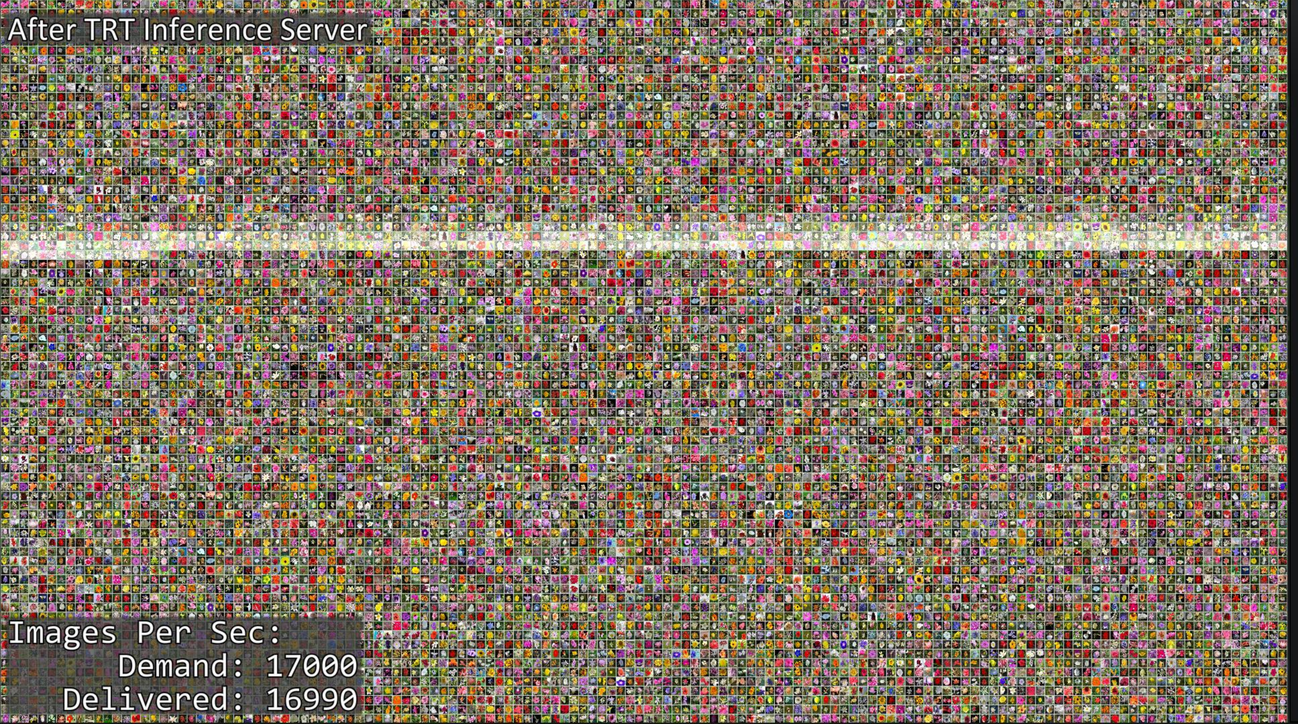


TensorRT 5.0

Support for Tensor Core



TensorRT Hyperscale



After

now-5m to now-4s Refresh every 2s

Inference Per Second By Model

20 K
15 K
10 K
5 K

20:14:00 20:14:30 20:15:00 20:15:30 20:16:00 20:16:30 20:17:00 20:17:30 20:18:00 20:18:30

Utilization

91%

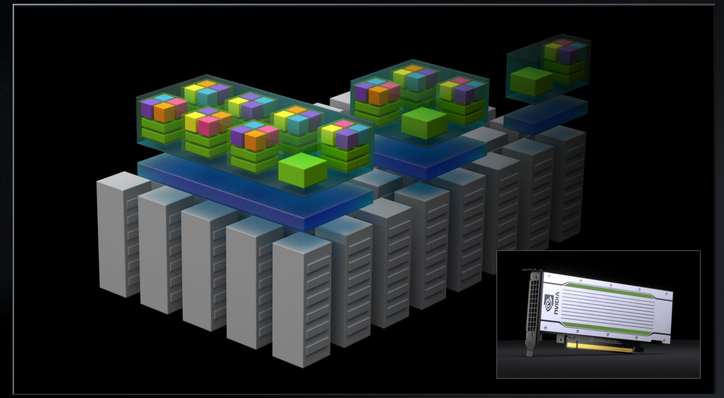
Device Utilization	Device Utilization	Device Utilization	Device Utilization	Device Utilization	Device Utilization	Device Utilization	Device Utilization
92%	92%	87%	93%	91%	91%	94%	91%

Models Used	Models Used	Models Used	Models Used	Models Used	Models Used	Models Used	Models Used
100% 50% 0%	100% 50% 0%	100% 50% 0%	100% 50% 0%	100% 50% 0%	100% 50% 0%	100% 50% 0%	100% 50% 0%

TURING — A GIANT LEAP

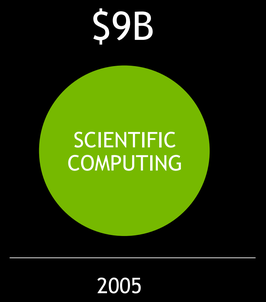
GAMING REINVENTED

WORLD'S FIRST RAY TRACING GPU

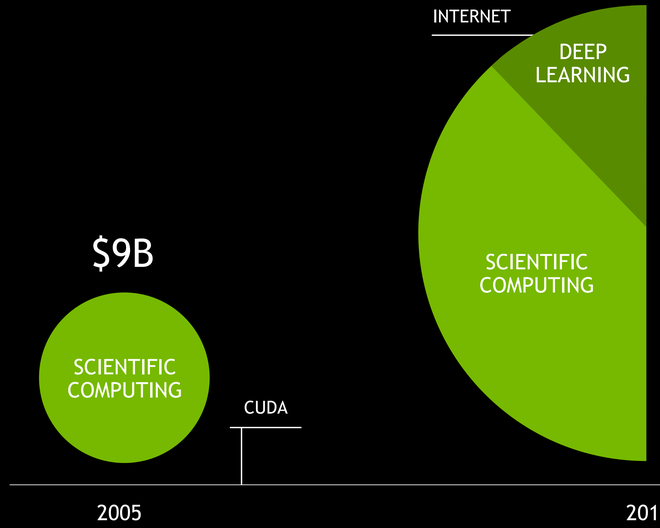


UNIVERSAL DEEP LEARNING ACCELERATOR

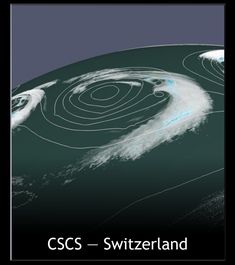
THE NEW HPC MARKET

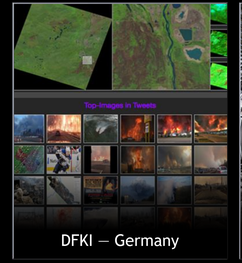


THE NEW HPC MARKET



CERN — Switzerland

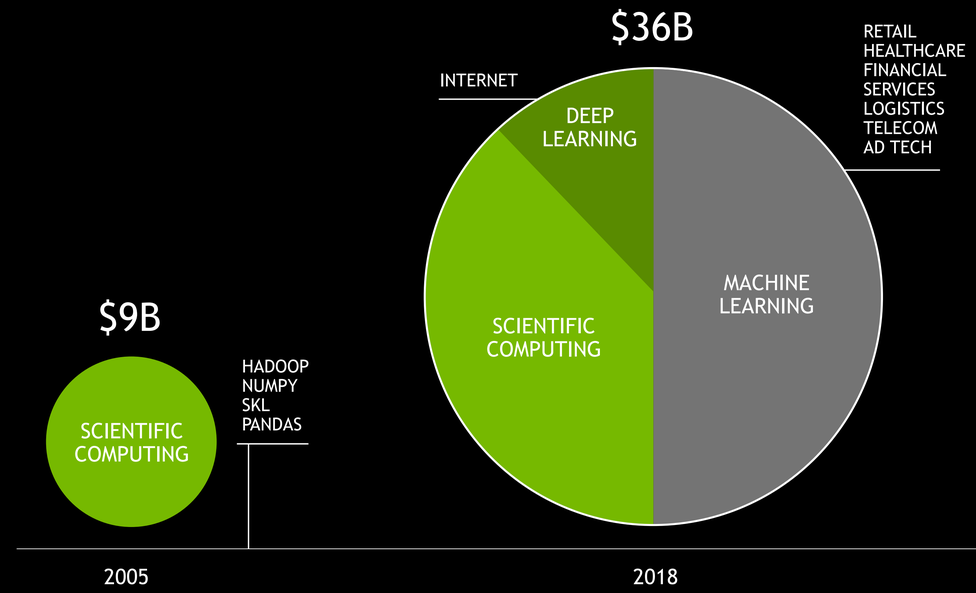

CSCS — Switzerland

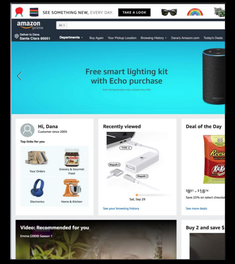


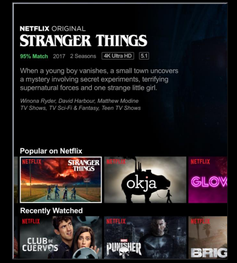
DFKI — Germany

Técnico Lisboa — Portugal

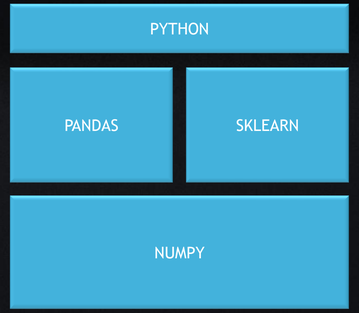
THE NEW HPC MARKET







THE DEFACTO DATA SCIENCE PLATFORM



python™

1991 | Guido van Rossum
Interpreted language emphasizing readability

NumPy

2006 | Travis Oliphant
Multi-dimensional arrays, math functions

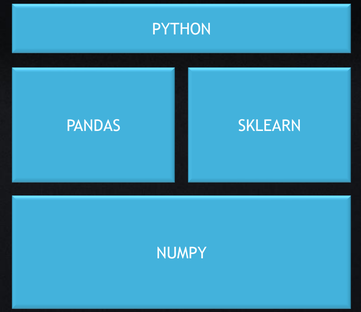
pandas

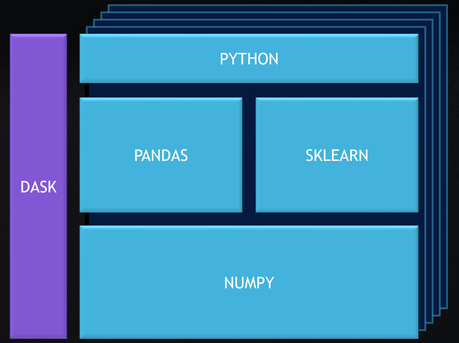
2008 | Wes McKinney
Data manipulation and analysis

scikit learn

2010 | Inria
Machine learning library

THE DEFACTO DATA SCIENCE PLATFORM



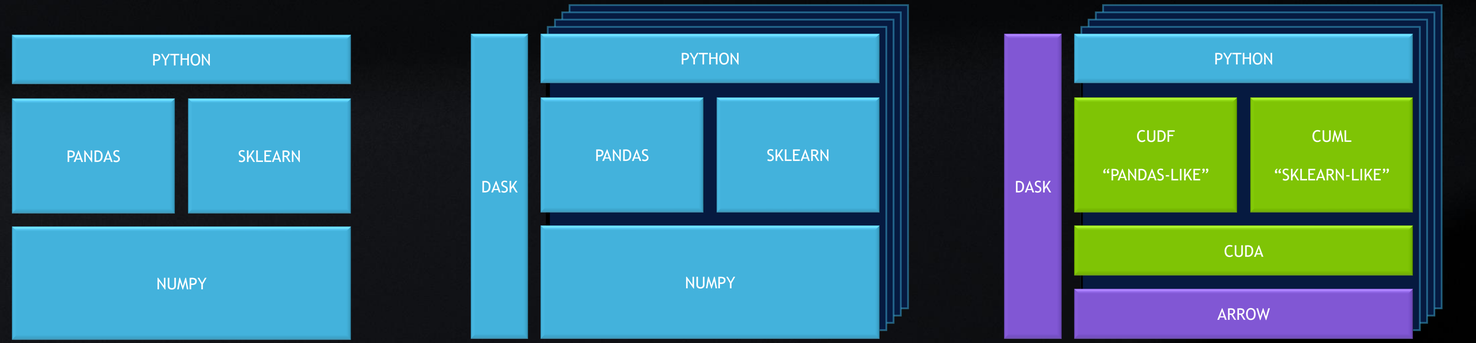


DASK

Matthew Rocklin

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

ANNOUNCING RAPIDS — ACCELERATED DATA SCIENCE

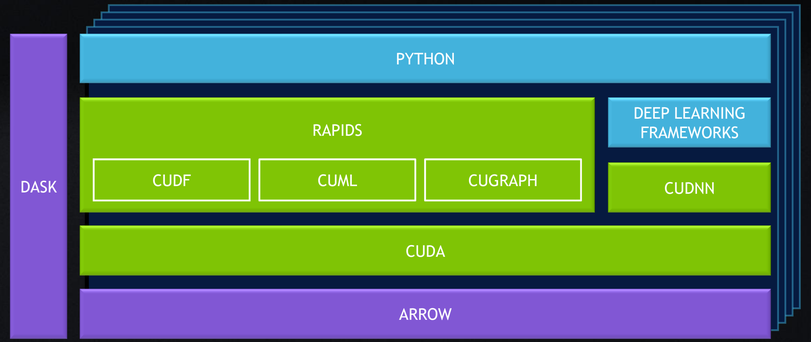


2016 | Wes McKinney

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

ANNOUNCING RAPIDS — ACCELERATED DATA SCIENCE

ANNOUNCING RAPIDS — ACCELERATED DATA SCIENCE



Spark Streaming GPU Support

cuML and cuGraph with Spark 2.4

Batch cuDF with Spark 2.4

APACHE Spark™

2010 | Matei Zaharia, CTO, Databricks

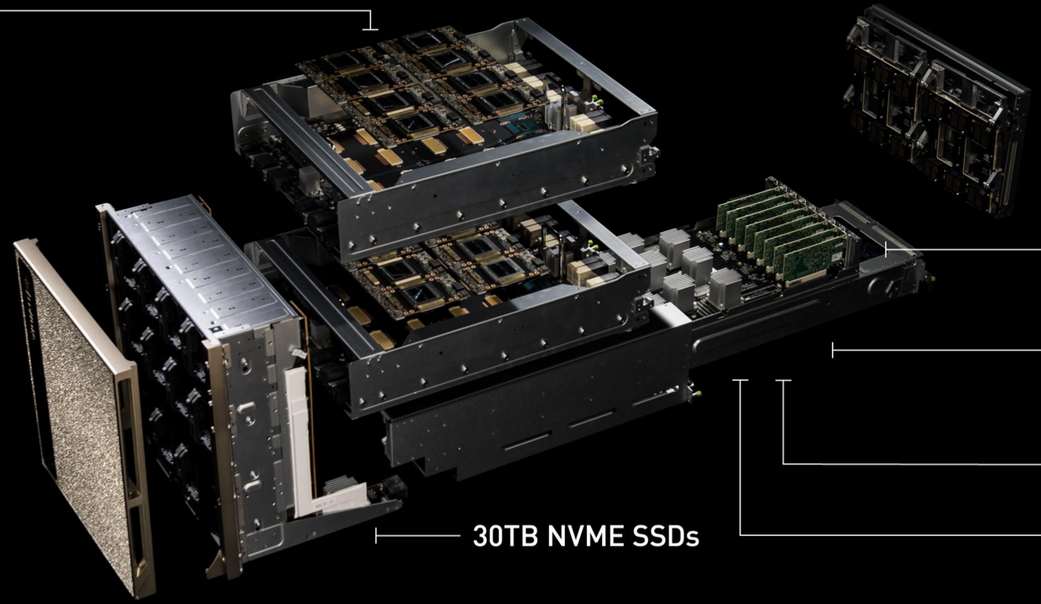
Unified analytics cluster-computing framework for big data and machine learning

ENTERPRISE-SCALE DATA SCIENCE

NEW
NVIDIA DGX-2

THE LARGEST GPU EVER CREATED

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



PLAY AUTO QUERY

OPACITY 100

SCALE 50

DTI BINS 10

RISK BINS 10

CREDIT BINS 10

PORTFOLIO RISK SCORE
- 0.00 - 0.05
- 0.05 - 0.10
- 0.10 - 0.15
- 0.15 - 0.20
- 0.20 - 0.2+

TOTAL UNPAID BALANCE

BAR HEIGHT

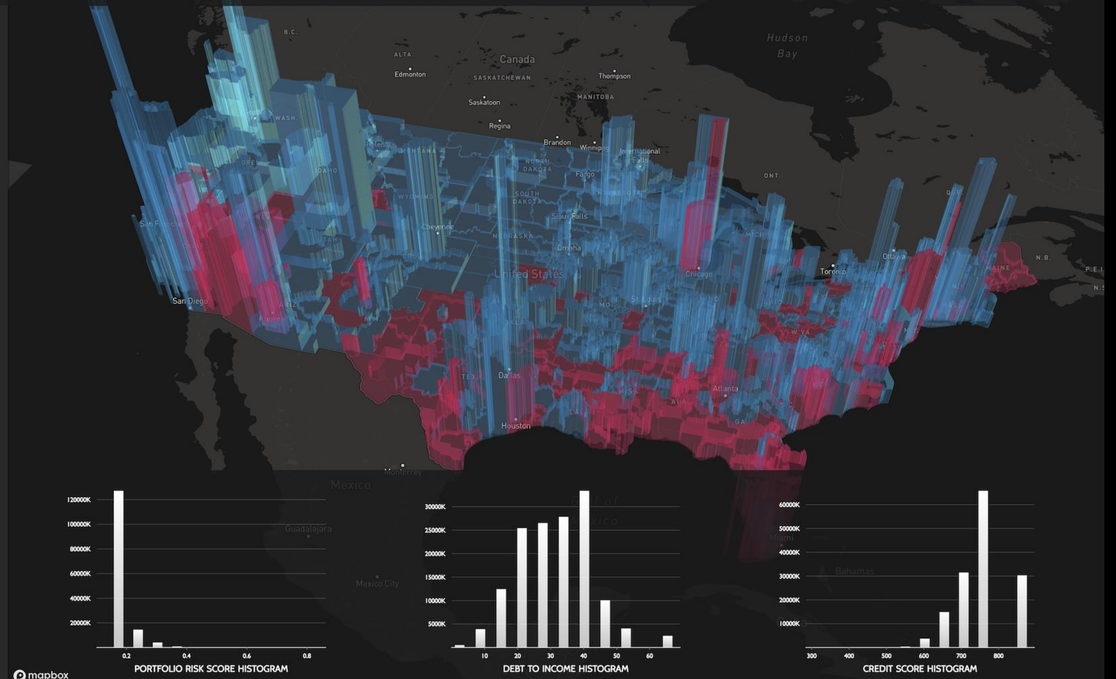
SELECTED BANK

All Banks

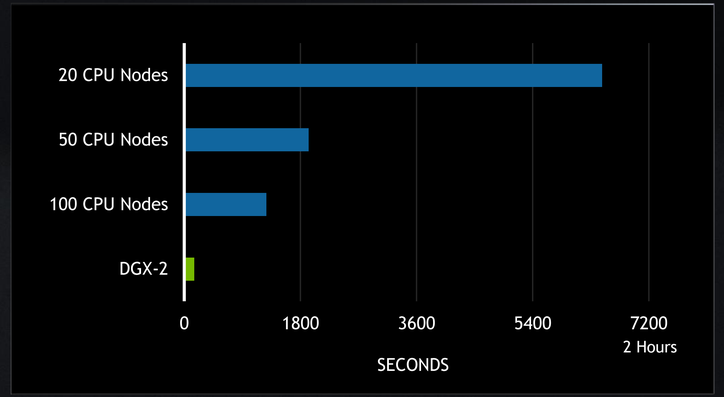
AREA DETAILS

NONE

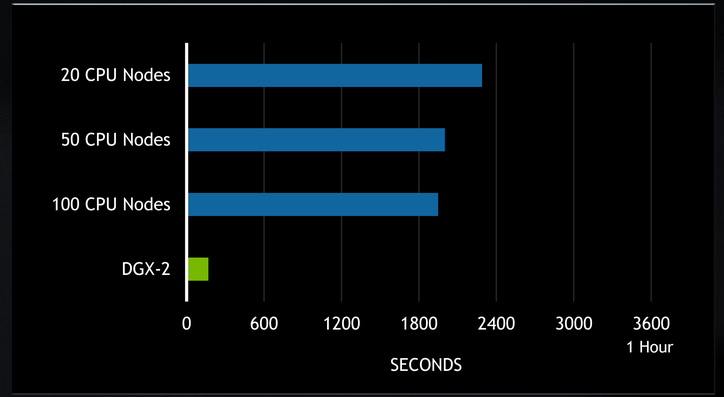
GPU AGGREGATED 100.00% OF 146,594,623 MORTGAGES



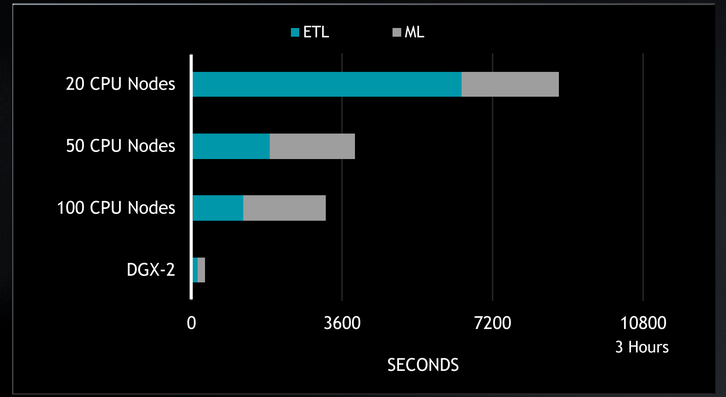
FASTER SPEEDS, REAL WORLD BENEFITS



ETL

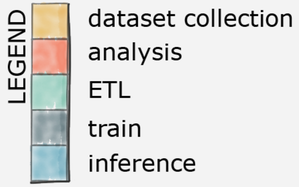


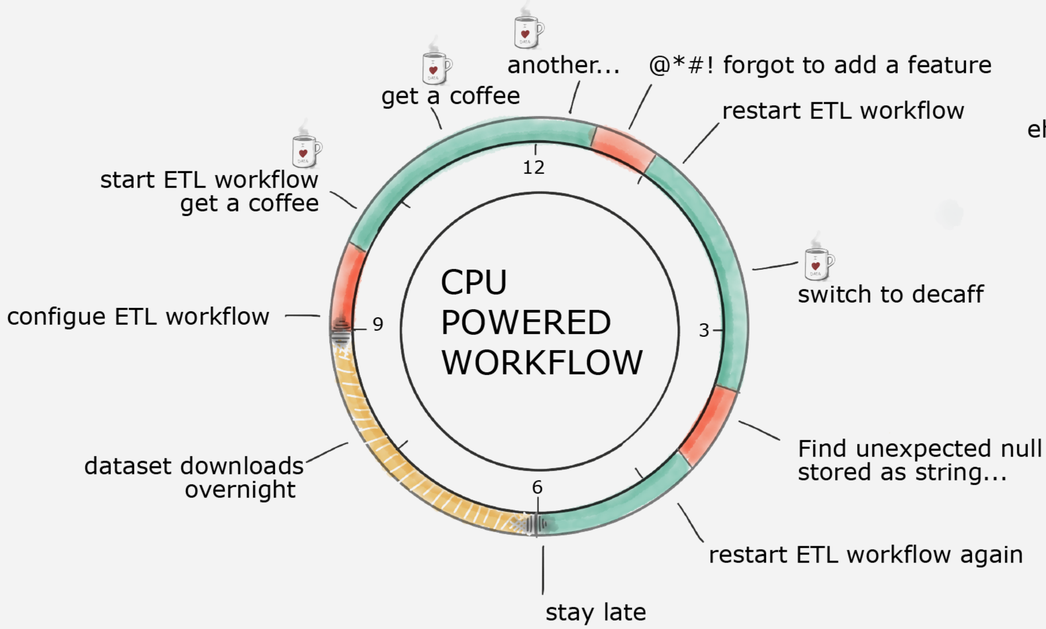
ML

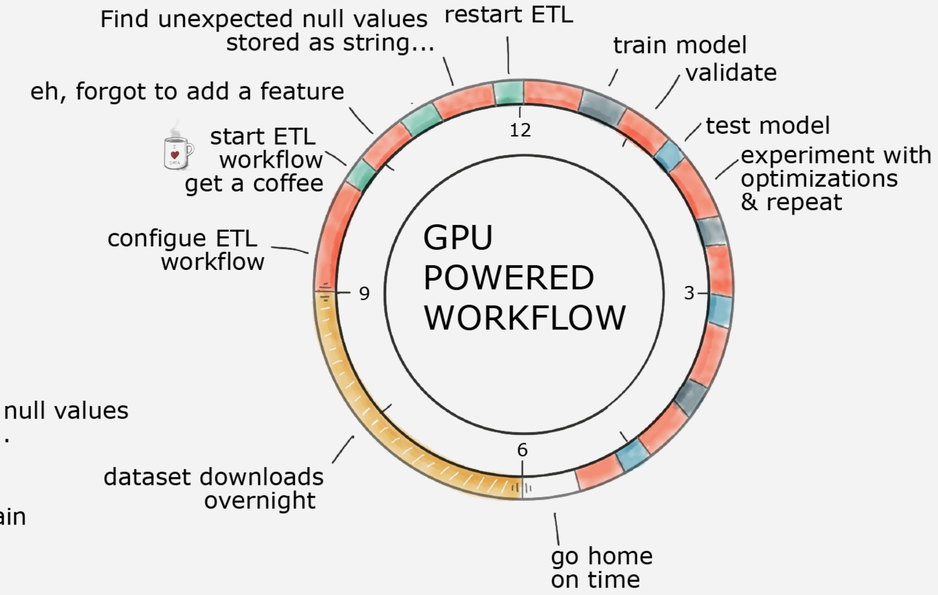


End-to-End

DAY IN THE LIFE OF A DATA SCIENTIST







TRADITIONAL DATA SCIENCE 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

ACCELERATING MACHINE LEARNING WITH RAPIDS

"RAPIDS software has immensely improved how we use data — enabling the most complex models to run at scale and deliver even more accurate forecasting."

- Jeremy King, EVP & CTO

BRINGING RAPIDS-OPTIMIZED SYSTEMS TO ENTERPRISES WORLDWIDE

"We are excited to partner with NVIDIA on RAPIDS to accelerate the application of data science and machine learning to help our customers drive faster and more insightful outcomes."

- Antonio Neri, CEO

Hewlett Packard Enterprise

INTEGRATING RAPIDS TO ACCELERATE DATA SCIENCE AND MACHINE LEARNING PLATFORMS

"We look forward to extending our successful partnership with NVIDIA, leveraging RAPIDS to provide new machine learning tools for our clients."

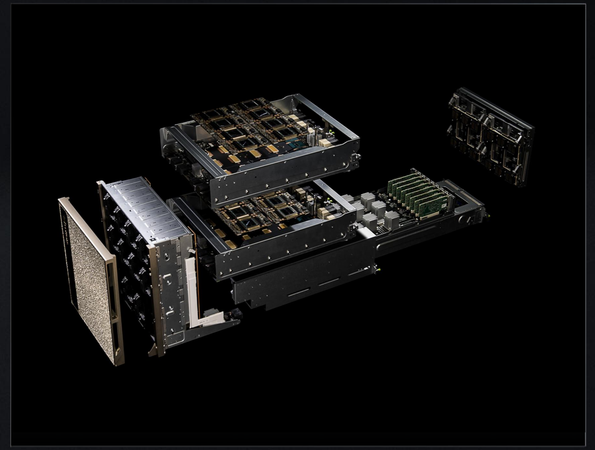
- Arvind Krishna, SVP, Hybrid Cloud and Director of IBM Research

ACCELERATING ORACLE CLOUD INFRASTRUCTURE WITH NVIDIA HGX-2 AND RAPIDS

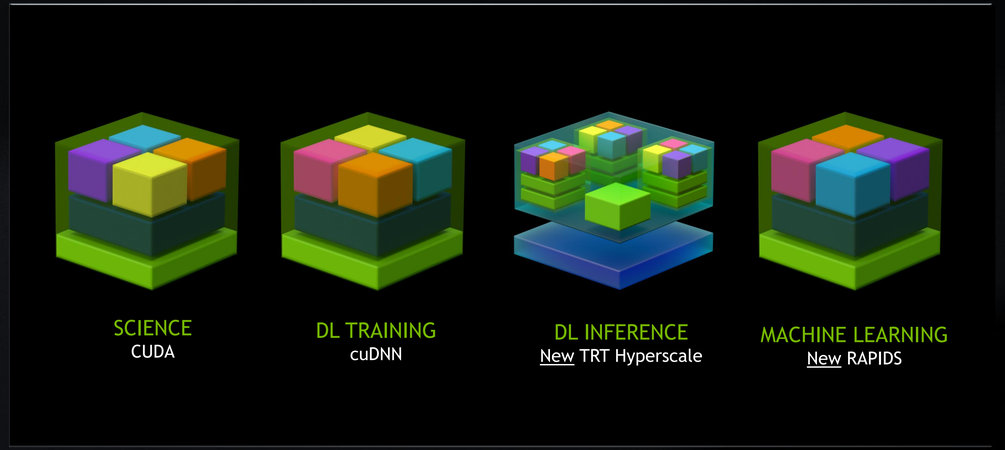
"RAPIDS software runs seamlessly on the Oracle Cloud, allowing customers to support their HPC, AI and data science needs, all while taking advantage of the portfolio of GPU instances available on Oracle Cloud Infrastructure."

- Clay Magouyrk, SVP, Software Development, Oracle Cloud Infrastructure

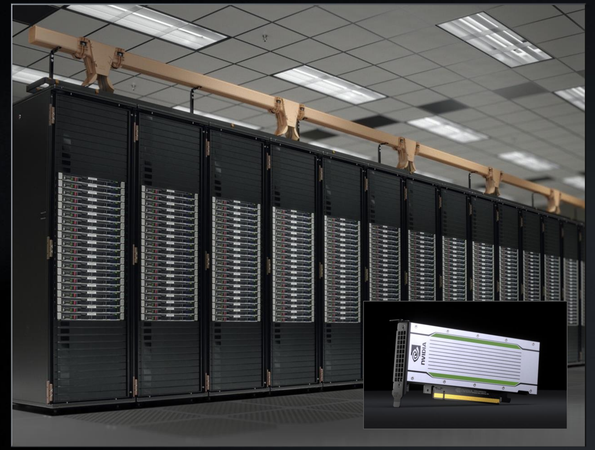
NVIDIA ACCELERATED HPC PLATFORM



DENSE HPC



NVIDIA HPC ACCELERATION STACKS

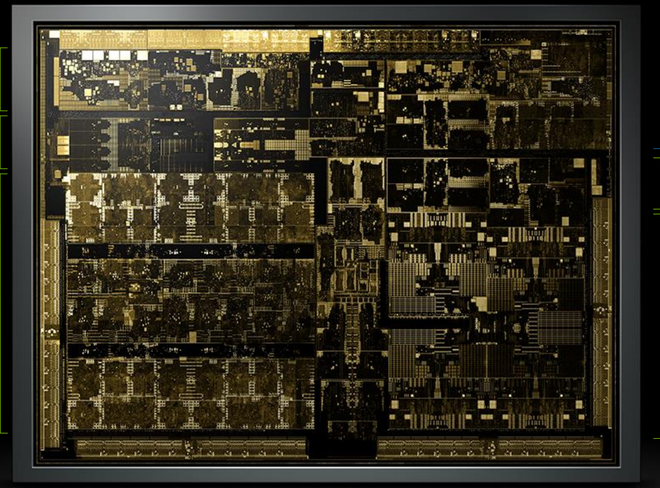


HYPERSCALE HPC

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

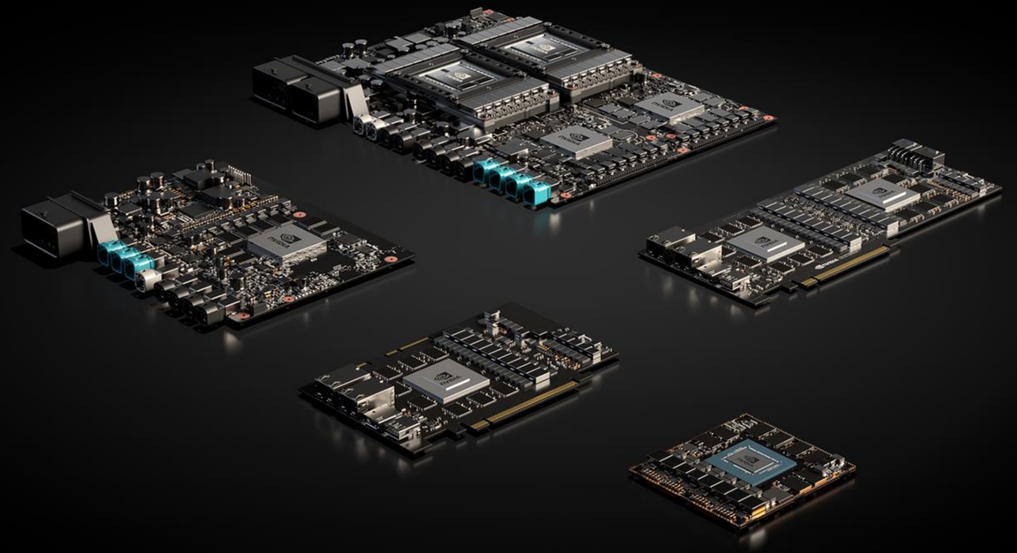
Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6

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



NEW CLARA AGX

ULTRASOUND

ENDOSCOPY

MAMMOGRAPHY

3D: CT, MRI, PET

RADIATION THERAPY

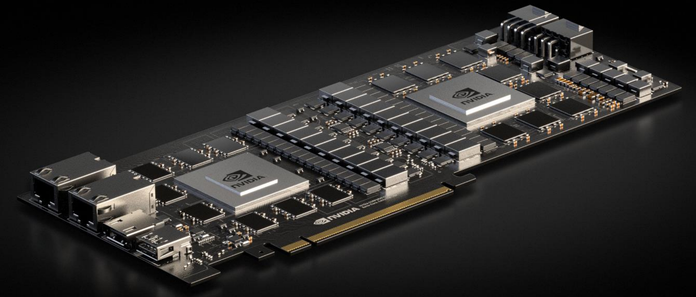
SEQUENCERS

DIGITAL PATHOLOGY

CRYO-EM

LIQUID BIOPSY

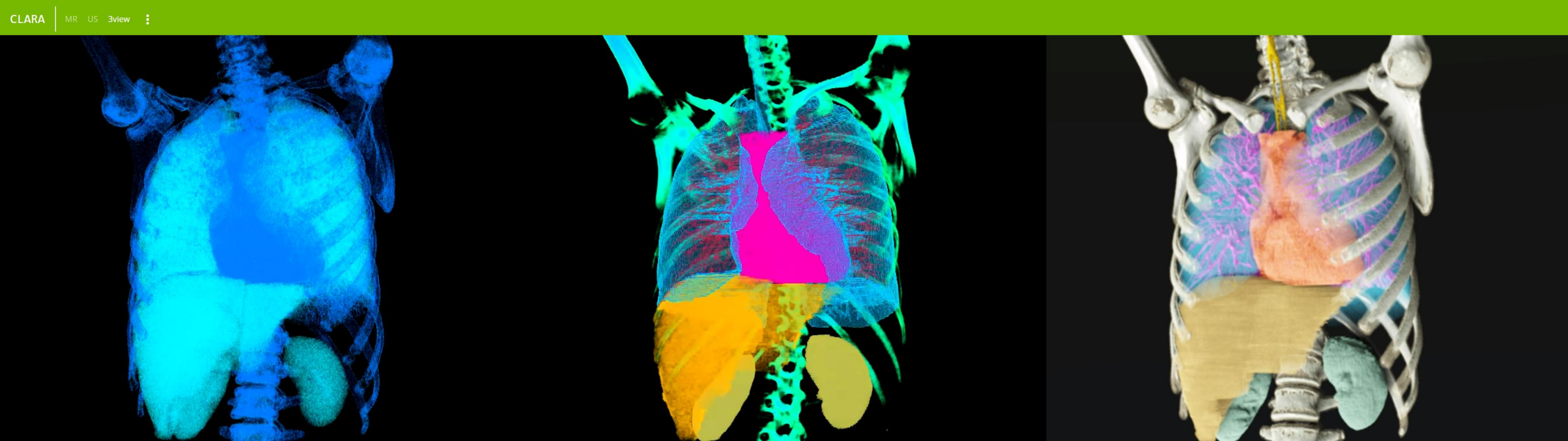
ROBOTIC SURGERY



Scale Up to 200 TOPS DL Processing

8 GIGA Rays

200W


CLARA
MR US 3view

ANNOUNCING
KING'S COLLEGE LONDON & NVIDIA BRING AI TO RADIOLOGY

Develop AI Training Tools for Imaging, NLP, Research | Deploy NVIDIA Clara to 4 NHS Hospitals, 8M Patients

KING'S College LONDON

OXFORD NANOPORE SELECTS NVIDIA AGX FOR PERSONAL DNA SEQUENCER

MinIT Powered by Jetson AGX
PromethION Powered by 4 V100 GPUs

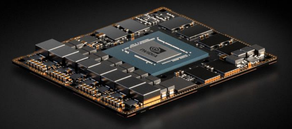
NVIDIA ISAAC

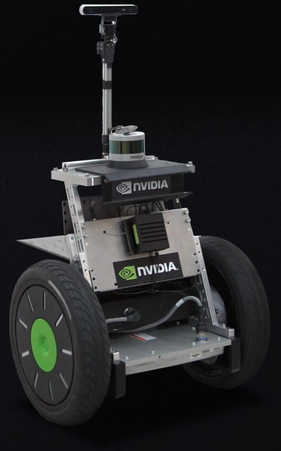
SENSOR PROCESSING

MAPPING & LOCALIZATION | PATH & TASK PLANNING

PERCEPTION | SITUATION UNDERSTANDING

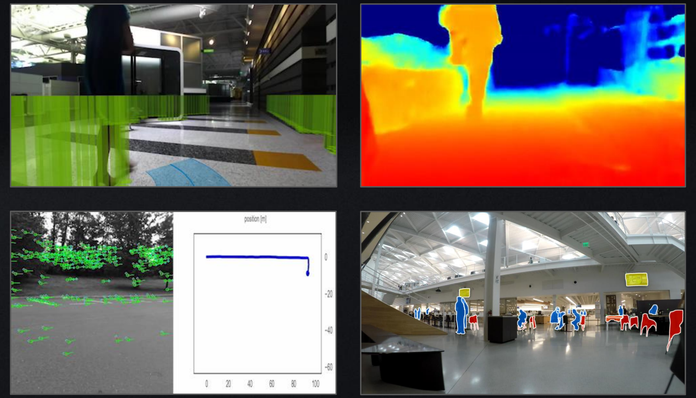
DIVERSITY & REDUNDANCY





Reality
Simulation

NVIDIA JETSON AGX — WORLD'S FIRST EDGE AI COMPUTER



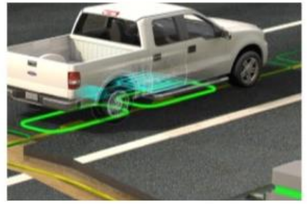
ISAAC GEMS

JETSON AGX XAVIER DEVELOPER KIT
Available Now

ISAAC SIM

NVIDIA AND ARROW BRING NEW JETSON XAVIER TO WORLD'S LARGEST INDUSTRIAL MARKETS

Robotics and Drones



Smart Infrastructure

Autonomous Assistants

Augmented Industry

Public Safety & Security

Experiential Retail

Five Years Out

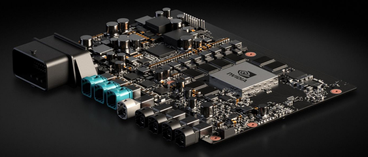
NVIDIA DRIVE — WORLD'S FIRST AUTONOMOUS VEHICLE PLATFORM

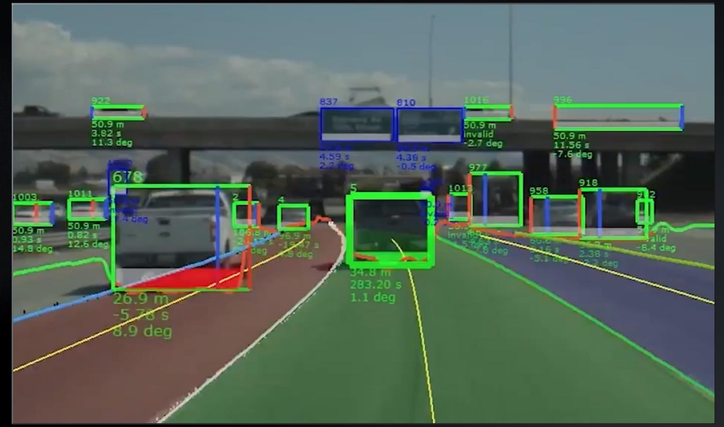
SENSOR PROCESSING

PERCEPTION | MAPPING & LOCALIZATION

SITUATION UNDERSTANDING | PATH & TASK PLANNING

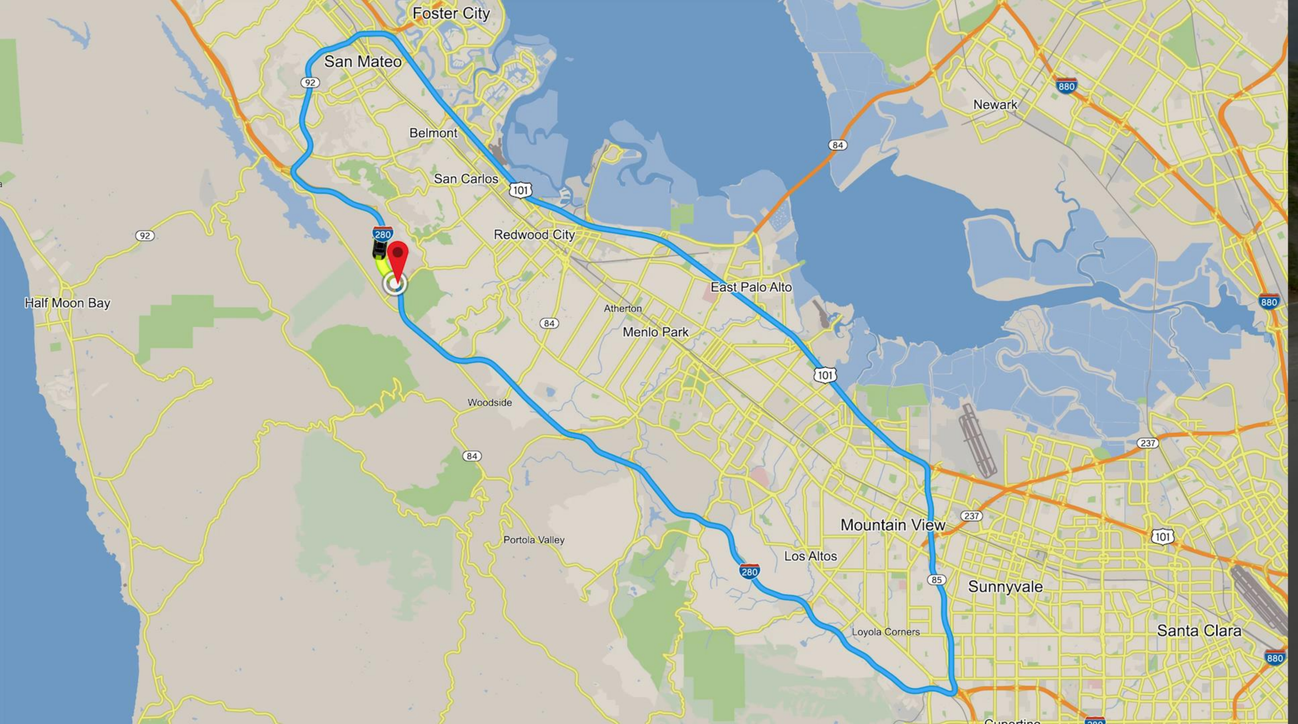
DIVERSITY & REDUNDANCY



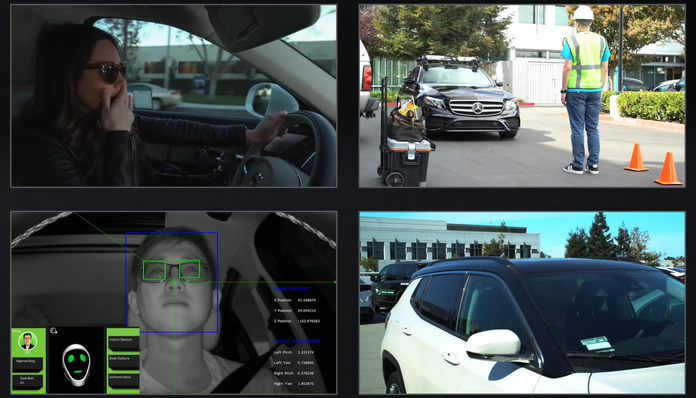


BRAKE
ACC
0 mph

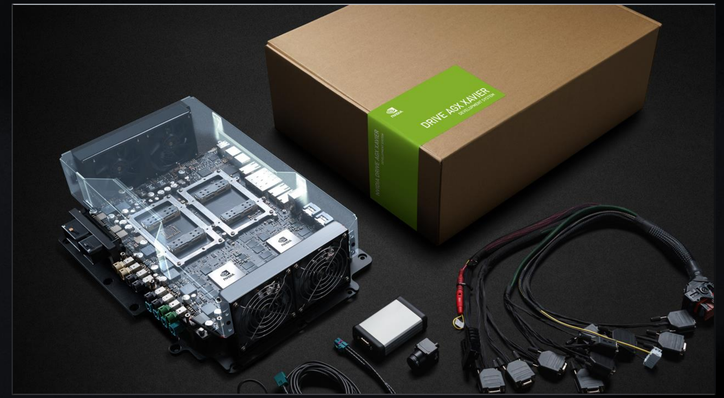
0 mph
ACC


Foster City
San Mateo
Belmont
San Carlos
Redwood City
Half Moon Bay
Atherton
Menlo Park
East Palo Alto
Woodside
Portola Valley
Los Altos
Mountain View
Sunnyvale
Loyola Corners
Santa Clara
Newark
92
101
280
84
880
237
85

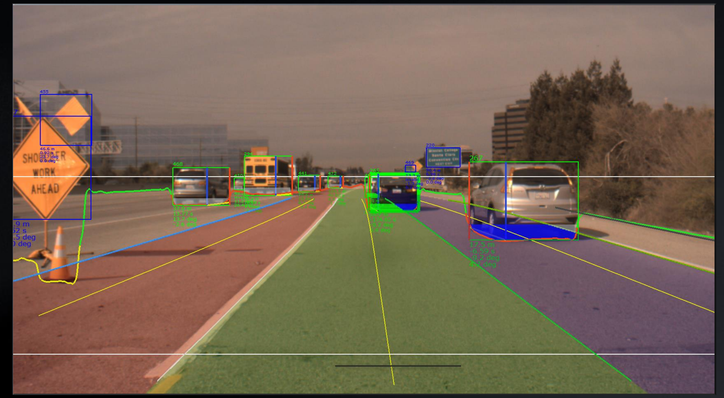
NVIDIA DRIVE — WORLD'S FIRST AUTONOMOUS VEHICLE PLATFORM



DRIVE IX
Available Now



DRIVE AGX XAVIER DEVELOPER KIT
Available Now



DRIVE AV
Available Now

ANNOUNCING VOLVO CARS SELECTS NVIDIA DRIVE AGX

DRIVE AGX Xavier to Pilot Next-generation Production Cars

ANNOUNCING VOLVO CARS SELECTS NVIDIA DRIVE AGX

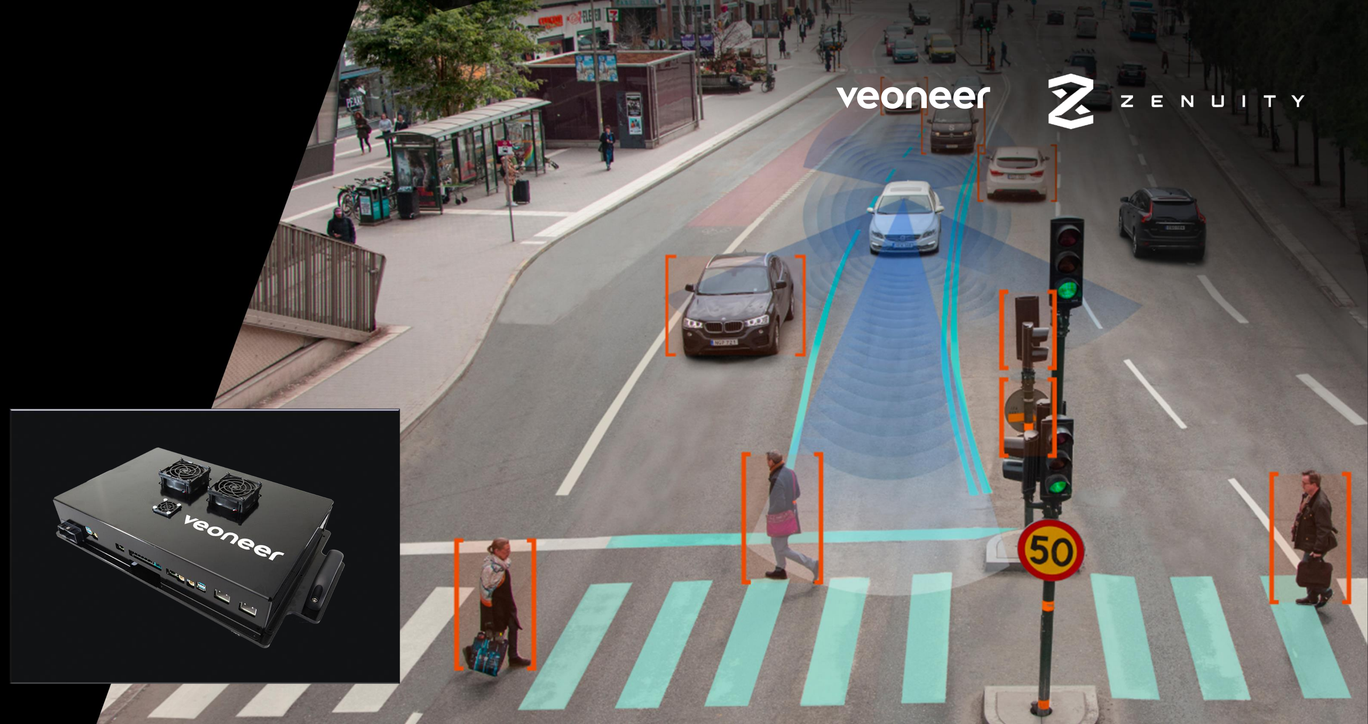
Henrik Green, Head of R&D, Volvo Cars

ANNOUNCING CONTINENTAL ADOPTS NVIDIA DRIVE AGX

Scalable Platform in Production 2021

ANNOUNCING VEONEER AND ZENUITY ADOPT NVIDIA DRIVE AGX

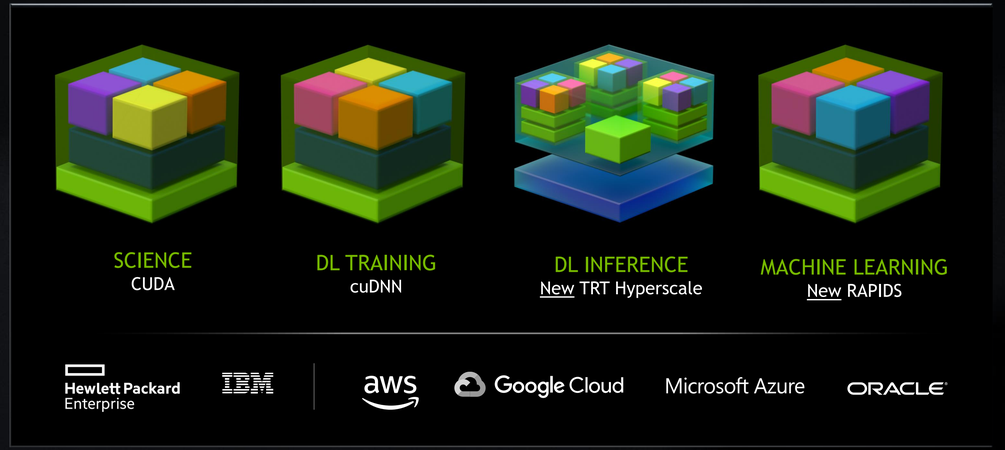
NVIDIA Xavier-based System Ready for Production by 2021



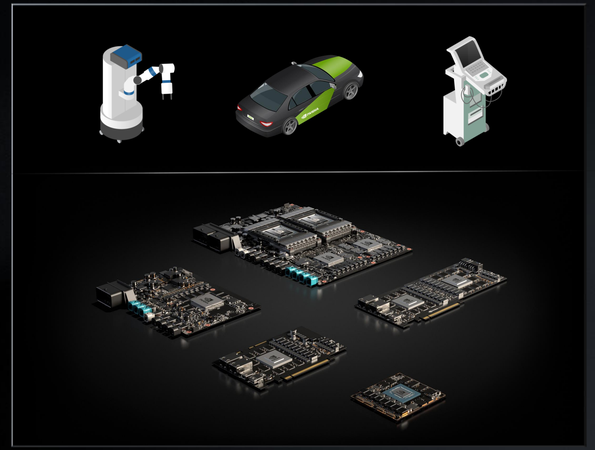
NVIDIA DRIVE PLATFORM ADOPTION ACROSS TRANSPORTATION

NEW NVIDIA PLATFORMS

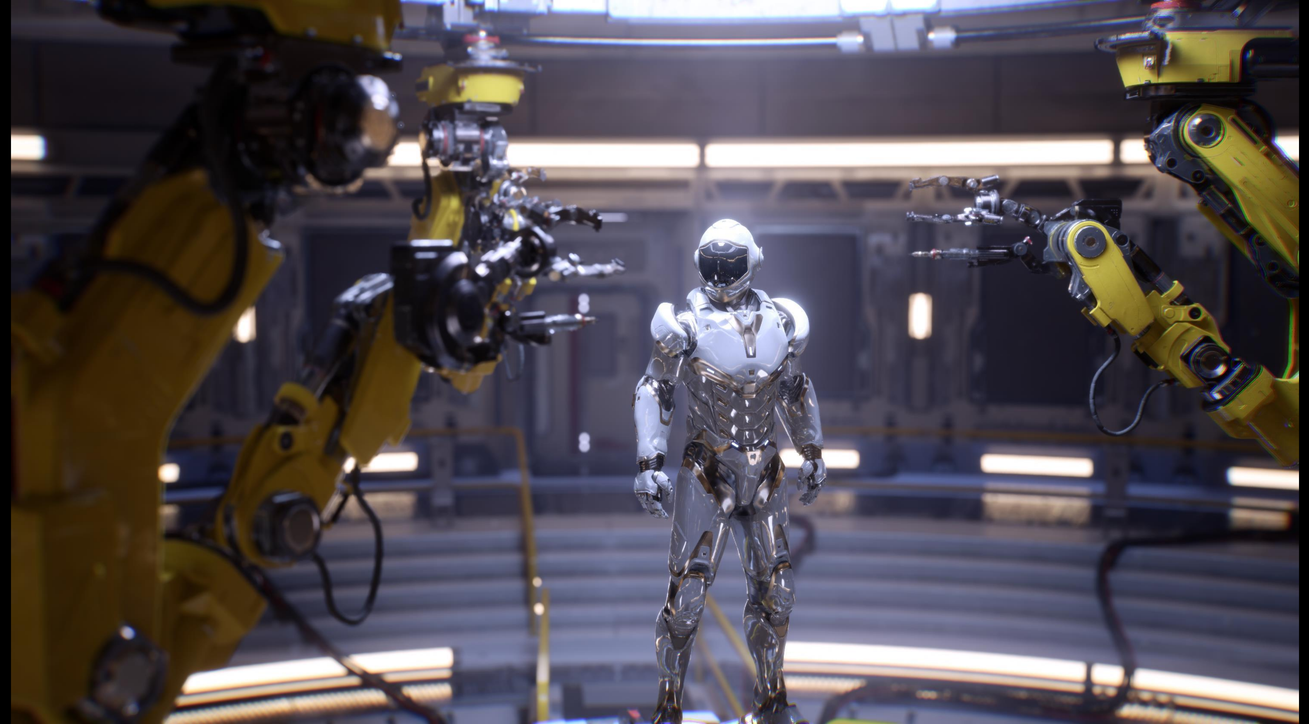
TURING



NVIDIA HPC ACCELERATION STACKS



NVIDIA AGX



NVIDIA