FORCES SHAPING COMPUTING

BEYOND MOORE'S LAW — 1000X EVERY 10 YEARS

ACCELERATED COMPUTING

DEEP NEURAL NETWORK

PROGRAM
RISE OF GPU COMPUTING

GTC Attendees — 7X in 5 Yrs
CUDA Downloads — 5X in 5 Yrs

2013 → 2018

IMAGING AND COMPUTER VISION
DATA SCIENCE
DEEP LEARNING
RAY TRACING
COMPUTATIONAL CHEMISTRY
MEDICAL IMAGING
BIOINFORMATICS
COMPUTATIONAL STRUCTURAL MECHANICS
NUMERICAL ANALYTICS
WEATHER AND CLIMATE
MATERIALS
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
Universals Deep Learning Accelerator

Tesla T4

Multi-precision Tensor Core
Support for Tensor Core

TensorRT 5.0
DNN Models
TensorRT Hyperscale

TensorRT Inference Server
Kubernetes
NV Docker
NV DL SDK

TENSOR CORE OPS

RESNET-50 INFERENCE (I/S)

TURING INT4
Volta
Pascal
Kepler
Maxwell

TURING INT8
TURNTEN	INT8
TURNTEN	INT4
FLOAT
INT8
INT4

DNN Models
TURING — A GIANT LEAP

GAMING REINVENTED

WORLD'S FIRST RAY TRACING GPU

UNIVERSAL DEEP LEARNING ACCELERATOR
THE NEW HPC MARKET

$9B

2005

SCIENTIFIC COMPUTING
THE NEW HPC MARKET

$9B

INTERNET
DEEP LEARNING
SCIENTIFIC COMPUTING

CUDA

2005
2018

CERN — Switzerland
CSCS — Switzerland
DFKI — Germany
Técnico Lisboa — Portugal
THE NEW HPC MARKET

$9B

SCIENTIFIC COMPUTING

2005

$36B

DEEP LEARNING

MACHINE LEARNING

SCIENTIFIC COMPUTING

INTERNET

HADOOP
NUMPY
SK
PANDAS

2018

RETAIL
HEALTHCARE
FINANCIAL SERVICES
LOGISTICS
TELECOM
AD TECH
THE DEFACTO DATA SCIENCE PLATFORM

1991 | Guido van Rossum
Interpreted language emphasizing readability

2006 | Travis Oliphant
Multi-dimensional arrays, math functions

2010 | Inria
Machine learning library

2008 | Wes McKinney
Data manipulation and analysis
Parallel processing in Python data analytics
Dynamic task scheduling
Collection of parallel arrays, data frames, lists

THE DEFACTO DATA SCIENCE PLATFORM

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

DASK
ANNOUNCING RAPIDS — ACCELERATED DATA SCIENCE

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

Pandas
Sklearn
Numpy
Dask
Arrow
Cuda
Cuml
Pandas-like
Sklearn-like

2016 | Wes McKinney
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

2010 | Matei Zaharia, CTO, Databricks

Unified analytics cluster-computing framework for big data and machine learning
ENTERPRISE-SCALE DATA SCIENCE

DGX-2: 512GB
DGX-1: 256GB
DGX STATION: 128GB
RTX 8000: 96GB
TESLA V100: 32GB
NEW NVIDIA DGX-2
THE LARGEST GPU EVER CREATED

2 PFLOPS
512GB HBM2
16 TB/sec Memory Bandwidth
10 kW | 160 kg
FASTER SPEEDS, REAL WORLD BENEFITS

ETL

ML

End-to-End
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
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
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
XAVIER
WORLD’S FIRST AUTONOMOUS MACHINE PROCESSOR

Most Complex SOC Ever Made
9 Billion Transistors, 350mm², 12nFFN
~8,000 Engineering Years

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

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

Vision Accelerator
1.7 TOPS
Stereo & Optical Flow Engine
2x 3.1 TOPS

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

Stereo & Optical Flow Engine
2x 3.1 TOPS

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

Stereo & Optical Flow Engine
2x 3.1 TOPS

Capture ARMv8 CPU
8 Cores
10-wide Supernormal
21 SpecInt2K

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

Stereo & Optical Flow Engine
2x 3.1 TOPS

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

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

Vision Accelerator
1.7 TOPS
Stereo & Optical Flow Engine
2x 3.1 TOPS

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

Stereo & Optical Flow Engine
2x 3.1 TOPS

Capture ARMv8 CPU
8 Cores
10-wide Supernormal
21 SpecInt2K
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-EA  LIQUID BIOPSY  ROBOTIC SURGERY

High-speed I/O  Image Recon  AI  Visualization

Scale Up to 200 TOPS DL Processing
8 GIGA Rays
200W
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
OXFORD NANOPORE SELECTS NVIDIA AGX FOR PERSONAL DNA SEQUENCER

MinIT Powered by Jetson AGX
PromethiON Powered by 4 V100 GPUs
SENSOR PROCESSING
MAPPING & LOCALIZATION
PATH & TASK PLANNING
PERCEPTION
SITUATION UNDERSTANDING
DIVERSITY & REDUNDANCY

NVIDIA ISAAC
NVIDIA JETSON AGX — WORLD’S FIRST EDGE AI COMPUTER

JETSON AGX XAVIER DEVELOPER KIT
Available Now

ISAAC GEMS

ISAAC SIM
NVIDIA AND ARROW BRING NEW JETSON XAVIER TO WORLD’S LARGEST INDUSTRIAL MARKETS
NVIDIA DRIVE — WORLD’S FIRST AUTONOMOUS VEHICLE PLATFORM

SENSOR PROCESSING

PERCEPTION

MAPPING & LOCALIZATION

SITUATION UNDERSTANDING

PATH & TASK PLANNING

DIVERSITY & REDUNDANCY
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

CARS
- Mercedes-Benz
- Ford
- Audi
- BMW
- Toyota
- Honda
- Tesla
- Uber
- ZENITY

MOBILITY SERVICES
- Uber
- ZENITY

TRUCKS
- Daimler
- DAF
- Volvo
- DAF
- ZENRUN

TIER ONES
- Bosch
- Continental
- TomTom

MAPPING
- TomTom
- ZENWAY

SPECIALTY
NEW NVIDIA PLATFORMS

TURING

NVIDIA HPC ACCELERATION STACKS

MACHINE LEARNING
New TRT Hyperscale

DL INFERENCE
New RAPIDS

DL TRAINING
cuDNN

SCIENCE
CUDA