
# ACCELERATING AI4EO

Carlo Nardone, Sr Solution Architect, NVIDIA

14/11/2018

# NVIDIA

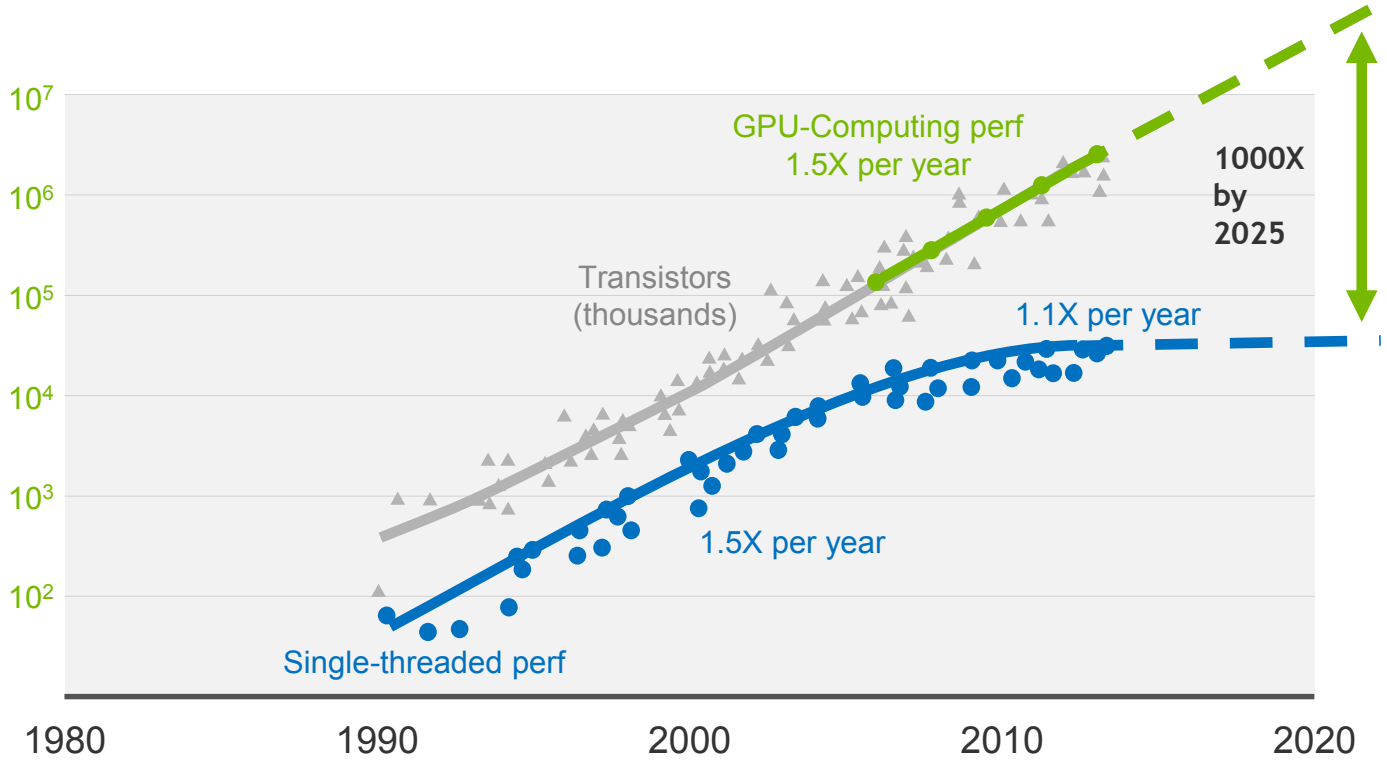
## 25 years long journey from computer gaming to AI Company

Computer Graphics

GPU Computing

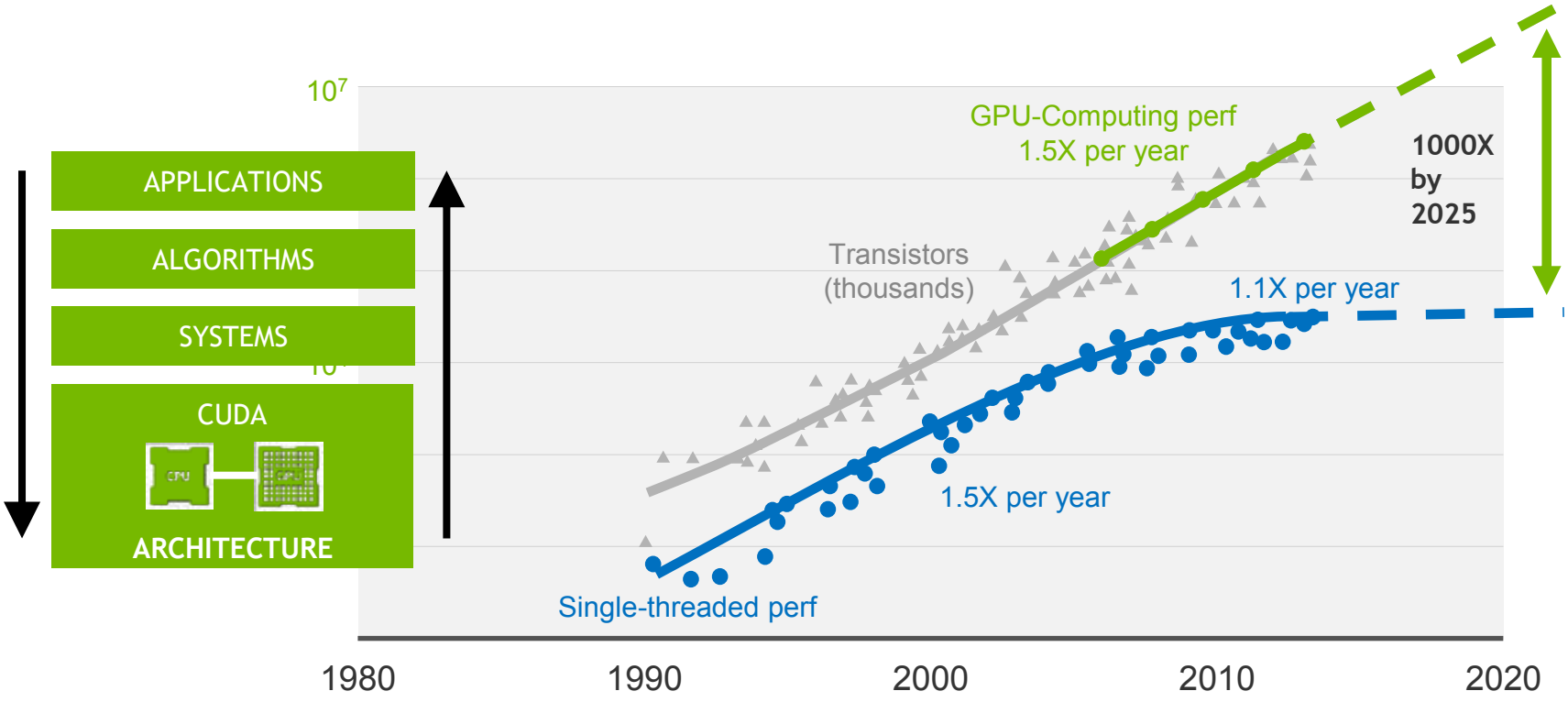
Artificial Intelligence

# RISE OF GPU COMPUTING



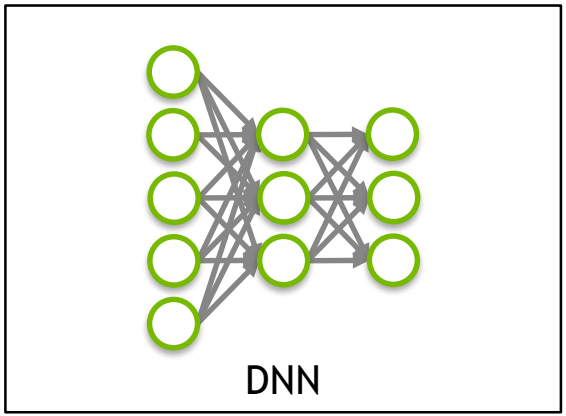

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

# RISE OF GPU COMPUTING




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

# THE BIG BANG IN MACHINE LEARNING

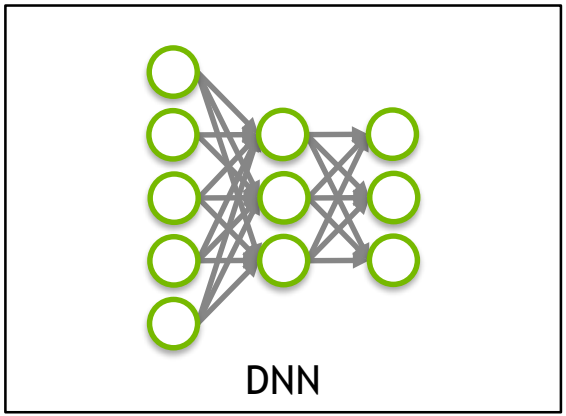


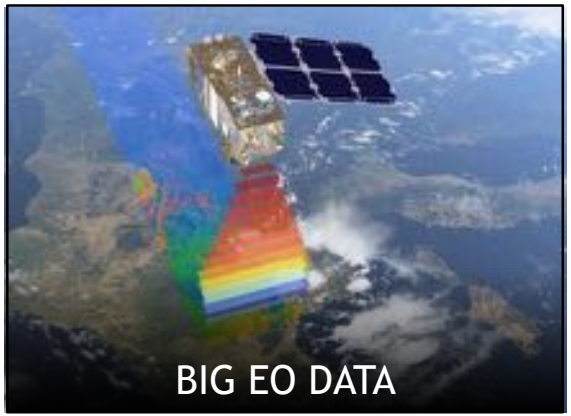



*"A.I. is the new electricity"*

*Andrew Ng – Stanford U. & Baidu Research*

# A BIG BANG IN EO ?





*AI4EO: A.I. for Earth Observation!*

# POWERING THE AI REVOLUTION

NVIDIA is advancing GPU computing for deep learning and AI at the speed of light. We create the entire stack. It starts with the most advanced GPUs and the systems and software we build on top of them. We integrate and optimize every deep learning framework. We work with the major systems companies and every major cloud service provider to make GPUs available in datacenters and in the cloud. And we create computers and software to bring AI to the edge, from self-driving cars to autonomous robots to medical devices.

ONE ARCHITECTURE

# TESLA V100 32GB

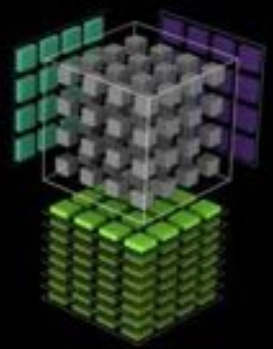
World's Most Advanced Data Center GPU
For Scale-up Environments

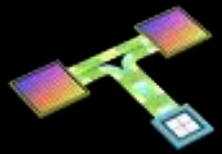
5,120 CUDA Cores
**640 NEW** Tensor Cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
**32GB** HBM2 @ 900GB/s | 300GB/s NVLink

# TESLA V100 32GB

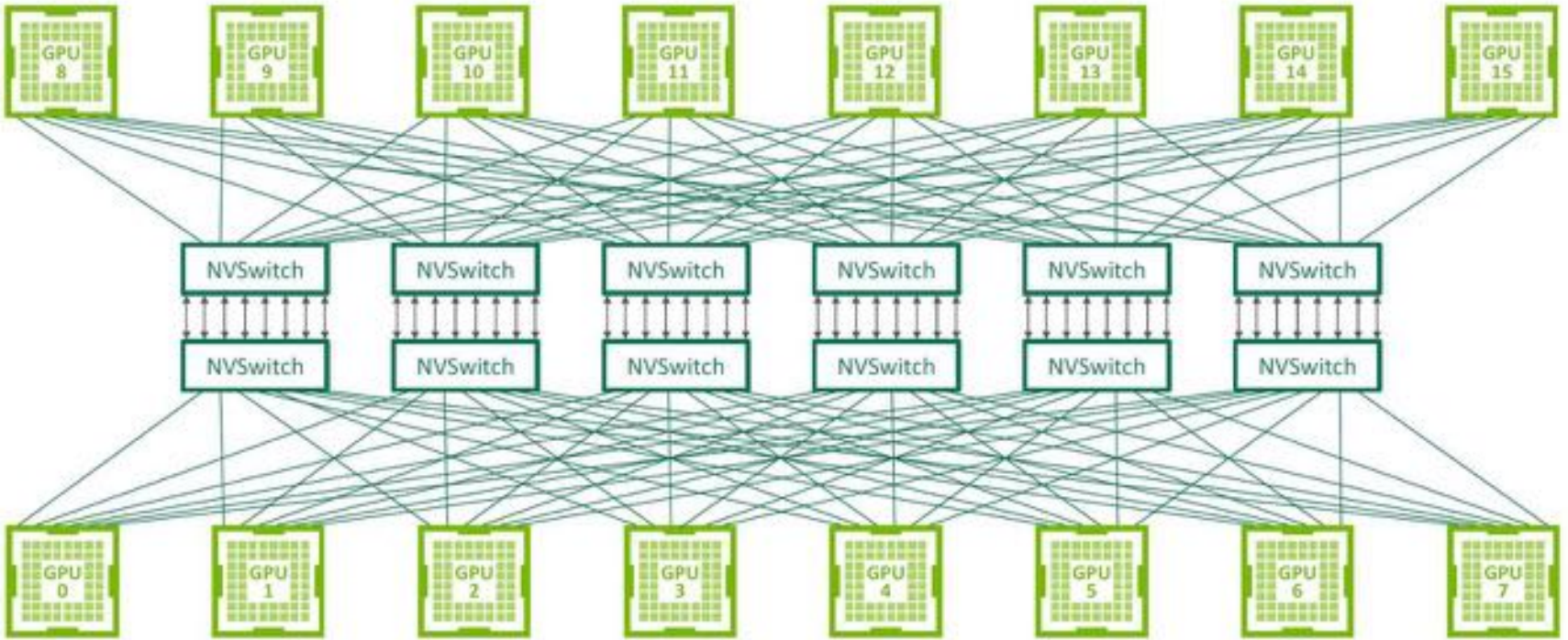
World's Most Advanced Data Center GPU
For Scale-up Environments

5,120 CUDA Cores
**640 NEW** Tensor Cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
**32GB** HBM2 @ 900GB/s | 300GB/s NVLink





# DGX-2: FULL NON-BLOCKING 16-WAY FABRIC



# TESLA T4

## WORLD'S MOST ADVANCED INFERENCE GPU

Universal Inference Acceleration
320 Turing Tensor cores
2,560 CUDA cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s

# JETSON AGX XAVIER

## World's first AI computer for Autonomous Machines

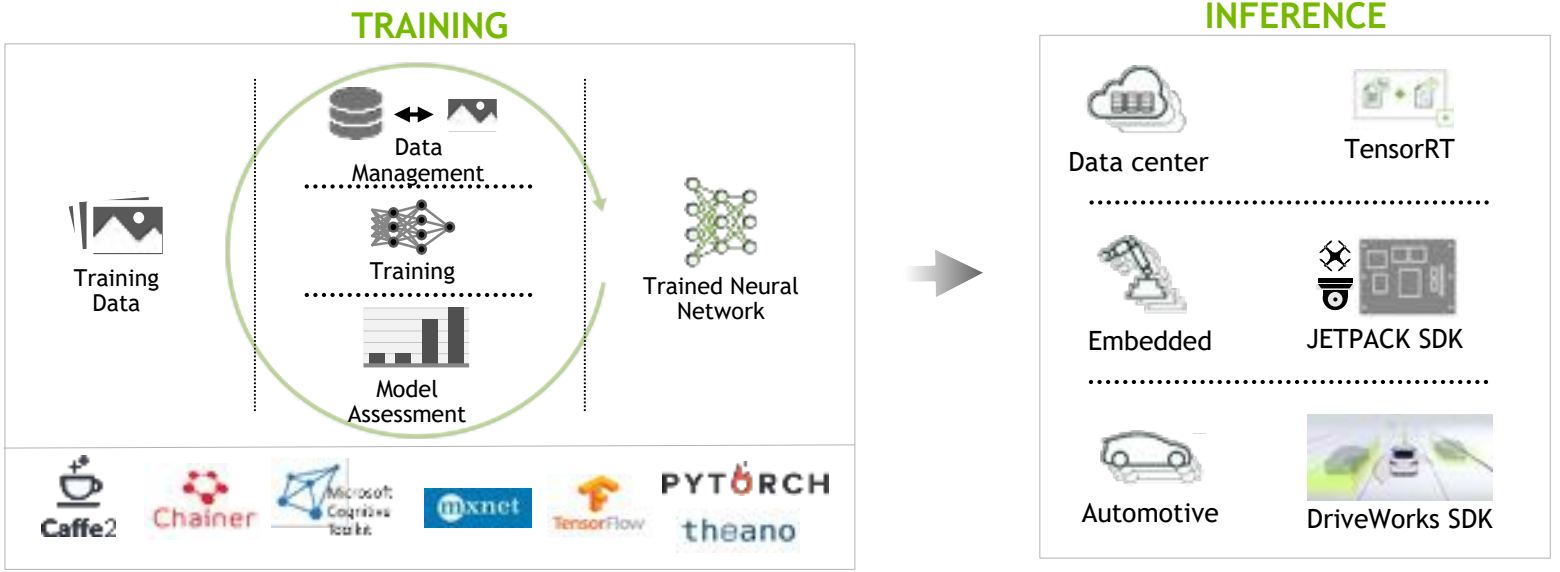
AI Server Performance in 30W • 15W • 10W

512 Volta CUDA Cores • 2x NVDLA

8 core CPU

32 DL TOPS

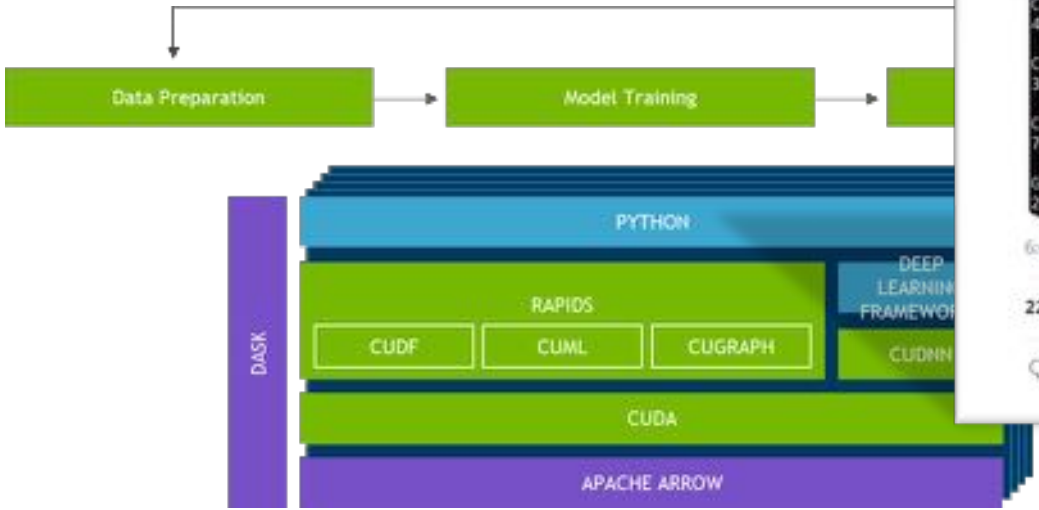
# NVIDIA DEEP LEARNING SOFTWARE STACK

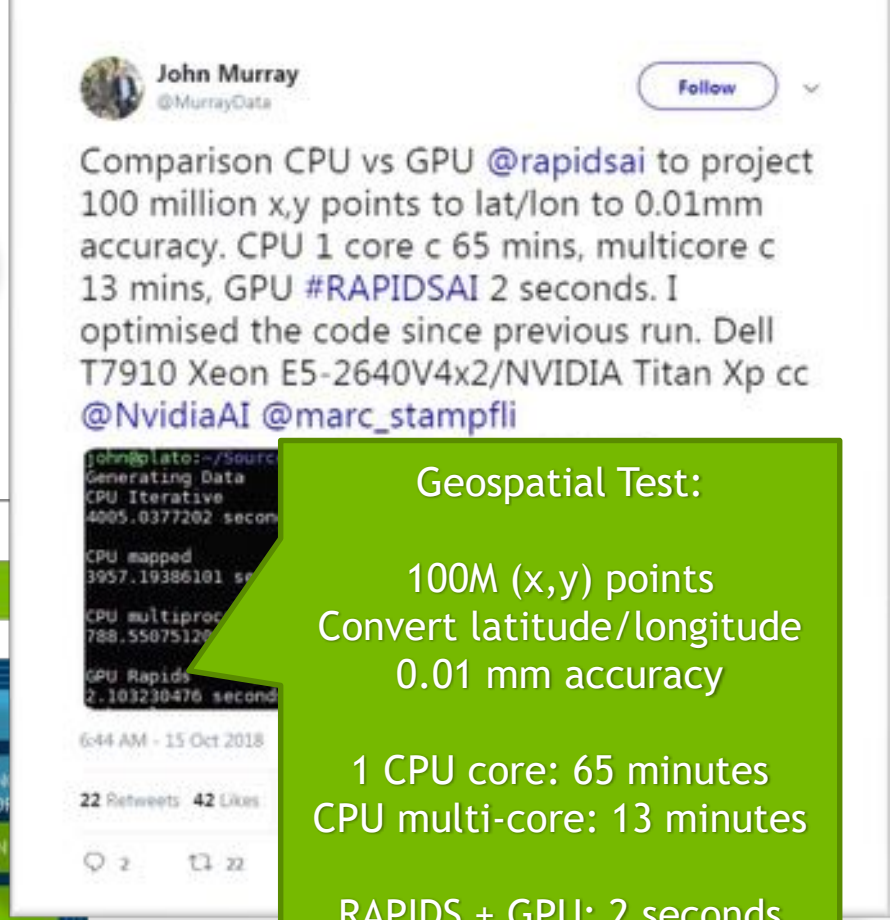



developer.nvidia.com/deep-learning-software

# RAPIDS

## A Geospatial use case




Geospatial Test:

100M (x,y) points
Convert latitude/longitude
0.01 mm accuracy

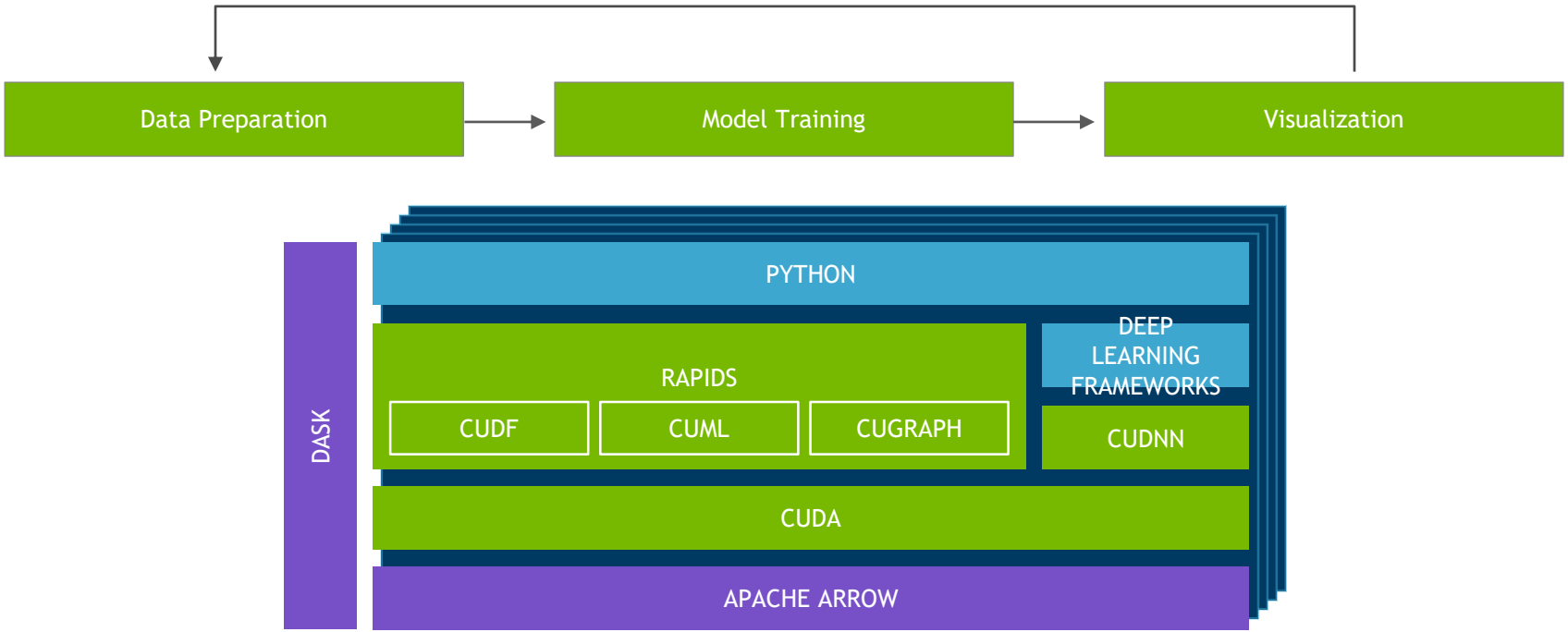
1 CPU core: 65 minutes
CPU multi-core: 13 minutes

RAPIDS + GPU: 2 seconds

Dell T7910 Xeon E5-2640V4x2 + NVIDIA Titan Xp

# RAPIDS — ACCELERATING DATA SCIENCE

## Open Software Stack for Python




THANK YOU!
cnardone @ nvidia.com
NVIDIA


BACKUP SLIDES

# NVIDIA POWERS WORLD'S FASTEST SUPERCOMPUTERS

## 127 Systems in Top500 (+48% YoY) | 22 of Top 25 Greenest


**ORNL Summit**
**World's Fastest**
**27,648 GPUs| 144 PF**

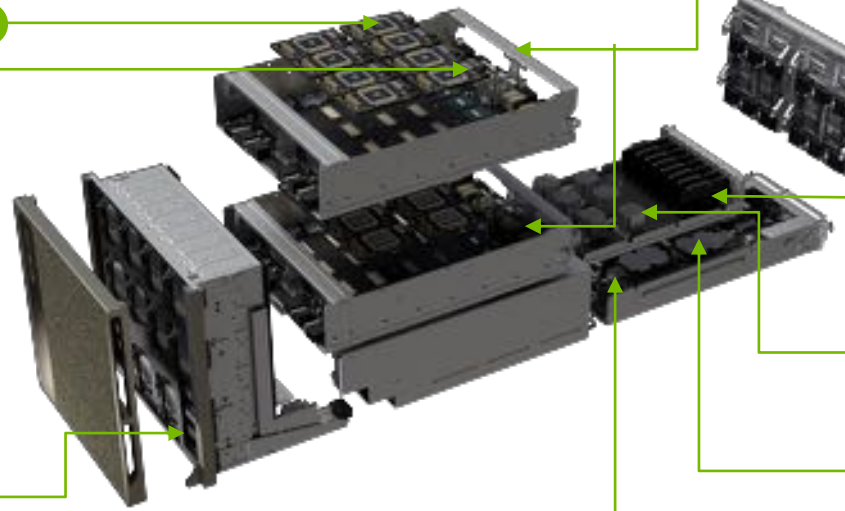
**LLNL Sierra**
**World's 2nd Fastest**
**17,280 GPUs| 95 PF**


**Piz Daint**
**Europe's Fastest**
**5,704 GPUs| 21 PF**

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

**ENI HPC4**
**Fastest Industrial**
**3,200 GPUs| 12 PF**

# NVIDIA DGX-2: “WORLD’S LARGEST GPU”



1 **NVIDIA Tesla V100 32GB**

2 **Two GPU Boards**
8 V100 32GB GPUs per board
6 NVSwitches per board
512GB Total HBM2 Memory
interconnected by
**Plane Card**

3 **Twelve NVSwitches**
2.4 TB/sec bi-section
bandwidth

4 **Eight EDR Infiniband/100 GigE**
1600 Gb/sec Total
Bi-directional Bandwidth

5 **PCIe Switch Complex**

6 **Two Intel Xeon Platinum CPUs**

7 **1.5 TB System Memory**

8 **30 TB NVME SSDs**
Internal Storage

# TESLA UNIVERSAL ACCELERATION PLATFORM

## Single Platform To Drive Utilization and Productivity

**CUSTOMER USECASES**


Speech


Translate


Recommender

CONSUMER INTERNET


Healthcare

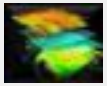

Manufacturing


Finance

INDUSTRIAL APPLICATIONS


Molecular Simulations


Weather Forecasting


Seismic Mapping

SUPERCOMPUTING

**APPS & FRAMEWORKS**

python

TensorFlow mxnet ONNX PYTORCH RAPIDS RALDI Chainer

Amber ANSYS +550 Applications NAMD SIMULIA

**NVIDIA SDK & LIBRARIES**

MACHINE LEARNING | RAPIDS

cuDF | cuML | cuGRAPH

DEEP LEARNING

cuDNN | cuBLAS | CUTLASS | NCCL | TensorRT

SUPERCOMPUTING

CuBLAS | CuFFT | OpenACC

CUDA

**TESLA GPUs & SYSTEMS**


TESLA GPU


VIRTUAL GPU


NVIDIA DGX FAMILY


NVIDIA HGX

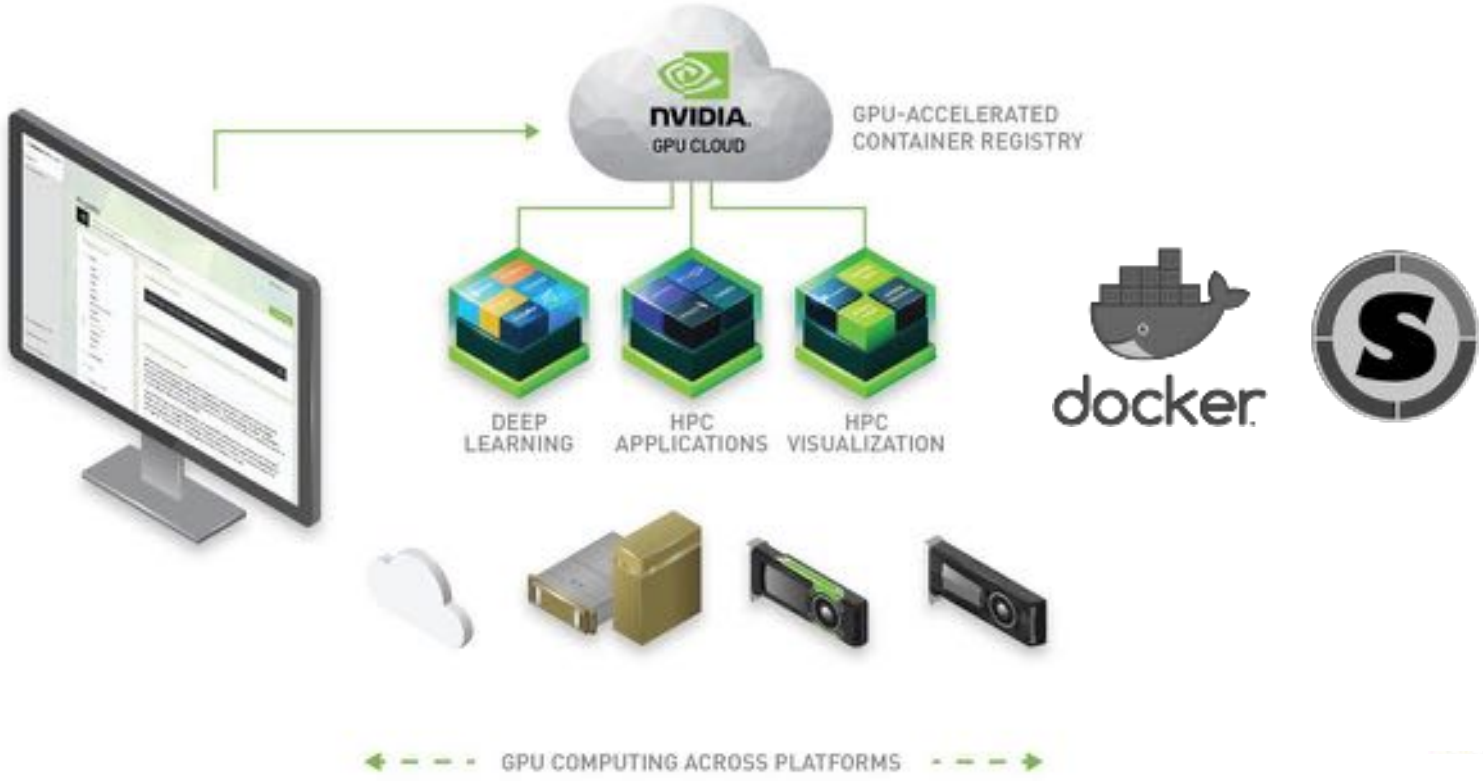
SYSTEM OEM

aws Google Cloud Platform Microsoft Azure

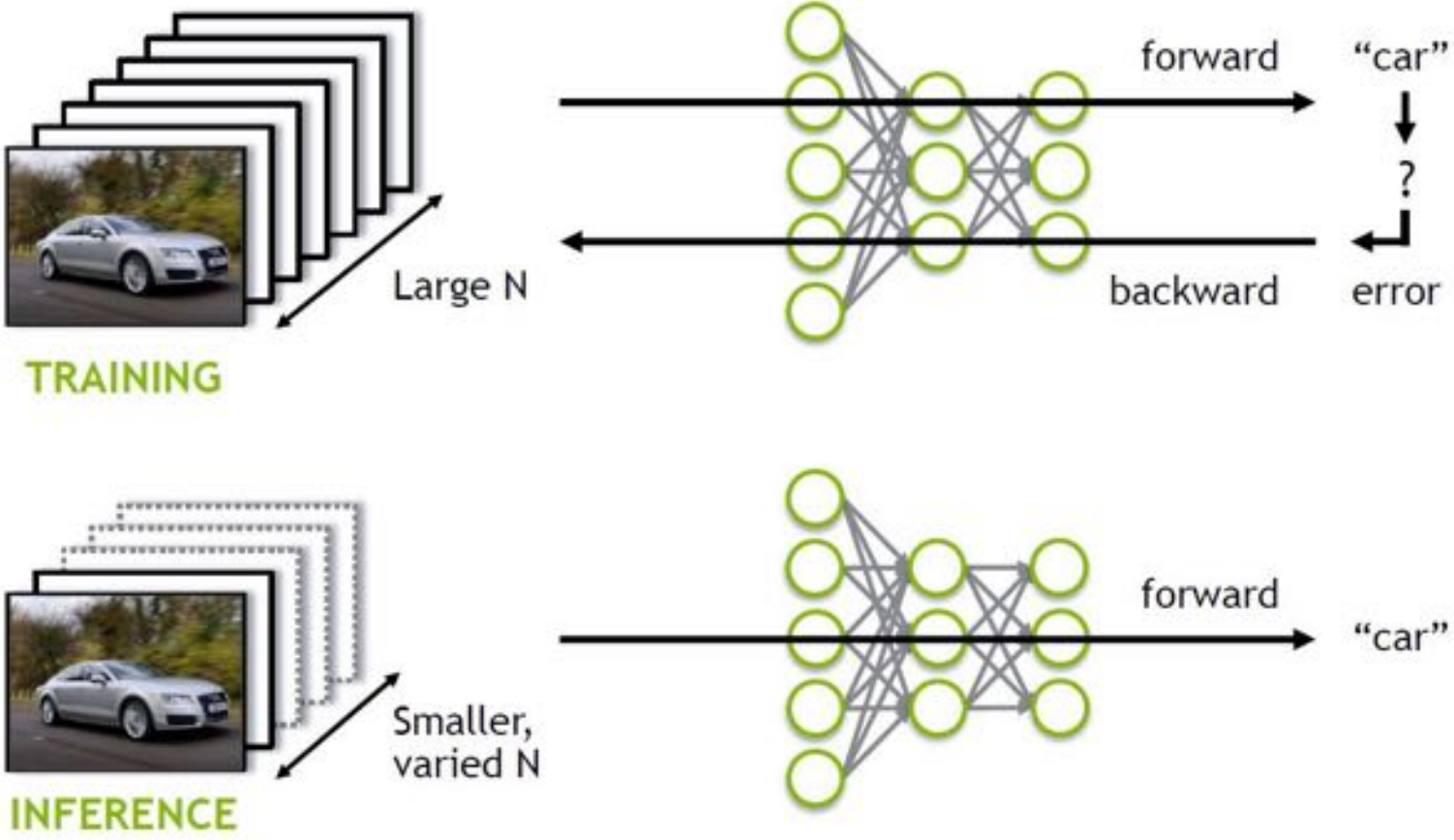
CLOUD

# NGC CONTAINER REGISTRY

Simple access to accelerated GPU software: ngc.nvidia.com

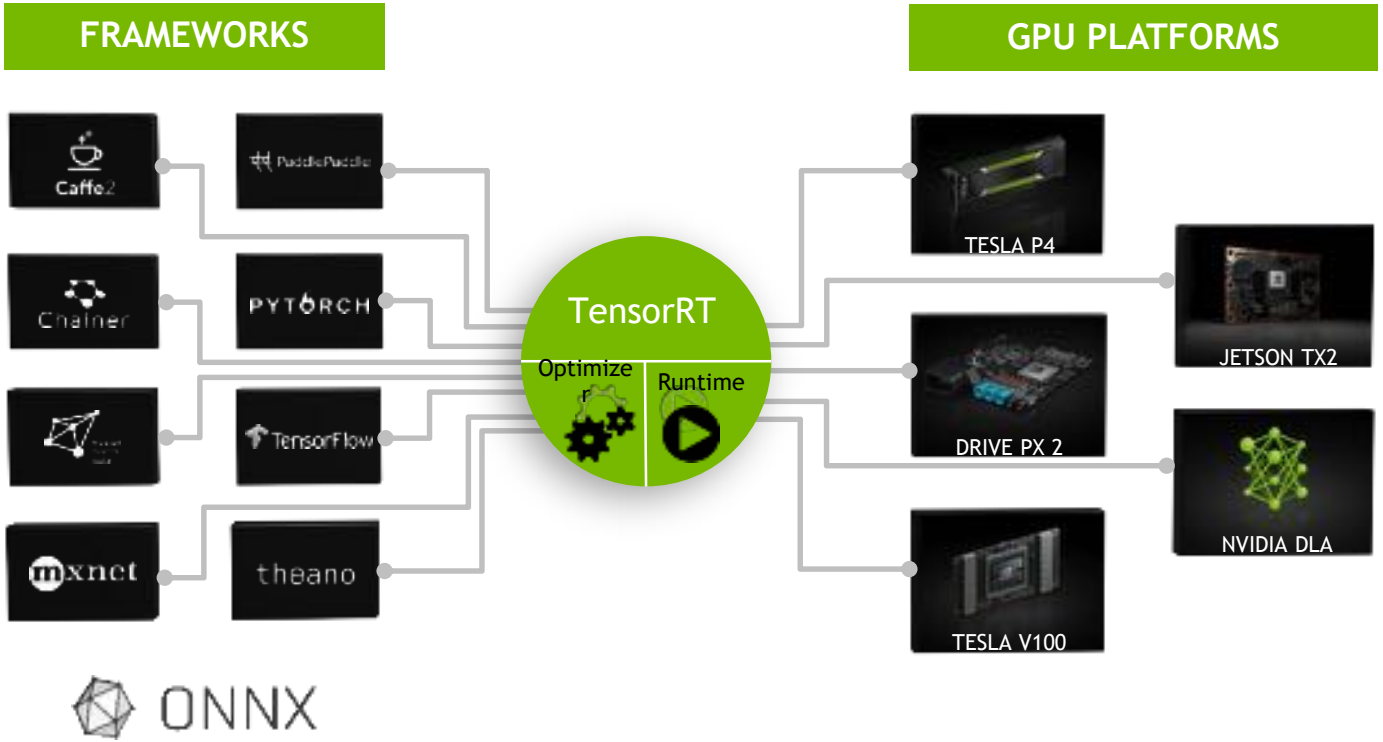


# TRAINING VS INFERENCE



# NVIDIA TENSORRT

## Programmable Inference Accelerator

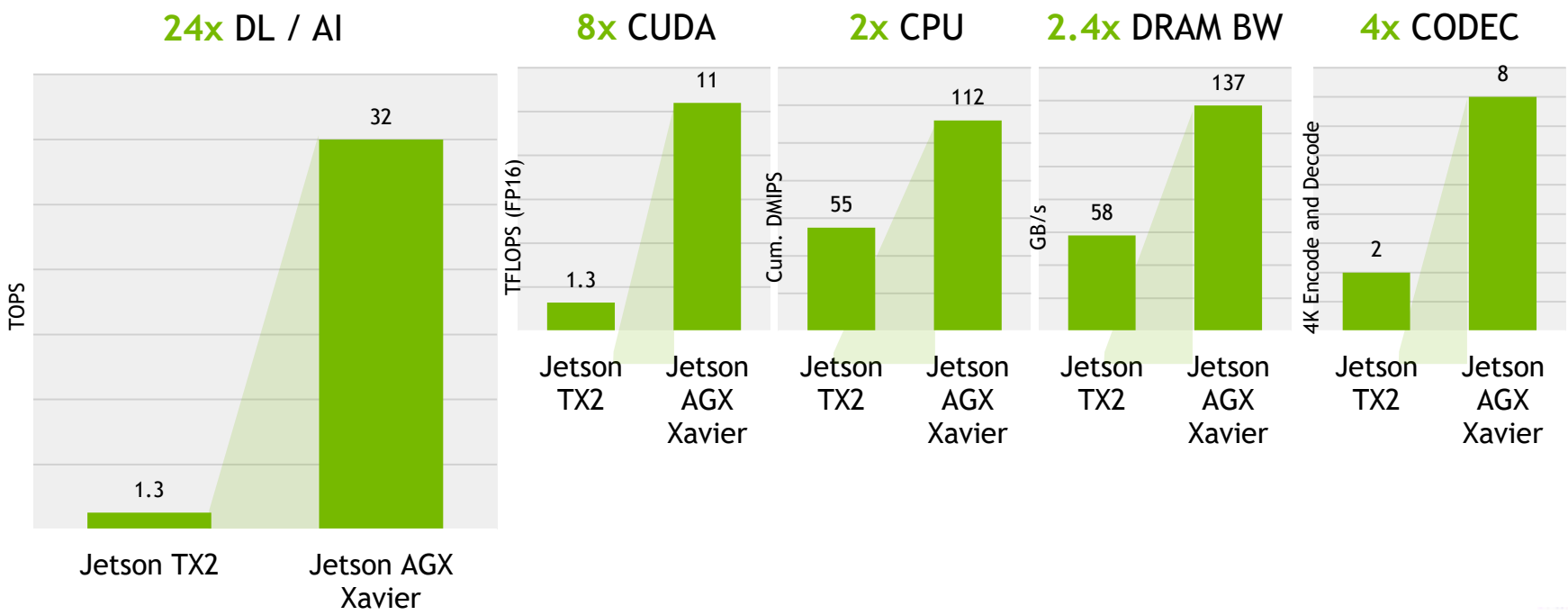



developer.nvidia.com/tensorrt

# JETSON AGX XAVIER

| | JETSON TX2 | JETSON AGX XAVIER | |
|---|---|---|---|
| **GPU** | 256 Core Pascal | 512 Core Volta | New! |
| **DL Accelerator** | - | NVDLA x 2 | New! |
| **Vision Accelerator** | - | VLA - 7 way VLIW Processor | New! |
| **CPU** | 6 core Denver and A57 CPUs | 8 core Carmel CPUs | New! |
| **Memory** | 8 GB 128 bit LPDDR4<br>58.4 GB/s | 16 GB 256 bit LPDDR4x<br>137 GB/s | x2 |
| **Storage** | 32 GB eMMC | 32 GB eMMC | |
| **Video Encode** | 2x 4K @30<br>HEVC | 2x 4K @ 60 / 4x 4K @30<br>HEVC | x2 |
| **Video Decode** | 2x 4K @30<br>12 bit support | 2x 8K @ 30 / 8x 4K @30<br>12 bit support | x4 |
| **Camera** | Up to 6 cameras<br>CSI2 D-PHY 1.2<br>2.5Gbps/lane | Up to 8 cameras<br>CSI2 D-PHY 1.2 2.5 Gbps/lane | +2 |
| **Mechanical** | 50mm x 87mm<br>400 pin connector | 100mm x 87mm<br>699 pin connector | |

# JETSON AGX XAVIER
# 20X PERFORMANCE IN 18 MONTHS



# JETSON AGX XAVIER
# GPU WORKSTATION PERF • 1/10TH POWER