ALISON B LOWNDES

AI DevRel | EMEA

@alisonblowndes

October 2018

www.FrontierDevelopmentLab.org

NASA FRONTIER DEVELOPMENT LAB

SETI INSTITUTE NASA esa NVIDIA IBM LOCKHEED MARTIN

FDL EUROPE

A MISSION CONTROL
FOR SPACESHIP EARTH

esa

EARTH OBSERVATION

KICK OFF WORKSHOP
ESA ESRIN, ROME
25th - 29th June '18

RESEARCH SPRINT
30th June - 17th August

IBM NVIDIA UNIVERSITY OF OXFORD CATAPULT Satellite Applications https://devblogs.nvidia.com/parallelforall/inside-volta/

TESLA V100 32GB

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

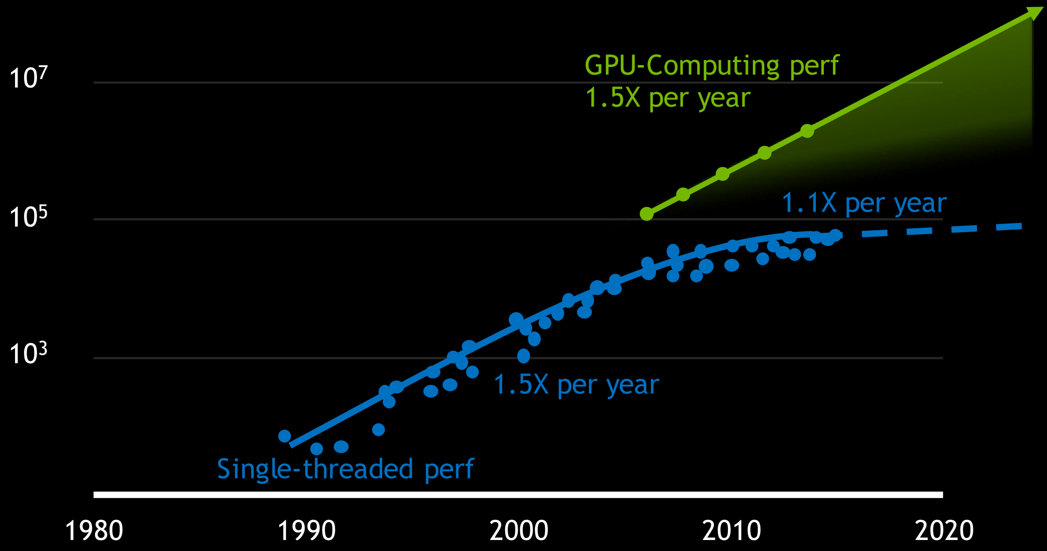
120 Tensor TFLOP

20MB SM RF | 16MB Cache | 32GB HBM2 @ 900 GB/s
300 GB/s NVLink

THE RISE OF GPU COMPUTING

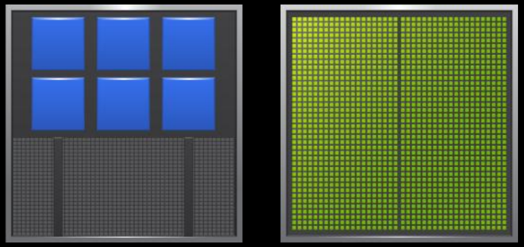
Big Data Needs Algorithms and Compute That Scales

END OF MOORE'S LAW

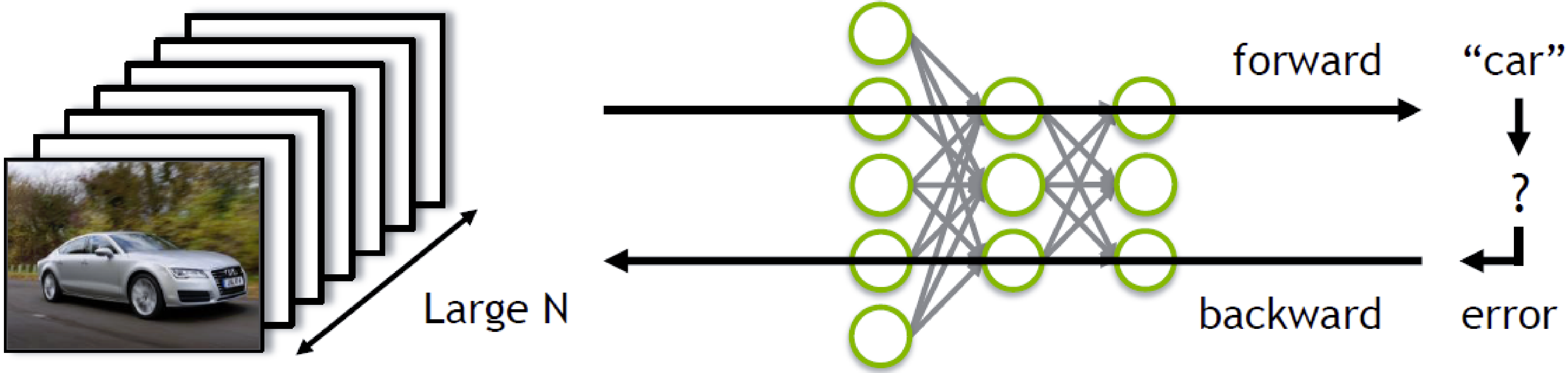


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

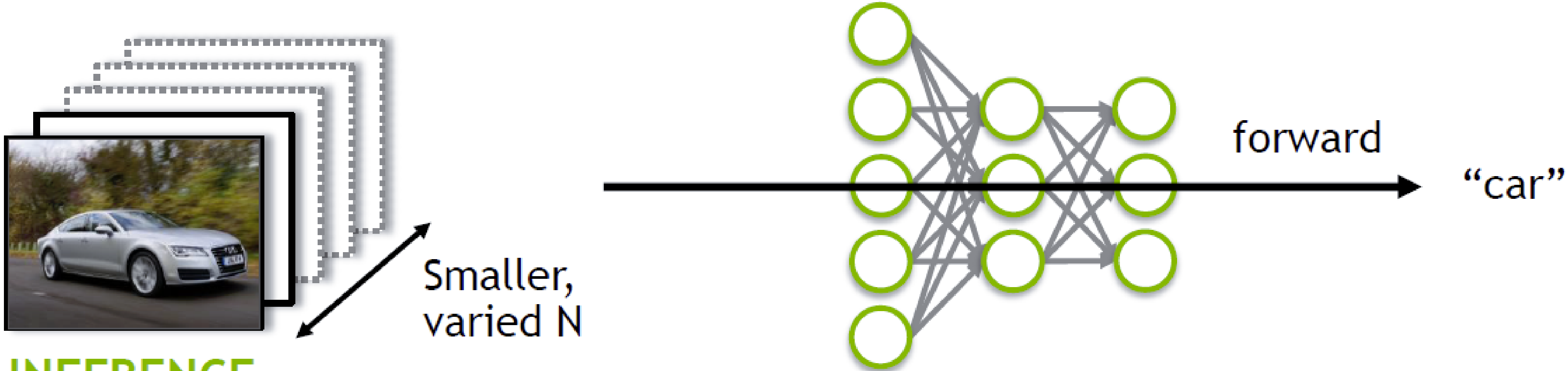
CPU vs. GPU



TRAINING VS INFERENCE

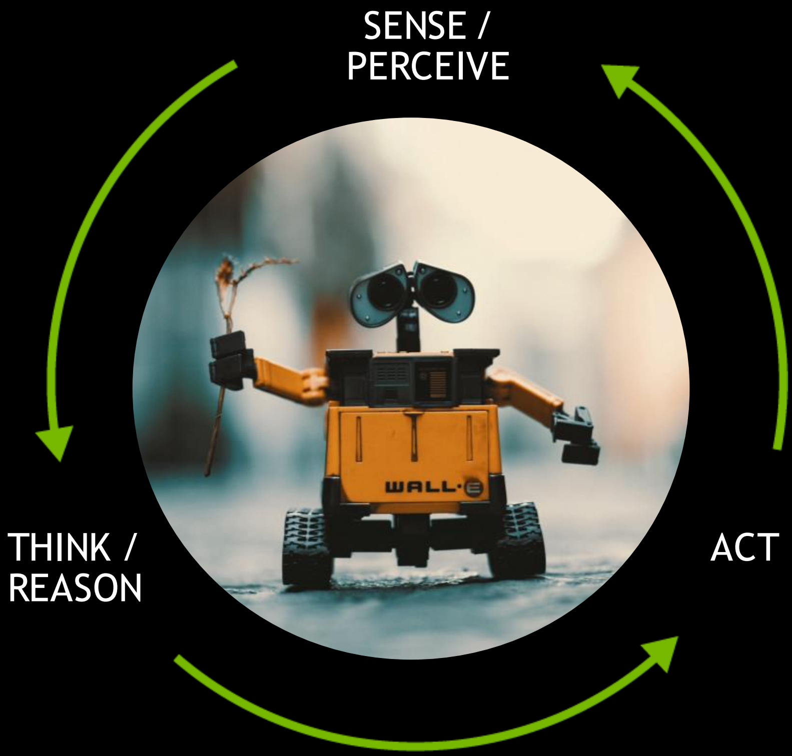


TRAINING



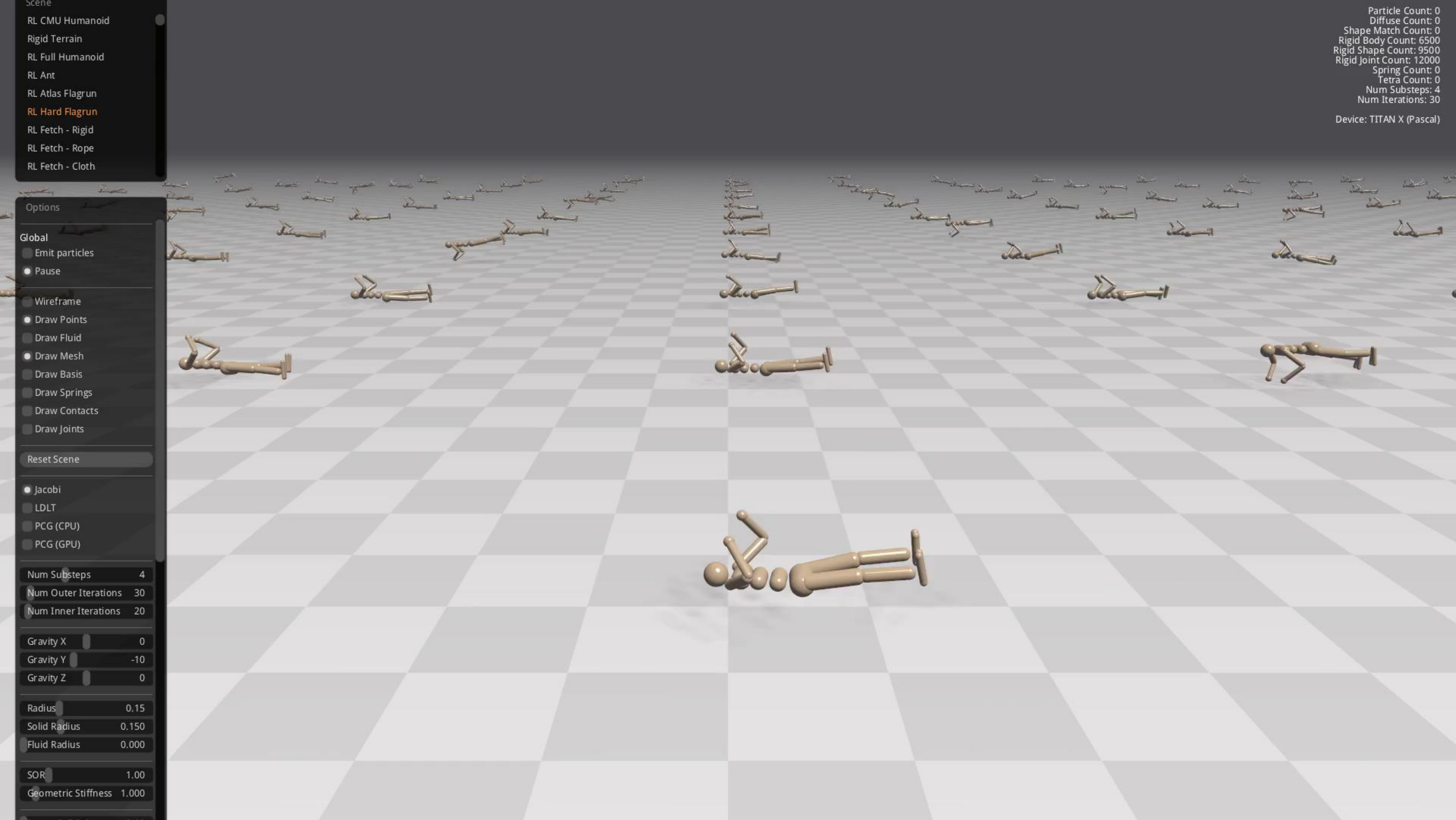
INFERENCE

ROBOTS


SENSE / PERCEIVE

THINK / REASON

ACT



Global

○ Emit particles
● Pause

○ Wireframe
● Draw Points
○ Draw Fluid
● Draw Mesh
○ Draw Basis
○ Draw Springs
○ Draw Contacts
○ Draw Joints

● Jacobi
○ LDLT
○ PCG (CPU)
○ PCG (GPU)

Num Substeps 4
Num Outer Iterations 30
Num Inner Iterations 20

Gravity X 0
Gravity Y -10
Gravity Z 0

Radius 0.15
Solid Radius 0.150
Fluid Radius 0.000

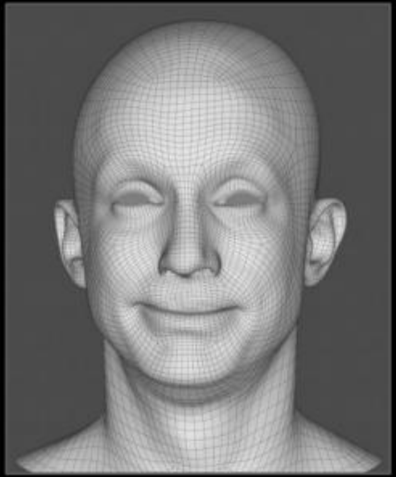
SOR 1.00
Geometric Stiffness 1.000

```
Particle Count: 0
Diffuse Count: 0
Shape Match Count: 0
Rigid Body Count: 6500
Rigid Shape Count: 9500
Rigid Joint Count: 12000
Spring Count: 0
Tetra Count: 0
Num Substeps: 4
Num Iterations: 30

Device: TITAN X (Pascal)
```

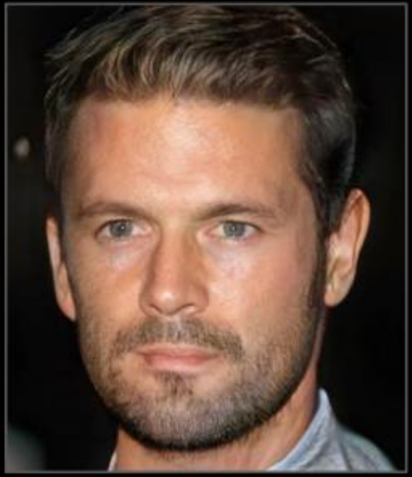
NVIDIA RESEARCH

NVIDIA Research
AI Autoencoder



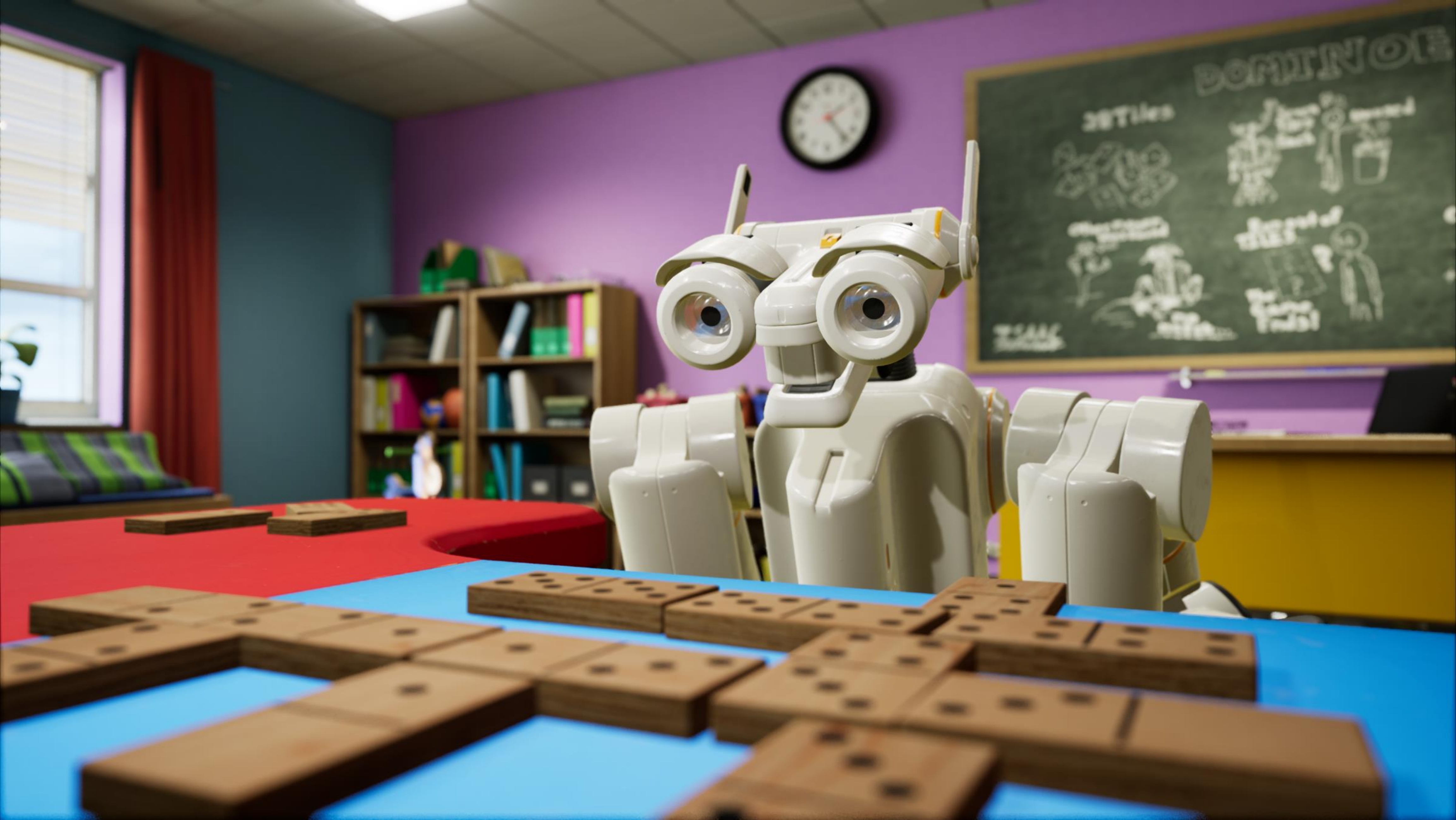
NVIDIA Research / Remedy
Audio-driven Facial Animation

NVIDIA Research
Semantic Manipulation with GANs



NVIDIA Research
Progressive GAN

NVIDIA Research / AIVA
RNNs for Music


DOMINOE
28Tiles

ISAAC

Simulate → Develop → Deploy

Simulate

World model
Warehouse · Office · Store · Home

Robot model
Carter · URDF loader

Simulation Engine
Photo-realistic Graphics · Physics · Soft bodies · · Procedural Generation · Massive parallelism · Unreal Engine 4 / Unity 3D

Virtual Sensors

Virtual Actuators

Develop

Navigation

Behaviors

Perception

Manipulation

Interactions with humans

ML
TensorRT · CUDA · Tensorflow · ...

Gems
Optimizers · Algebra · EKFs · Depth · ...

Isaac Framework
Codelets · Behaviors · 3D Poses · Distributed · Messaging · Synchronization · Record & Replay · Configuration · Visualization

Sensor Processing

Actuator Control

Deploy

Drivers
Lidar · Camera · IMU · Robot Base · ...

Jetson
Fully integrated with X2 and Xavier

Unified Message API
Use the same messages for simulation, actual hardware and across all apps

HW Sensor

HW Actuator

CUDA DEVELOPMENT ECOSYSTEM

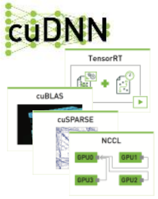
GPU Users

Applications

Domain Specialists

Frameworks

Problem Specialists



Libraries

New Algorithm Developers and Optimization Experts

Directives and Standard Languages

CUDA-C++
CUDA Fortran

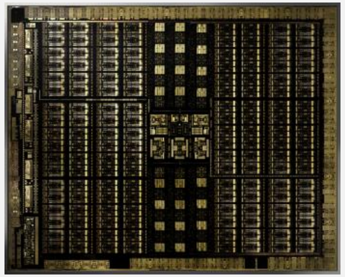
Extended Standard Languages

Ease of use → *Specialized Performance*

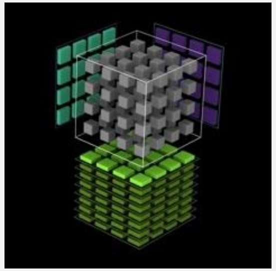
CUDA: Programming Model, GPU Architecture, System Architecture

CUDA 10 - TURING

Turing Architecture

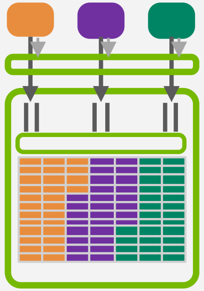


Multi-Precision Tensor Core

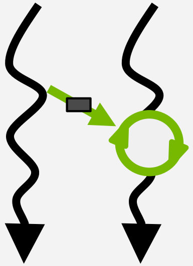


RT Core

Turing MPS



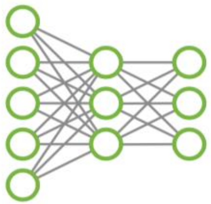
Independent Thread Scheduling



Inference Accelerated, Graphics Reinvented, Volta’s Programmability

NVIDIA DEEP LEARNING SDK UPDATE

GPU-accelerated DL Primitives



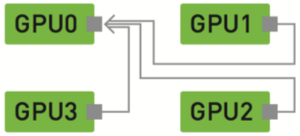
cuDNN 7.3

Support for Turing

Optimizations for RNNs

Leading frameworks support

Multi-GPU & Multi-node



NCCL 2

Multi-node distributed training (multiple machines)

Leading frameworks support

High-performance Inference Engine

TensorRT 5

TensorFlow model reader

Object detection

INT8 RNNs support

Jetson Xavier

developer.nvidia.com/jetson-xavier

30W • 15W • 10W

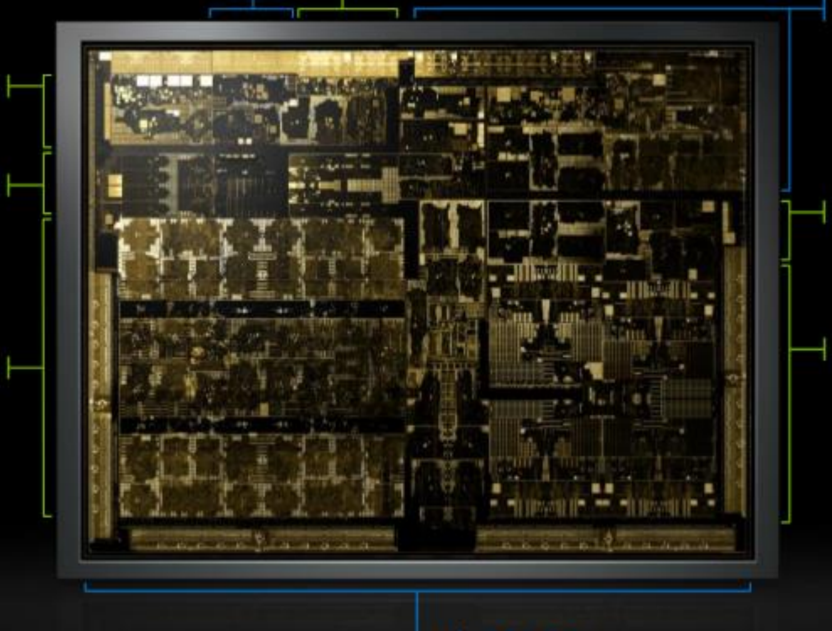
512 Volta CUDA Cores

8 core CPU

32 DL TOPS

XAVIER

World's First Autonomous Machines Processor



16 Lane CSI
109 Gbps CPHY 1.1
1Gb Ethernet

DLA
5.7 TFLOPS FP16
11.4 TOPS INT8

Industry Standard High-Speed IO
PCle 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

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

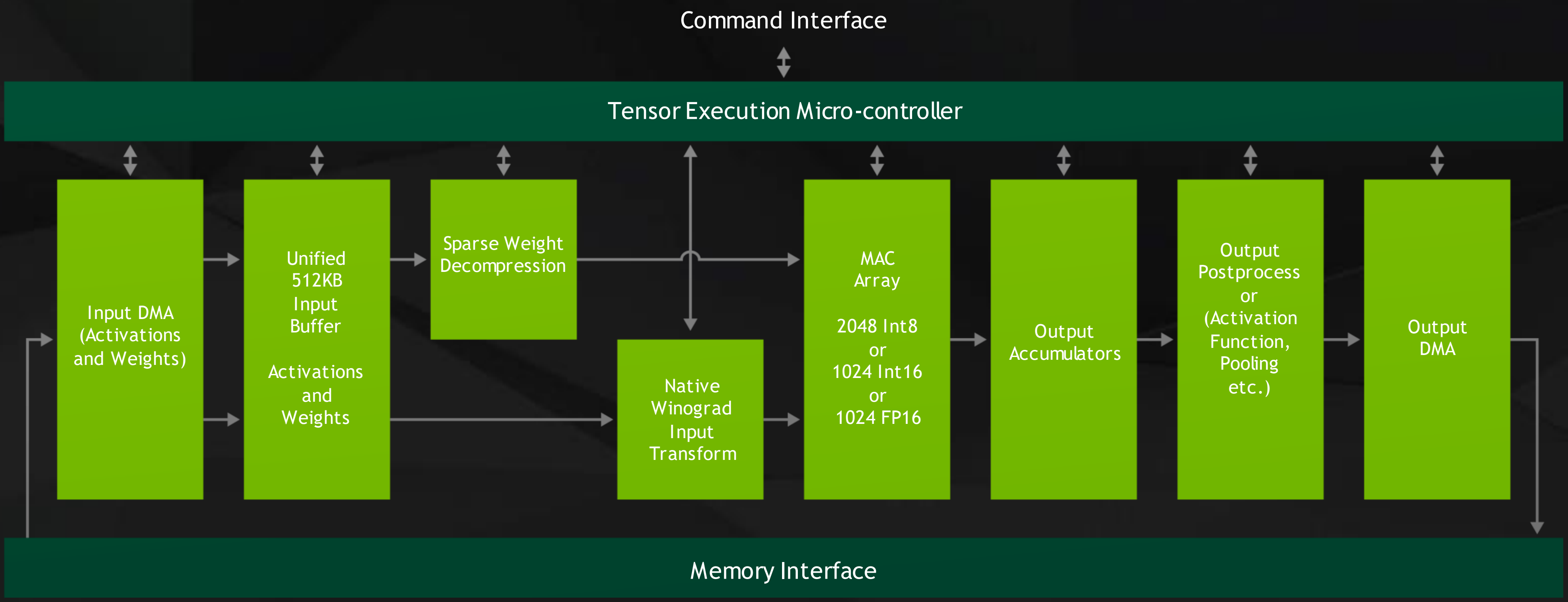
ISP
2.4 GPIX/s
Native Full-range HDR
Tile-based Processing

Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6 (est.)

256-Bit LPDDR4X
137 GB/s

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

XAVIER DLA
NOW OPEN SOURCE


Command Interface
Tensor Execution Micro-controller
Input DMA (Activations and Weights)
Unified 512KB Input Buffer
Activations and Weights
Sparse Weight Decompression
Native Winograd Input Transform
MAC Array
2048 Int8 or 1024 Int16 or 1024 FP16
Output Accumulators
Output Postprocess or (Activation Function, Pooling etc.)
Output DMA
Memory Interface

WWW.NVDLA.ORG

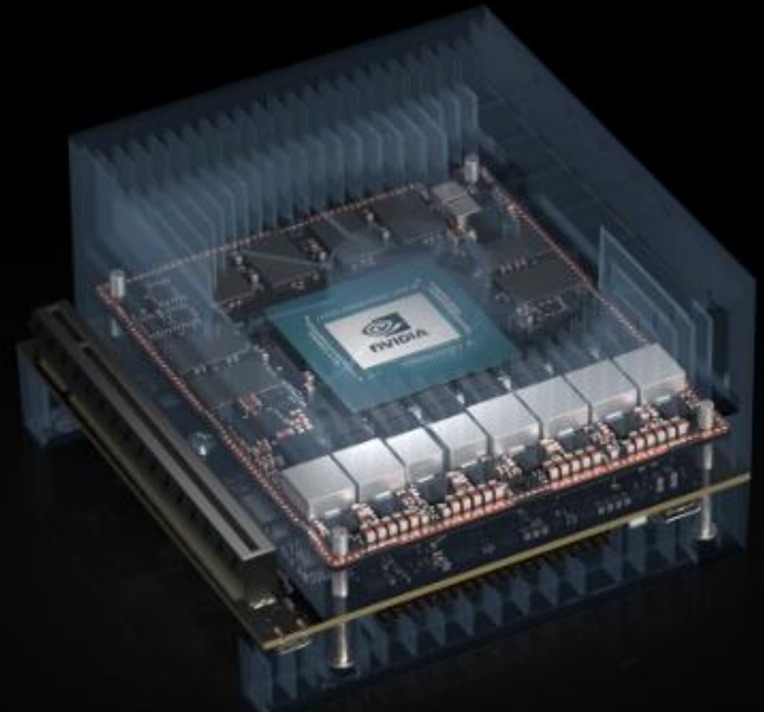
JETSON XAVIER DEVELOPER KIT

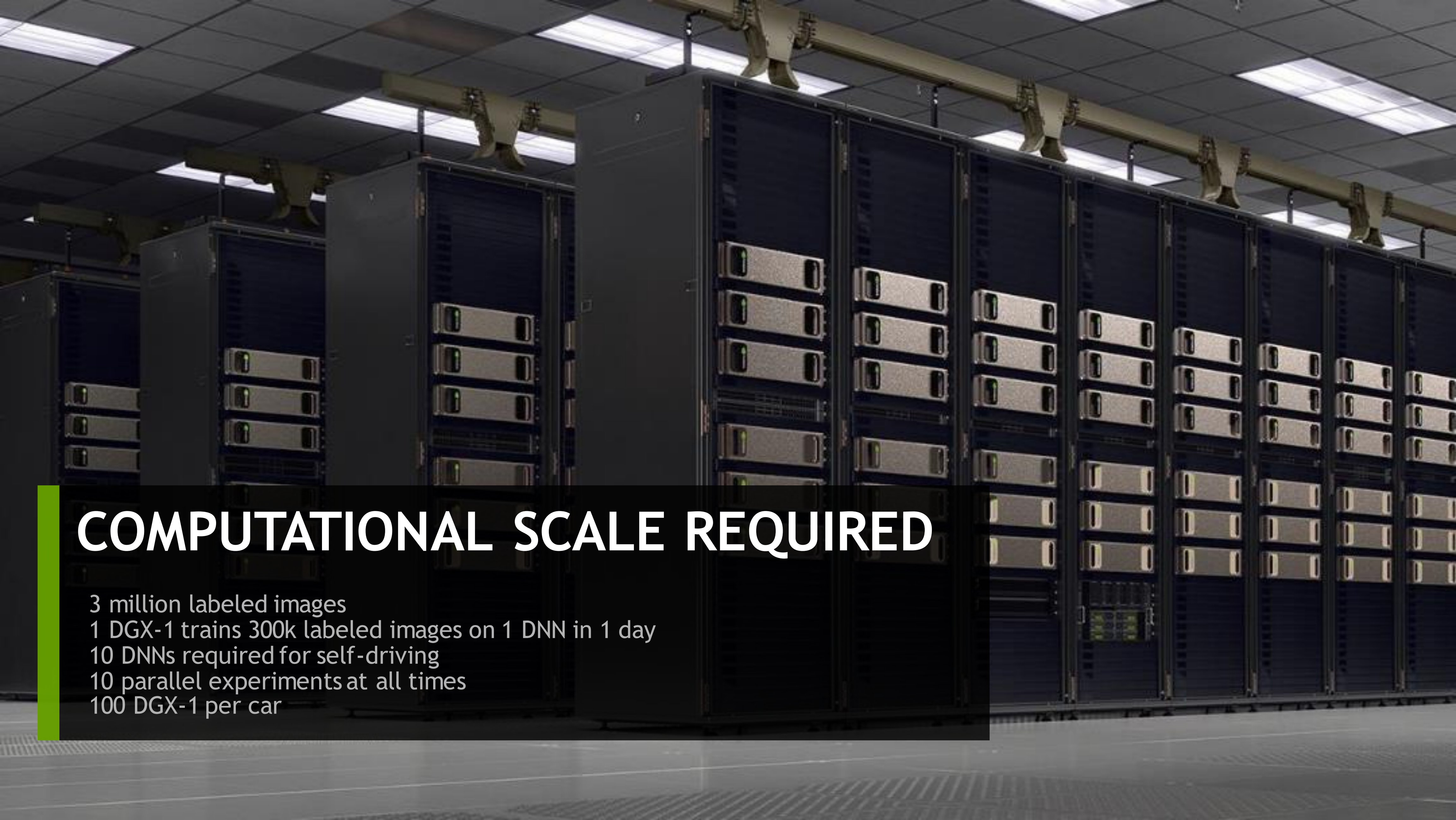
$1299 (pre-order only)

Available from distributors WW

Early access August 2018

JETSON XAVIER DEVELOPER KIT





COMPUTATIONAL SCALE REQUIRED

3 million labeled images
1 DGX-1 trains 300k labeled images on 1 DNN in 1 day
10 DNNs required for self-driving
10 parallel experiments at all times
100 DGX-1 per car

TensorRT SUPPORTS Xavier

Optimized Inference on the World's Most Powerful SoC

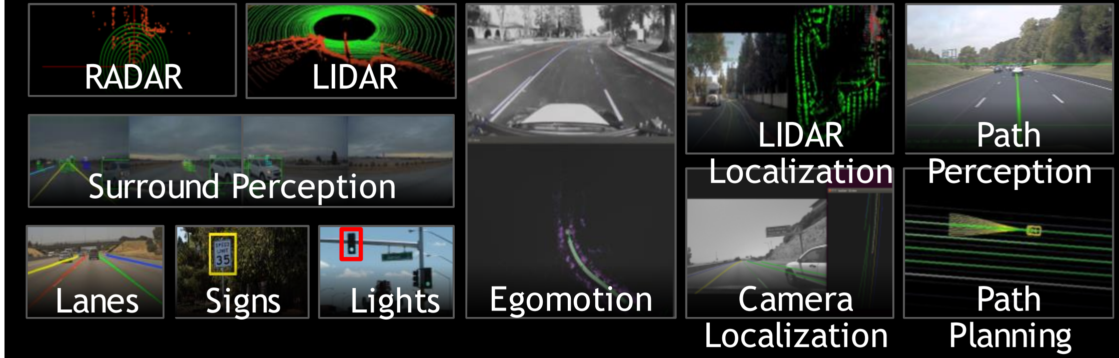
Deploy deep learning inference on Xavier platforms through NVIDIA DRIVE AI Platforms

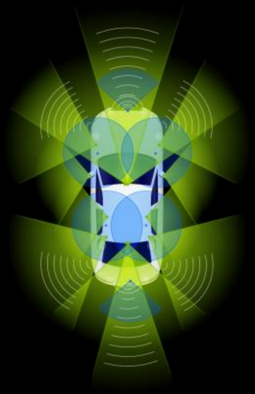
Import models in any framework (including TensorFlow, Caffe and Torch) through ONNX, Universal Framework Format or custom C/C++ API

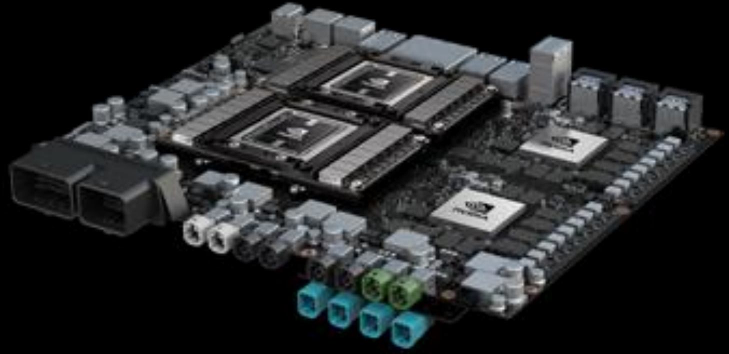
Optimize CNN, RNN and novel neural network layers and deploy reduced precision on Tensor Cores

Download for development or host environment today

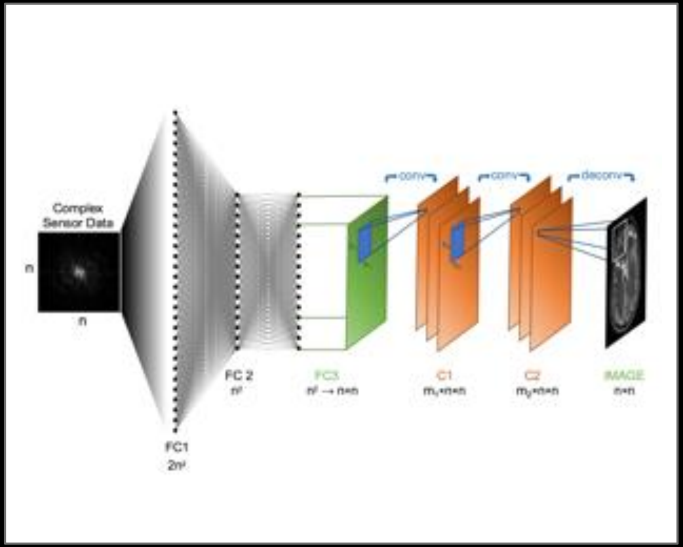
developer.nvidia.com/tensorrt





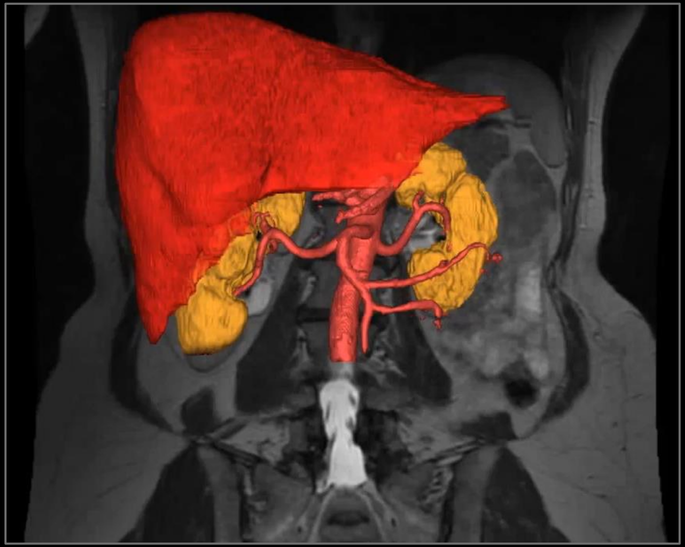


AI IS THE FUTURE OF INTELLIGENT INSTRUMENTS



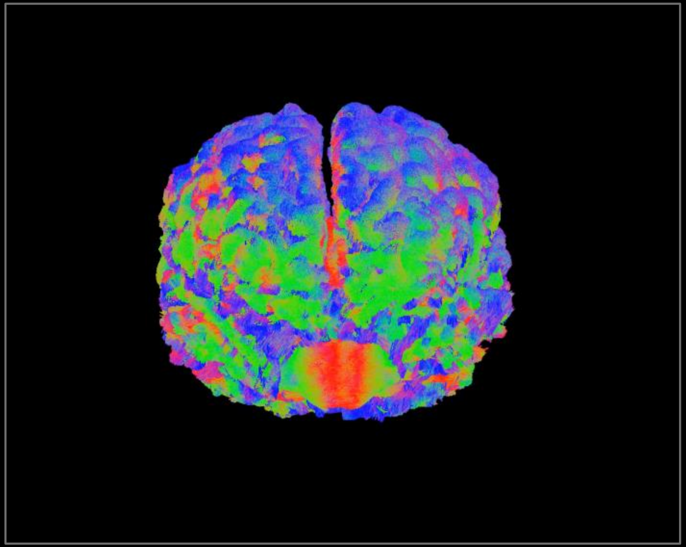
AUTOMAP RECONSTRUCTION

Source: https://arxiv.org/pdf/1704.08841.pdf



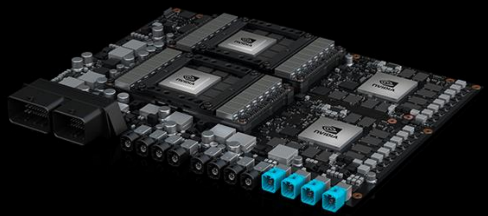
ANATOMY DETECTION

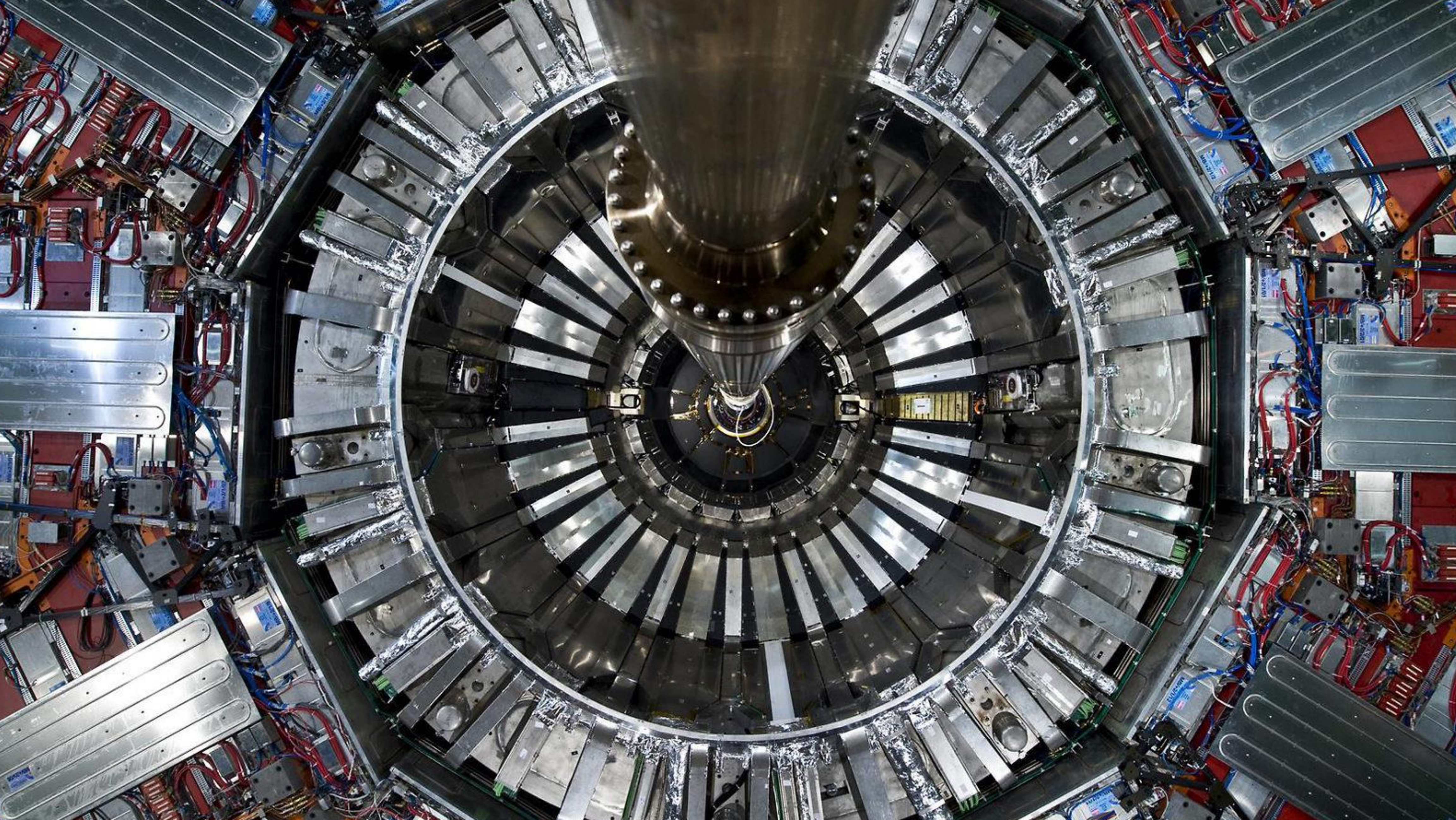
Source: GE Healthcare



INTELLIGENT RENDERING

Source: GE Healthcare





NVIDIA POWERS FASTEST SUPERCOMPUTERS IN US, EUROPE, JAPAN, INDUSTRY

17 of World's 20 Most Energy-efficient Supercomputers

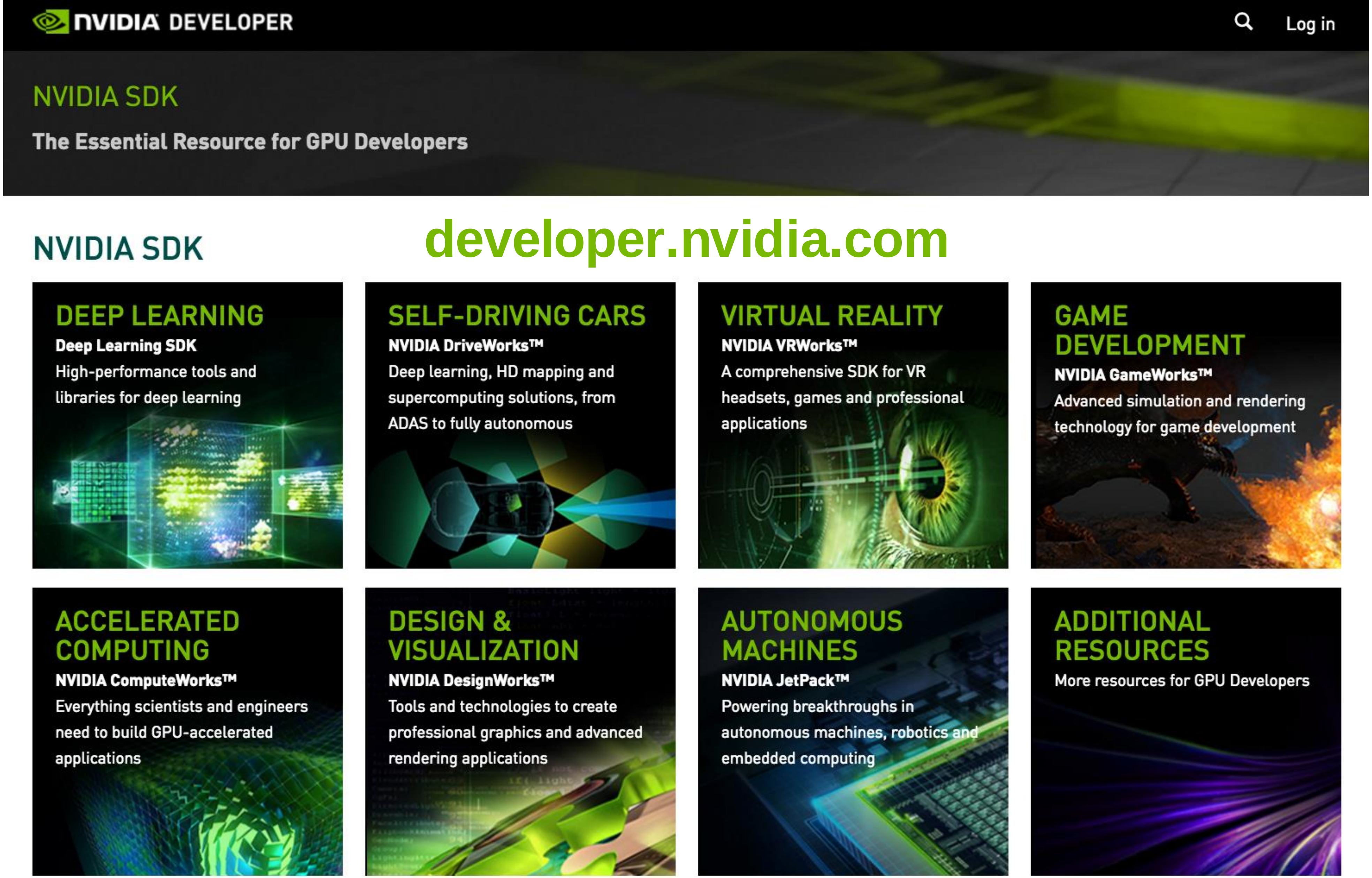
ORNL Summit
World's Fastest
27,648 GPUs| 122 PF

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



NVIDIA SDK

The Essential Resource for GPU Developers

developer.nvidia.com

NVIDIA SDK

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks™
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

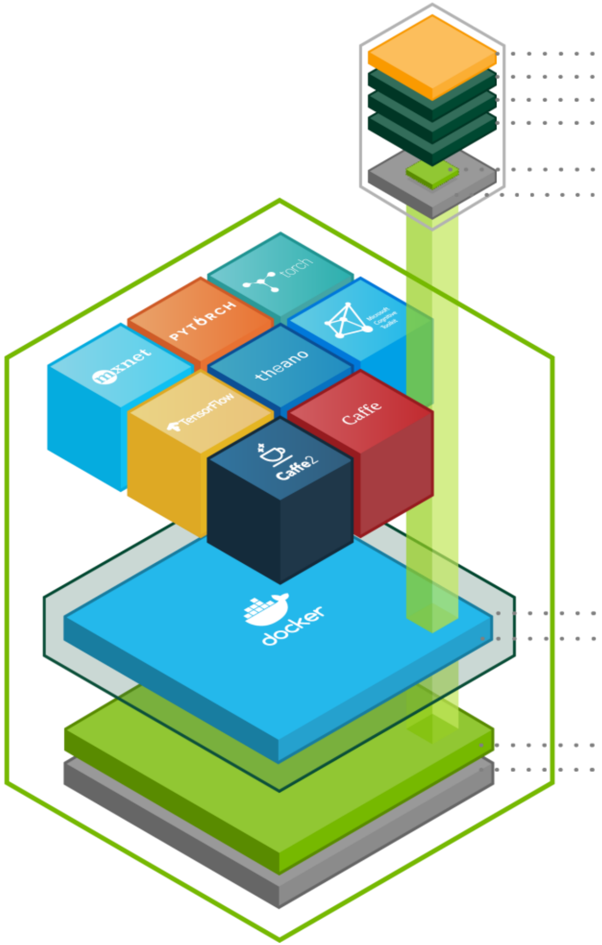
GAME DEVELOPMENT
NVIDIA GameWorks™
Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING
NVIDIA ComputeWorks™
Everything scientists and engineers need to build GPU-accelerated applications

DESIGN & VISUALIZATION
NVIDIA DesignWorks™
Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES
NVIDIA JetPack™
Powering breakthroughs in autonomous machines, robotics and embedded computing

ADDITIONAL RESOURCES
More resources for GPU Developers



CONTAINERIZED APPLICATION

DEEP LEARNING APPLICATIONS
DEEP LEARNING FRAMEWORKS
DEEP LEARNING LIBRARIES
CUDA TOOLKIT

MOUNTED NVIDIA DRIVER
CONTAINER OS

CONTAINERIZATION TOOL

NVIDIA CONTAINER RUNTIME FOR DOCKER
DOCKER ENGINE

NVIDIA DRIVER
HOST OS

NVIDIA GPU CLOUD SOFTWARE STACK

WHY CONTAINERS?

Benefits of Containers:

Simplify deployment of GPU-accelerated software, eliminating time-consuming software integration work

Isolate individual deep learning frameworks and applications

Share, collaborate, and test applications across different environments

To learn more:
nvidia.com/ngc
To sign up:
ngc.nvidia.com

CUDA CONTAINERS ON NVIDIA GPU CLOUD

CUDA containers available from NGC Registry at nvcr.io/nvidia/cuda

Three different flavors:

Base

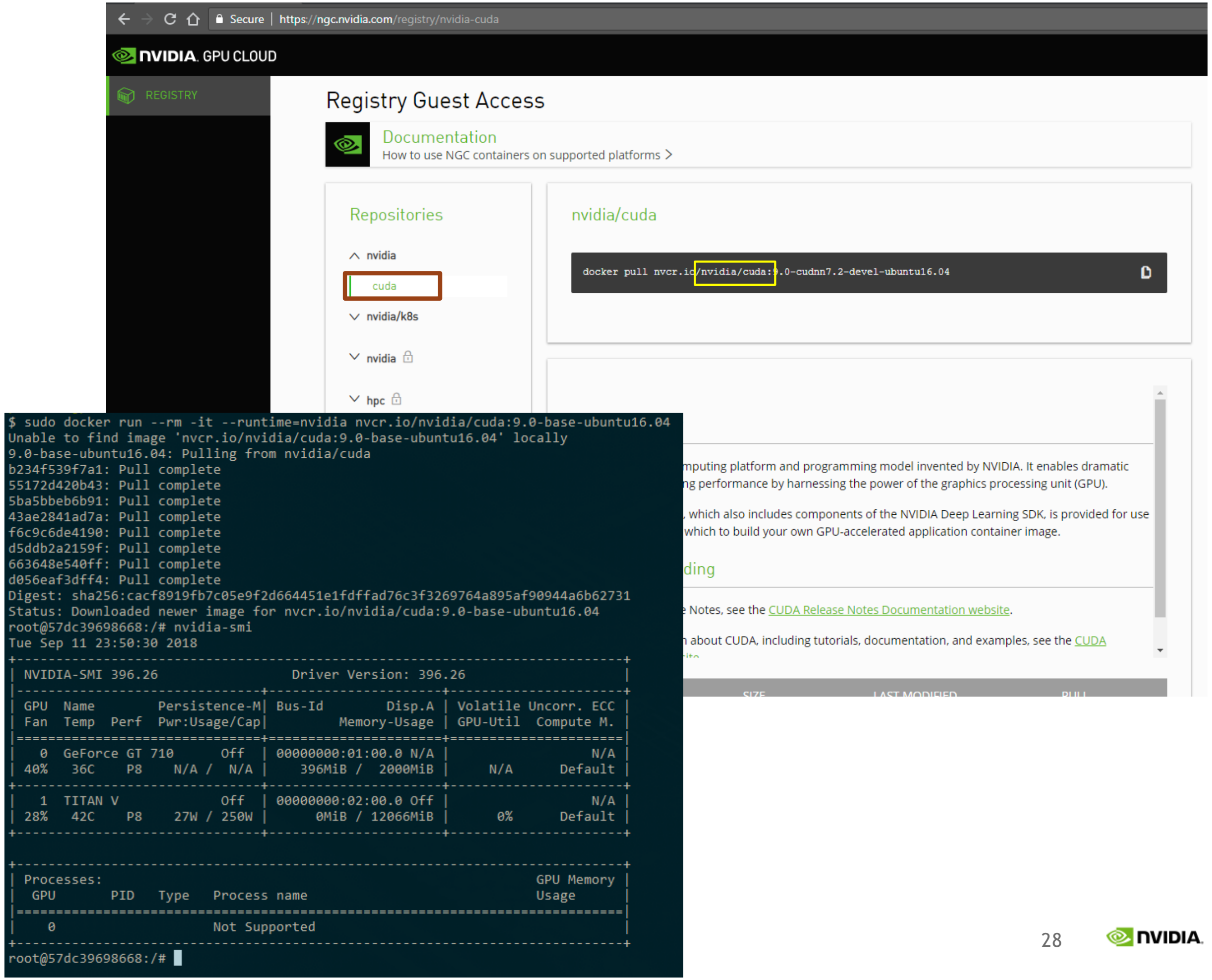
Contains the minimum components required to run CUDA applications

Runtime

Contains *base* + CUDA libraries (e.g. cuBLAS, cuFFT)

Devel

Contains *runtime* + CUDA command line developer tools. Some *devel* tags also include cuDNN



Registry Guest Access

Documentation
How to use NGC containers on supported platforms >

nvidia/cuda

```
docker pull nvcr.io/nvidia/cuda:9.0-cudnn7.2-devel-ubuntu16.04
```

mputing platform and programming model invented by NVIDIA. It enables dramatic ng performance by harnessing the power of the graphics processing unit (GPU).

, which also includes components of the NVIDIA Deep Learning SDK, is provided for use which to build your own GPU-accelerated application container image.

ding

e Notes, see the CUDA Release Notes Documentation website.

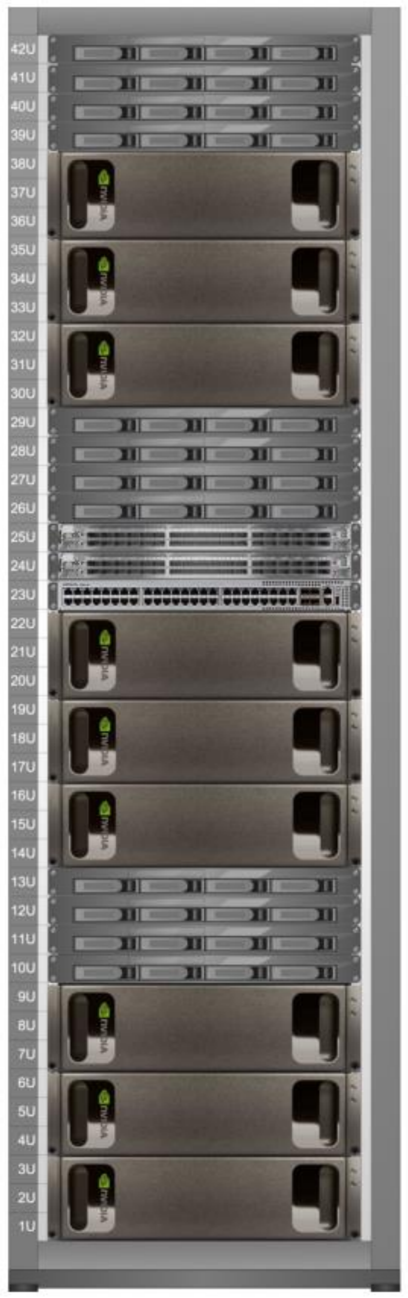
n about CUDA, including tutorials, documentation, and examples, see the CUDA

```
$ sudo docker run --rm -it --runtime=nvidia nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04
Unable to find image 'nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04' locally
9.0-base-ubuntu16.04: Pulling from nvidia/cuda
b234f539f7a1: Pull complete
55172d420b43: Pull complete
5ba5bbeb6b91: Pull complete
43ae2841ad7a: Pull complete
f6c9c6de4190: Pull complete
d5ddb2a2159f: Pull complete
663648e540ff: Pull complete
d056eaf3dff4: Pull complete
Digest: sha256:cacf8919fb7c05e9f2d664451e1fdffad76c3f3269764a895af90944a6b62731
Status: Downloaded newer image for nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04
root@57dc39698668:/# nvidia-smi
Tue Sep 11 23:50:30 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 396.26                 Driver Version: 396.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce GT 710      Off  | 00000000:01:00.0 N/A |                  N/A |
| 40%   36C    P8    N/A /  N/A |    396MiB /  2000MiB |     N/A      Default |
+-------------------------------+----------------------+----------------------+
|   1  TITAN V             Off  | 00000000:02:00.0 Off |                  N/A |
| 28%   42C    P8    27W / 250W |      0MiB / 12066MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0                    Not Supported                                       |
+-----------------------------------------------------------------------------+
root@57dc39698668:/# 
```

DGX POD ARCHITECTURE

a single data center rack containing up to 9x NVIDIA DGX-1 servers, storage, networking & NVIDIA AI software



Nine DGX-1 servers

12 storage servers

10 GbE (min) storage & management switch

Mellanox 100 Gpps intra-rack high speed network switches.

NVIDIA DGX-2

THE LARGEST GPU EVER CREATED

2 PFLOPS | 512GB HBM2 | 10 kW | 350 lbs

NEXT STEPS

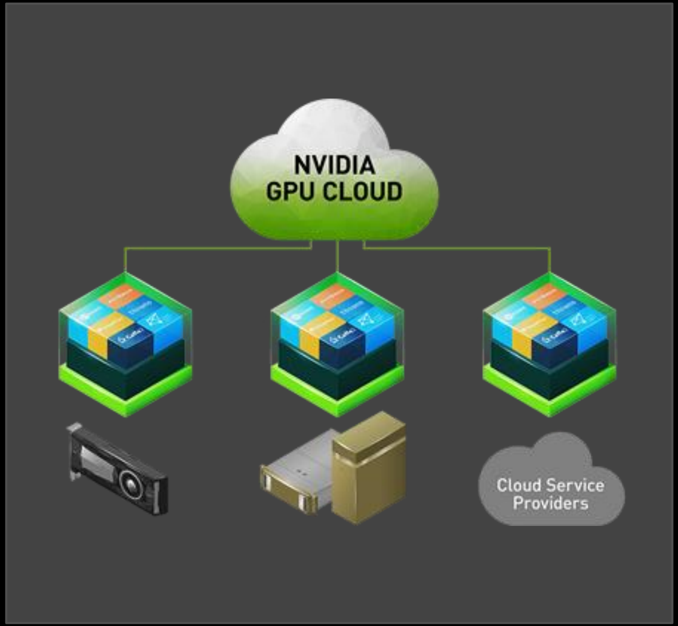
GTC Munich | October 9-11 2018

www.nvidia.com/en-eu/gtc/

25% discount: NVALOWNDES

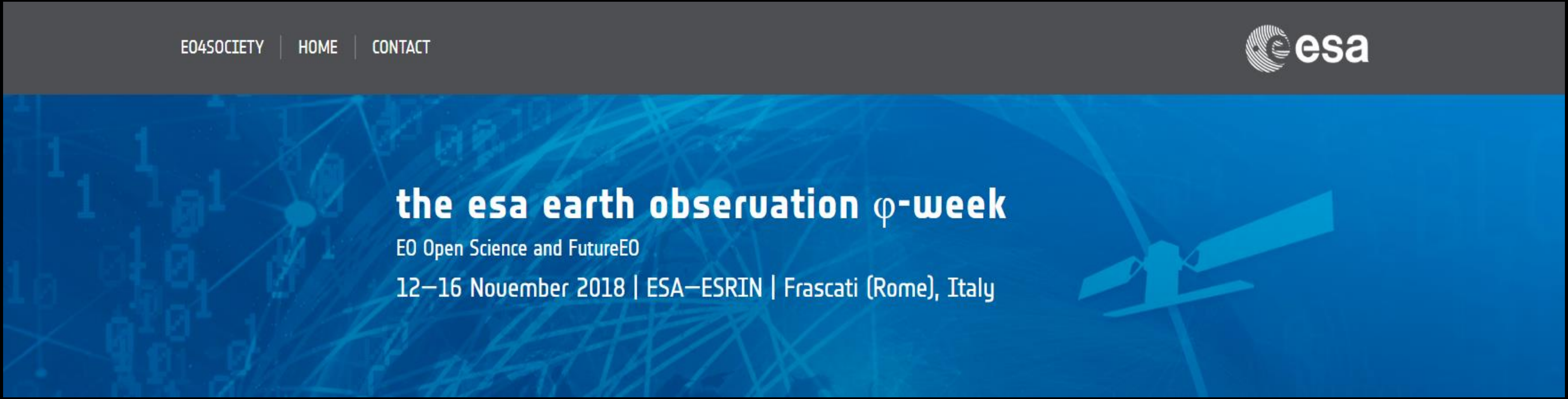
NVIDIA Deep Learning Institute

www.nvidia.com/en-us/deep-learning-ai/education



NGC

www.nvidia.com/en-us/gpu-cloud



the esa earth observation φ-week

EO Open Science and FutureEO

12–16 November 2018 | ESA–ESRIN | Frascati (Rome), Italy