# AI DEEP LEARNING

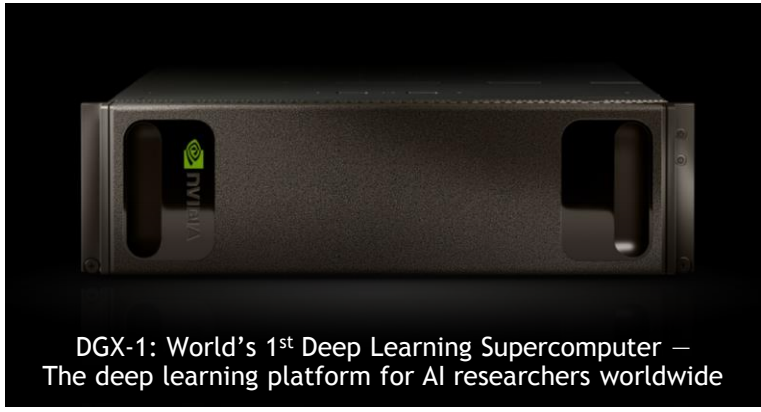
NVIDIA

# NVIDIA

## Pioneered GPU Computing | Founded 1993 | $7B | 9,500 Employees

100M NVIDIA GeForce Gamers —
The world's largest gaming platform


Pioneering AI computing for
self-driving cars


DGX-1: World's 1st Deep Learning Supercomputer —
The deep learning platform for AI researchers worldwide


GE Revolution —
The GPU choice when it really matters

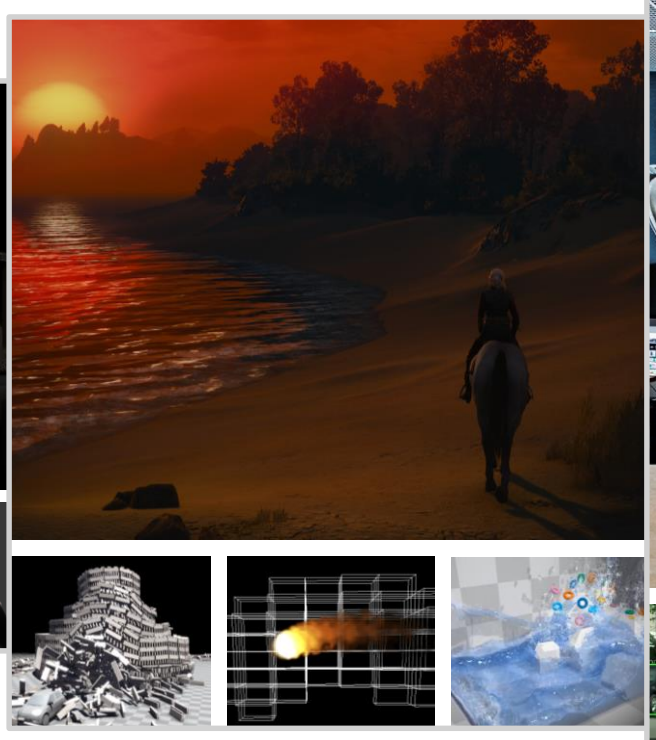

The visualization platform of every car company
and movie studio


The processor of #1 U.S. supercomputer and
9 of 10 of world's most energy-efficient supercomputers

# NVIDIA

## Computing for the Most Demanding Users


GPU Computing

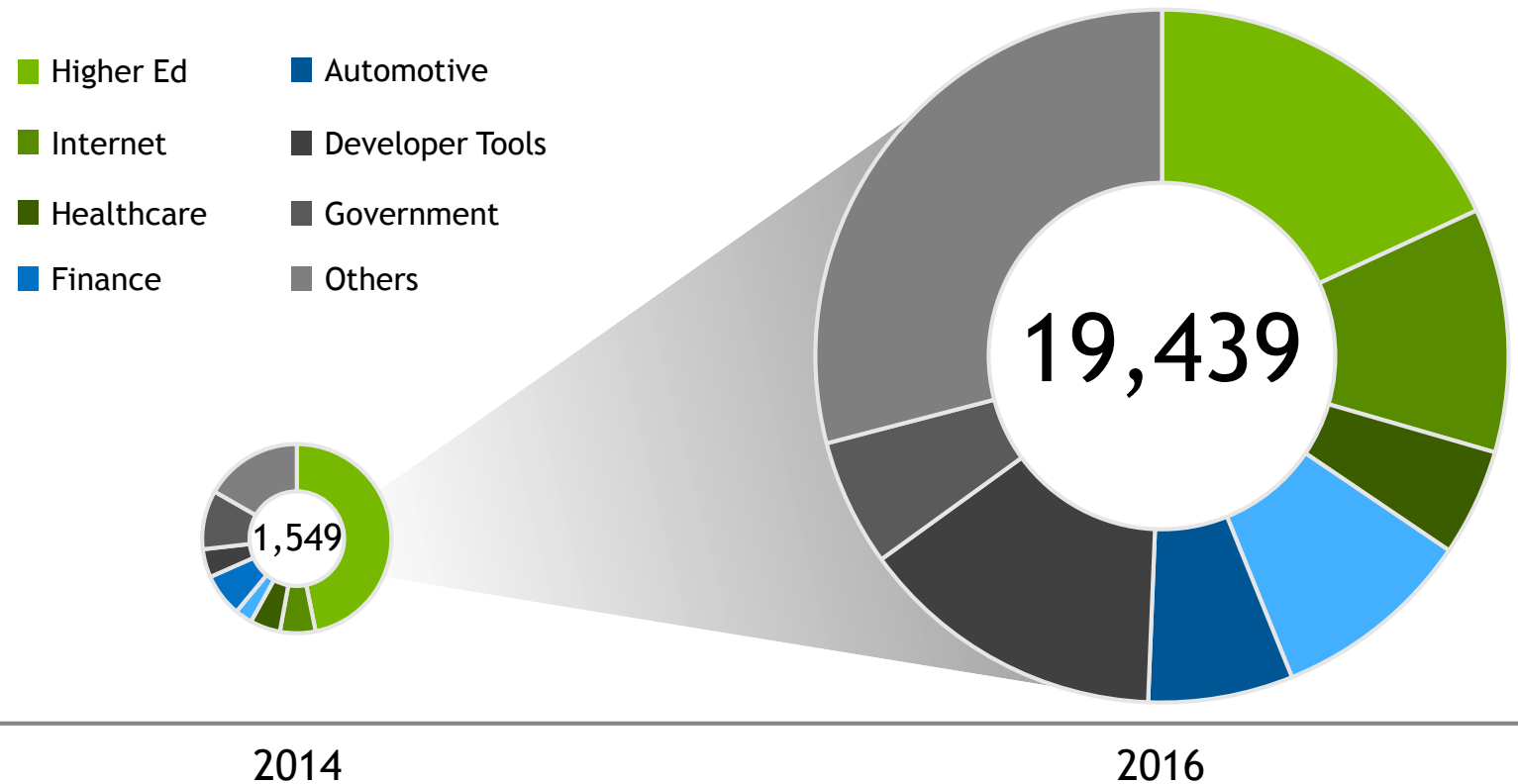


Computing Human Imagination

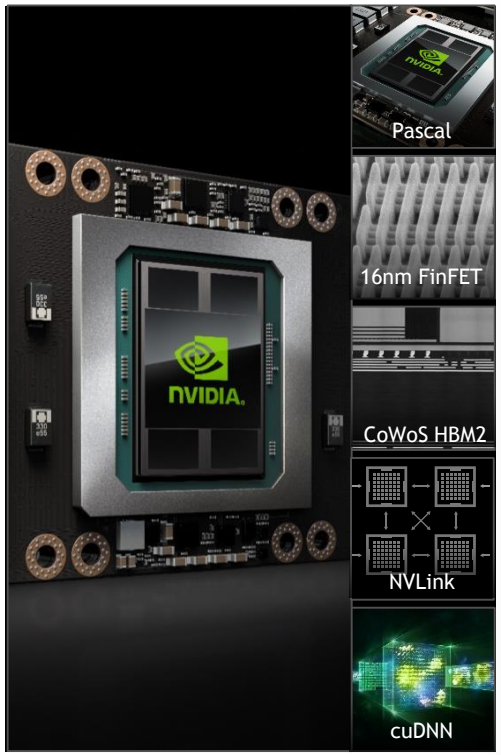

Computing Human Intelligence

# EVERY INDUSTRY HAS AWOKEN TO AI

## Organizations engaged with NVIDIA on Deep Learning

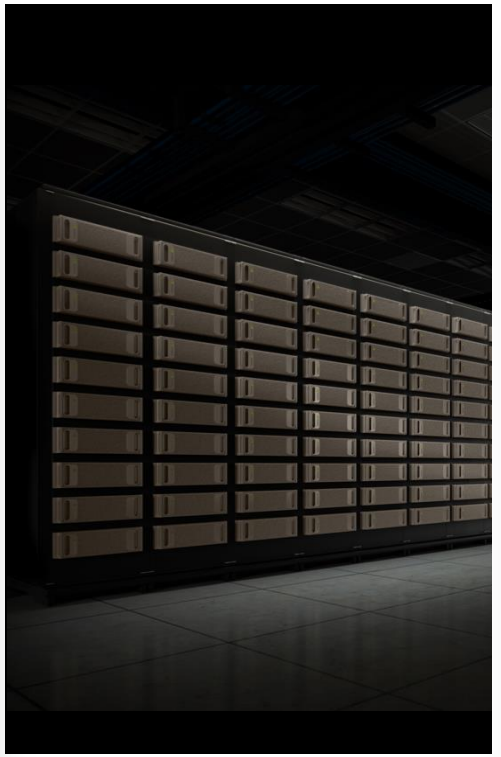


# NVIDIA IS DEEPLY INVESTED IN GPU COMPUTING

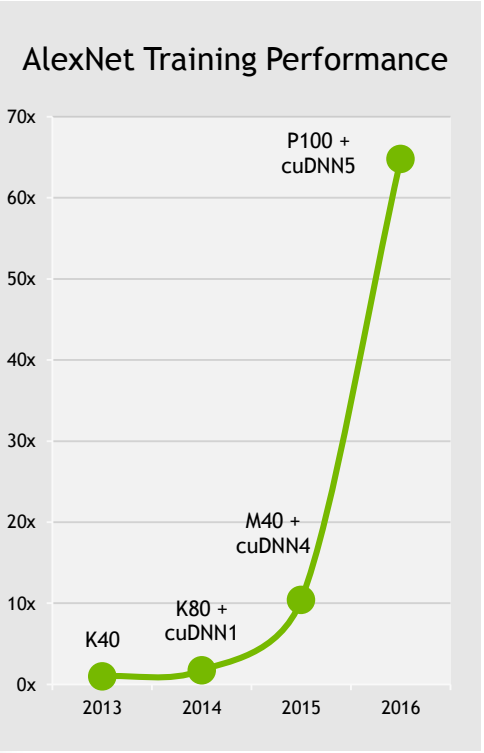



Pascal- 5 Miracles

NVIDIA DGX-1

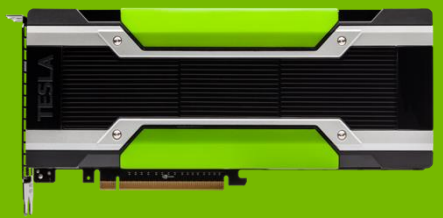


NVIDIA DGX SATURNV



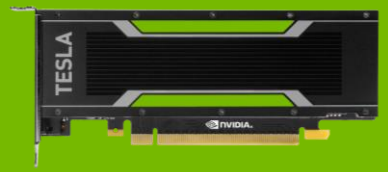

65x in 3 Years

# TESLA P100 ACCELERATORS

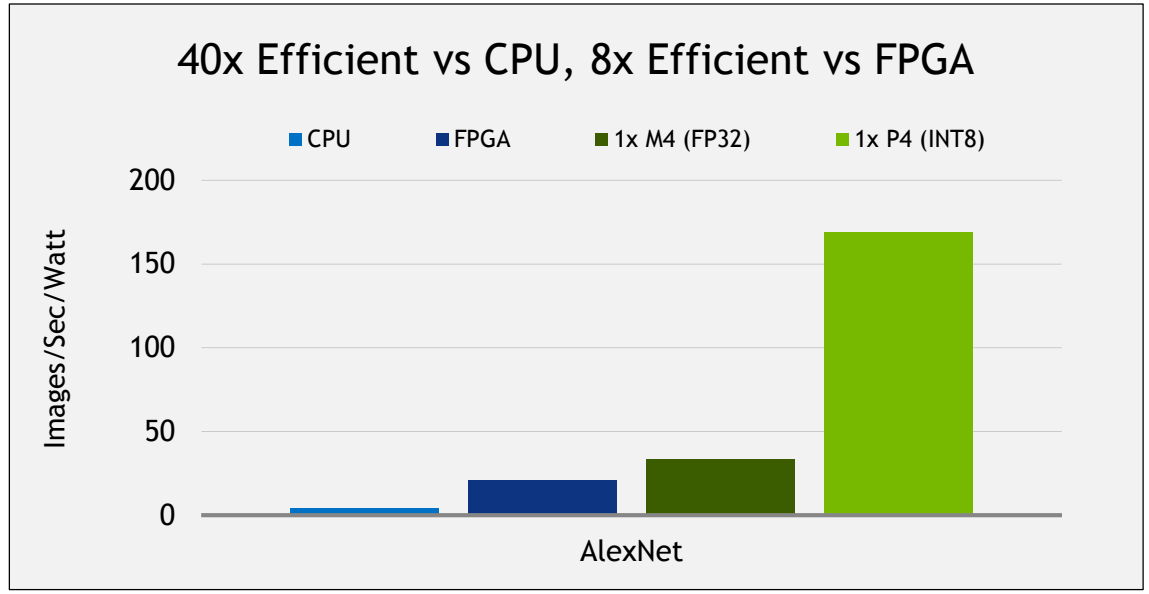
| |  Tesla P100 with NVLink |  Tesla P100 for PCIe |
|---|---|---|
| **Compute** | 5.3 TF DP · 10.6 TF SP · 21.2 TF HP | 4.7 TF DP · 9.3 TF SP · 18.7 TF HP |
| **Memory** | HBM2: 732 GB/s · 16 GB | HBM2 16GB: 732 GB/s<br>HBM2 12GB: 549 GB/s |
| **Interconnect** | NVLink (160 GB/s) + PCIe Gen3 (32 GB/s) | PCIe Gen3 (32 GB/s) |
| **Programmability** | Page Migration Engine<br>Unified Memory | Page Migration Engine<br>Unified Memory |
| **Power** | 300W | 250W |

Interconnect Speed is measured bi-directional

# TESLA P4



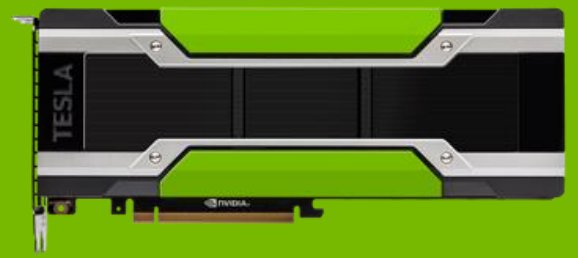

Maximum Efficiency for Scale-out Servers

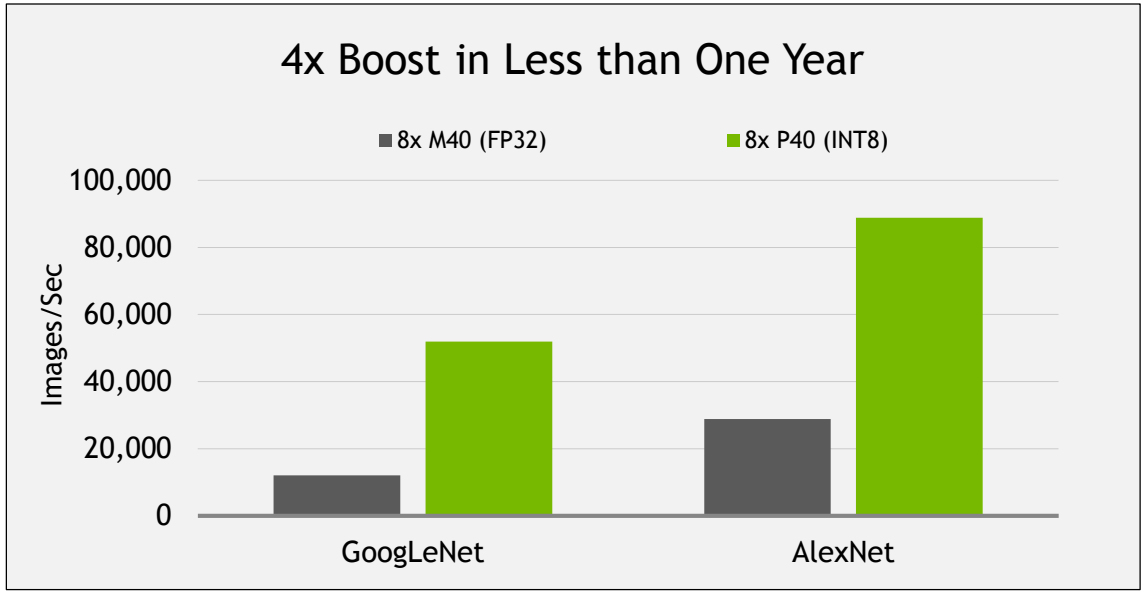

| P4 | |
|---|---|
| **# of CUDA Cores** | 2560 |
| **Peak Single Precision** | 5.5 TeraFLOPS |
| **Peak INT8** | 22 TOPS |
| **Low Precision** | 4x 8-bit vector dot product with 32-bit accumulate |
| **Video Engines** | 1x decode engine, 2x encode engine |
| **GDDR5 Memory** | 8 GB @ 192 GB/s |
| **Power** | 50W & 75 W |

*AlexNet, batch size = 128, CPU: Intel E5-2690v4 using Intel MKL 2017, FPGA is Arria10-115*
*1x M4/P4 in node, P4 board power at 56W, P4 GPU power at 36W, M4 board power at 57W, M4 GPU power at 39W, Perf/W chart using GPU power*

# TESLA P40




Highest Throughput for Scale-up Servers


| P40 | |
|---|---|
| **# of CUDA Cores** | 3840 |
| **Peak Single Precision** | 12 TeraFLOPS |
| **Peak INT8** | 47 TOPS |
| **Low Precision** | 4x 8-bit vector dot product with 32-bit accumulate |
| **Video Engines** | 1x decode engine, 2x encode engines |
| **GDDR5 Memory** | 24 GB @ 346 GB/s |
| **Power** | 250W |

*GoogLeNet, AlexNet, batch size = 128, CPU: Dual Socket Intel E5-2697v4*

# TESLA PLATFORM

## Leading Data Center Platform for HPC and AI

**APPLICATIONS & SERVICES**

amazon alexa · IBM Watson · Microsoft Cognitive Services · NVIDIA DIGITS · Google Cloud Platform Machine Learning Services

AI TRAINING & INFERENCE

ANSYS · SIMULIA · GROMACS FAST. FLEXIBLE. FREE. · VASP · +400 More Applications

HPC

**INDUSTRY TOOLS**

Caffe · mxnet · theano · Microsoft CNTK · torch · TensorFlow

FRAMEWORKS

ResNet · GoogleNet · AlexNet · DeepSpeech · Inception · BigLSTM

MODELS

allinea · ArrayFire · CONTINUUM ANALYTICS · WOLFRAM · MathWorks

ECOSYSTEM TOOLS & LIBRARIES

**NVIDIA SDK**

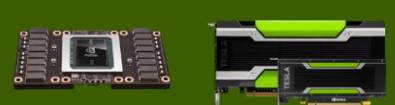
cuDNN · TensorRT · cuBLAS · NCCL · DeepStream SDK

DEEP LEARNING SDK

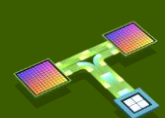
C/C++ · Fortran · CUDA C/C++ · PGI OpenACC Directives for Accelerators

COMPUTEWORKS

**TESLA GPU & SYSTEMS**



TESLA GPU



NVLINK

DELL · Hewlett Packard Enterprise · IBM

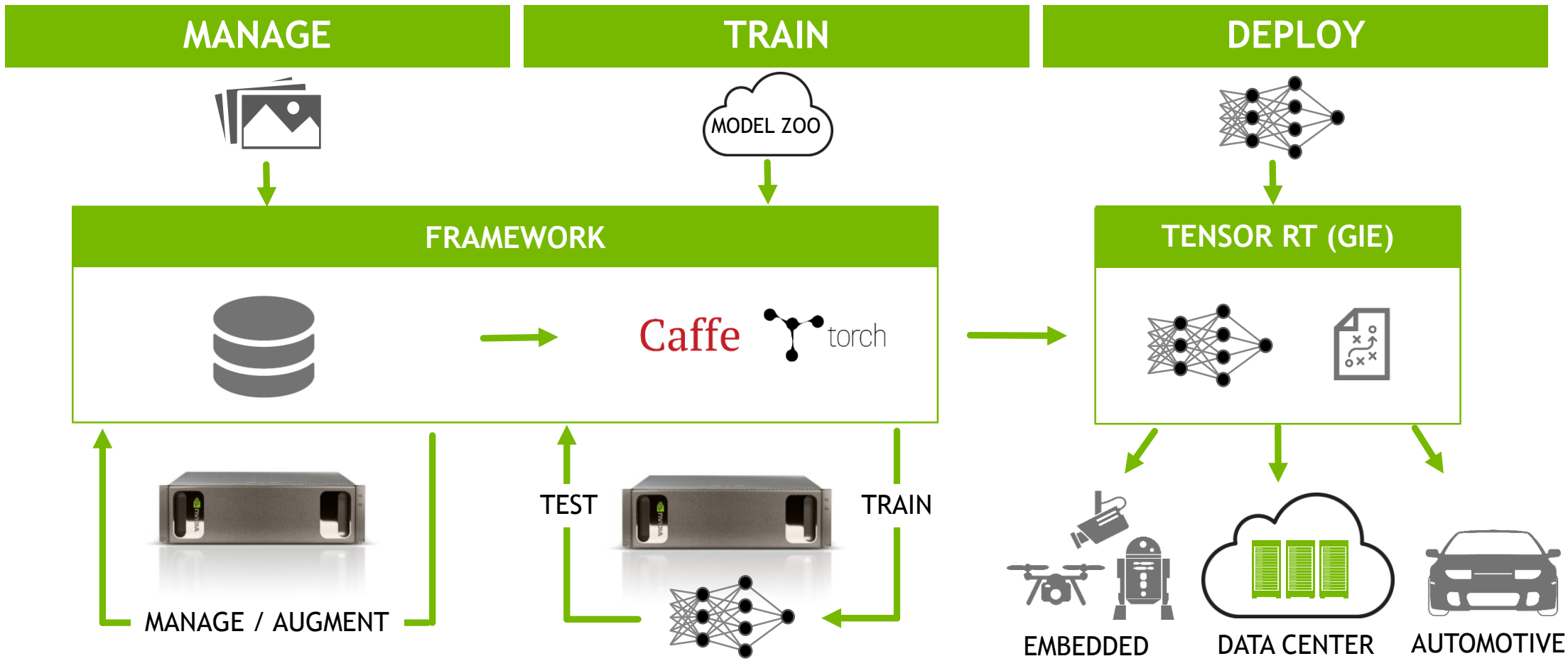
SYSTEM OEM

amazon web services · Google Cloud Platform · Microsoft Azure

CLOUD

# THE WORKFLOW

## A complete GPU-accelerated deep learning workflow

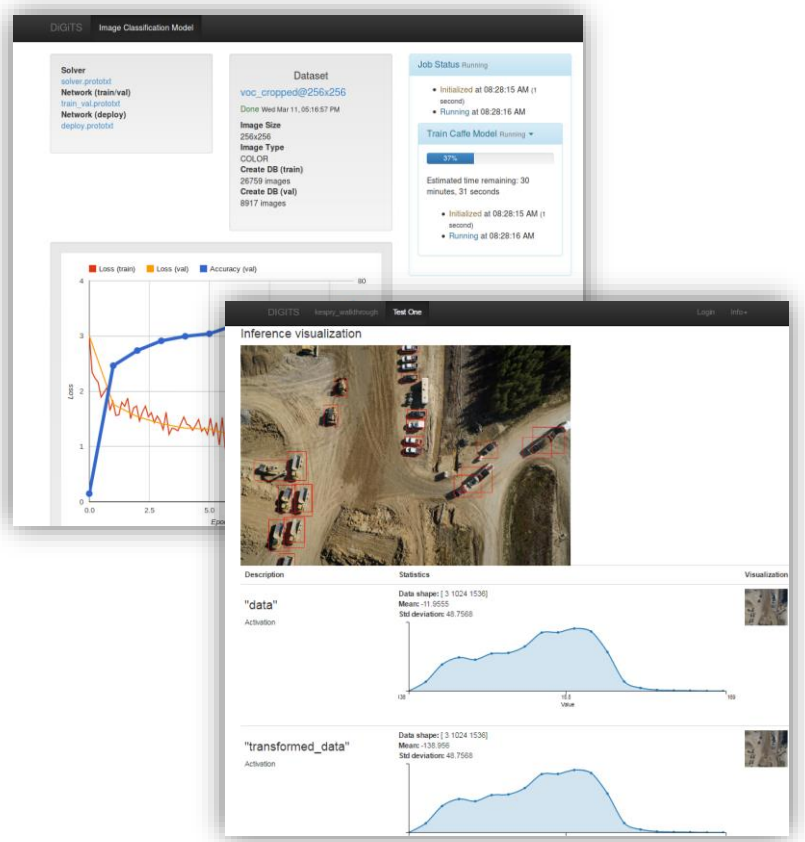


# NVIDIA DIGITS

## Interactive Deep Learning GPU Training System

Interactive deep neural network development environment for image classification and object detection

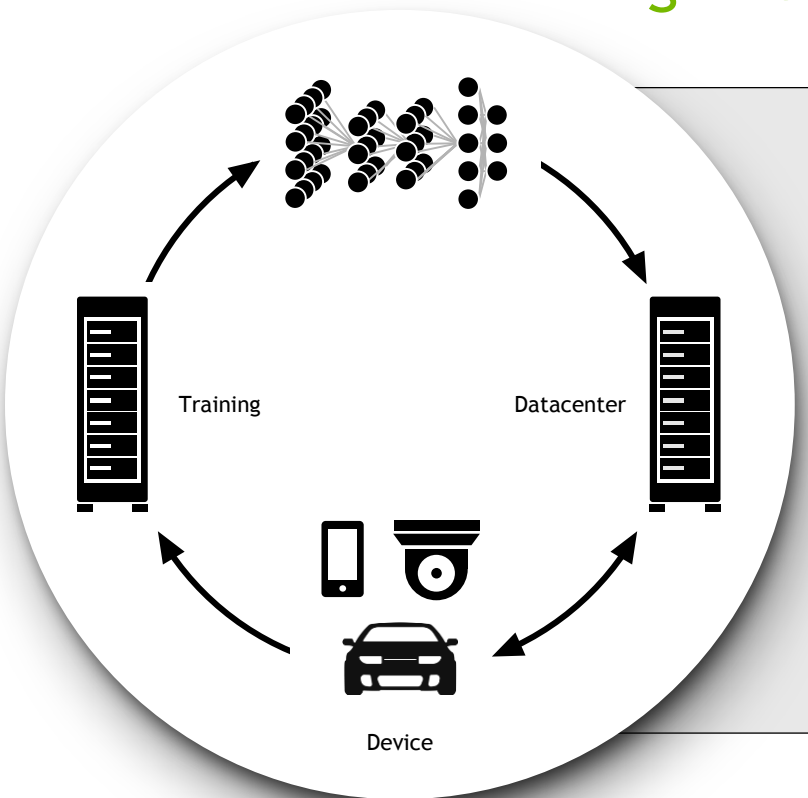
- Schedule, monitor, and manage neural network training jobs
- Analyze accuracy and loss in real time
- Track datasets, results, and trained neural networks
- Scale training jobs across multiple GPUs automatically

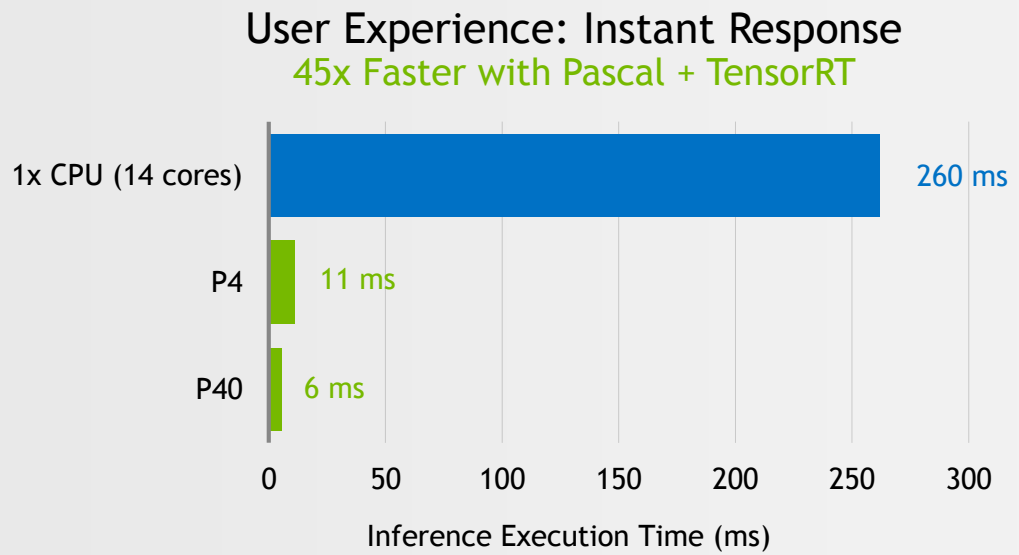



developer.nvidia.com/digits

# INTRODUCING NVIDIA TensorRT

## High Performance Inference Engine




Faster, more responsive AI-powered services such as voice recognition, speech translation

Efficient inference on images, video, & other data in hyperscale production data centers

*Based on VGG-19 from IntelCaffe Github: https://github.com/intel/caffe/tree/master/models/mkl2017_vgg_19*
*CPU: IntelCaffe, batch size = 4, Intel E5-2690v4, using Intel MKL 2017 | GPU: Caffe, batch size = 4, using TensorRT internal version*

# NVIDIA'S GPU EDUCATORS PROGRAM

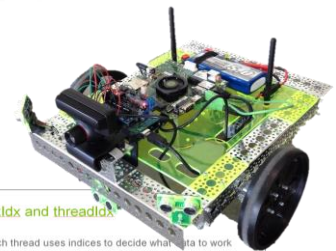
## Advancing STEM Education with Accelerated Computing

The Flagship Offering: **GPU Teaching Kits** - Breaking the barriers of GPU education in academia:

- Lecture slides
- Lecture videos
- Hands-on labs/solutions
- Larger coding projects/solutions
- Quiz/exam questions/solutions
- Text and e-books

Different kits for different courses

- **Accelerated/Parallel Computing** *(available now!)*
- **Robotics** *(available now!)*
- Machine/Deep Learning ***(coming soon!)***
- Computer Vision, Computer Architecture, Computational Domain Sciences, Mathematics, etc. ***(future)***




**Get started today!**

developer.nvidia.com/educators

# DEEP LEARNING INSTITUTE

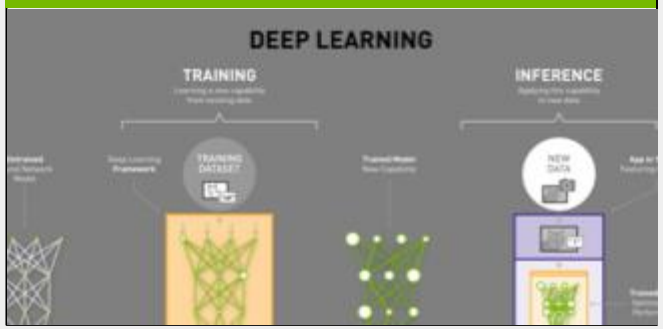
## Overview

DLI Mission: Helping the world to solve the most challenging problems using AI and deep learning.

We help developers, data scientists and engineers to get started in training, optimizing, and deploying neural networks to solve real-world problems in diverse disciplines such as self-driving cars, healthcare, consumer services and robotics.

## For Everyone

### INTRO MATERIALS

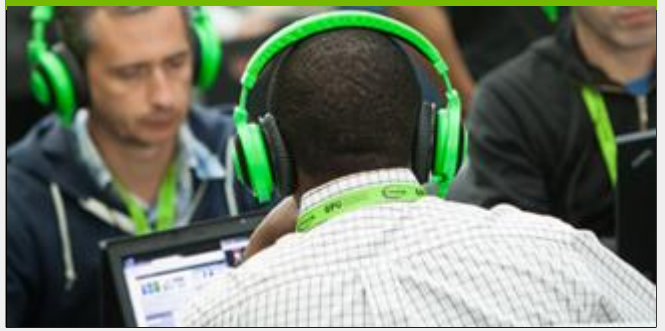



### CASE STUDIES


## For Developers, Data Scientists, Researchers

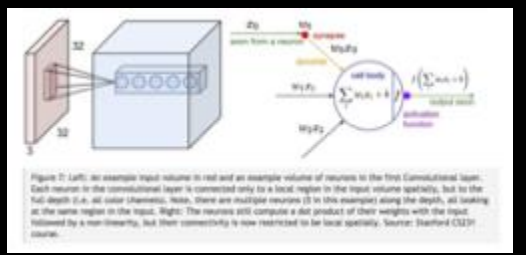
### ON-SITE WORKSHOPS

### SELF-PACED LABS



### PARTNER COURSES


### TECHNICAL BLOGS



# DLI TRAINING OFFERINGS

## SELF-PACED LABS

Online labs offer on-demand 24/7 access to introductory concepts

Prerequisites: Developer, Data Scientist or Researcher. Not designed for non-developers

## INSTRUCTOR-LED WORKSHOPS

Both beginner and intermediate labs are offered - typically “Getting Started” and two more advanced labs constitute a day-long workshop

Prerequisites - same as self-paced labs.

## 5-DAY COURSES

Coming later in 2017

Industry-specific courses teach students to fine-tune a neural network to deploy on a specific platform (e.g. NVIDIA Drive PX2)

Prerequisites - varies by industry

# GET STARTED TODAY

## LEARN THE BASICS

Watch “Deep Learning Demystified”

Listen to the NVIDIA AI Podcast

Review examples of AI in action

## LEARN HOW

**Take a self-paced lab online**

- or-

**Request an On-Site Workshop**

Contact your NVIDIA account mgr

or

DeepLearningInstitute-core@nvidia.com

GPU TECHNOLOGY CONFERENCE



MUNICH 10-12 OCT 2017

# GTC EUROPE 2017

# MUNICH

GTC Europe is the must-attend event of the year in Europe for developers, data scientists and senior decision makers. The 3-day conference will feature over a hundred new sessions on AI and deep learning, self-driving cars, big data analytics, virtual reality, and much more.

SAVE THE DATE

## THE POWER OF GTC

GTC and the global GTC event series offer valuable training and a showcase of the most vital work in the computing industry today – including artificial intelligence and deep learning, virtual reality, and self-driving cars.

Subscribe for updates

# THANK YOU

**PIERO ALTOE'**
*BDM HPC/DL*
*PALTOE@NVIDIA.COM*

NVIDIA