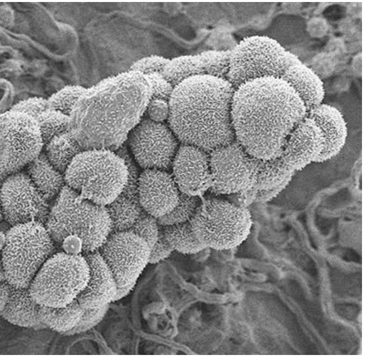
DEEP LEARNING

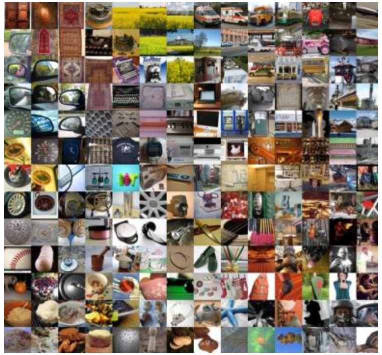
ALISON B LOWNDES

Deep Learning Solutions Architect &
Community Manager | EMEA

THE GPU-ACCELERATED WORLD

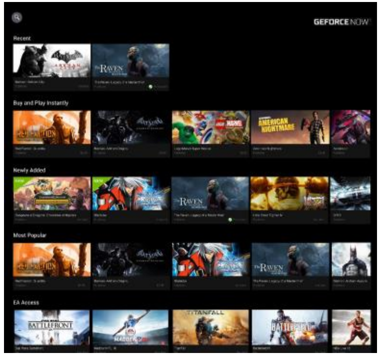


HPC



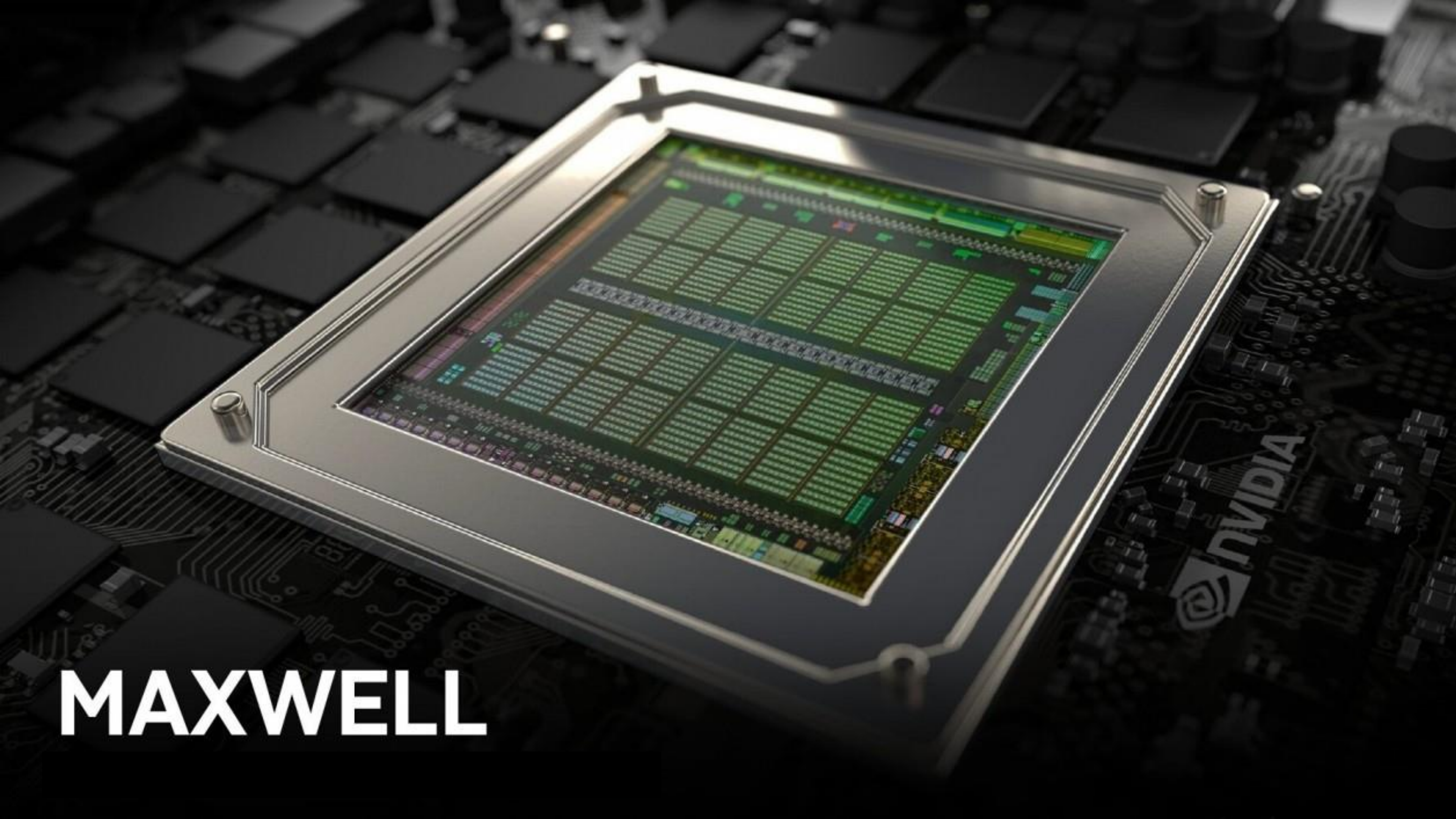
DEEP LEARNING

PC VIRTUALIZATION



CLOUD GAMING

RENDERING



MAXWELL

Why is Deep Learning Hot Now?

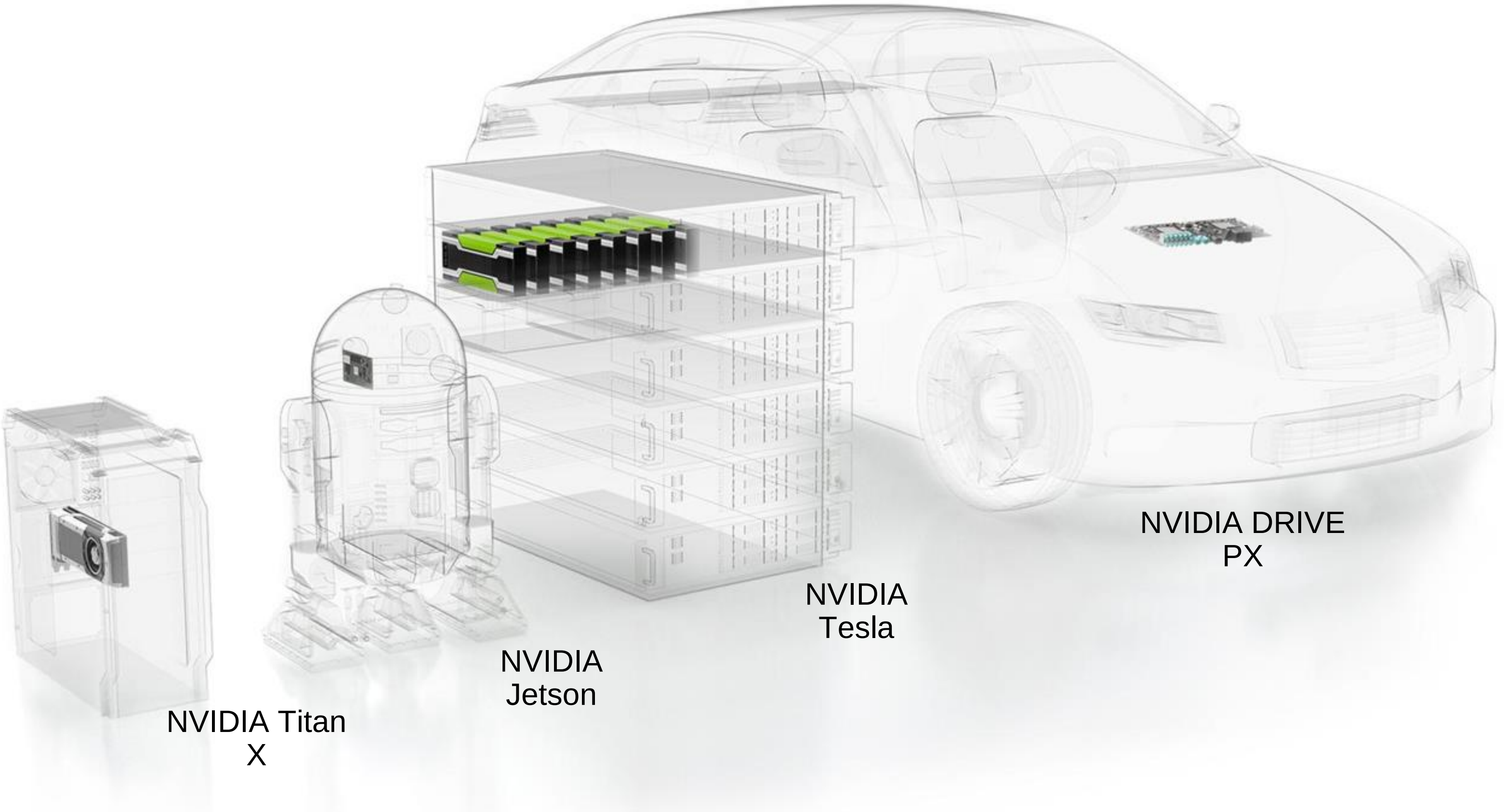
Big Data Availability

facebook	350 millions images uploaded per day
Walmart	2.5 Petabytes of customer data hourly
YouTube	300 hours of video uploaded every minute

New ML Techniques

GPU Acceleration

DEEP LEARNING EVERYWHERE



NVIDIA Titan X

NVIDIA Jetson

NVIDIA Tesla

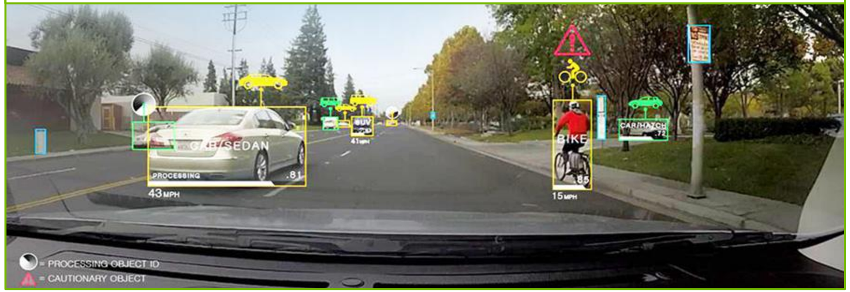
NVIDIA DRIVE PX

Practical Examples of Deep Learning

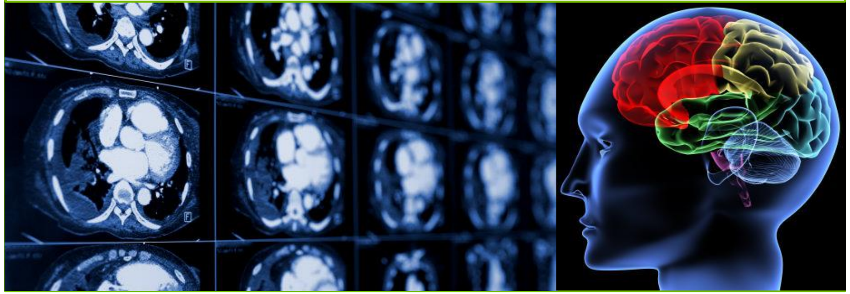
Image Classification, Object Detection, Localization, Action Recognition

Speech Recognition, Speech Translation, Natural Language Processing

Pedestrian Detection, Lane Detection, Traffic Sign Recognition

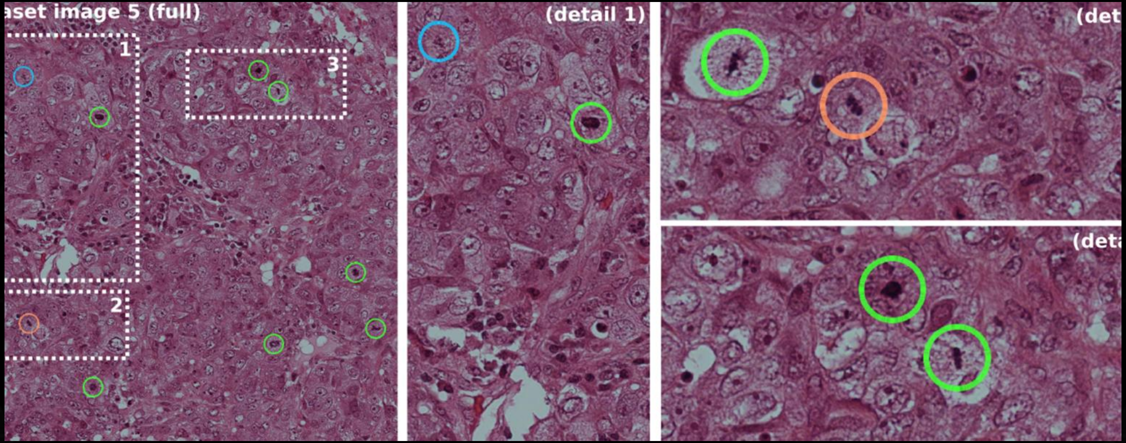


Breast Cancer Cell Mitosis Detection, Volumetric Brain Image Segmentation



CANCER SCREENING

Mitosis Detecion



Ciresan et al. *Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks*, 2013

GPUs and Deep Learning

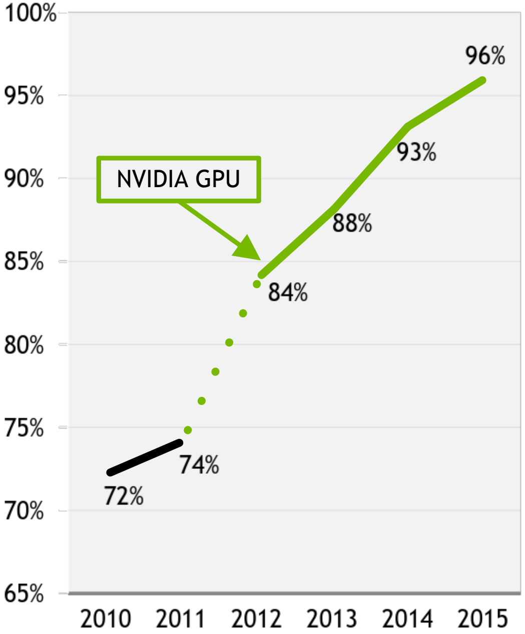
	NEURAL NETWORKS	GPUS
Inherently Parallel	✓	✓
Matrix Operations	✓	✓
FLOPS	✓	✓
Bandwidth	✓	✓

GPUs deliver --

- *same or better prediction accuracy*
- *faster results*
- *smaller footprint*
- *lower power*

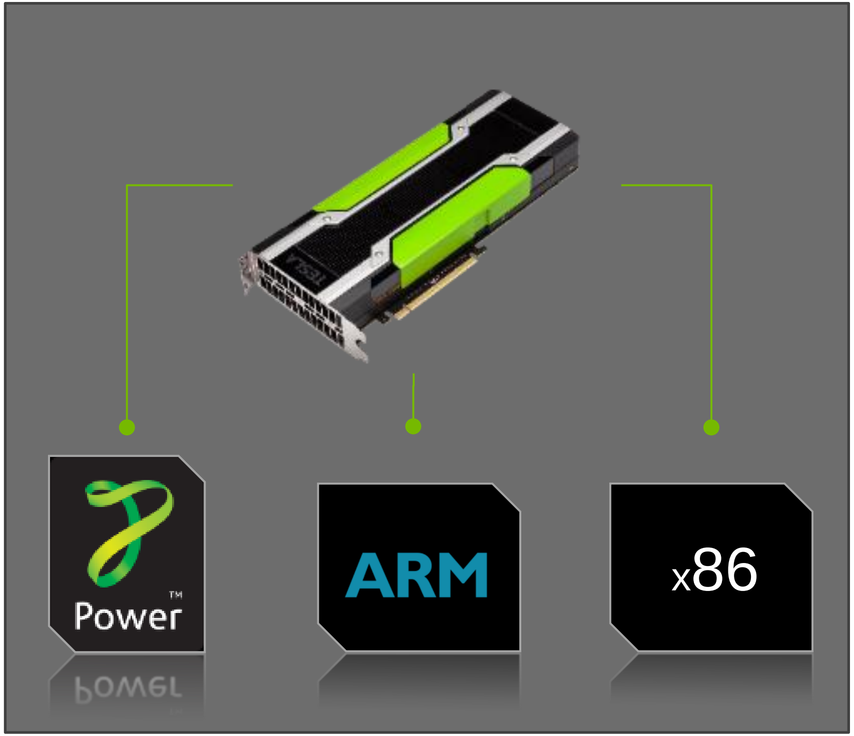
Image Recognition

IMAGENET



Deep Learning Platform Update

GPU Computing



CUDA

Framework to Program NVIDIA GPUs

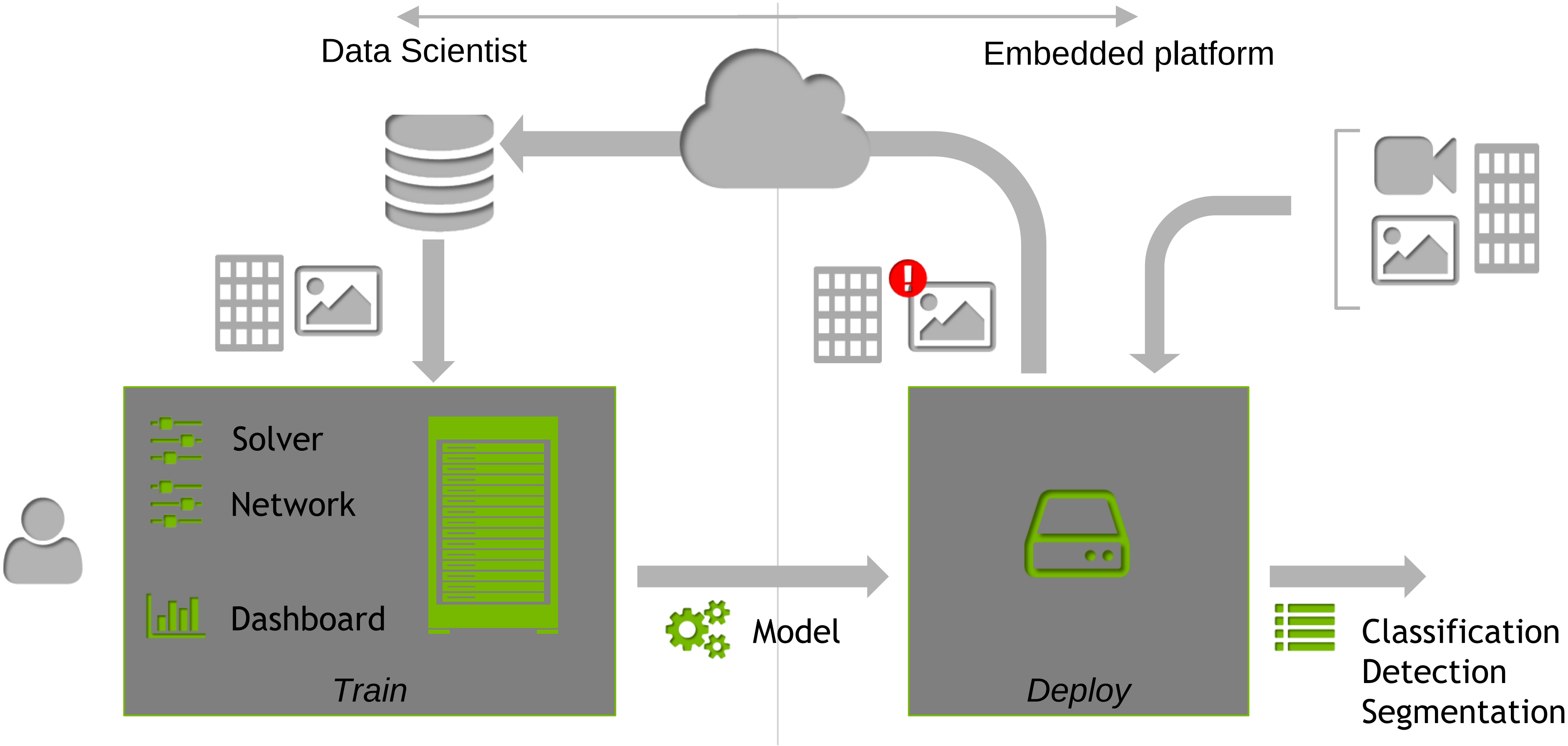
A simple sum of two vectors (arrays) in C

```
void vector_add(int n, const float *a, const float *b, float *c)
{
  for( int idx = 0 ; idx < n ; ++idx )
    c[idx] = a[idx] + b[idx];
}
```

GPU friendly version in CUDA

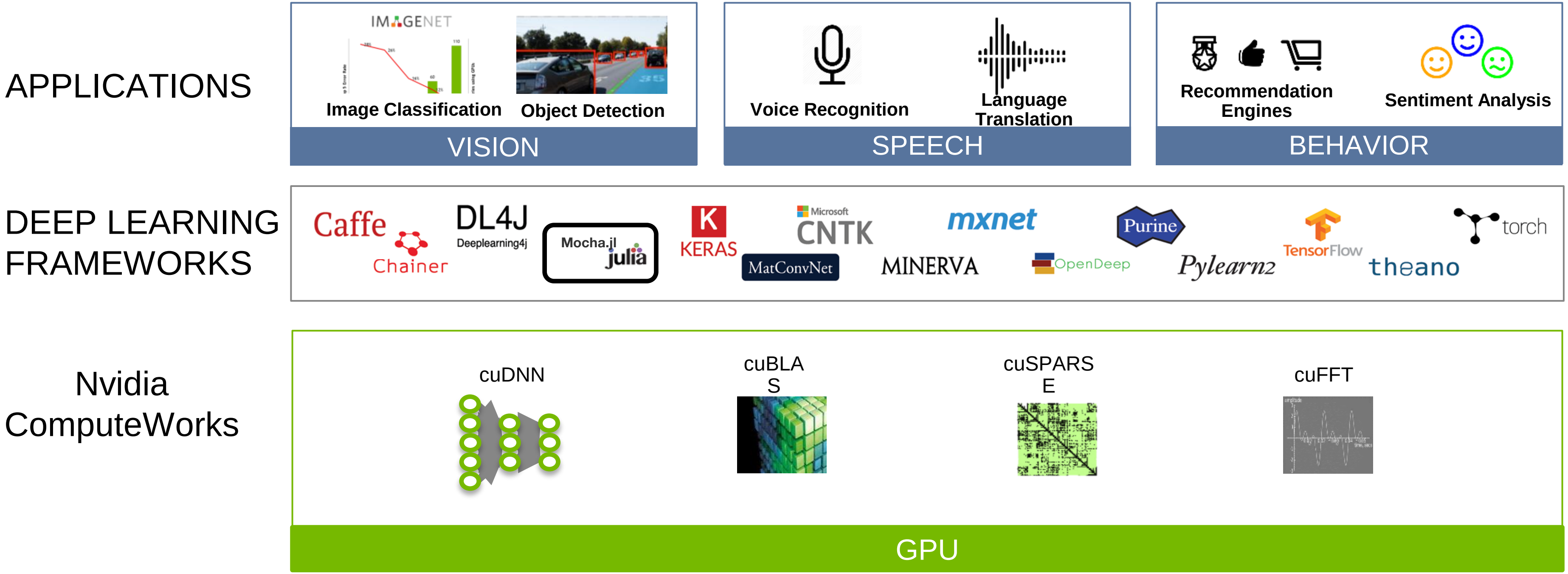
```
__global__ void vector_add(int n, const float *a, const float *b, float *c)
{
  int idx = blockIdx.x*blockDim.x + threadIdx.x;
  if( idx < n )
    c[idx] = a[idx] + b[idx];
}
```

An end-to-end solution



DEEP LEARNING ECOSYSTEM

Deep Learning Frameworks Enable Deep Learning Applications



APPLICATIONS

Image Classification

Object Detection

VISION

Voice Recognition

Language Translation

SPEECH

Recommendation Engines

Sentiment Analysis

BEHAVIOR

DEEP LEARNING FRAMEWORKS

Caffe

Chainer

DL4J
Deeplearning4j

Mocha.jl
julia

KERAS

Microsoft CNTK

MatConvNet

mxnet

MINERVA

Purine

OpenDeep

Pylearn2

TensorFlow

torch

theano

Nvidia ComputeWorks

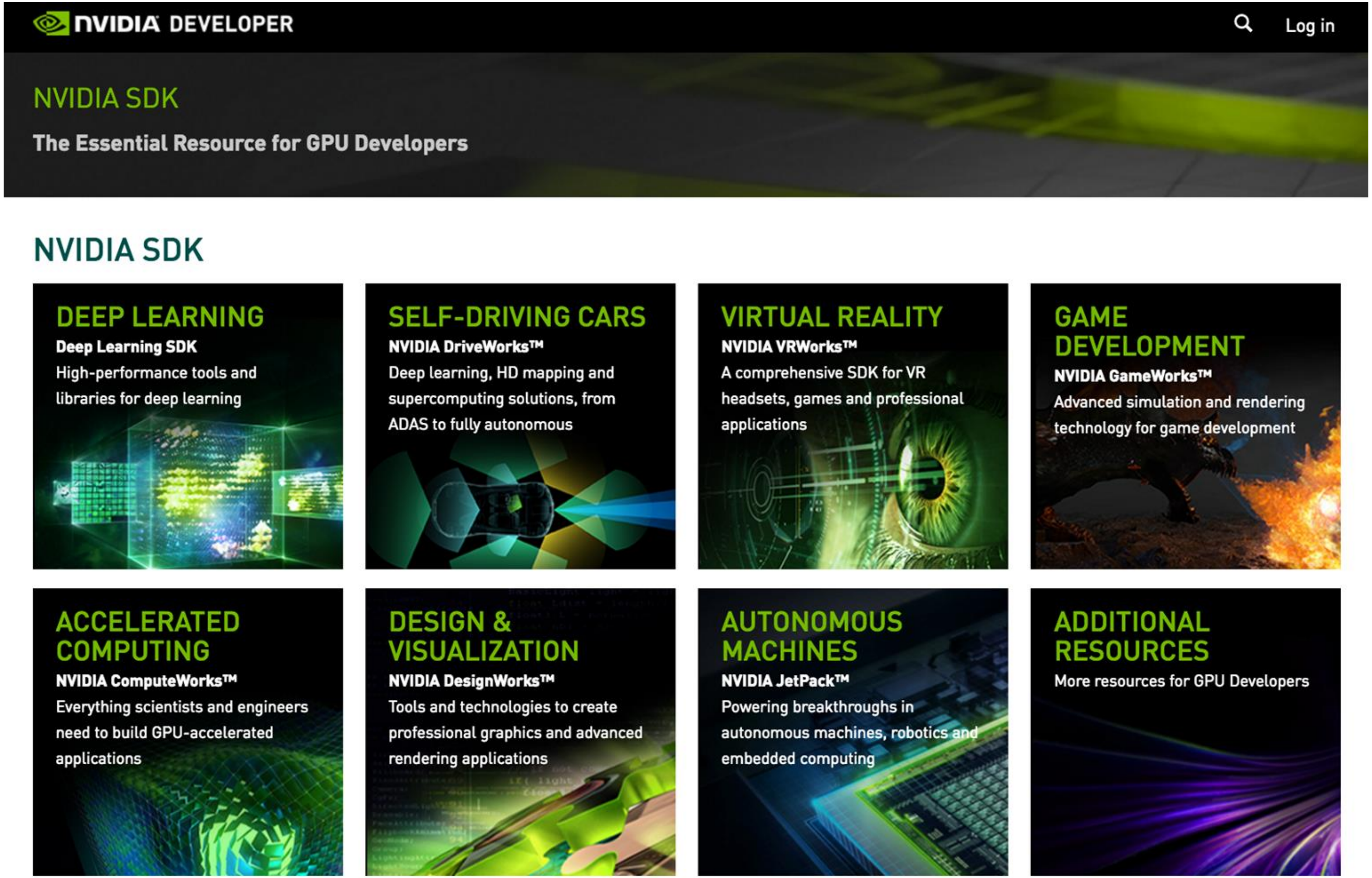
cuDNN

cuBLA
S

cuSPARS
E

cuFFT

GPU



NVIDIA DEVELOPER

Log in

NVIDIA SDK

The Essential Resource for GPU Developers

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

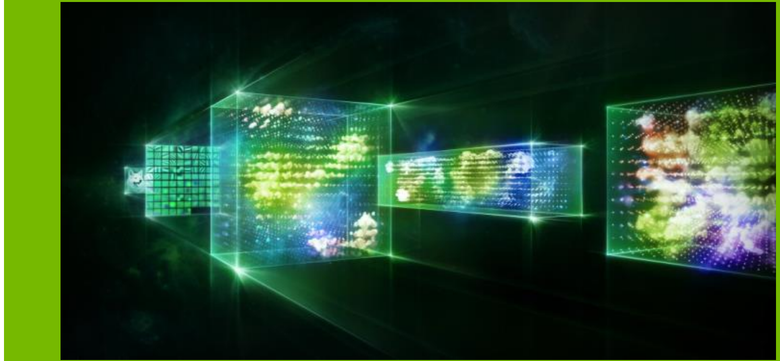
NVIDIA Deep Learning SDK

High performance GPU-acceleration for deep learning

Powerful tools and libraries for designing and deploying GPU-accelerated deep learning applications

- High performance building blocks for training deep neural networks on NVIDIA GPUs
- Accelerated linear algebra subroutines for developing novel deep learning algorithms
- Multi-GPU scaling that accelerates training on up to eight GPU

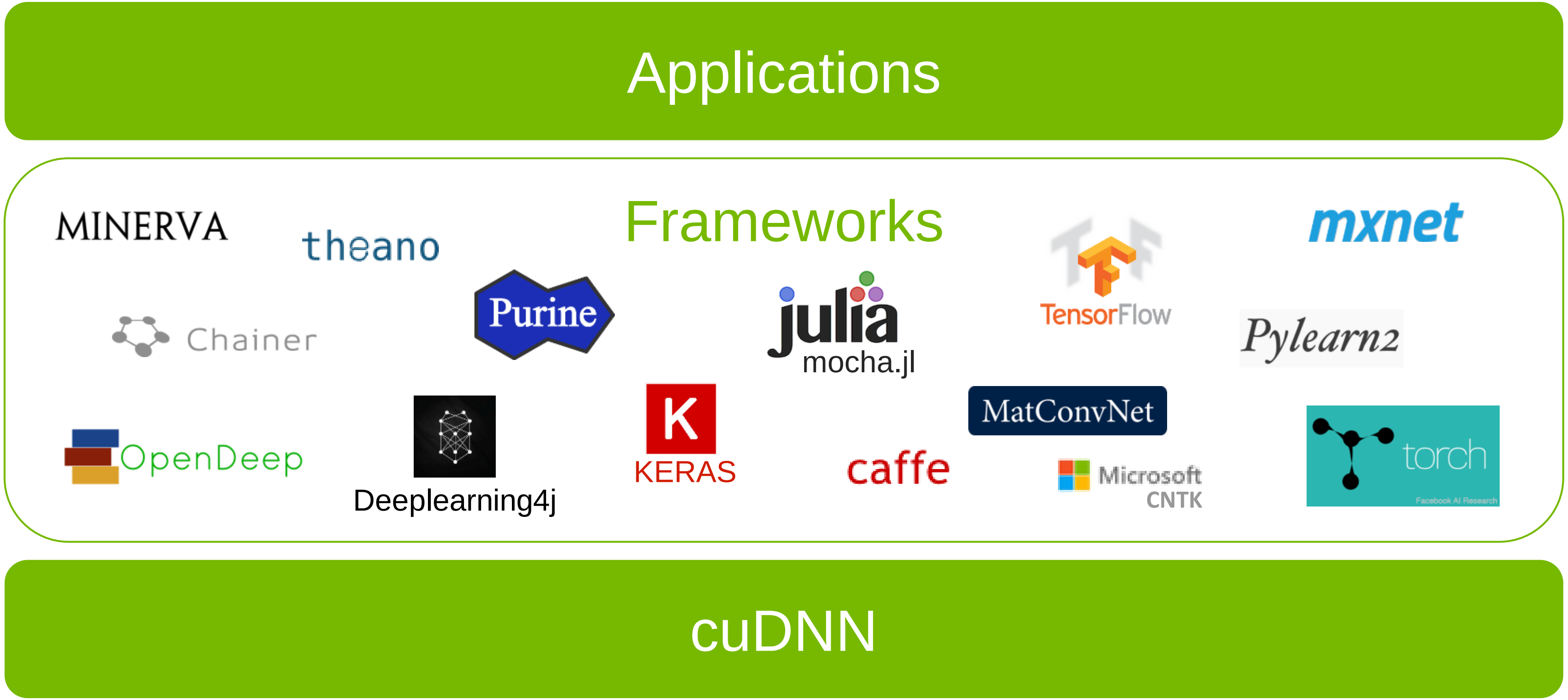
developer.nvidia.com/deep-learning-software



> “We are amazed by the steady stream of improvements made to the NVIDIA Deep Learning SDK and the speedups that they deliver”

— Frédéric Bastien, Team Lead (Theano) MILA

cuDNN: Powering Deep Learning



Applications

Frameworks

MINERVA

theano

mxnet

TensorFlow

Chainer

Purine

julia

mocha.jl

Pylearn2

OpenDeep

Deeplearning4j

K

KERAS

caffe

MatConvNet

Microsoft

CNTK

torch

Facebook AI Research

cuDNN

MOST POPULAR FRAMEWORKS

		CAFFE	TORCH	THEANO	TENSORFLOW
Applications		Image, Video	Image, Video, Speech	Image, Video, Speech	Image, Video, Speech
cuDNN		v5	v5	v5	v5
Multi-GPU		✓	✓	✓	✓
Neural Network		CNN, RNN	CNN, RNN(cuDNN accelerated)	CNN, RNN	CNN, RNN
Programming Interface(s)		C++, Python, MATLAB	Lua, LuaJIT, C++	Python	C++, Python
Platforms		Linux, Android, MacOS, Windows	Linux, Android, MacOS, iOS	Linux, Windows, MacOS	Linux, MacOS
Product Support	**Train**	Geforce, Tesla, DGX-1			
	Infer	Tesla, TX1	Tesla	Tesla	Tesla

OTHER NOTABLE FRAMEWORKS

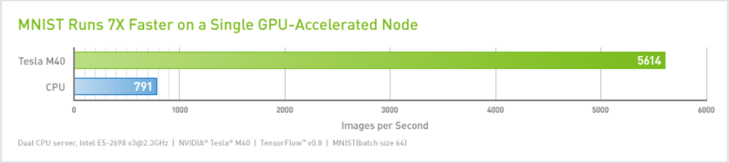
		CNTK	DSSTNE	CHAINER	MXNET	KALDI
Applications		Speech	Recommender	IoT	Image, Video, Speech	Speech
cuDNN		v4	v5	v5	v5	x
Multi-GPU		✓	✓	✓	✓	x
Neural Network		CNN, RNN	FC		CNN, RNN	RNN
Programming Interface(s)		C++, Python	C++	Python	C++, Python, Matlab, JavaScript	C++
Platforms		Windows, Linux	Linux	Linux	Windows, iOS, Android, Linux	Linux
Product Support	**Training**	Geforce, Tesla, DGX-1				
	Inference	Tesla, TX1	Tesla	Tesla	Tesla	Tesla

TENSORFLOW BY GOOGLE

Benchmarks & Highlights

TensorFlow

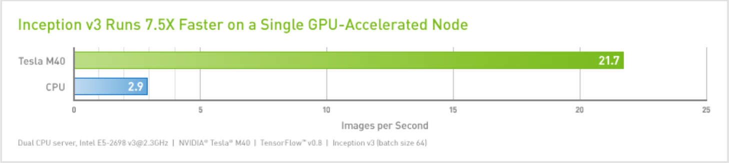
IMAGES PER SECOND FOR MNIST



MNIST Runs 7X Faster on a Single GPU-Accelerated Node

Dual CPU server, Intel E5-2698 v3@2.3GHz | NVIDIA® Tesla® M40 | TensorFlow™ v0.8 | MNIST(batch size 64)

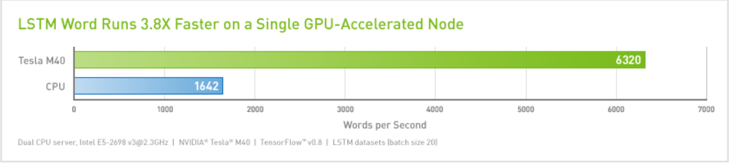
IMAGES PER SECOND FOR INCEPTION V3



Inception v3 Runs 7.5X Faster on a Single GPU-Accelerated Node

Dual CPU server, Intel E5-2698 v3@2.3GHz | NVIDIA® Tesla® M40 | TensorFlow™ v0.8 | Inception v3 (batch size 64)

WORDS PER SECOND WITH LSTM WORD



LSTM Word Runs 3.8X Faster on a Single GPU-Accelerated Node

Dual CPU server, Intel E5-2698 v3@2.3GHz | NVIDIA® Tesla® M40 | TensorFlow™ v0.8 | LSTM datasets (batch size 20)

- Fastest Growing
- Flexible – any computation as a data flow graph
- Distributed
- SyntaxNet

TensorFlow

FEATURES

Deep Flexibility

Express any computation as a data flow graph

Auto-Differentition

Just define the computation architecture and feed data

True Portability

GPUs, CPUs, Desktops, Servers, Mobiles

Language Options

Python, C++, Java, JavaScript, R

Connect Research and Production

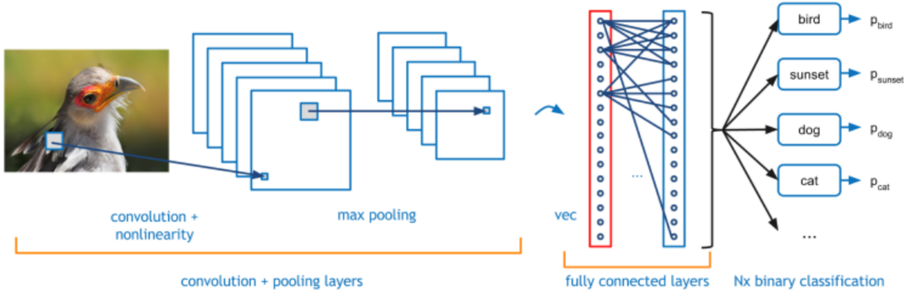
Allows researchers to push ideas to products faster

Maximize Performance

Threads, queues ad asynchronous computation to use GPUs and CPUs

Caffe ABOUT

- Released in 2014 by Yangqing while at UC Berkeley, Caffe is the most popular open source Deep Learning framework to date
- It has been the de facto framework for image classification.
- It's known for its massive collection of different neural networks in the Model Zoo
- It is a foundation for many other frameworks such as CaffeOnSpark by Yahoo.



Caffe FEATURES

Expressive Architecture

Models, optimization, and GPU/CPU are defined by configuration instead of coding

Speed

Designed for massive deployment, Caffe can process over 60M images per day with a single K40 GPU

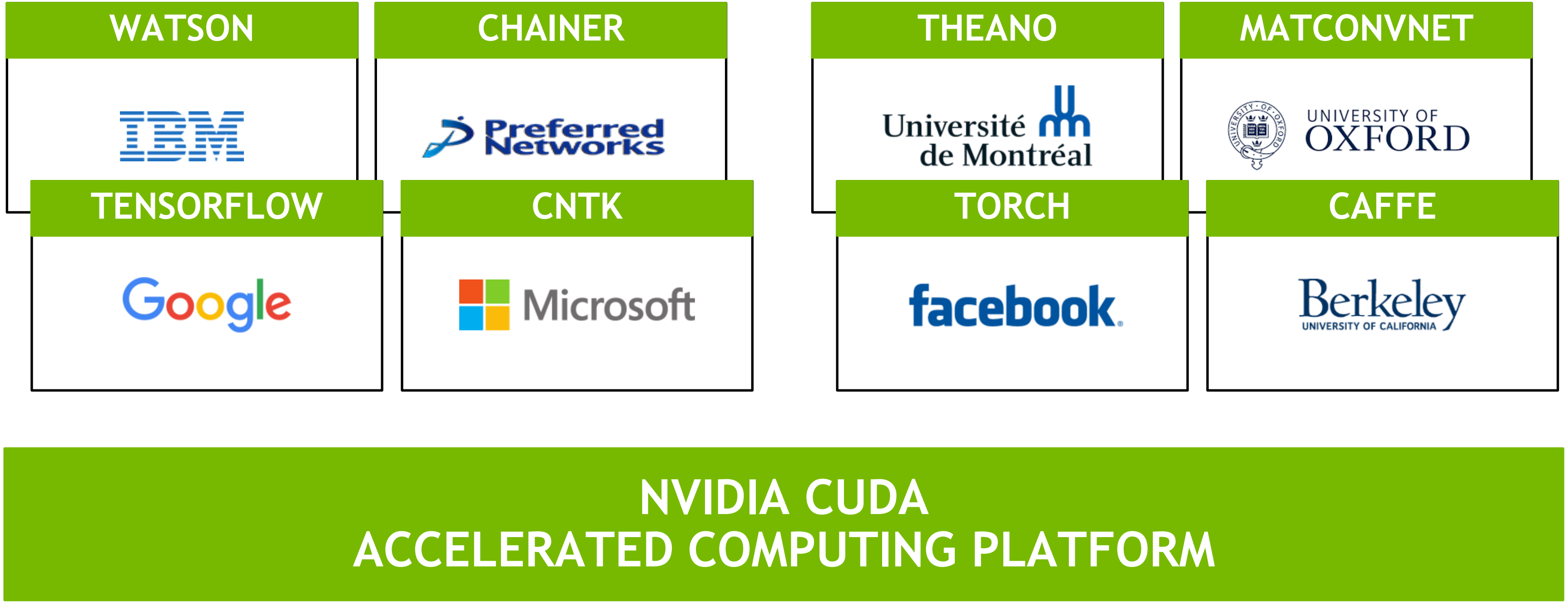
Extensible Code

Coding style fosters active development to stay innovative

Community

Powers academic research projects, startup prototypes, and large-scale industrial applications

NVIDIA GPU: the engine of deep learning



WATSON

IBM

CHAINER

Preferred Networks

THEANO

Université de Montréal

MATCONVNET

UNIVERSITY OF OXFORD

TENSORFLOW

Google

CNTK

Microsoft

TORCH

facebook

CAFFE

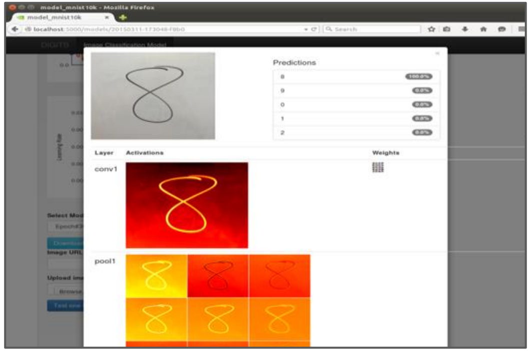
Berkeley UNIVERSITY OF CALIFORNIA

NVIDIA CUDA
ACCELERATED COMPUTING PLATFORM

Deep Learning Performance Doubles

For Data Scientist and Researchers

Train Models up to 2x Faster with Automatic Multi-GPU Scaling & Object Detection



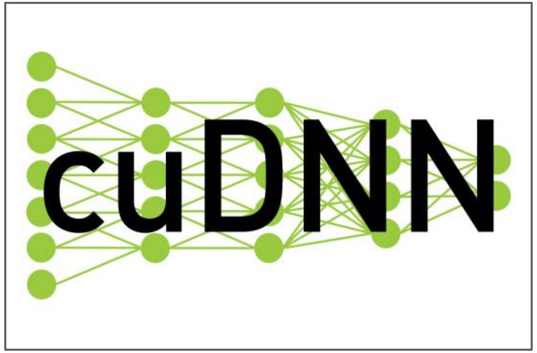
model_mnist10k - Mozilla Firefox
model_mnist10k
localhost:5000/models/20150311-173048-f8b0
Search
DIGITS
Image Classification Model
Layer
Activations
Weights
conv1
pool1

Predictions

8	100.0%
9	0.0%
0	0.0%
1	0.0%
2	0.0%

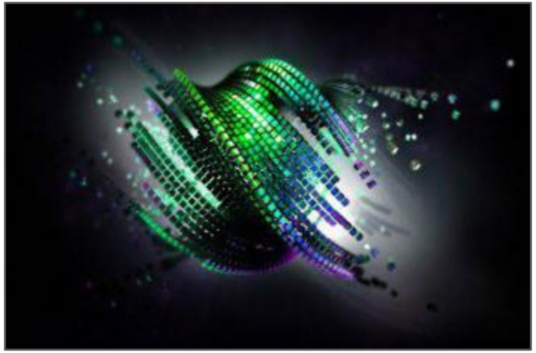
DIGITS 4

2x Faster Single GPU Training
Support for Larger Models,
support for RNN LSTM



cuDNN 5.1

2x Larger Datasets
Instruction-level Profiling



CUDA 7.5

DIGITS™

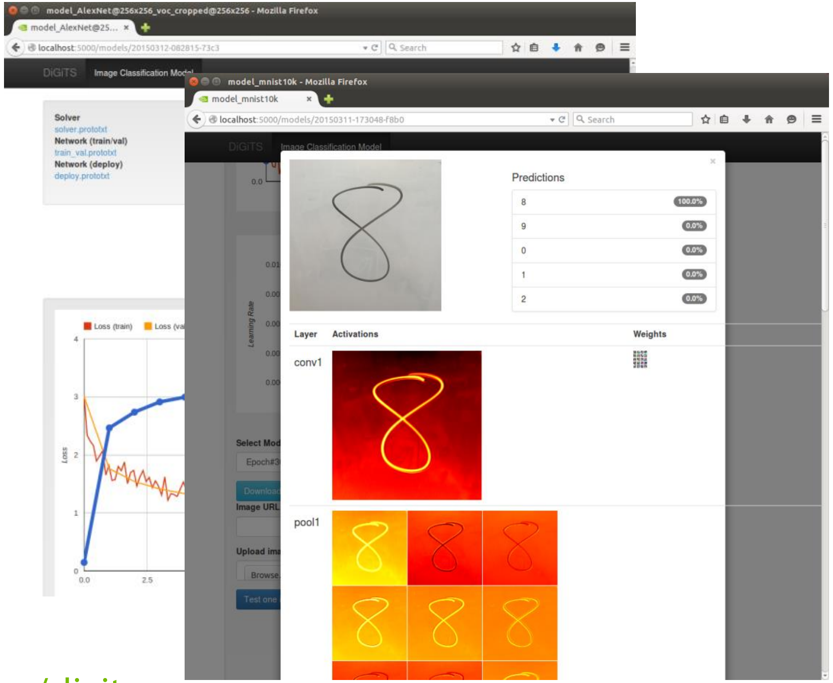
Interactive Deep Learning GPU Training System

Quickly design the best deep neural network (DNN) for your data

Train on multi-GPU (automatic)

Visually monitor DNN training quality in real-time

Manage training of many DNNs in parallel on multi-GPU systems

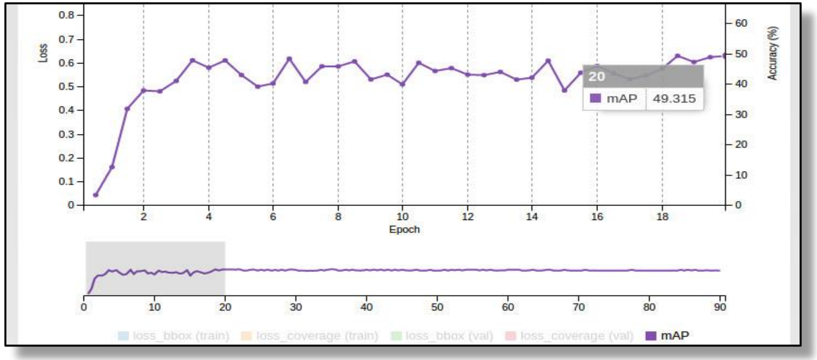


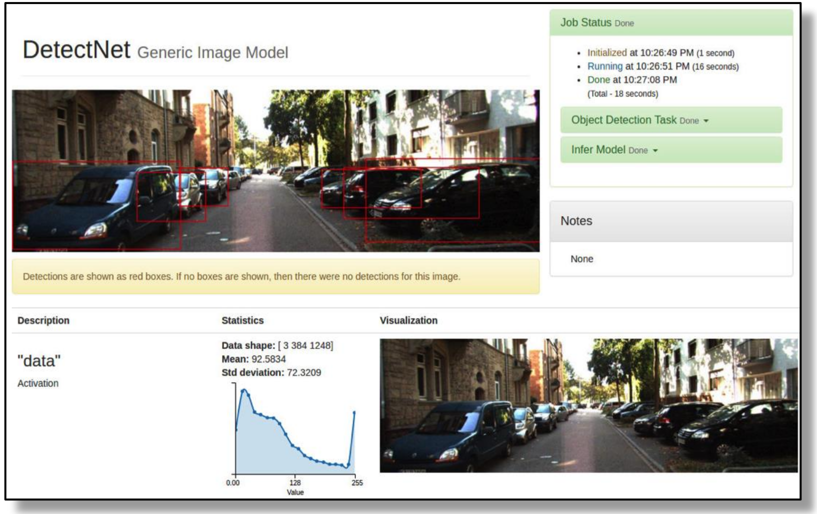
developer.nvidia.com/digits

Preview DIGITS Future

Object Detection Workflow

- Object Detection Workflows for Automotive and Defense
- Targeted at Autonomous Vehicles, Remote Sensing





DetectNet Generic Image Model

Job Status Done

- Initialized at 10:26:49 PM (1 second)
- Running at 10:26:51 PM (16 seconds)
- Done at 10:27:08 PM (Total - 18 seconds)

Object Detection Task Done ▾

Infer Model Done ▾

Notes

None

Detections are shown as red boxes. If no boxes are shown, then there were no detections for this image.

Description

"data"

Activation

Statistics

Data shape: [3 384 1248]
Mean: 92.5834
Std deviation: 72.3209

Visualization

developer.nvidia.com/digits

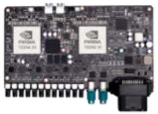
GPU INFERENCE ENGINE (GIE)

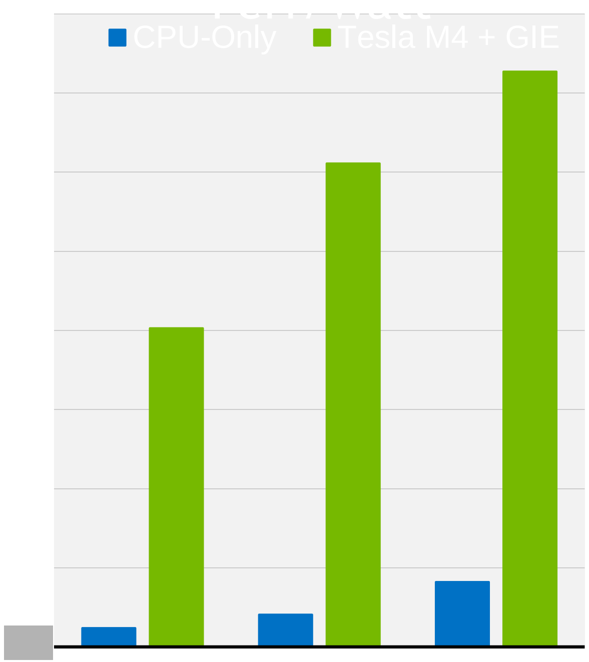
High-performance deep learning inference for production deployment

DATA CENTER

AUTOMOTIVE

EMBEDDED





GoogLenet, CPU-only vs Tesla M4 + GIE on Single-socket Haswell E5-2698 v3@2.3GHz with HT

developer.nvidia.com/gie

CUDNN 5.1 – WHAT’S NEW

LSTM RNNs, Pascal GPU support, Improved Performance

High-performance deep learning primitives

LSTM recurrent neural networks deliver up to 6x speedup in Torch

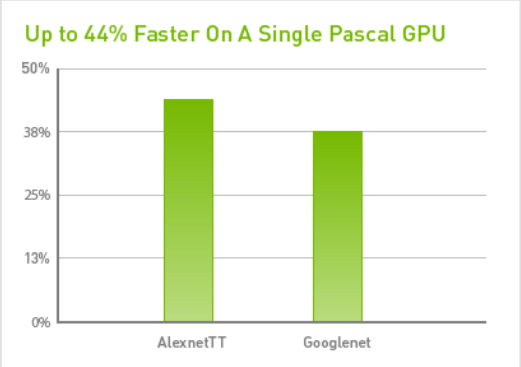
Up to 44% faster training on a single NVIDIA® Pascal™ GPU

Improved performance and reduced memory usage with FP16 routines on Pascal GPUs

developer.nvidia.com/cudnn

5.9

Speedup of Torch with cuDNN 5



cuDNN 4 + CUDA 7.5 on M40 vs cuDNN 5 RC + CUDA 8 EA on P100, Intel® Xeon® Processor E5-2698

Optimising RNNs with cuDNN v5.1

ParallelForAll

devblogs.nvidia.com/parallelforall/optimizing-recurrent-neural-networks-cudnn-5/

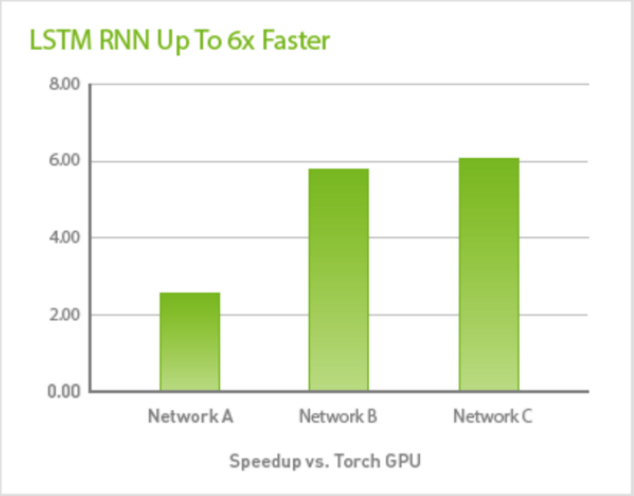


Figure 1: cuDNN 5 + Torch speedup vs. Torch-rnn implementation, M40, Intel® Xeon® Processor E5-2698

Supports:

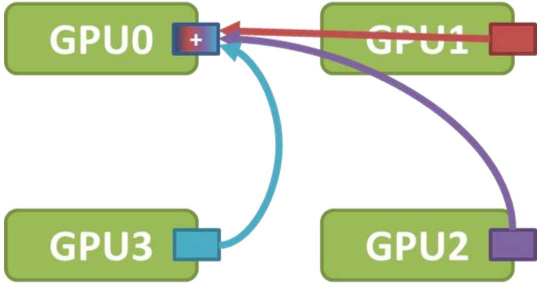
- **ReLU & tanh activation functions**
- **Gated Recurrent Units (GRU)**
- **Long Short-Term Memory (LSTM)**

NCCL

Accelerating Multi-GPU Communications

A topology-aware library of accelerated collectives to improve the scalability of multi-GPU applications

- Patterned after MPI’s collectives: includes all-reduce, all-gather, reduce-scatter, reduce, broadcast
- Optimized intra-node communication
- Supports multi-threaded and multi-process applications



github.com/NVIDIA/nccl

developer.nvidia.com/nvgraph

nvGRAPH

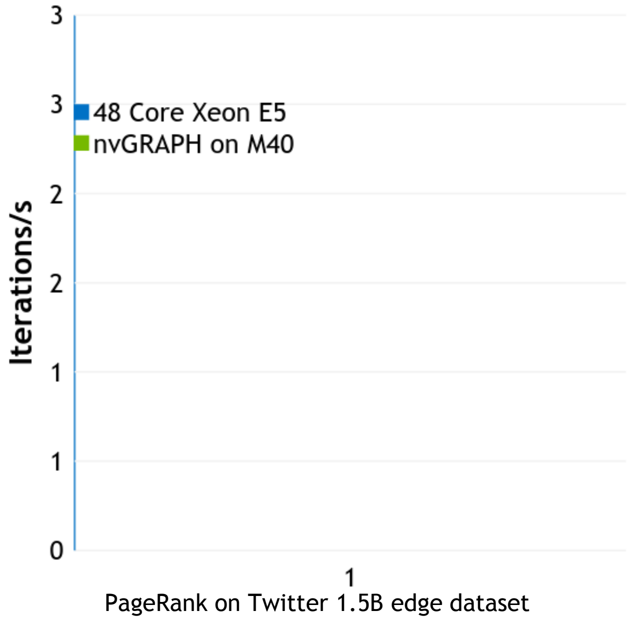
Accelerated Graph Analytics

nvGRAPH for high performance graph analytics

- Deliver results up to 3x faster than CPU-only
- Solve graphs with up to 2.5 Billion edges on 1x M40
- Accelerates a wide range of graph analytics apps:

PageRank	Single Source Shortest Path	Single Source Widest Path
Search	Robotic Path Planning	IP Routing
Recommendation Engines	Power Network Planning	Chip Design / EDA
Social Ad Placement	Logistics & Supply Chain Planning	Traffic sensitive routing

nvGRAPH: 3x Speedup



CPU System:
4U server w/ 4x12-core Xeon E5-2697 CPU,
30M Cache, 2.70 GHz, 512 GB RAM

cuSPARSE: (DENSE MATRIX) X (SPARSE VECTOR)

Speeds up Natural Language Processing

cusparse<T>gemvi()

$y = \alpha * op(A) * x + \beta * y$

A = dense matrix

x = sparse vector

y = dense vector

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \alpha \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \end{bmatrix} \begin{bmatrix} - \\ 2 \\ - \\ - \\ 1 \end{bmatrix} + \beta \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Sparse vector could be frequencies of words in a text sample

cuSPARSE provides a full suite of accelerated sparse matrix functions

developer.nvidia.com/cusparse

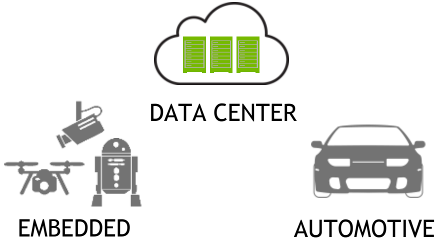
What’s new in deep learning software

DIGITS 4

Objection Detection

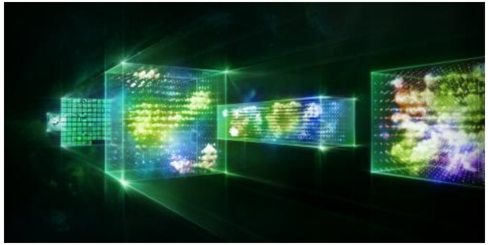
GIE

High performance deep learning inference



cuDNN 5.1

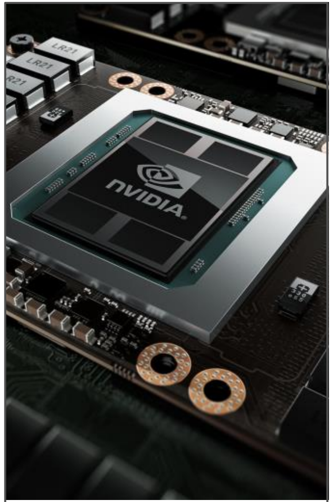
Improved performance for VGG, ResNet style networks



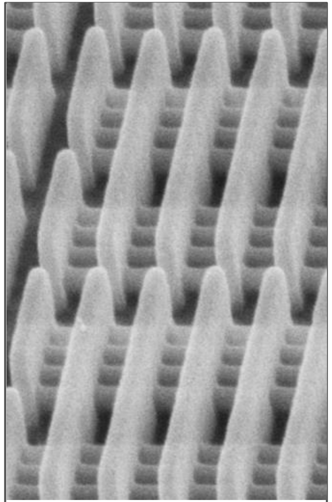
Deep Learning Hardware

INTRODUCING TESLA P100

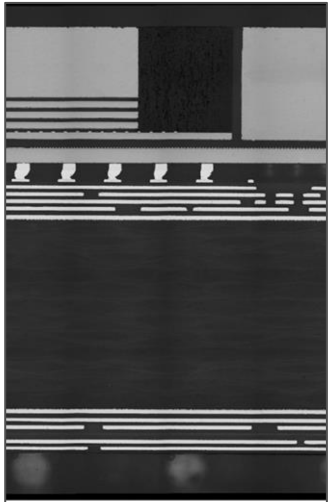
Five Technology Breakthroughs Made it Possible



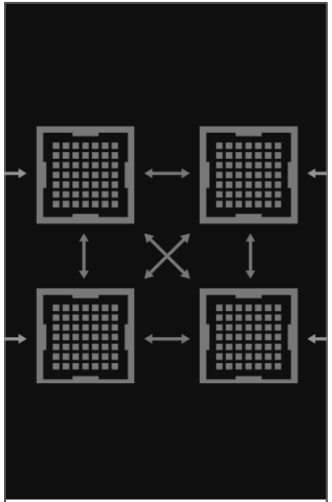
Pascal Architecture



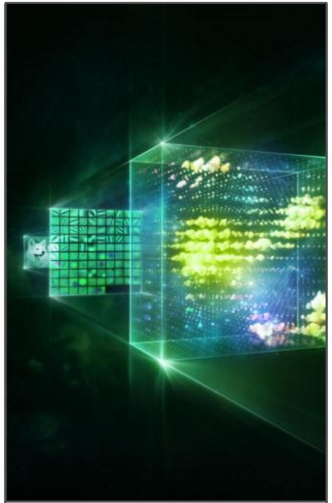
16nm
FinFET



COWOS with
HBM2 Stacked Memory



NVLink



New AI
Algorithms

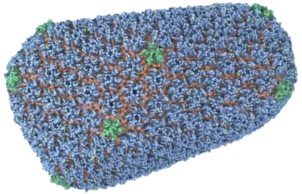
VISUALIZATION-ENABLED SUPERCOMPUTERS

Simulation + Visualization

CSCS Piz Daint

Galaxy Formation

NCSA Blue Waters



Molecular Dynamics

ORNL Titan

Cosmology

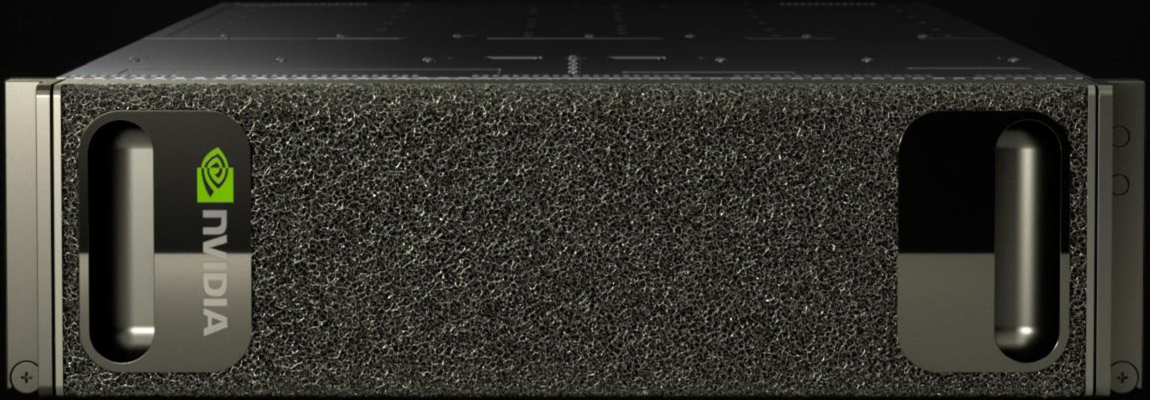
NVIDIA DGX-1

WORLD’S FIRST DEEP LEARNING SUPERCOMPUTER

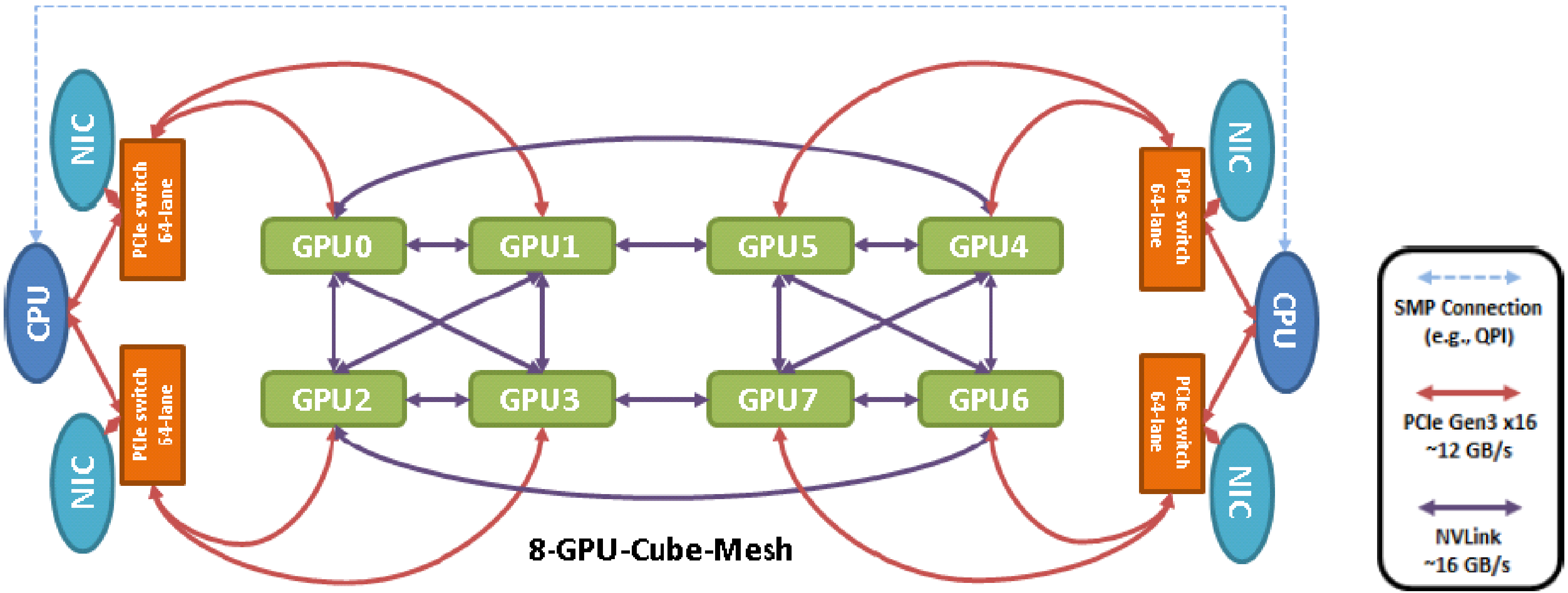
Engineered for deep learning | 170TF FP16 | 8x Tesla P100
NVLink hybrid cube mesh | Accelerates major AI frameworks

8x Tesla P100 16GB, Dual Xeon, NVLink Hybrid Cube Mesh
7 TB SSD, Dual 10GbE, Quad IB 100Gb
3RU – 3200W

NVIDIA® DGX-1™



DGX-1 SYSTEM TOPOLOGY



For the 8-GPU-Cube-Mesh topology, there is no need to use PCIe for any GPU-to-GPU communications (whether point-to-point or collective).

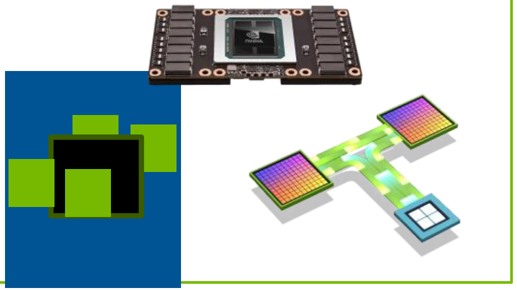
CUDA 8 – WHAT'S NEW

P100 Support

Stacked Memory

NVLINK

FP16 math



Unified Memory

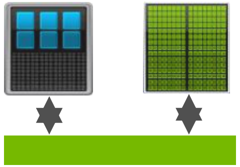
Larger Datasets

Demand Paging

New Tuning APIs

Standard C/C++ Allocators

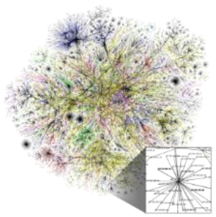
CPU/GPU Data Coherence & Atomics



Libraries

New nvGRAPH library

cuBLAS improvements for Deep Learning



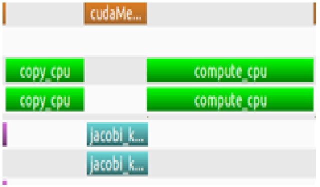
Developer Tools

Critical Path Analysis

2x Faster Compile Time

OpenACC Profiling

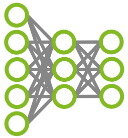
Debug CUDA Apps on Display GPU



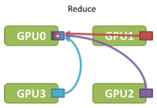
NVIDIA DGX-1 SOFTWARE STACK

Optimized for Deep Learning Performance

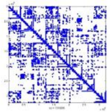
Accelerated Deep Learning



cuDNN

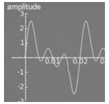


NCCL



cuSPARSE

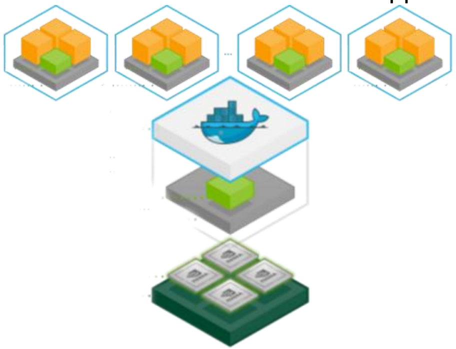
cuBLAS



cuFFT

Container Based Applications

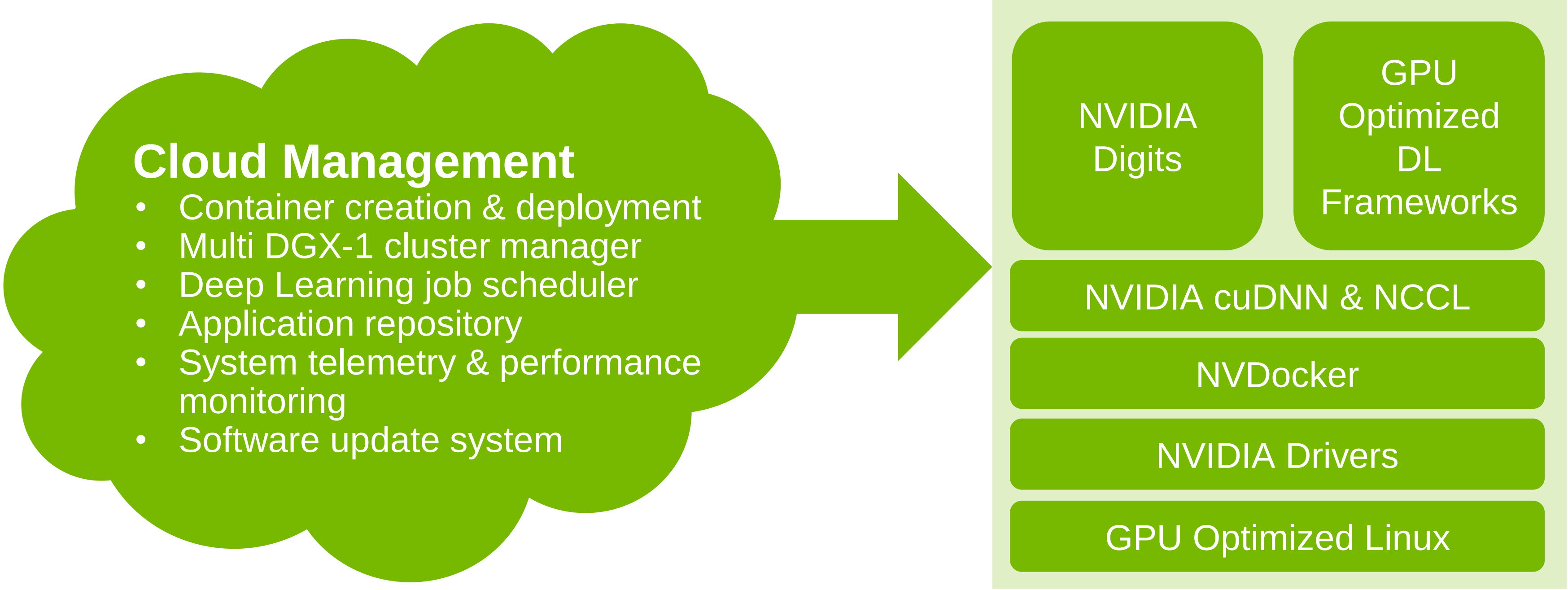
Digits DL Frameworks GPU Apps



NVIDIA Cloud Management

NVIDIA DGX-1 SOFTWARE STACK

Optimized for Deep Learning Performance



Cloud Management

- Container creation & deployment
- Multi DGX-1 cluster manager
- Deep Learning job scheduler
- Application repository
- System telemetry & performance monitoring
- Software update system

NVIDIA Digits

GPU Optimized DL Frameworks

NVIDIA cuDNN & NCCL

NVDocker

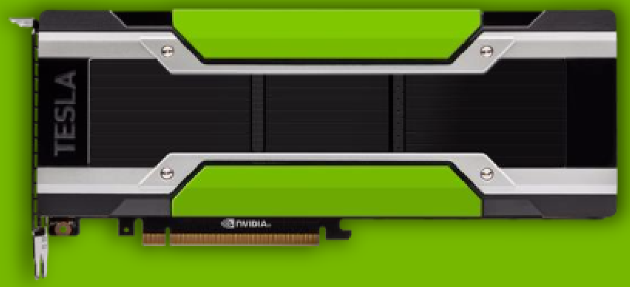
NVIDIA Drivers

GPU Optimized Linux

NVIDIA DGX-1

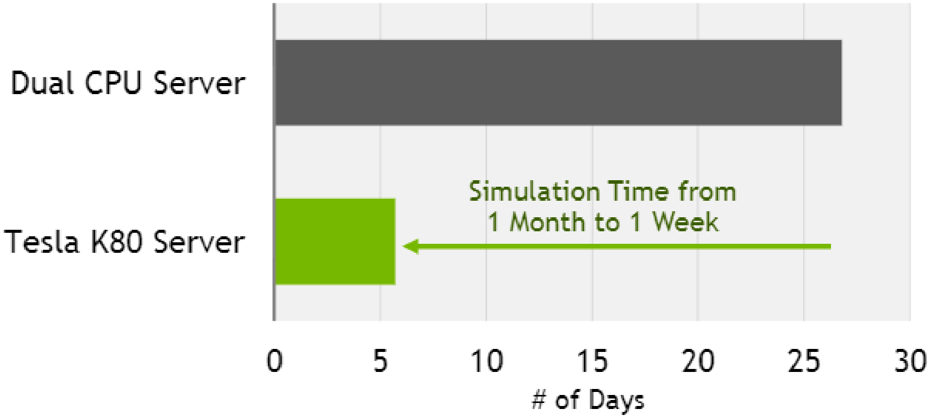
TESLA K80

World's Fastest Accelerator for HPC & Data Analytics



5x Faster

AMBER Performance

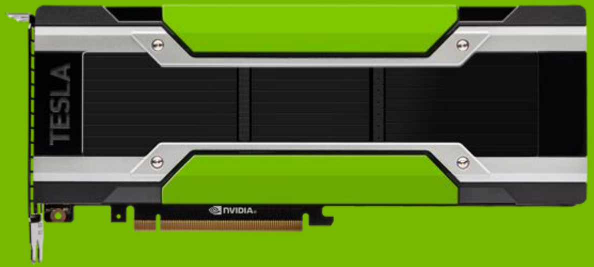


CUDA Cores	2496
Peak DP	1.9 TFLOPS
Peak DP w/ Boost	2.9 TFLOPS
GDDR5 Memory	24 GB
Bandwidth	480 GB/s
Power	300 W
GPU Boost	Dynamic

AMBER Benchmark: PME-JAC-NVE Simulation for 1 microsecond

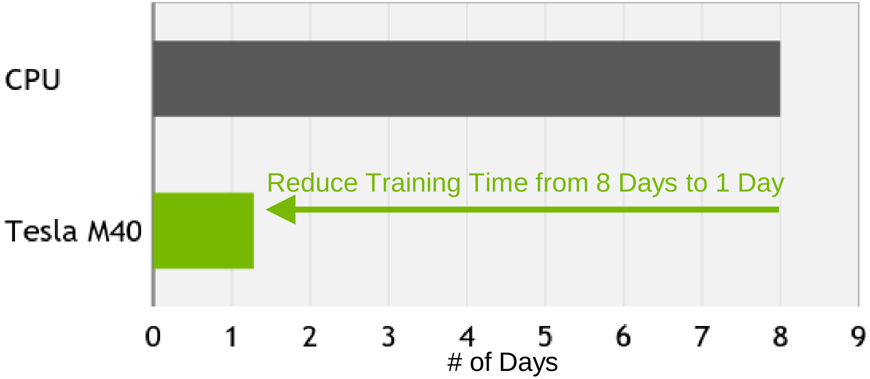
TESLA M40

World's Fastest Accelerator for Deep Learning



8x Faster

Caffe Performance

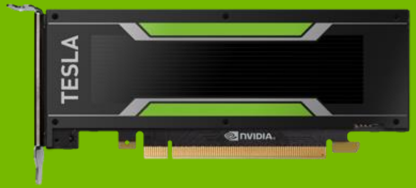


CUDA Cores	3072
Peak SP	7 TFLOPS
GDDR5 Memory	12 GB
Bandwidth	288 GB/s
Power	250W

Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2

TESLA M4

Highest Throughput Hyperscale Workload Acceleration



Video Processing

4x

Stabilization and Enhancements

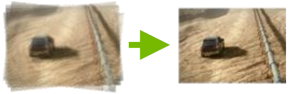


Image Processing

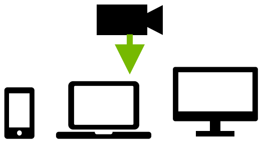
5x

Resize, Filter, Search, Auto-Enhance

Video Transcode

2x

H.264 & H.265, SD & HD



Machine Learning Inference

2x

CUDA Cores	1024
Peak SP	2.2 TFLOPS
GDDR5 Memory	4 GB
Bandwidth	88 GB/s
Form Factor	PCIe Low Profile
Power	50 – 75 W

Preliminary specifications. Subject to change.

A SUPERCOMPUTER FOR AUTONOMOUS MACHINES

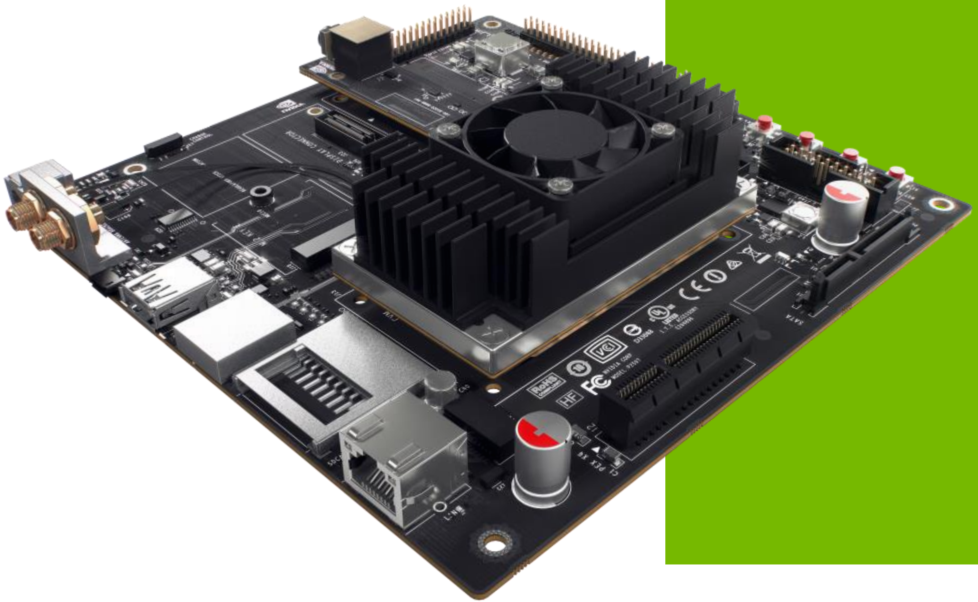
Bringing AI and machine learning to a world of robots and drones

Jetson TX1 is the first embedded computer designed to process deep neural networks

1 TeraFLOPS in a credit-card sized module

Jetson TX1

	JETSON TX1
GPU	1 TFLOP/s 256-core Maxwell
CPU	64-bit ARM A57 CPUs
Memory	4 GB LPDDR4 | 25.6 GB/s
Video decode	4K 60Hz
Video encode	4K 30Hz
CSI	Up to 6 cameras | 1400 Mpix/s
Display	2x DSI, 1x eDP 1.4, 1x DP 1.2/HDMI
Wifi	802.11 2x2 ac
Networking	1 Gigabit Ethernet
PCIE	Gen 2 1x1 + 1x4
Storage	16 GB eMMC, SDIO, SATA
Other	3x UART, 3x SPI, 4x I2C, 4x I2S, GPIOs



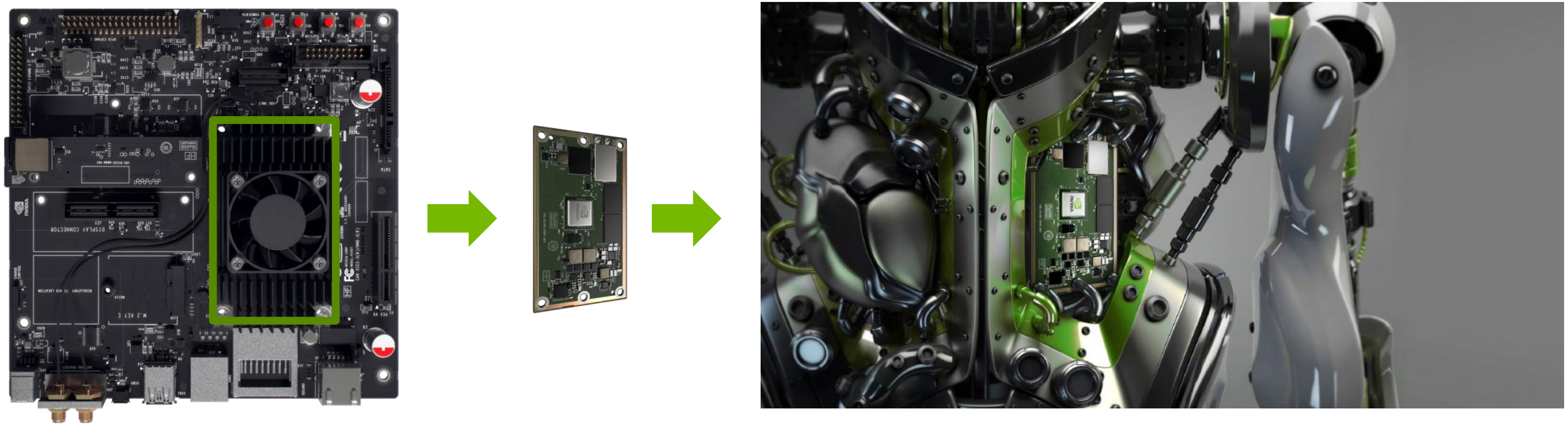
Jetson TX1 Developer Kit

Jetson TX1

Developer Board

5MP Camera

Jetson SDK



Develop and deploy

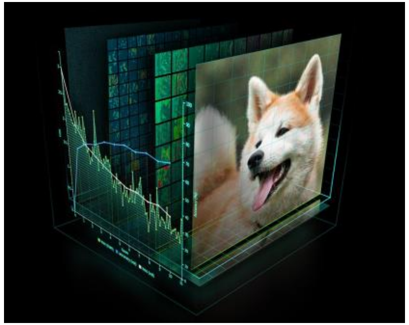
Jetson TX1 and Jetson TX1 Developer Kit

GPU TECHNOLOGY CONFERENCE

Sep 28-29, 2016 | Amsterdam
www.gputechconf.eu #GTC16EU

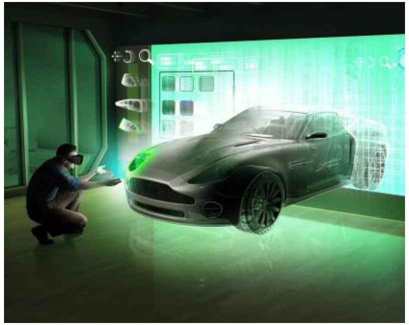
EUROPE’S BRIGHTEST MINDS & BEST IDEAS

GET A 20% DISCOUNT WITH CODE *ALLOGTCEU2016*

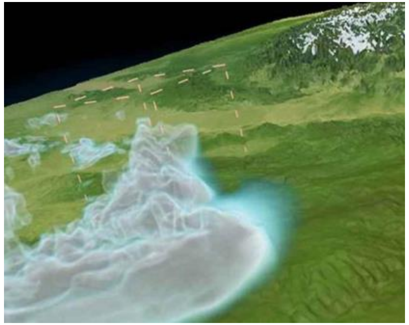


DEEP LEARNING & ARTIFICIAL INTELLIGENCE

AUTONOMOUS VEHICLES



VIRTUAL REALITY & AUGMENTED REALITY



SUPERCOMPUTING & HPC

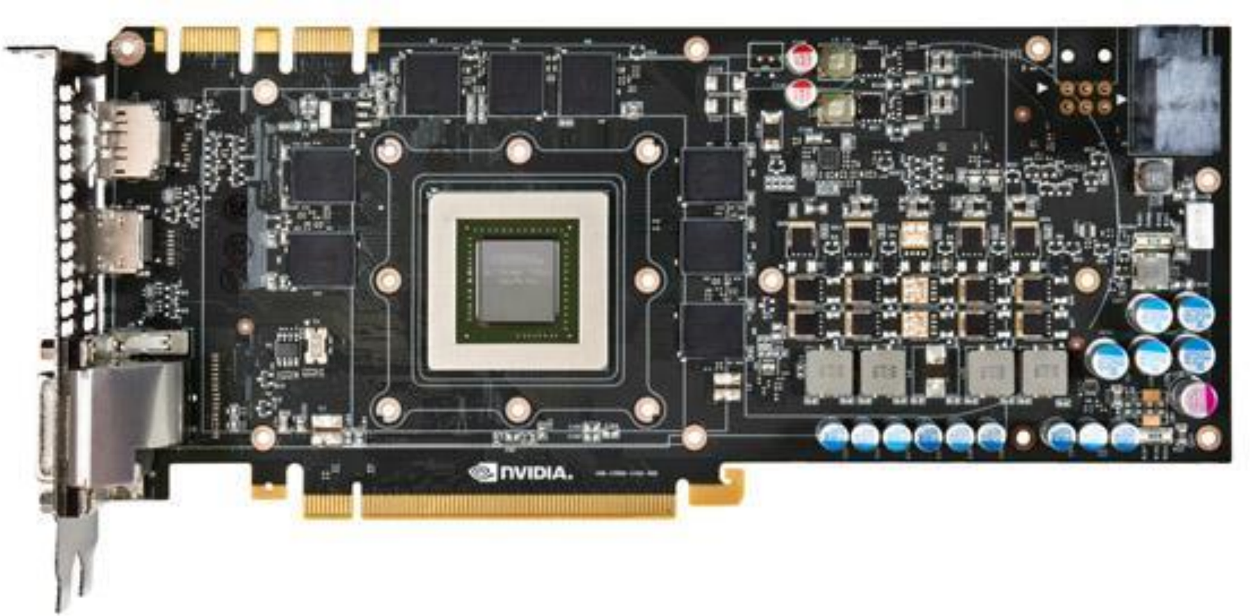
GTC Europe is a two-day conference designed to expose the innovative ways developers, businesses and academics are using parallel computing to transform our world.

2 Days | 1,000 Attendees | 50+ Exhibitors | 50+ Speakers | 10+ Tracks | 15+ Hands-on Labs| 1-to-1 Meetings

Deep Learning in the Cloud

NVIDIA in AWS

currently 2.2GFlops - g2.2xlarge - soon to be upgraded



Deep Learning Lab

http://nvlabs.qwiklab.com

qwikLABS