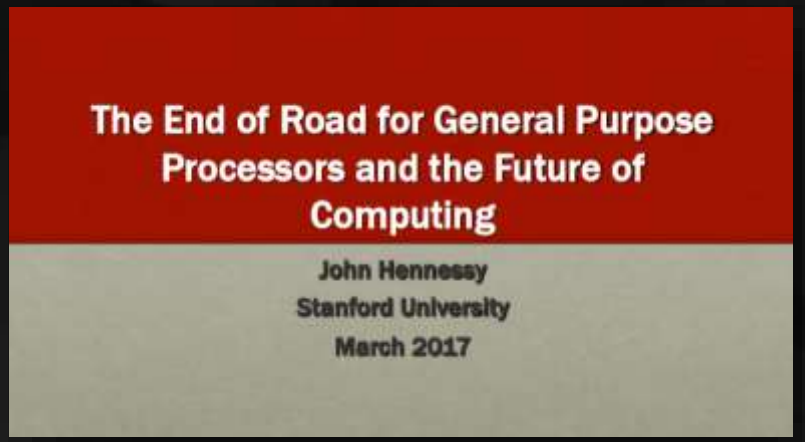
NEW NVIDIA PLATFORM FOR AI

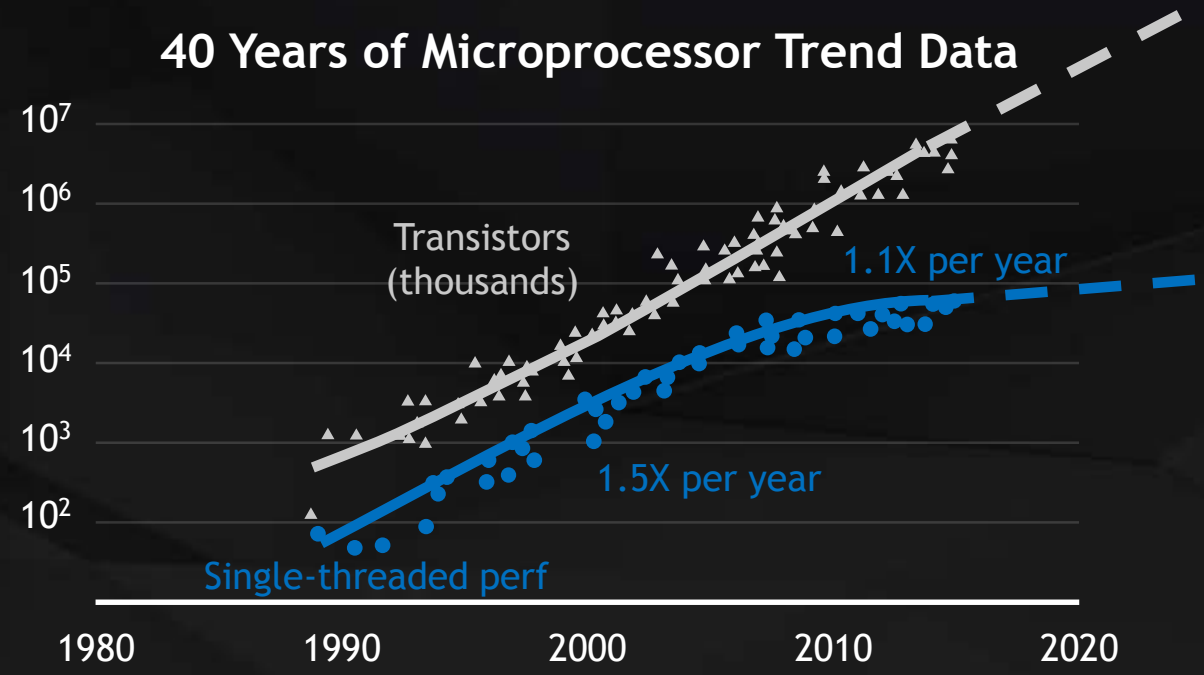
Pedro Mario Cruz e Silva (pcruzesilva@nvidia.com) LinkedIn
Solution Architect Manager
Enterprise Latin America
Global Oil & Gas Team

NVIDIA.

"GTC 2017: 'I AM AI' OPENING IN KEYNOTE"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=SUNPRR4O5ZA

LIFE AFTER MOORE'S LAW

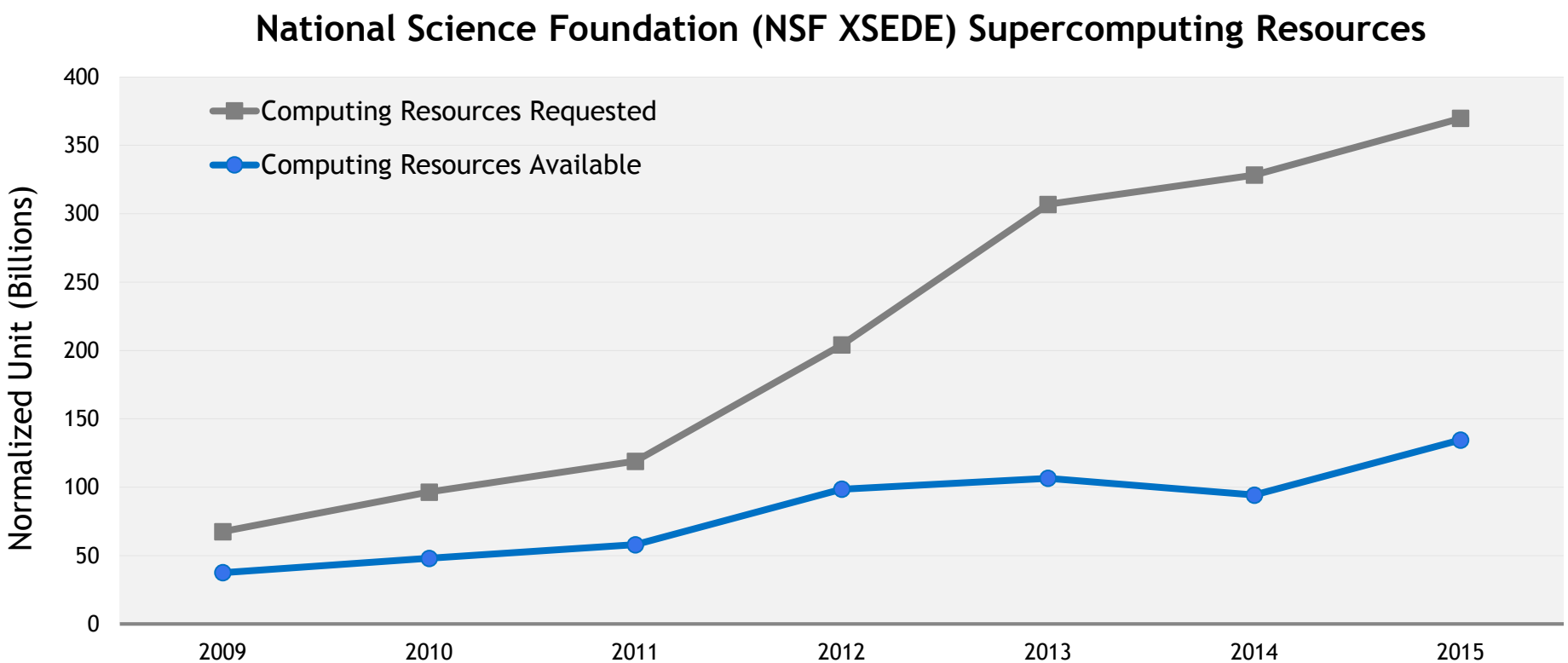




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

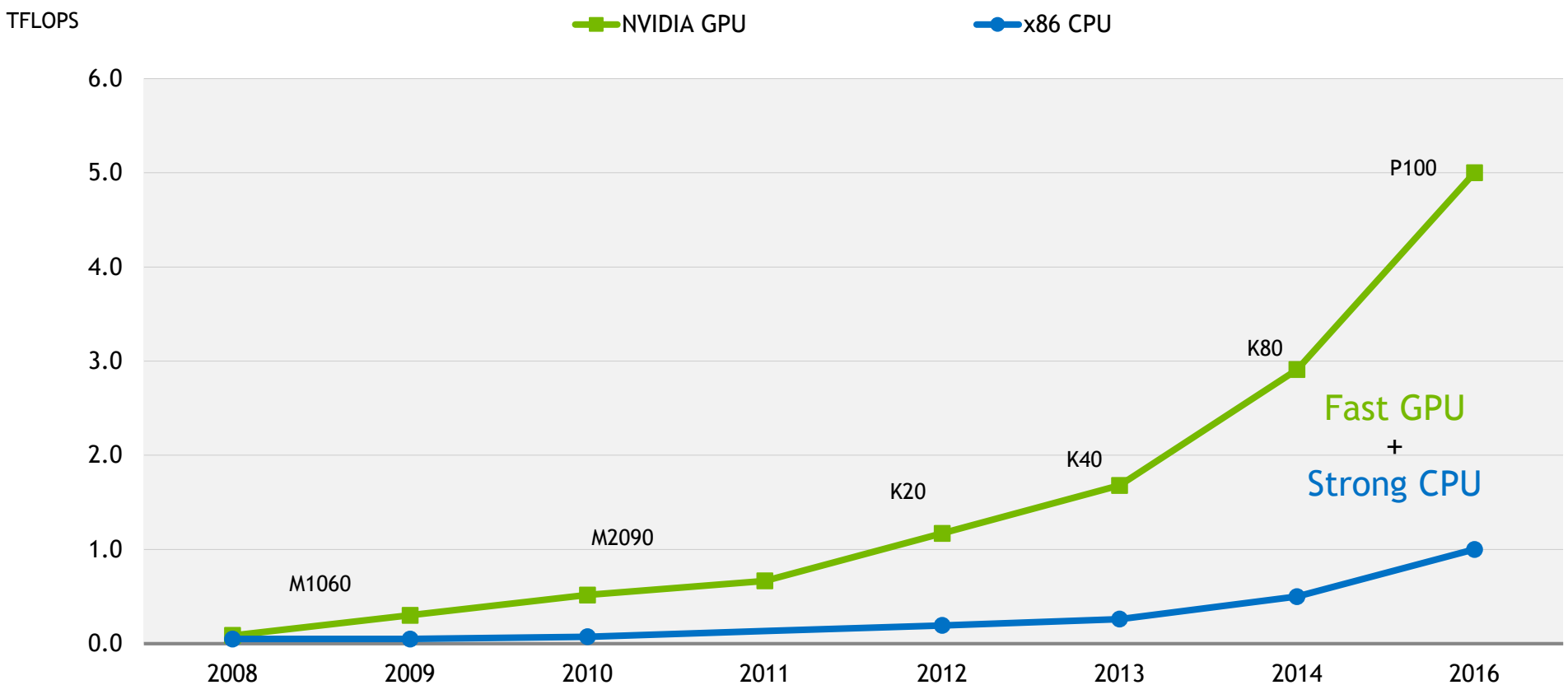
200B CORE HOURS OF LOST SCIENCE

Data Center Throughput is the Most Important Thing for HPC

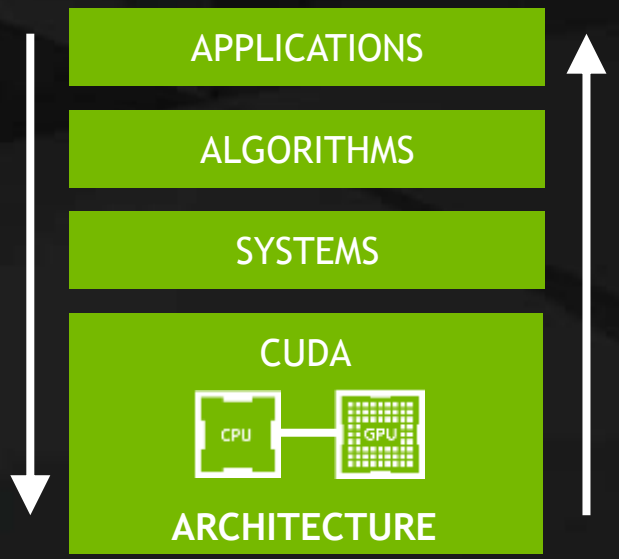


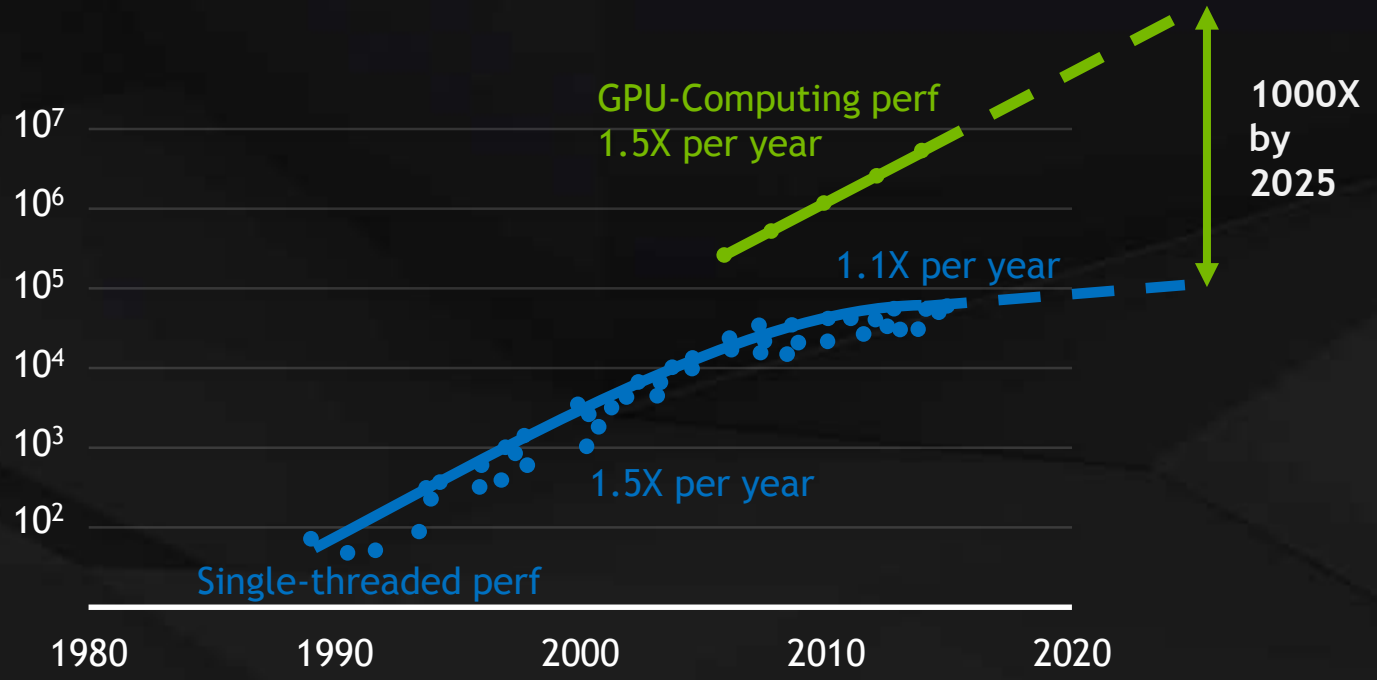
Source: NSF XSEDE Data: https://portal.xsede.org/#/gallery
NU = Normalized Computing Units are used to compare compute resources across supercomputers and are based on the result of the High Performance LINPACK benchmark run on each system

THE ADVANTAGES OF GPU-ACCELERATED DATA CENTER



RISE OF GPU COMPUTING



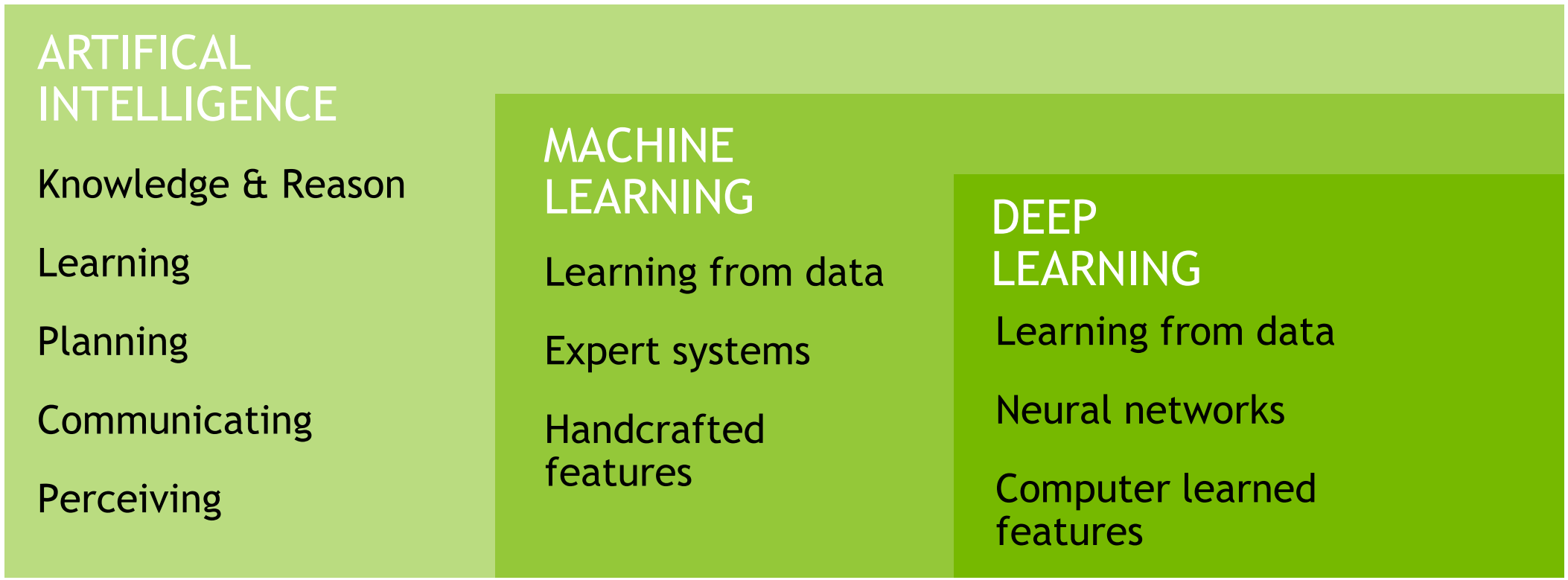


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

DEEP LEARNING

LEARNING FROM DATA

AND SOME BUZZ WORDS



A NEW COMPUTING MODEL

TRAINING

Training Data

Input

"Label"
Output

Trained Neural Network

INFERENCE

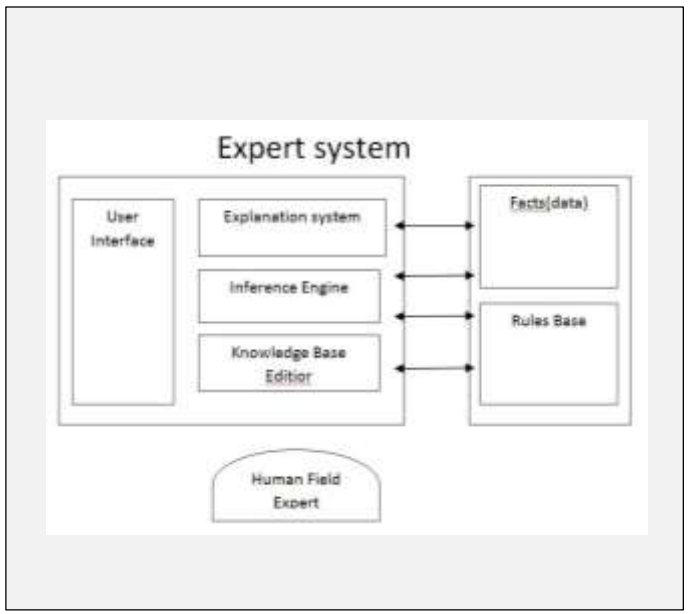
Input

Trained Neural Network

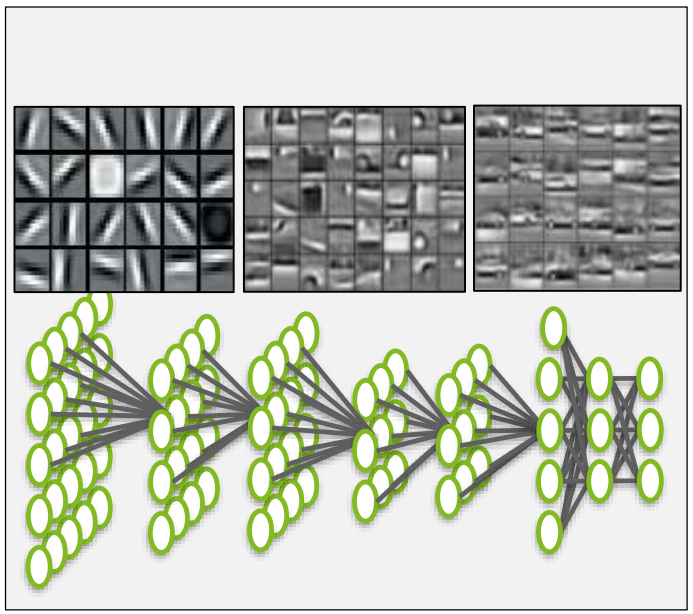
"Label"
Output

A NEW COMPUTING MODEL

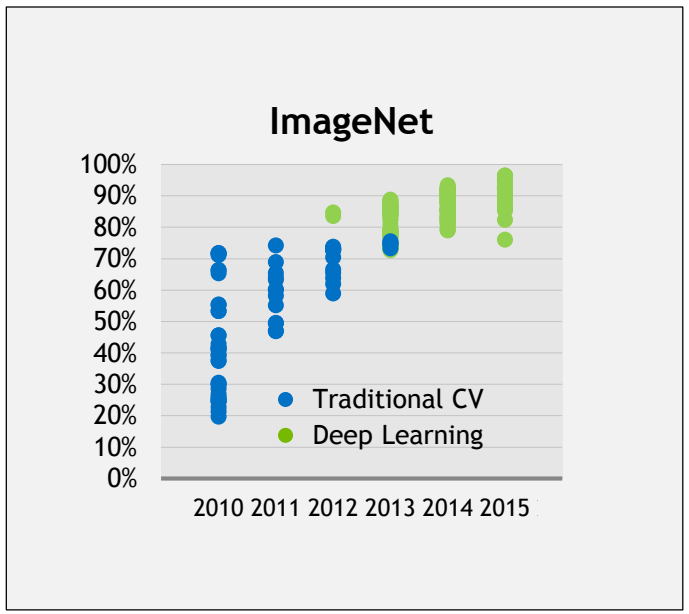
Outperform experts, facts, rules with software that writes software



Traditional Computer Vision
Experts + Time



Deep Learning Object Detection
DNN + Data + GPU



Deep Learning Achieves
"Superhuman" Results

“ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS”

Tompson, J., Schlachter, K., Sprechmann, P., & Perlin, K. (2016). Accelerating Eulerian Fluid Simulation With Convolutional Networks. *arXiv preprint arXiv:1607.03597*.

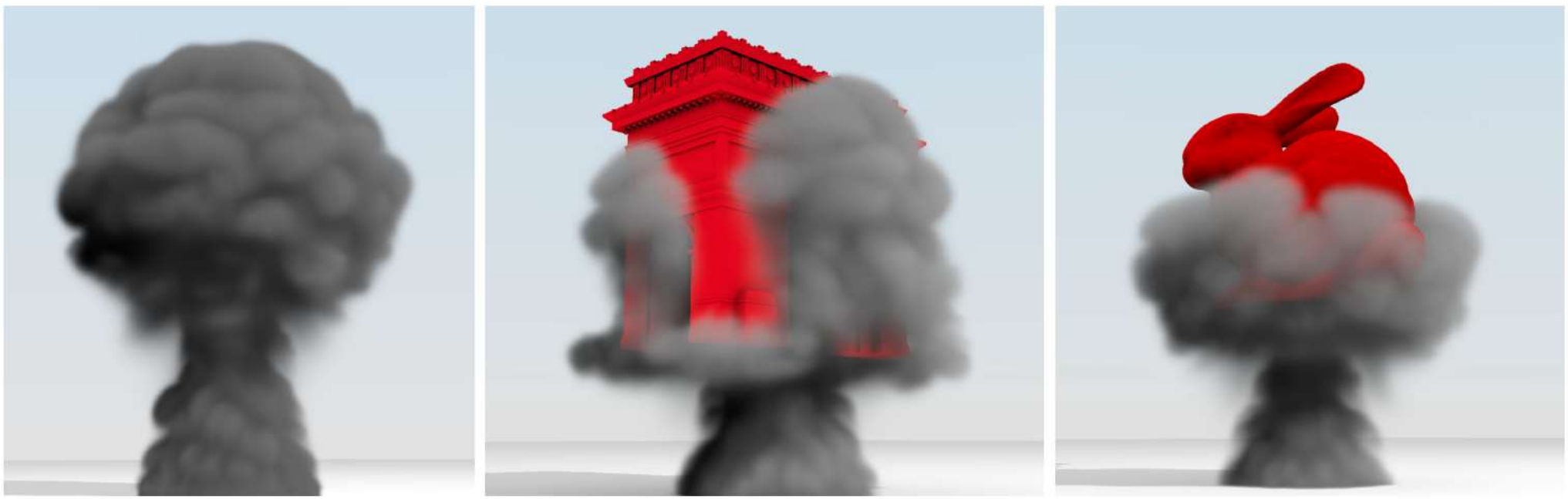


Fig. 1: Smoke simulation using our system - our method is capable of fast and accurate simulation of the Euler Equations for incompressible fluid flow at interactive frame-rates. Videos can be found at: `http://cims.nyu.edu/~schlacht/CNNFluids.htm`.

"ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=W71ZXKNIJFO

DEEP LEARNING SOFTWARE

POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

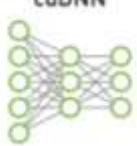
SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

Caffe
Chainer
DL4J
Deeplearning4j
Mocha.jl
julia
K
KERAS
Microsoft
CNTK
mxnet
Purine
TensorFlow
torch
MatConvNet
MINERVA
OpenDeep
Pylearn2
theano

NVIDIA DEEP LEARNING SDK

cuDNN



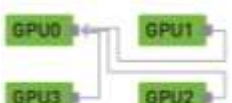
TensorRT

DeepStream SDK

cuBLAS

cuSPARSE

NCCL



developer.nvidia.com/deep-learning-software

DEEP LEARNING WORKFLOWS

New in DIGITS 5

IMAGE CLASSIFICATION

98% Dog

2% Cat

Classify images into classes or categories

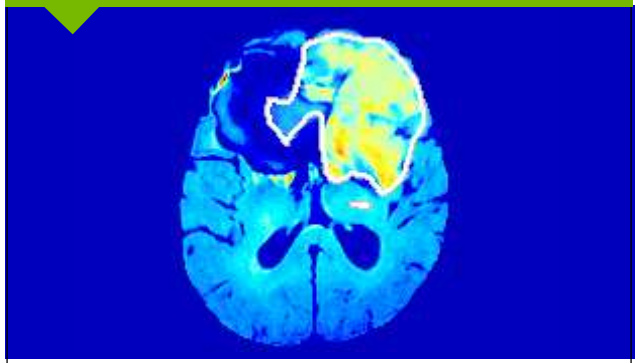
Object of interest could be anywhere in the image

OBJECT DETECTION

Find instances of objects in an image

Objects are identified with bounding boxes

IMAGE SEGMENTATION



Partition image into multiple regions

Regions are classified at the pixel level

WHAT'S NEW IN DIGITS?

TENSORFLOW SUPPORT

Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS

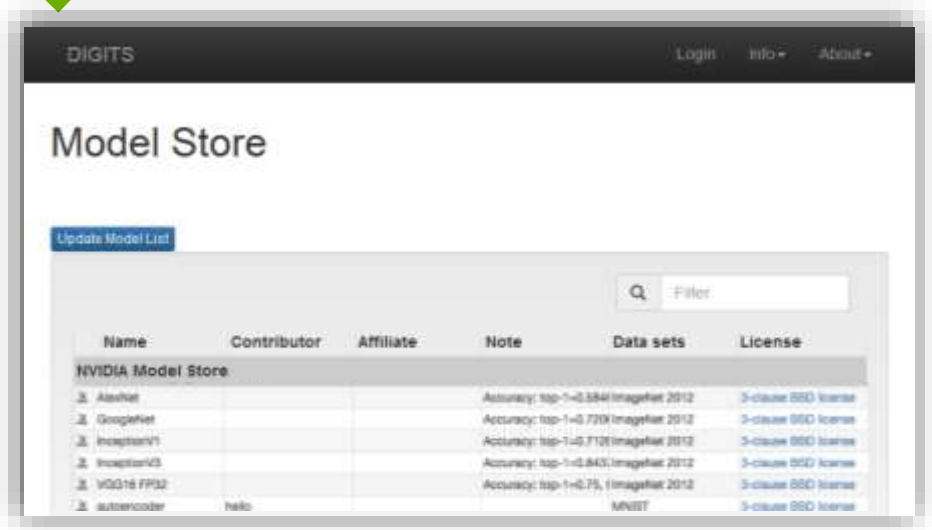
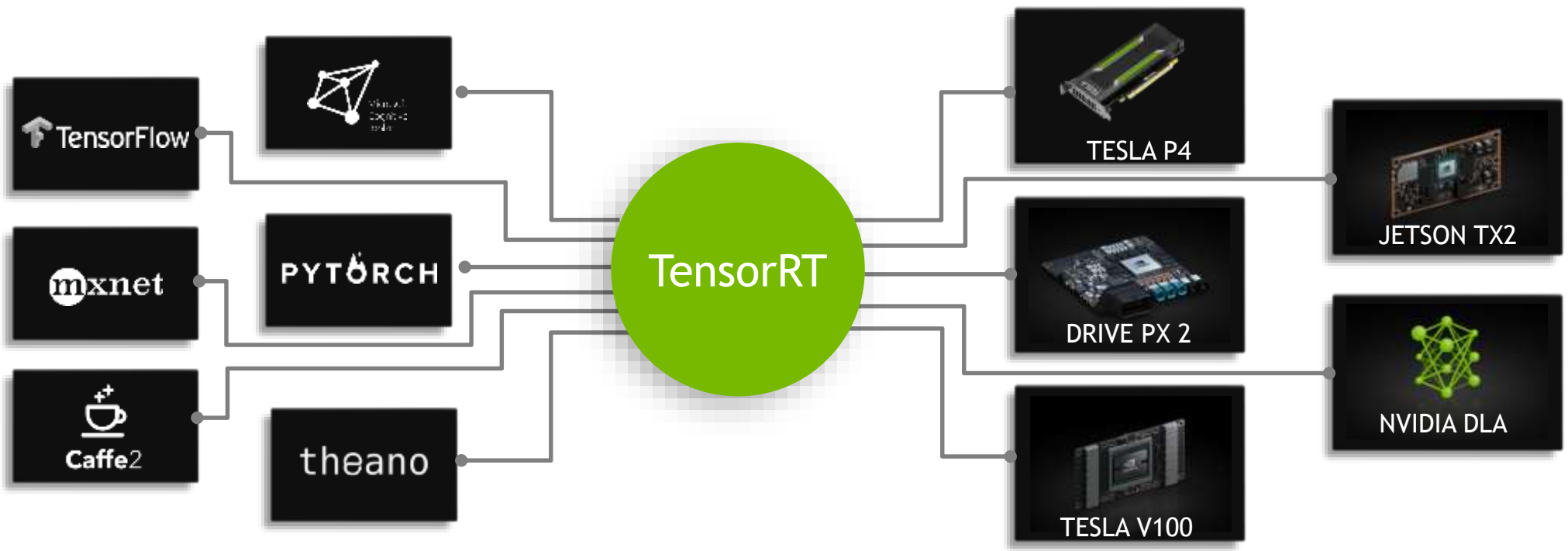


Image Classification: VGG-16, ResNet50
Object Detection: DetectNet

NVIDIA TENSORRT PROGRAMMABLE INFERENCE ACCELERATOR



DEEP LEARNING APPLICATIONS

AI TOOL BOOSTS CUSTOMER SERVICE

KLM's 235 social media service agents engage in 15K conversations a week, 24/7. To contend with the overwhelming volume of messages, KLM uses GPU-accelerated deep learning to predict the best response to an incoming message and shows it to a contact center agent for approval or personalization before sending it to the customer. The resulting time savings for KLM service agents means they can focus on customers with more pressing needs and handle a greater volume of questions while still maintaining a high degree of customer satisfaction.

KLM

DigitalGenius
Human+AI Customer Service

Fraud Prevention (ML & DL)

10,000s of features make up todays fraudulent behavior. AI can detect patterns faster and more accurate than humans

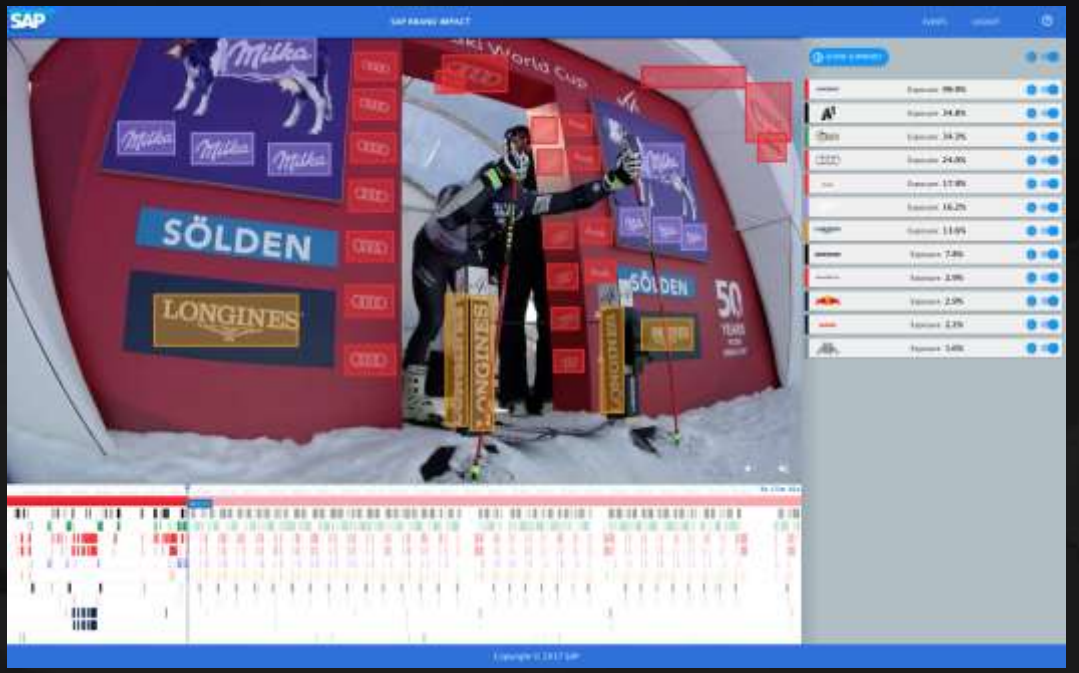
-Hui Wang, Senior Director of Global Risk Sciences, Pay Pal

SAP AI FOR THE ENTERPRISE

First commercial AI offerings from SAP

Brand Impact, Service Ticketing, Invoice-to-Record applications

Powered by NVIDIA GPUs on DGX-1 and AWS

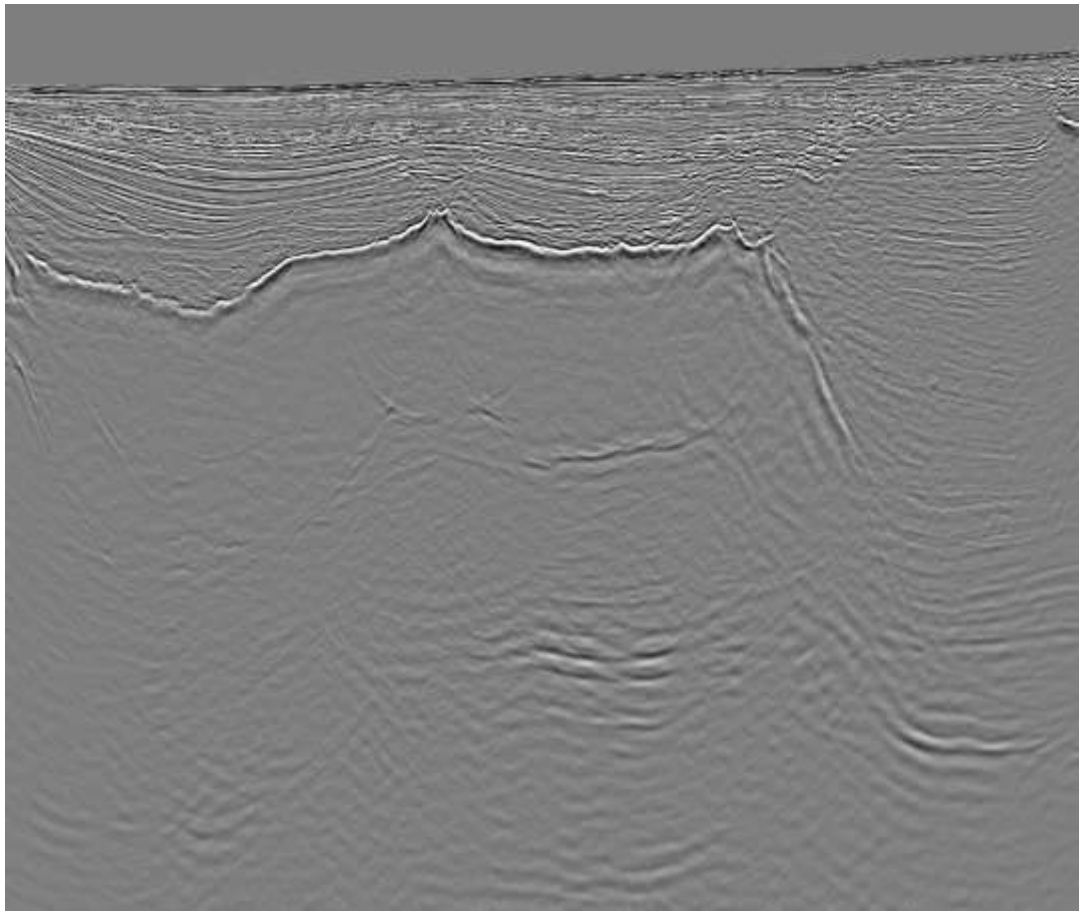


MICROSOFT - "BUILD 2017: WORKPLACE SAFETY DEMONSTRATION"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=PL-C00M2CNI

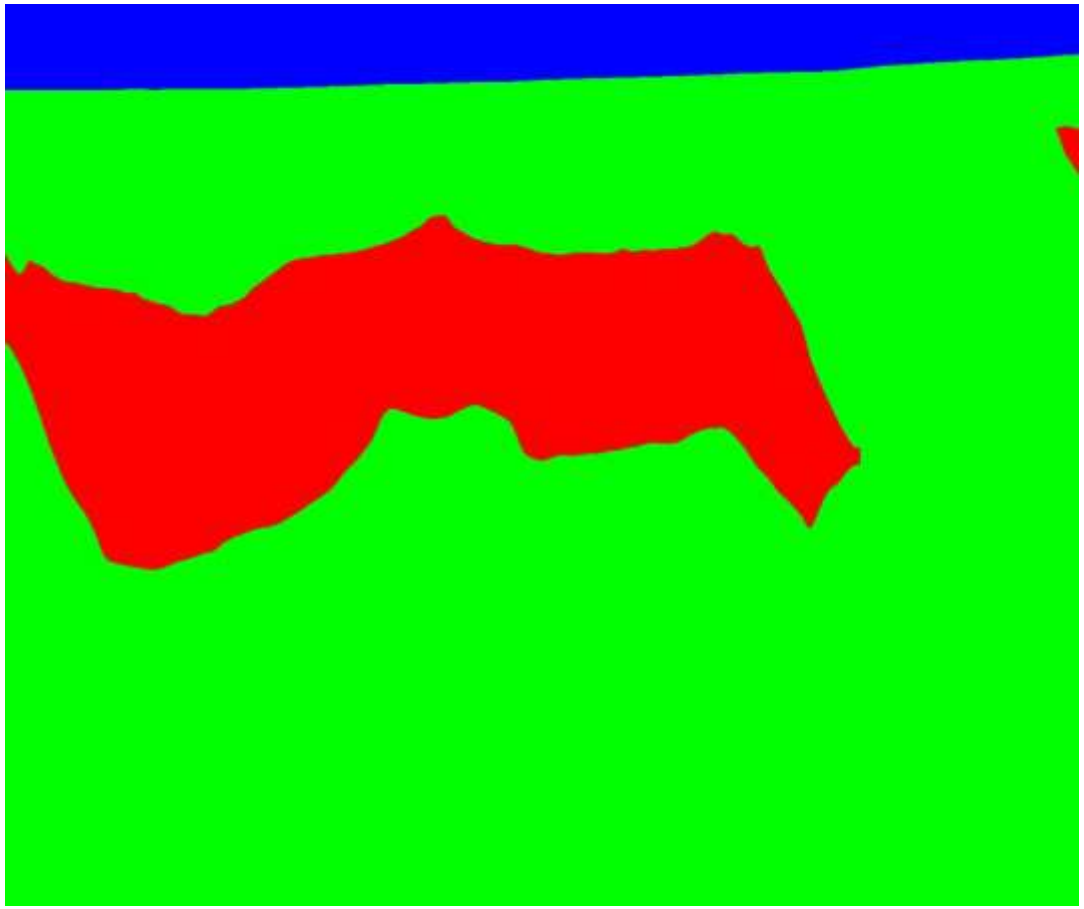
"FUTURE OF AI CITIES ON DISPLAY AT ISC WEST 2017"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=ZHBV34KOWHM

"LIP READING SENTENCES IN THE WILD"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=5AOGZAUPILE

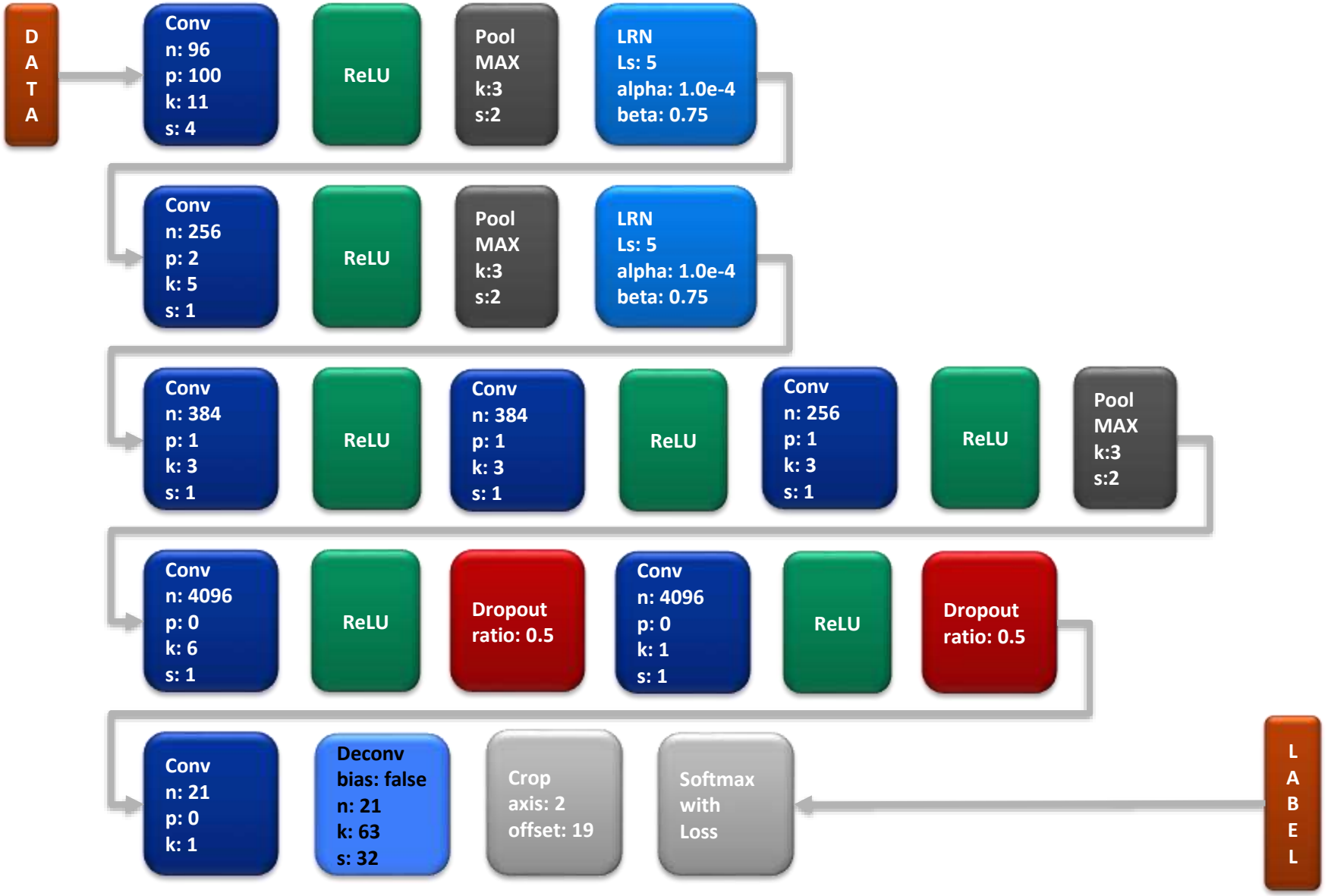
TRAINING SET



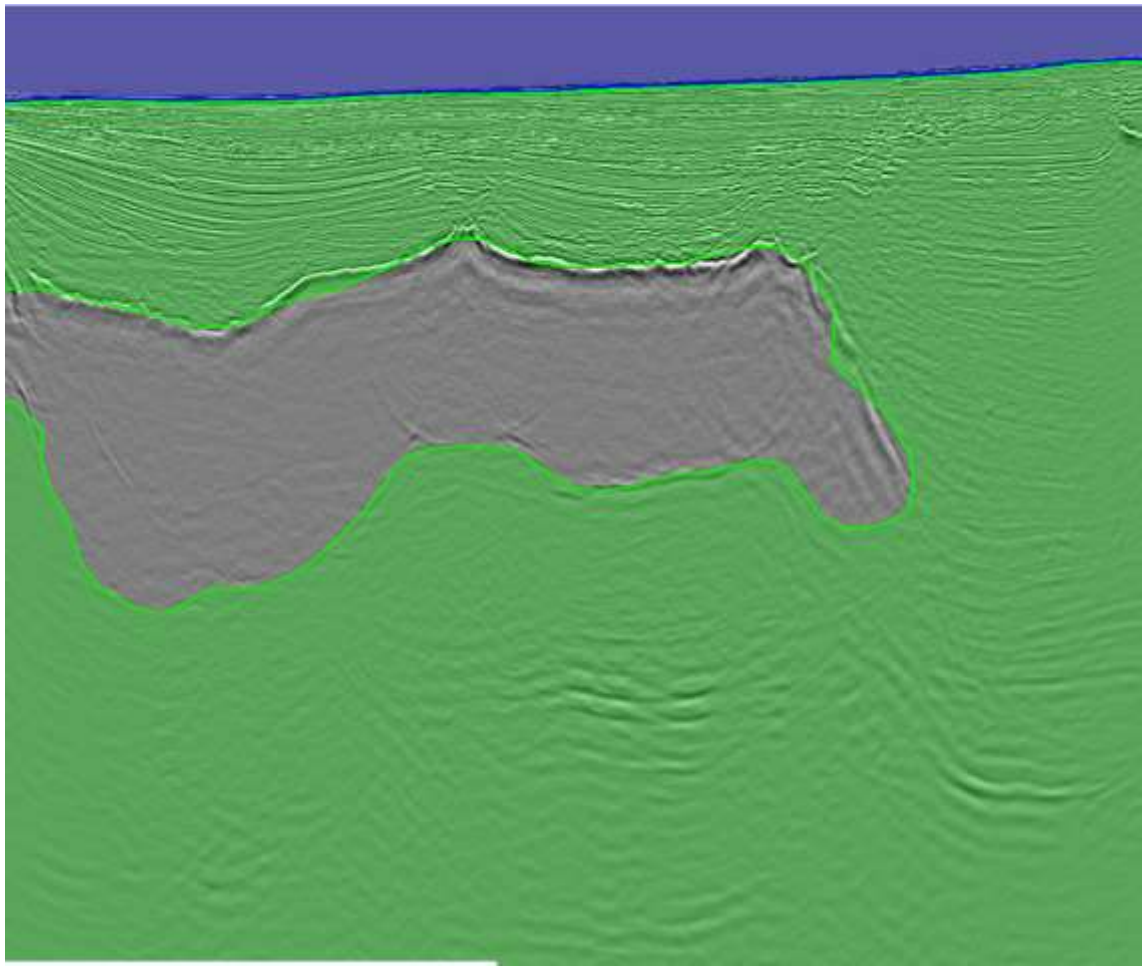
Features (Seismic)

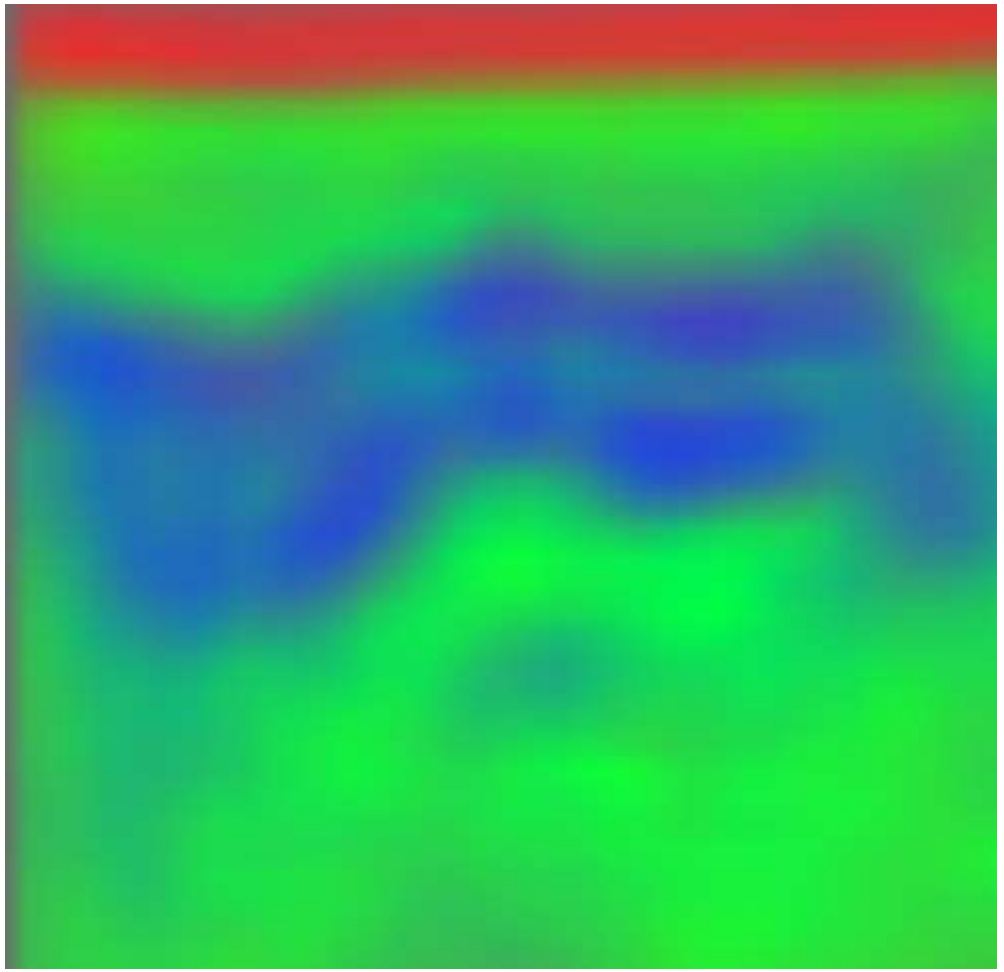


Labels


DATA
Conv
n: 96
p: 100
k: 11
s: 4
ReLU
Pool
MAX
k:3
s:2
LRN
Ls: 5
alpha: 1.0e-4
beta: 0.75
Conv
n: 256
p: 2
k: 5
s: 1
ReLU
Pool
MAX
k:3
s:2
LRN
Ls: 5
alpha: 1.0e-4
beta: 0.75
Conv
n: 384
p: 1
k: 3
s: 1
ReLU
Conv
n: 384
p: 1
k: 3
s: 1
ReLU
Conv
n: 256
p: 1
k: 3
s: 1
ReLU
Pool
MAX
k:3
s:2
Conv
n: 4096
p: 0
k: 6
s: 1
ReLU
Dropout
ratio: 0.5
Conv
n: 4096
p: 0
k: 1
s: 1
ReLU
Dropout
ratio: 0.5
Conv
n: 21
p: 0
k: 1
Deconv
bias: false
n: 21
k: 63
s: 32
Crop
axis: 2
offset: 19
Softmax
with
Loss
LABEL

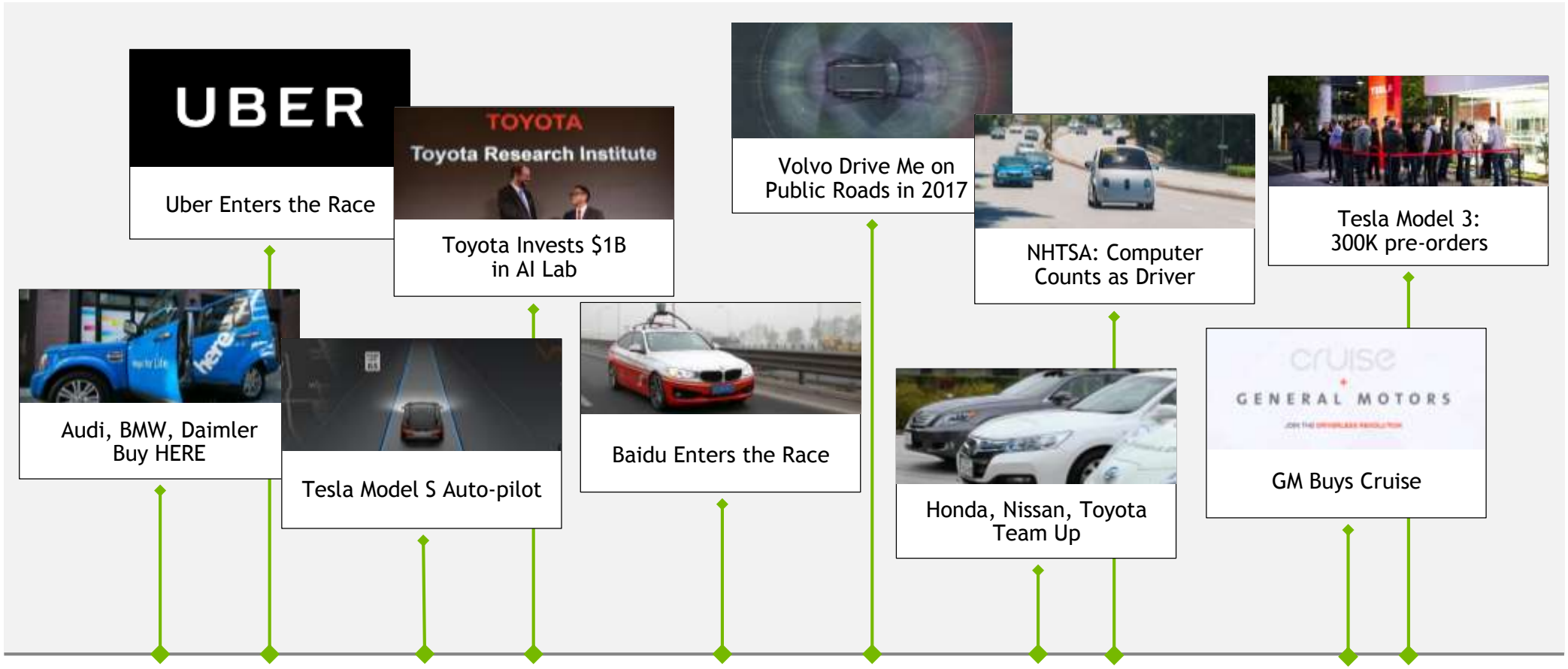
RESULT



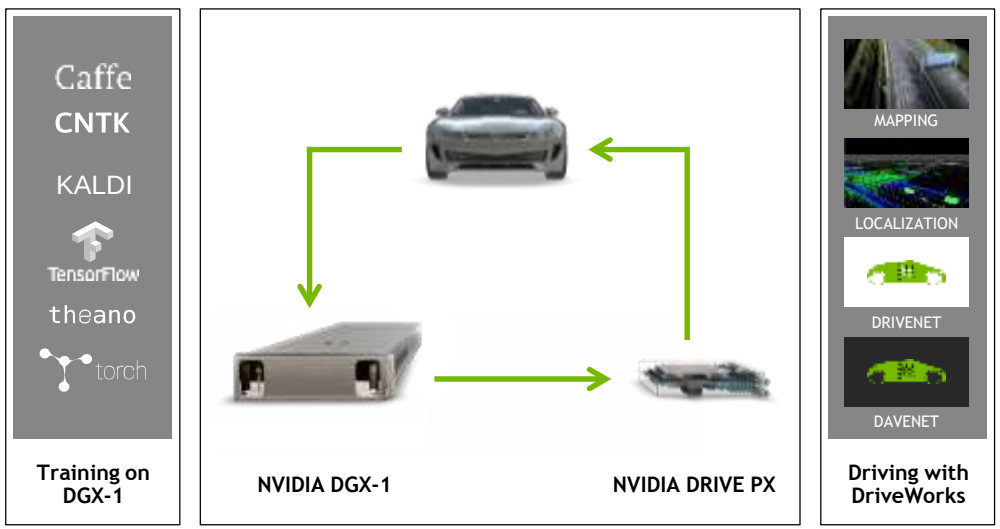


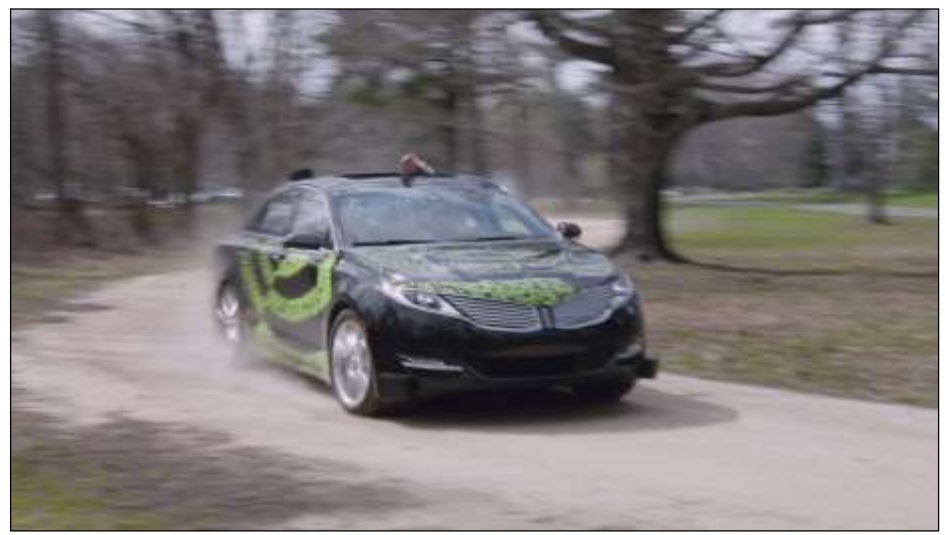
AUTONOMOUS CARS

AN AMAZING YEAR FOR SELF-DRIVING CARS



NEW AI DRIVING





"NVIDIA DRIVE AUTONOMOUS VEHICLE PLATFORM"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=0RC4RQYLTEU

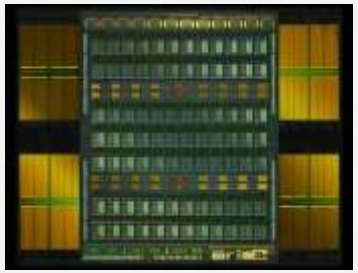
"NVIDIA SELF-DRIVING CAR DEMO AT CES 2017"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=FMVWLR0X1SK

TESLA PLATFORM

TESLA V100

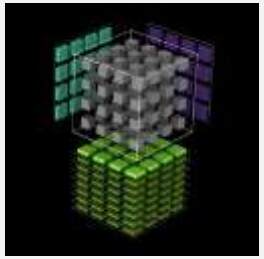
The Fastest and Most Productive GPU for AI and HPC

Volta Architecture



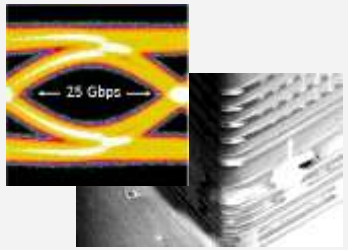
Most Productive GPU

Tensor Core



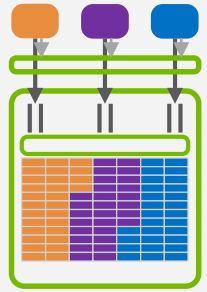
125 Programmable TFLOPS Deep Learning

Improved NVLink & HBM2



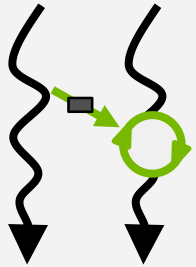
Efficient Bandwidth

Volta MPS



Inference Utilization

Improved SIMT Model



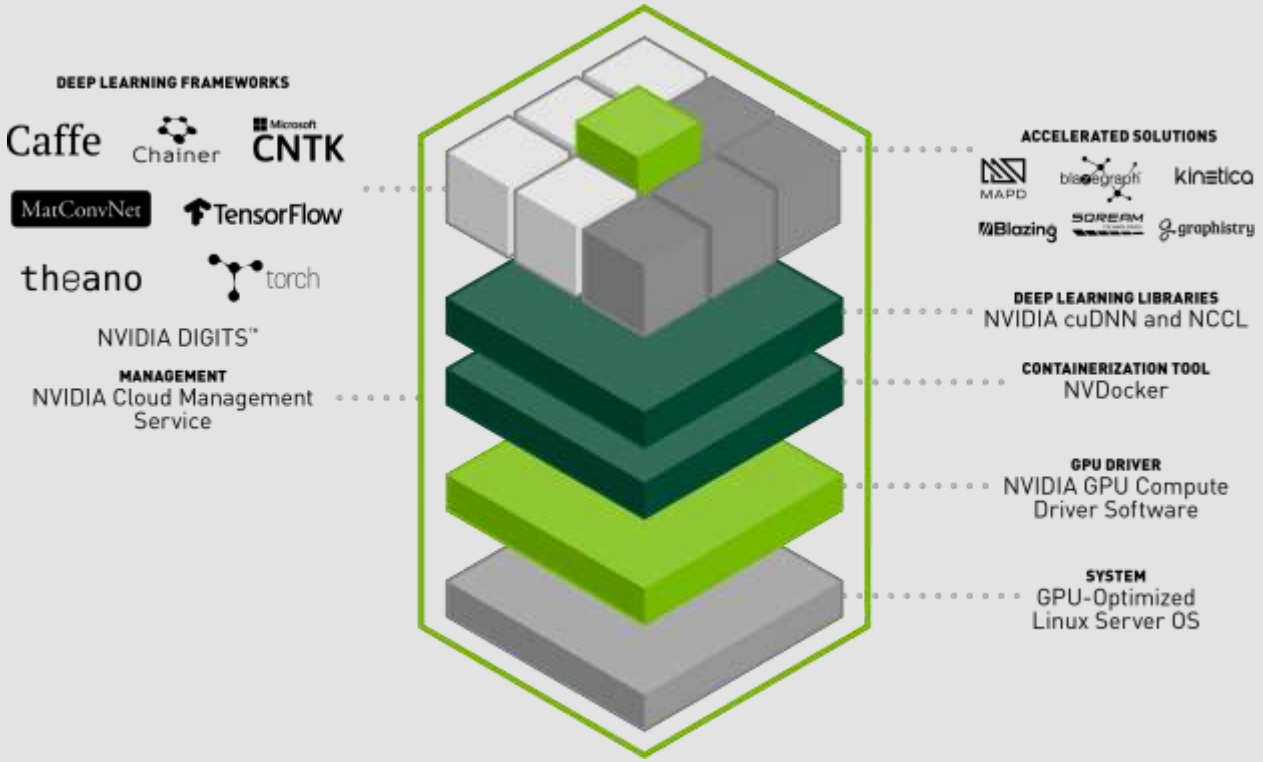
New Algorithms

"DGX-1: WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER IN A BOX"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=FAZS4V2AOLI&T

NVIDIA® DGX-1™

DGX STACK

Complete Analytics and Deep Learning platform



Instant productivity — plug-and-play, supports every AI framework and accelerated analytics software applications

Performance optimized across the entire stack

Always up-to-date via the cloud

Mixed framework environments — baremetal and containerized

Direct access to NVIDIA experts

GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm^2

5,120 CUDA cores

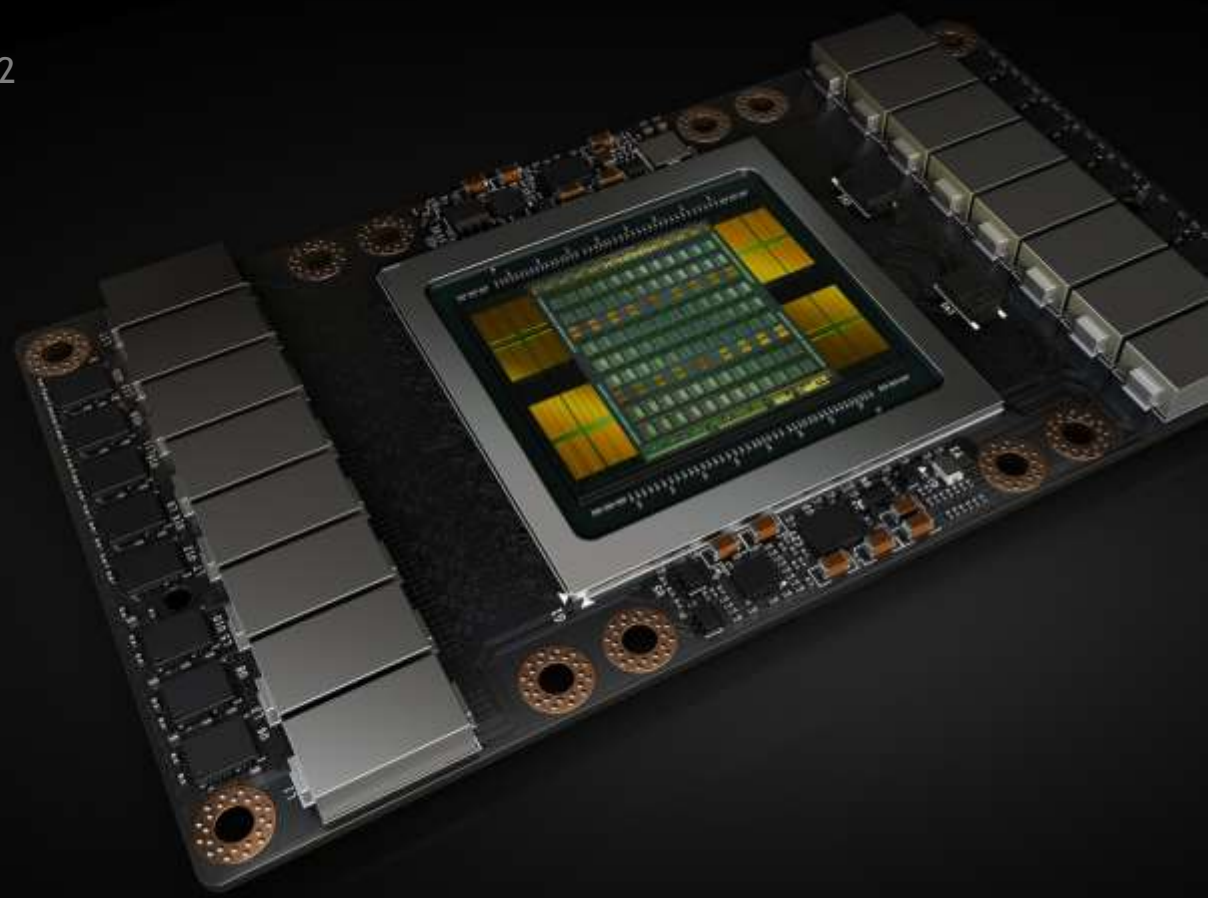
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

NEW 120 Tensor TFLOPS

20MB SM RF | 16MB Cache

16GB HBM2 @ 900 GB/s

300 GB/s NVLink

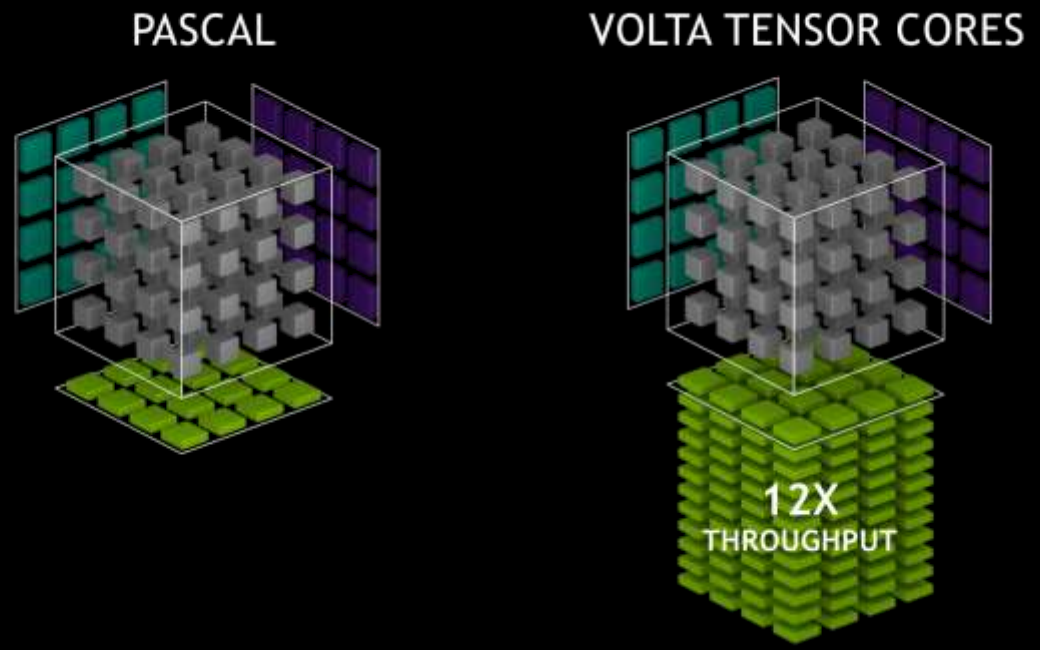


New CUDA TensorOp instructions
& data formats

4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

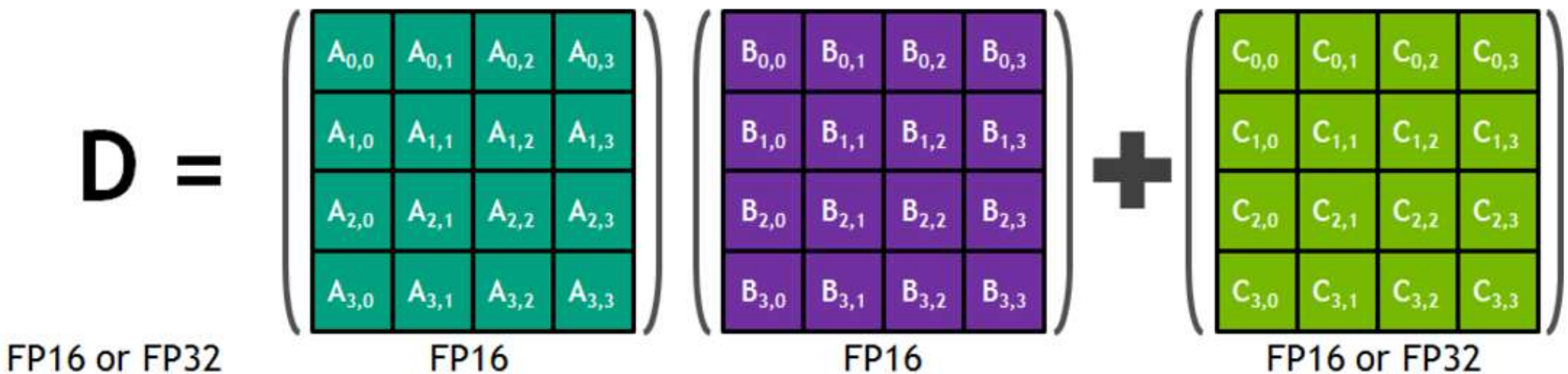
Optimized for deep learning



Activation Inputs Weights Inputs Output Results

TENSOR CORE

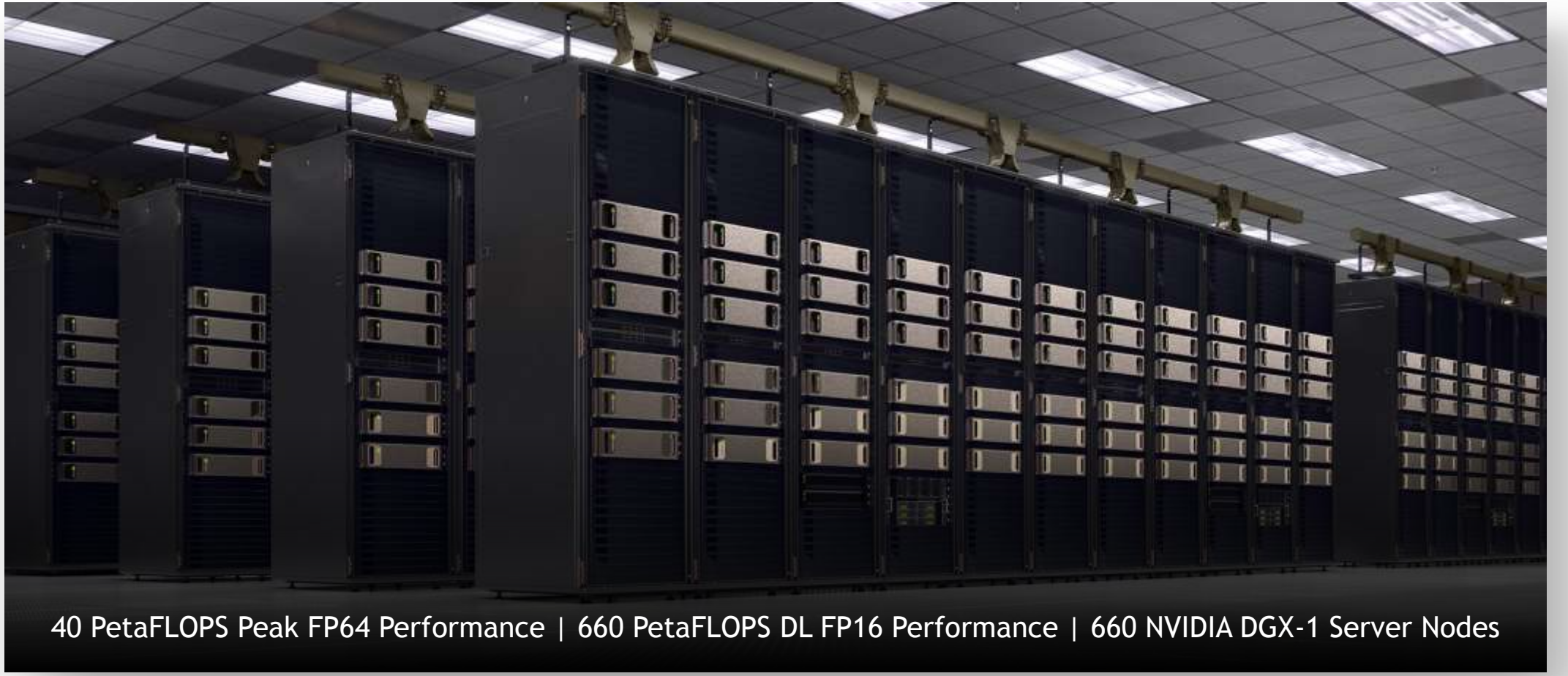
4x4x4 matrix multiply and accumulate



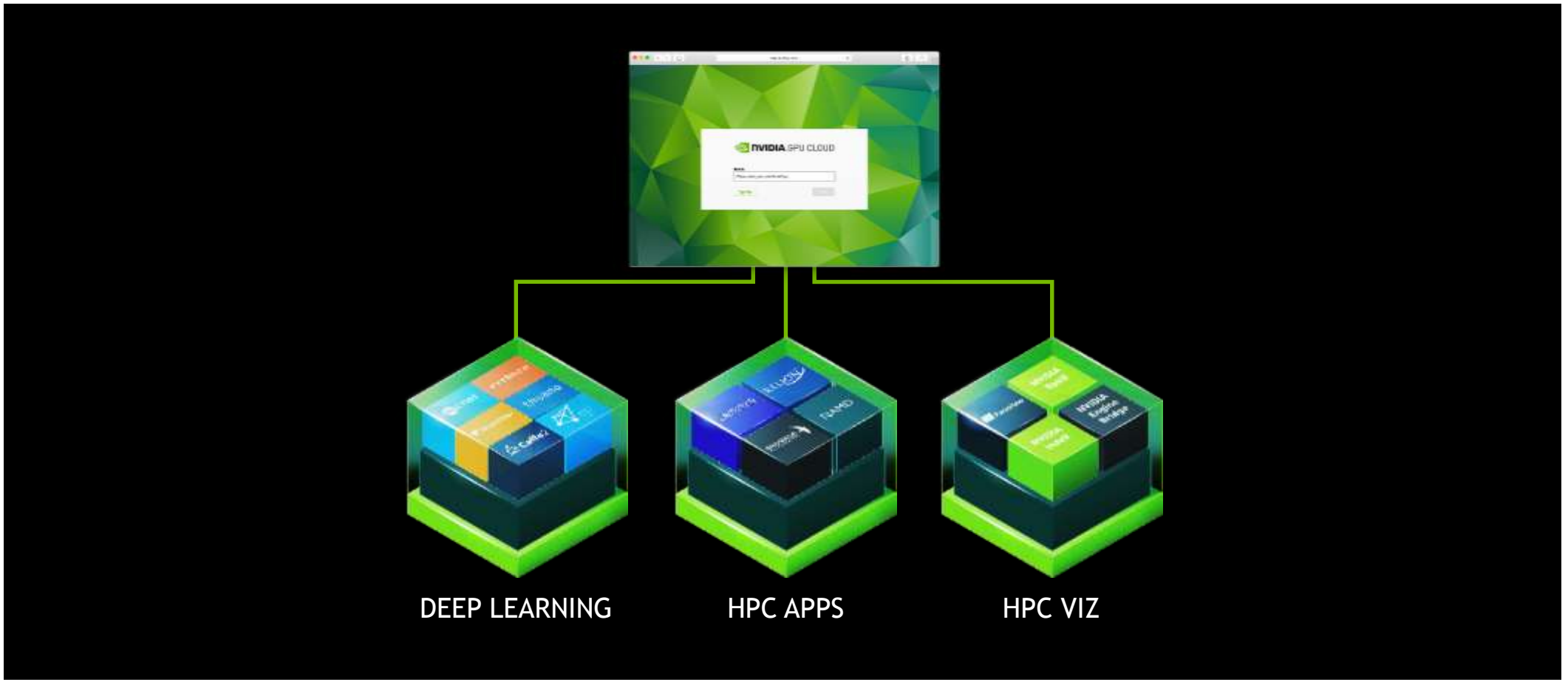
Tesla P100 vs Tesla V100

	Tesla P100 (Pascal)	Tesla V100 (Volta)
Memory	16 GB (HBM2)	16 GB (HMB2)
Memory Bandwidth	720 GB/s	900 GB/s
NVLINK	160 GB/s	300 GB/s
CUDA Cores (FP32)	3584	5120
CUDA Cores (FP64)	1792	2560
Tensor Cores (TC)	NA	640
Peak TFLOPS/s (FP32)	10.6	15
Peak TFLOPS/s (FP64)	5.3	7.5
Peak TFLOPS/s (TC)	NA	120
Power	300 W	300 W

ANNOUNCING
NVIDIA SATURNV WITH VOLTA



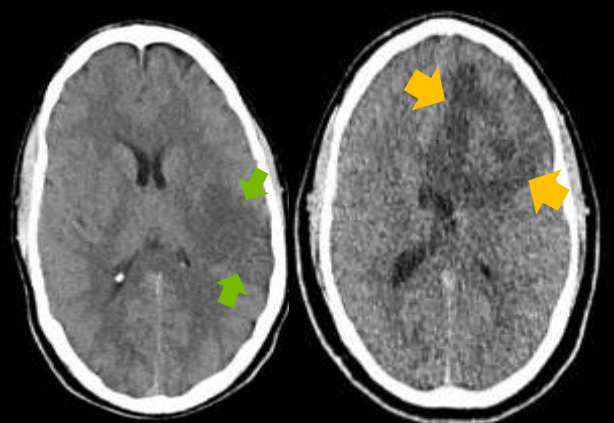
NVIDIA GPU CLOUD
SIMPLIFYING AI & HPC



THE VALUE OF GPU COMPUTING PLATFORM FOR AI

AI TO TRANSFORM EVERY INDUSTRY

HEALTHCARE



>80% Accuracy & Immediate Alert to Radiologists

INFRASTRUCTURE

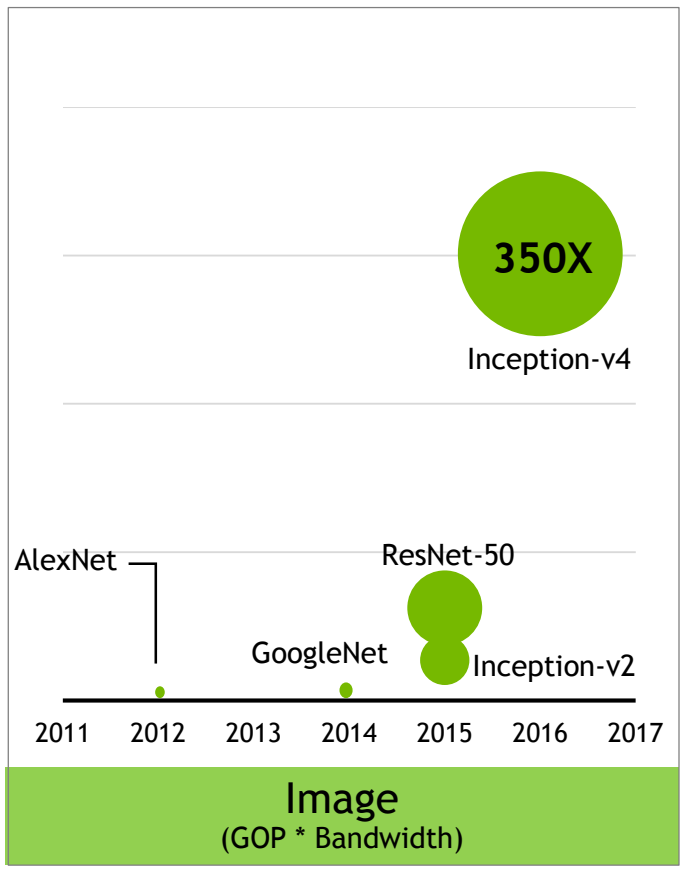
50% Reduction in Emergency Road Repair Costs

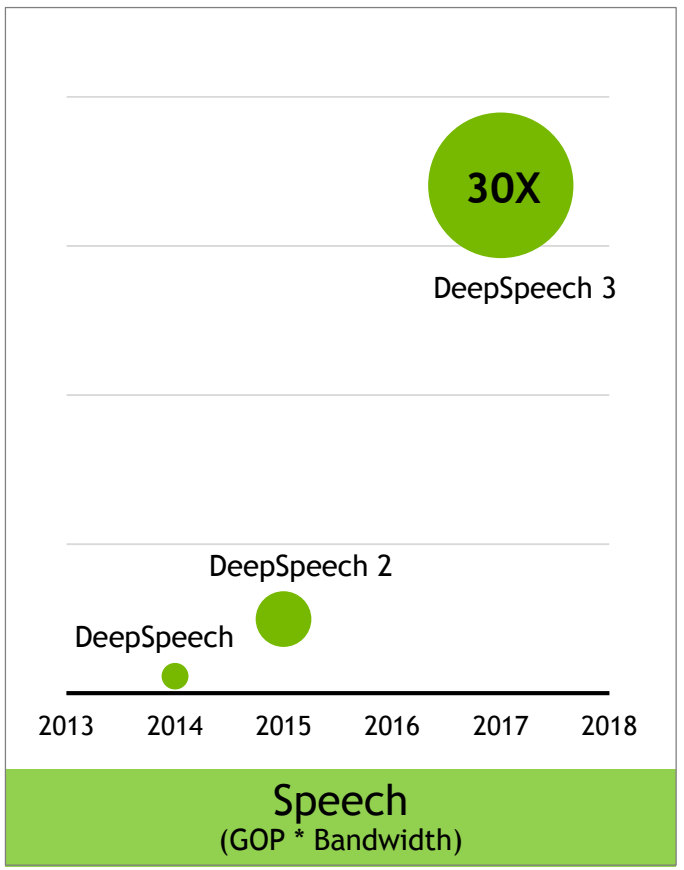
IOT

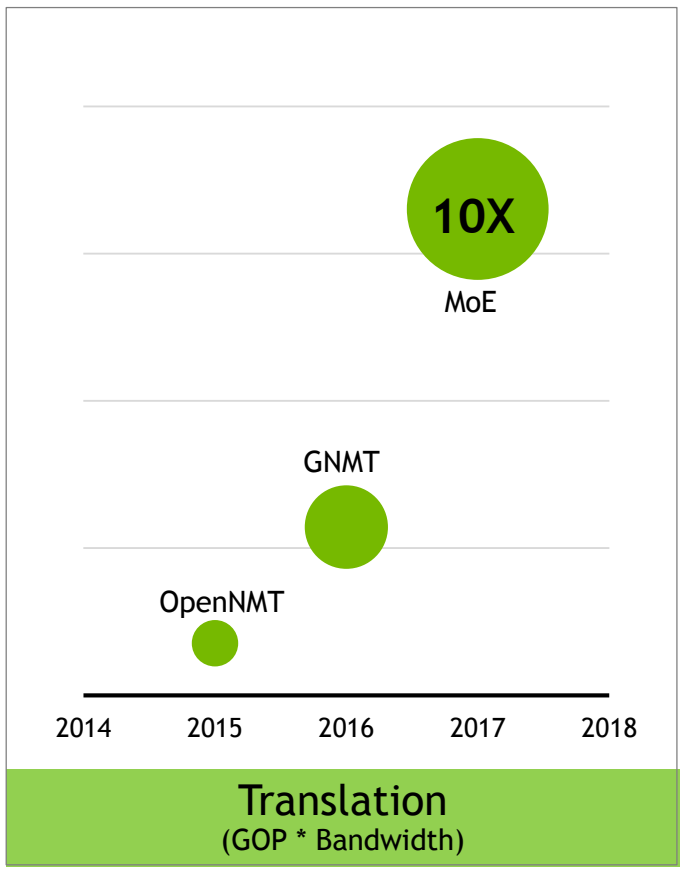
>$6M / Year Savings and Reduced Risk of Outage

NEURAL NETWORK COMPLEXITY IS EXPLODING

Bigger and More Compute Intensive







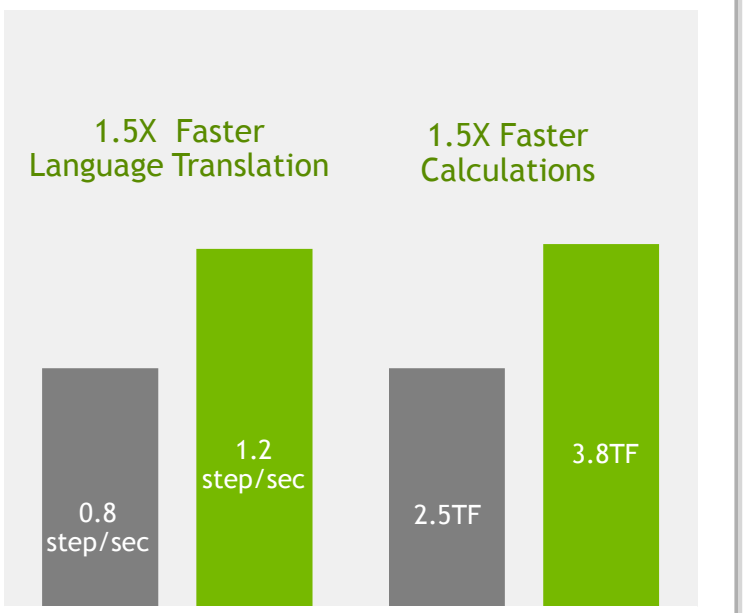
WORLD'S MOST ADVANCED DATA CENTER GPU
NOW WITH 2X THE MEMORY

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

FASTER RESULTS ON COMPLEX DL AND HPC

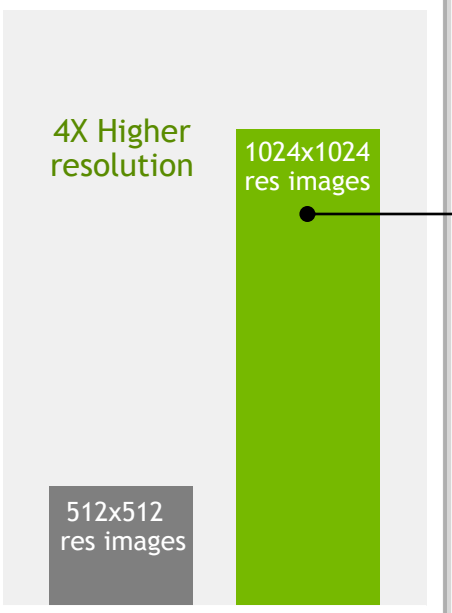
Up to 50% Faster Results With 2x The Memory

FASTER RESULTS



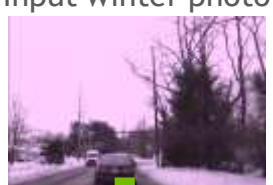
HIGHER ACCURACY

HIGHER RESOLUTION



Unsupervised Image Translation

Input winter photo



AI converts it to summer

V100 16GB V100 32GB

Dual E5-2698v4 server, 512GB DDR4, Ubuntu 16.04, CUDA9, cuDNN7| NMT is GNMT-like and run with TensorFlow NGC Container 18.01 (Batch Size= 128 (for 16GB) and 256 (for 32GB) | FFT is with cufftbench 1k x 1k x 1k and comparing 2 V100 16GB (DGX1V) vs. 2 V100 32GB (DGX1V)

R-CNN for object detection at 1080P with Caffe | V100 16GB uses VGG16| V100 32GB uses Resnet-152

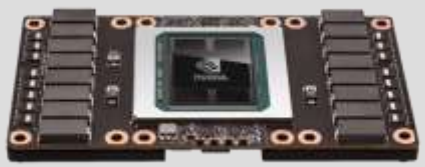
GAN by NVRESEARCH (https://arxiv.org/pdf/1703.00848.pdf) | V100 16GB and V100 32GB with FP32

TESLA V100: CHOOSING BETWEEN 32GB & 16GB

Application	Today's Products	New Deployments	Benefits
DL Training	P100, P40, V100 16GB	V100 32GB	Faster Result with Larger More Complex Models
Memory-Constrained HPC (Seismic, Graphs, CFD, Physics, Climate, Signal Processing, Finance)	K80, P100, V100 16GB	V100 32GB	Faster Results with Large Datasets
Compute-Bound HPC (Life Science, Molecular Dynamics)	K80, P100, V100 16GB	V100 16GB	Higher TCO

V100 WITH 32GB HBM2

Maintain Form Factor Compatibility

Form Factor		
Performance	7.8TF DP, 15.7TF SP, 125TF FP16	7TF DP, 14TF SP, 112TF FP16
Memory Size	32GB HBM2	32GB HBM2
Memory Bandwidth	900GB/s	900GB/s
GPU Peer to Peer	NVLink	PCIe Gen3
Power	300W	250W
Available From All Major OEMs	cisco CRAY DELL EMC Hewlett Packard Enterprise IBM inspur Lenovo SUPERMICRO TYAN	

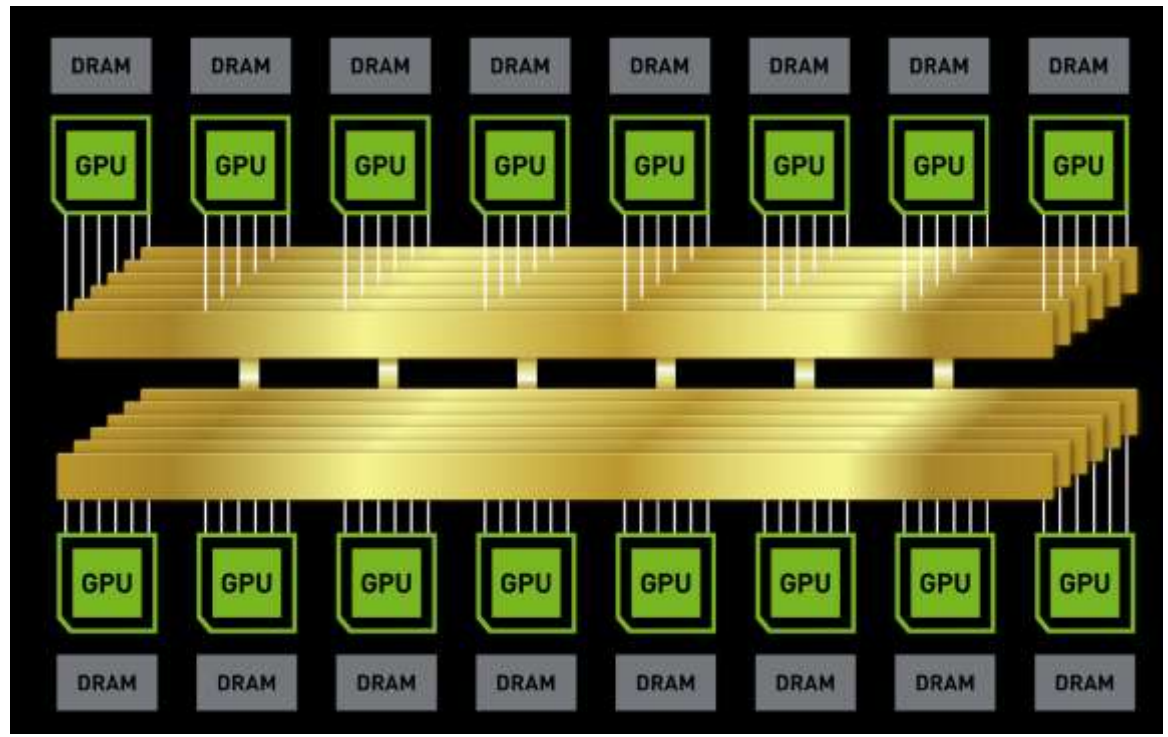
SXM2 32GB P/N = 900-2G503-0010-000, PCIE 32GB P/N = 900-2G500-0010-000

WORLD'S HIGHEST BANDWIDTH ON-NODE SWITCH

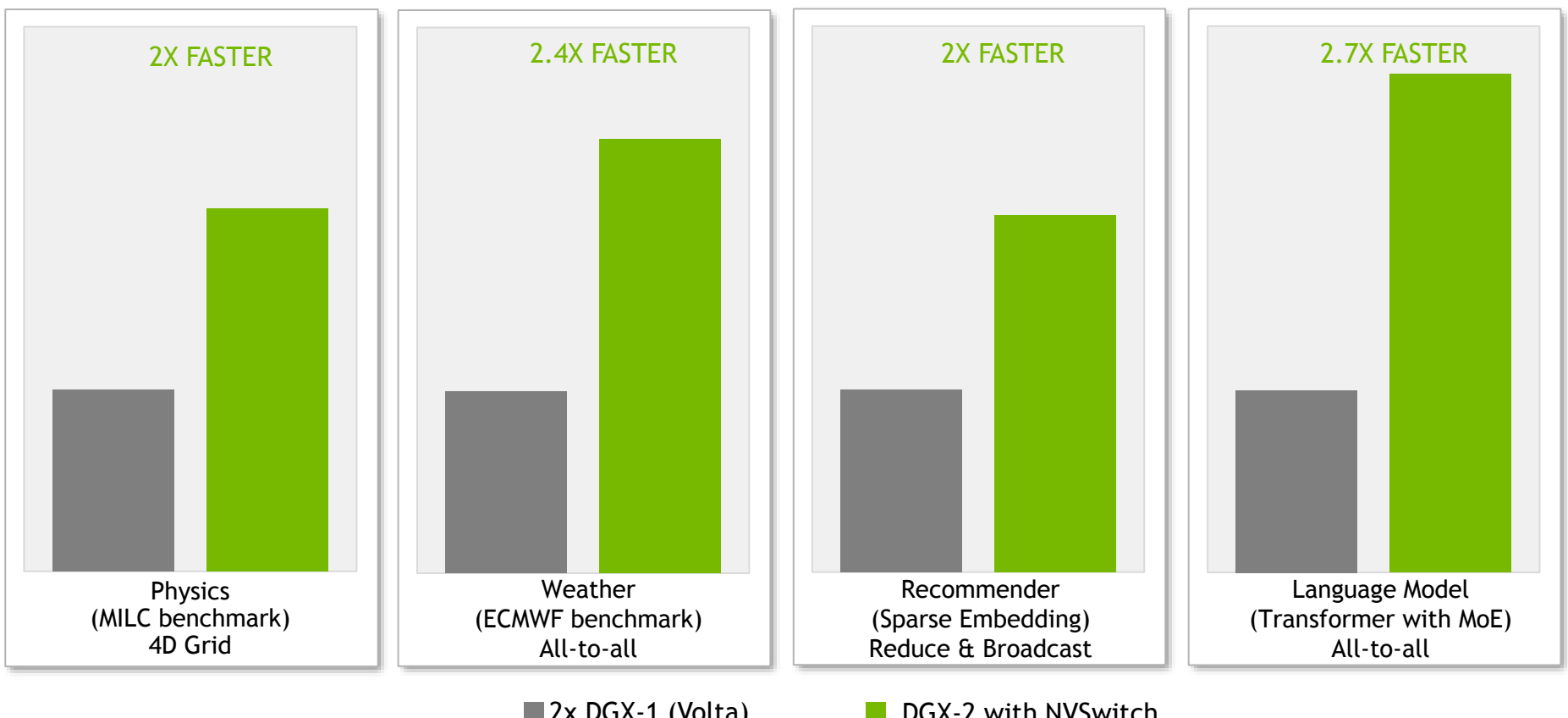
7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per port bi-directional
Fully-connected crossbar
2 billion transistors | 47.5mm x 47.5mm package

ENABLES THE WORLD'S LARGEST GPU

16 Tesla V100 32GB Connected by New NVSwitch
2 petaFLOPS of DL Compute
Unified 512GB HBM2 GPU Memory Space
300GB/sec Every GPU-to-GPU
2.4TB/sec of Total Cross-section Bandwidth



2X HIGHER PERFORMANCE WITH NVSWITCH



2 DGX-1V servers have dual socket Xeon E5 2698v4 Processor. 8 x V100 GPUs. Servers connected via 4X 100Gb IB ports | DGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 V100 GPUs

THE WORLD'S FIRST 2 PETAFLOPS SYSTEM

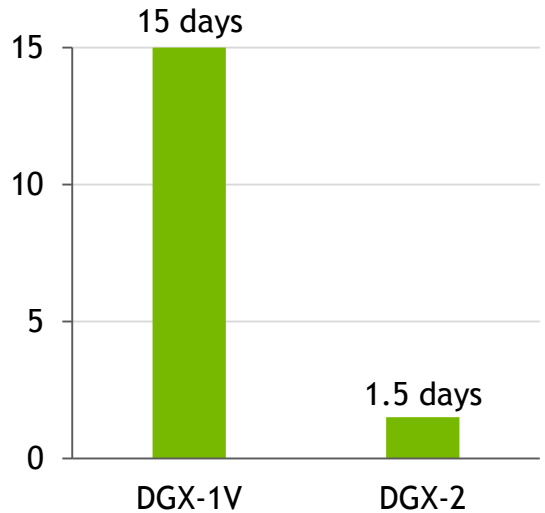
INTRODUCING NVIDIA DGX-2

THE WORLD'S MOST POWERFUL AI SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

- DGX-2 is the newest addition to the DGX family, powered by DGX software
- Deliver accelerated AI-at-scale deployment and simplified operations
- Step up to DGX-2 for unrestricted model parallelism and faster time-to-solution

10X PERFORMANCE GAIN LESS THAN A YEAR

DGX-1, SEP'17

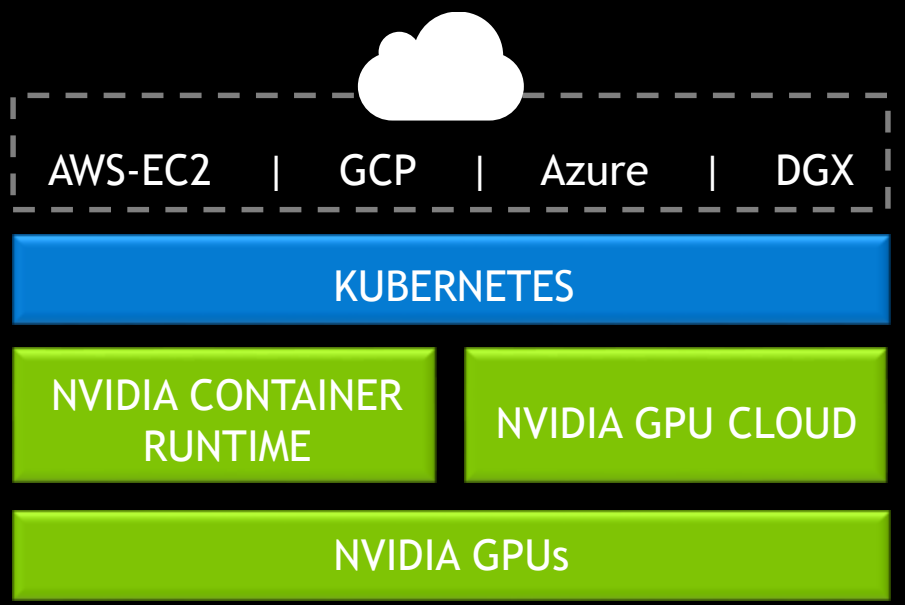


DGX-2, Q3'18

PyTorch Stack: Time to Train FAIRSEQ

software improvements across the stack including NCCL, cuDNN, etc.

Container Orchestration for DL Training & Inference



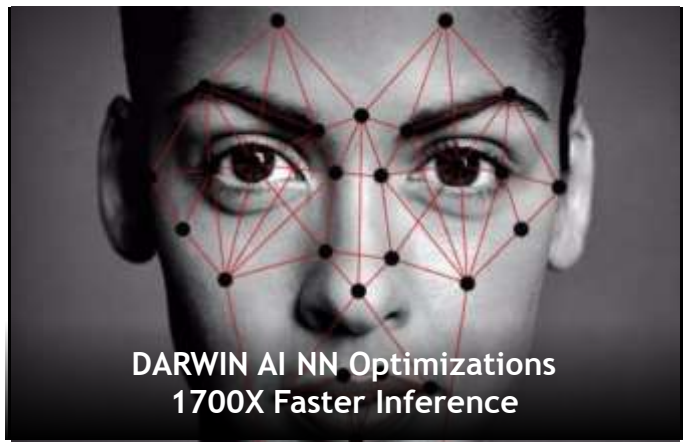
KUBERNETES on NVIDIA GPUs

- Scale-up Thousands of GPUs Instantly
- Self-healing Cluster Orchestration
- GPU Optimized Out-of-the-Box
- Powered by NVIDIA Container Runtime
- Included with Enterprise Support on DGX
- Available end of April 2018

INFERENCE ON GPUS TODAY

BING Object Detection
60X Improved Latency

iFLYTEK Speech Recognition
10X Concurrent Requests Per Server



DARWIN AI NN Optimizations
1700X Faster Inference

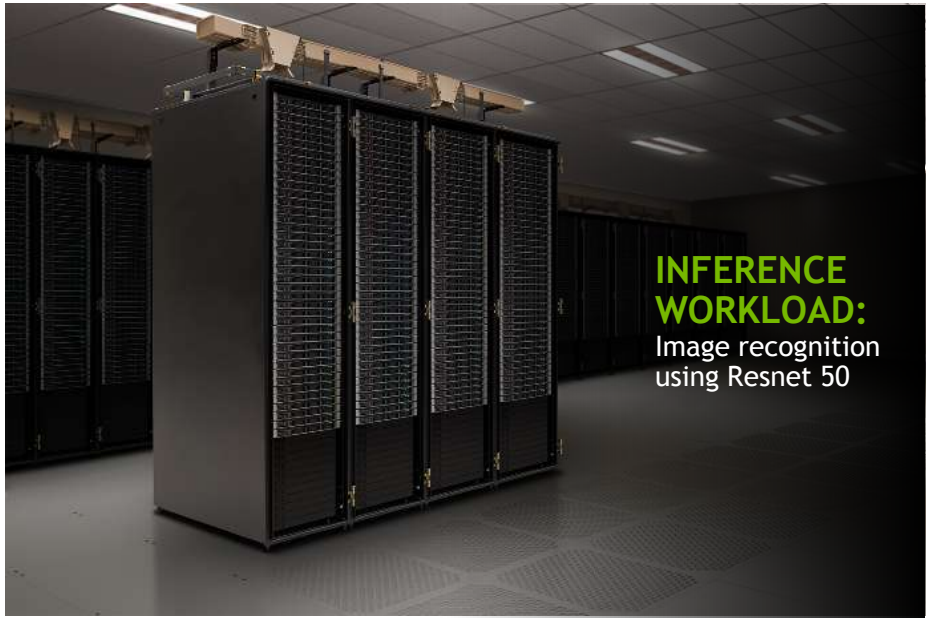
VALOSSA Video Intelligence
6X Faster Video Processing

ALIBABA Neural Machine Translation
3X Requests Per Server

KLM Social Media Engagement
10X increase in customer responses

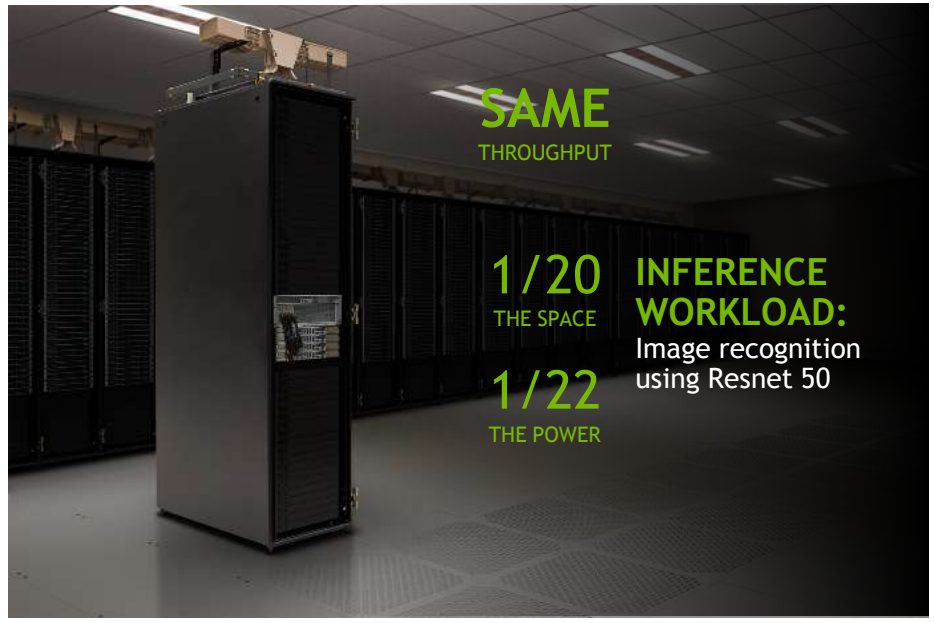
NVIDIA TESLA PLATFORM SAVES MONEY

Game-Changing Inference Performance



160 CPU Servers
45,000 images/sec
65 KWatts

=



1 HGX Server
45,000 images/sec
3 KWatts

NVIDIA AI INFERENCE

30M
HYPERSCALE SERVERS

TensorRT 4

TensorFlow Integration

Kaldi Optimization

ONNX WinML

190X
IMAGE
ResNet-50 with TensorFlow Integration

50X
NLP
GNMT

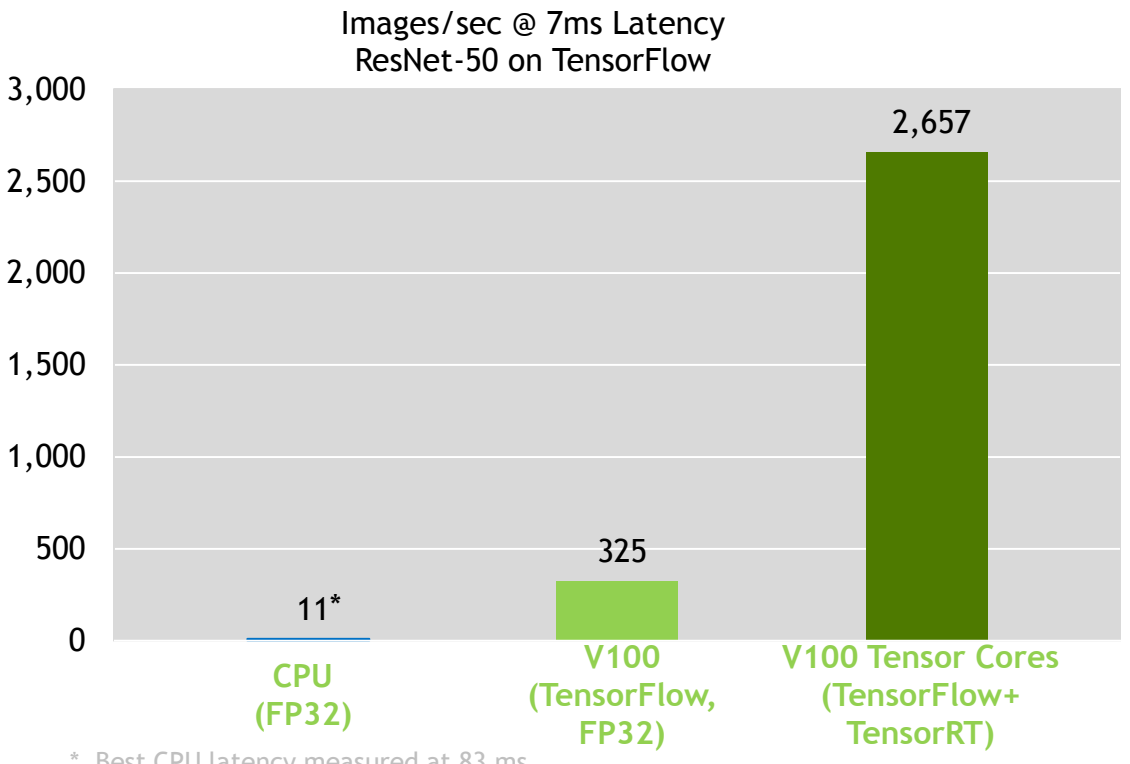
45X
RECOMMENDER
Neural Collaborative Filtering

36X
SPEECH SYNTH
WaveNet

60X
SPEECH RECOG
DeepSpeech 2 DNN

TensorRT INTEGRATED WITH TensorFlow

Delivers 8x Faster Inference with TensorFlow + TRT



* Best CPU latency measured at 83 ms

CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads.
Volta V100 SXM; CUDA (384.111; v9.0.176);
Batch size: CPU=1, TF_GPU=2, TF-TRT=16 w/ latency=6ms

TensorFlow
TensorRT

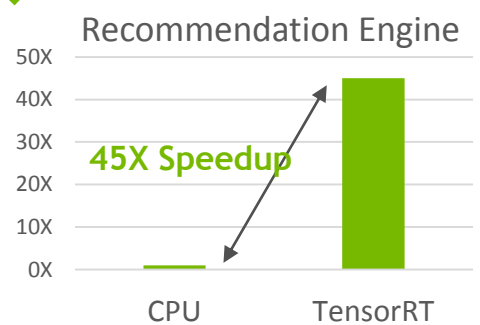
Available in TensorFlow 1.7

https://github.com/tensorflow

NVIDIA TensorRT 4 RC NOW AVAILABLE

RNN and MLP Layers • ONNX Import • NVIDIA DRIVE Support

Maximize RNN and MLP Throughput



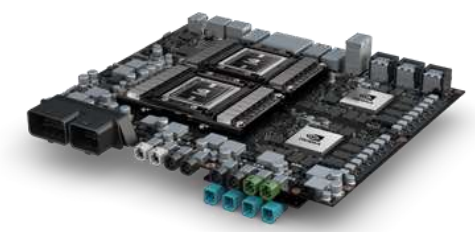
Speed up speech, audio and recommender app inference performance through new layers and optimizations

Optimize and Deploy ONNX Models

ONNX

Easily import and accelerate inference for ONNX frameworks (PyTorch, Caffe 2, CNTK, MxNet and Chainer)

Support for NVIDIA DRIVE Xavier

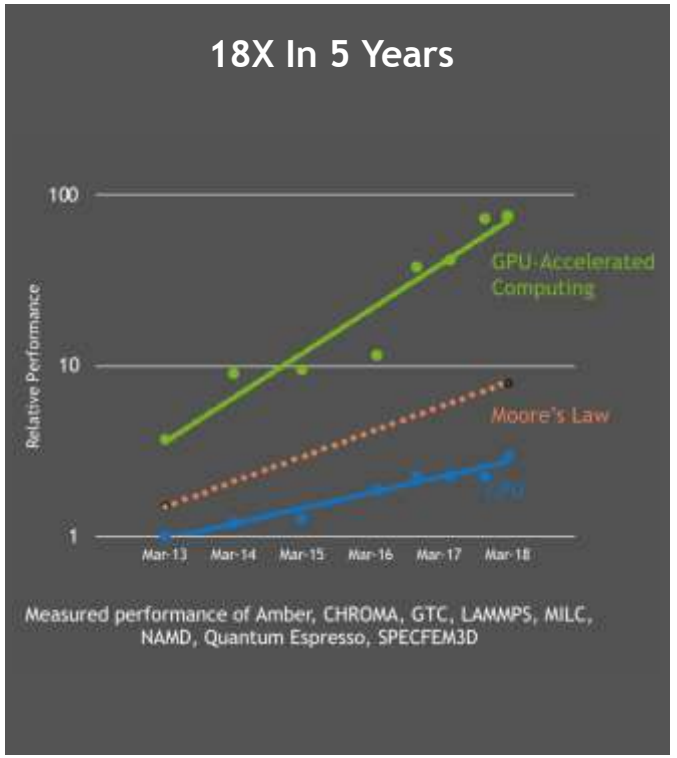


Deploy optimized deep learning inference models NVIDIA DRIVE Xavier

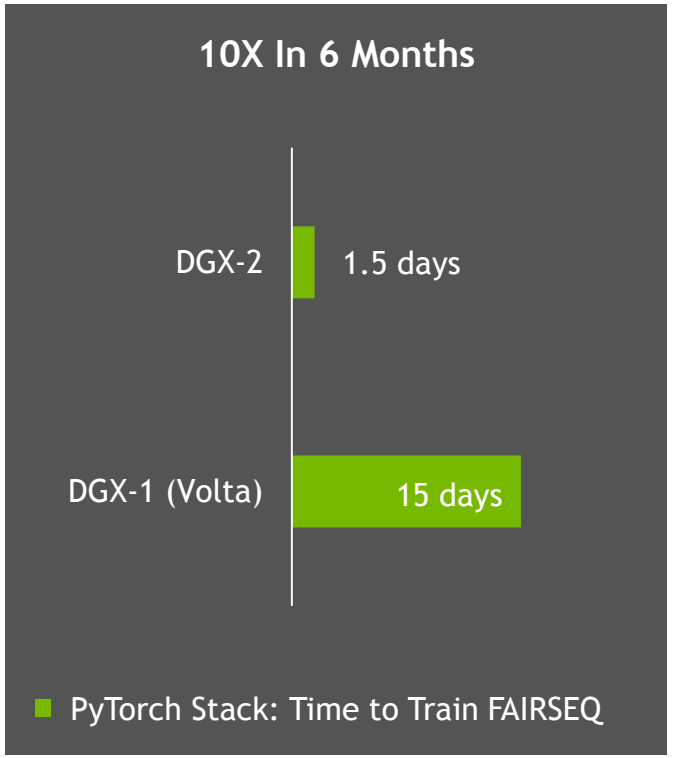
Free download to members of NVIDIA Developer Program

developer.nvidia.com/tensorrt

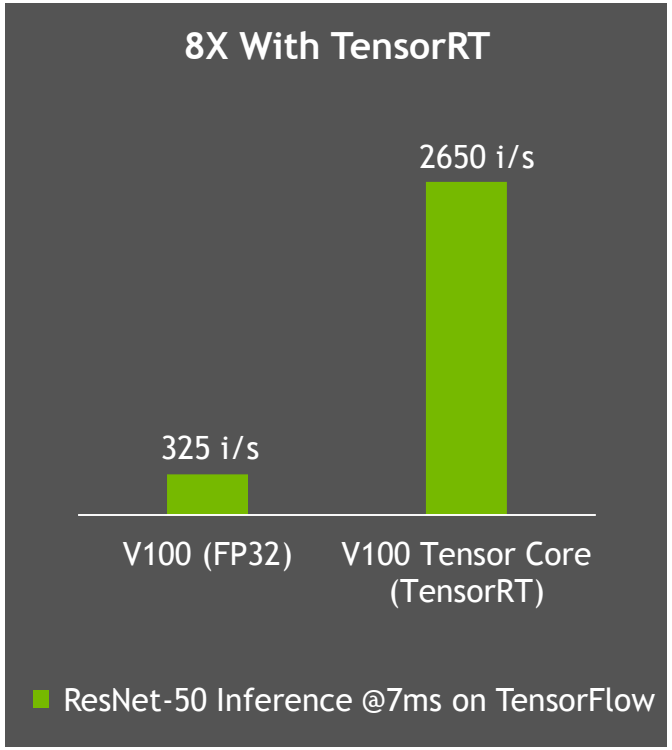
GPU COMPUTING STACK CONTINUES TO EVOLVE



Beyond Moore's Law For HPC



Accelerating Time To Solution For AI Training



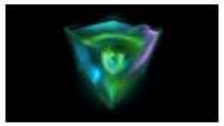
Compelling Platform For AI Inference

NVIDIA SUPPORT PROGRAMS

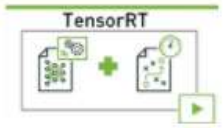
What's New

OptiX 5.0 Available OptiX 5 SDK features built-in support for motion blur, deep-learning based denoiser and more.

VRWorks 360 Video 1.1 Available This release bring many improvements listed below including new calibration technologies.



CUDA 9.1 Available New NPP functions for augmentation, multi-GPU enhancement in cuFFT, cuBLAS updates for Volta GPUs and more.



TensorRT 3 Available New TensorFlow model optimization, faster mixed-precision for CNNs and RNNs used for vision, speech & NLP.

See More

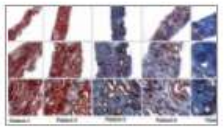
Developer News

AI Helps Farmers Distinguish Crop Data in Real Time
April 9, 2018

MIT Researchers Use AI to Capture Silent Speech
April 9, 2018

Drink up! Beer Tasting Robot Uses AI to Assess Quality
April 6, 2018

Researchers Develop AI System for License Plate Recognition
April 5, 2018



Boston University Researchers Use AI to Detect Kidney Disease
April 4, 2018

See More

developer.nvidia.com

NVIDIA DEEP LEARNING INSTITUTE

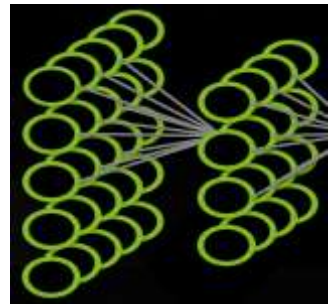
Hands-on self-paced and instructor-led training in deep learning and accelerated computing for developers

Request onsite instructor-led workshops at your organization: **www.nvidia.com/requestdli**

Take self-paced labs online: **www.nvidia.com/dlilabs**

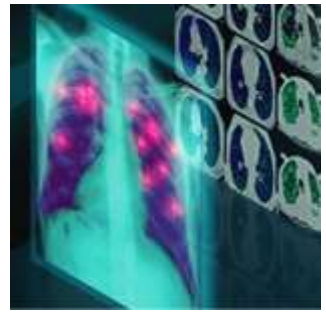
Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program: **www.nvidia.com/dli**

Caffe2 · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH

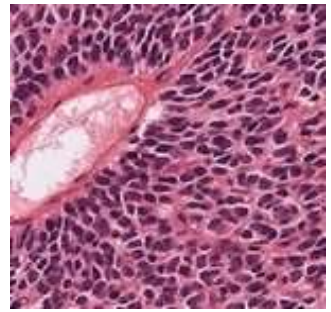


Deep Learning Fundamentals

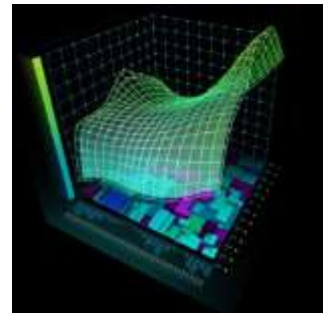
Autonomous Vehicles



Medical Image Analysis



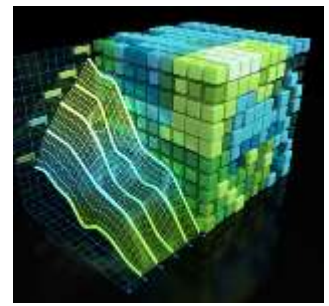
Genomics



Finance

Intelligent Video Analytics

Game Development & Digital Content



Accelerated Computing Fundamentals

More industry-specific training coming soon…

NVIDIA HW GRANT PROGRAM

Titan X Pascal

- Scientific Computing
- HPC
- Deep Learning

Quadro P5000

- Scientific Visualization
- Virtual Reality

Jetson TX2 (Dev Kit)

- Robotics
- Autonomous Machines

http://www.nvidia.com/object/inception-program.html

Pedro Mario Cruz e Silva (pcruzesilva@nvidia.com)
Solution Architect Manager
Enterprise Latin America
Global Oil & Gas Team
LinkedIn