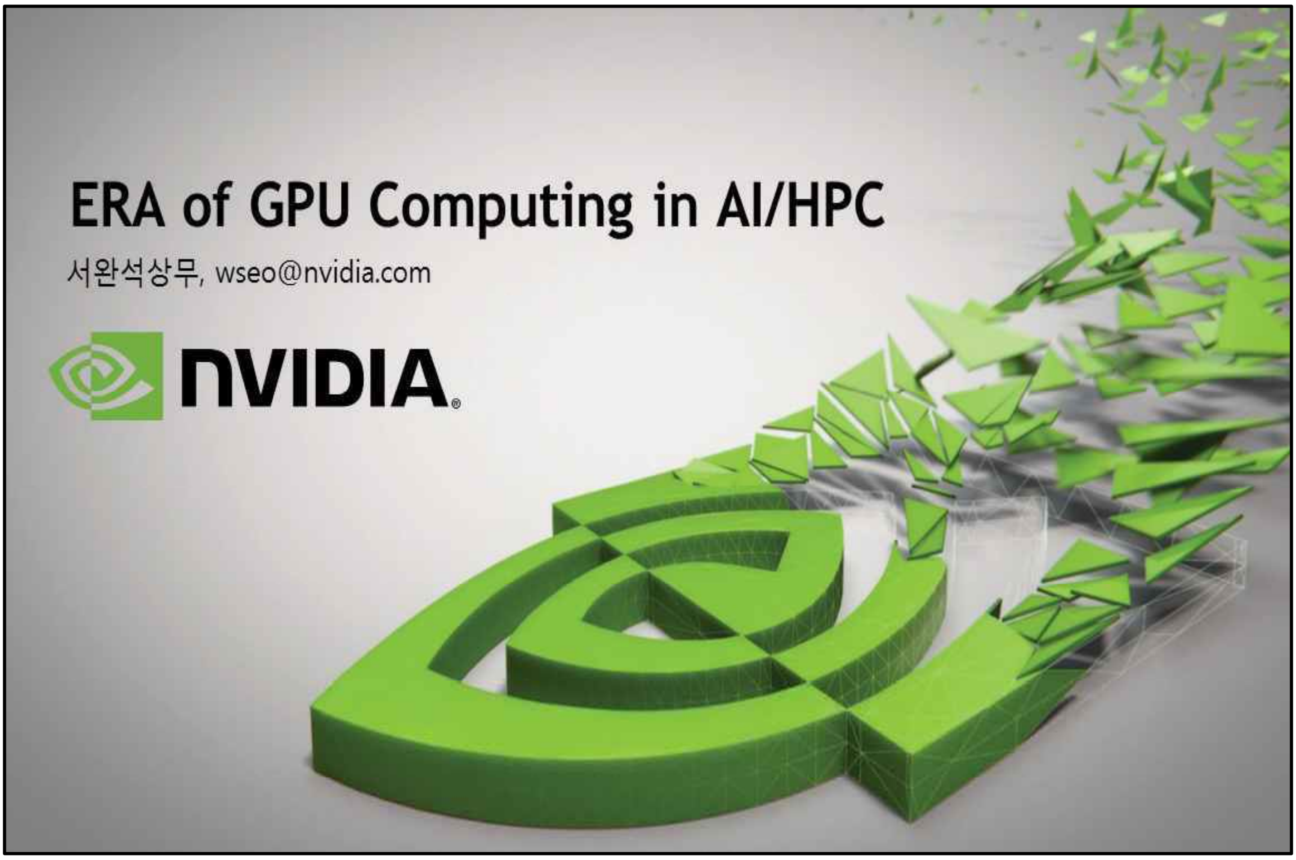
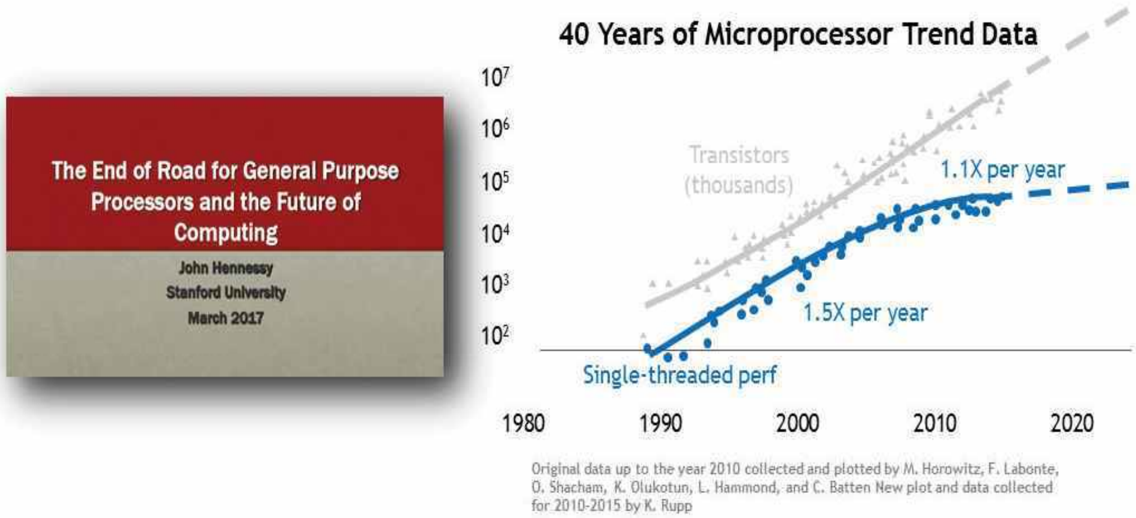
Era of GPU Computing in AI HPC

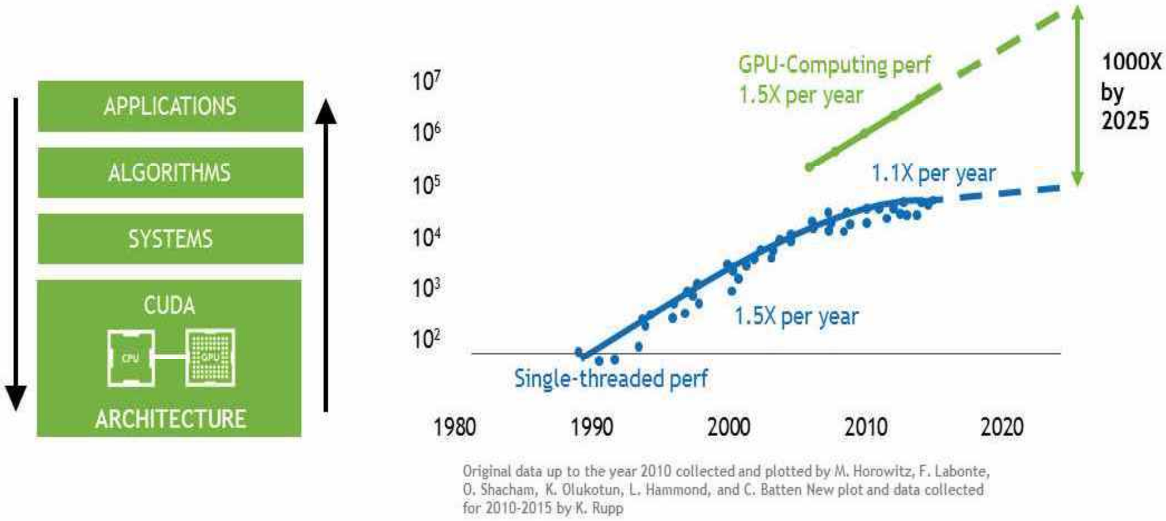
서완석 상무 (엔비디아)

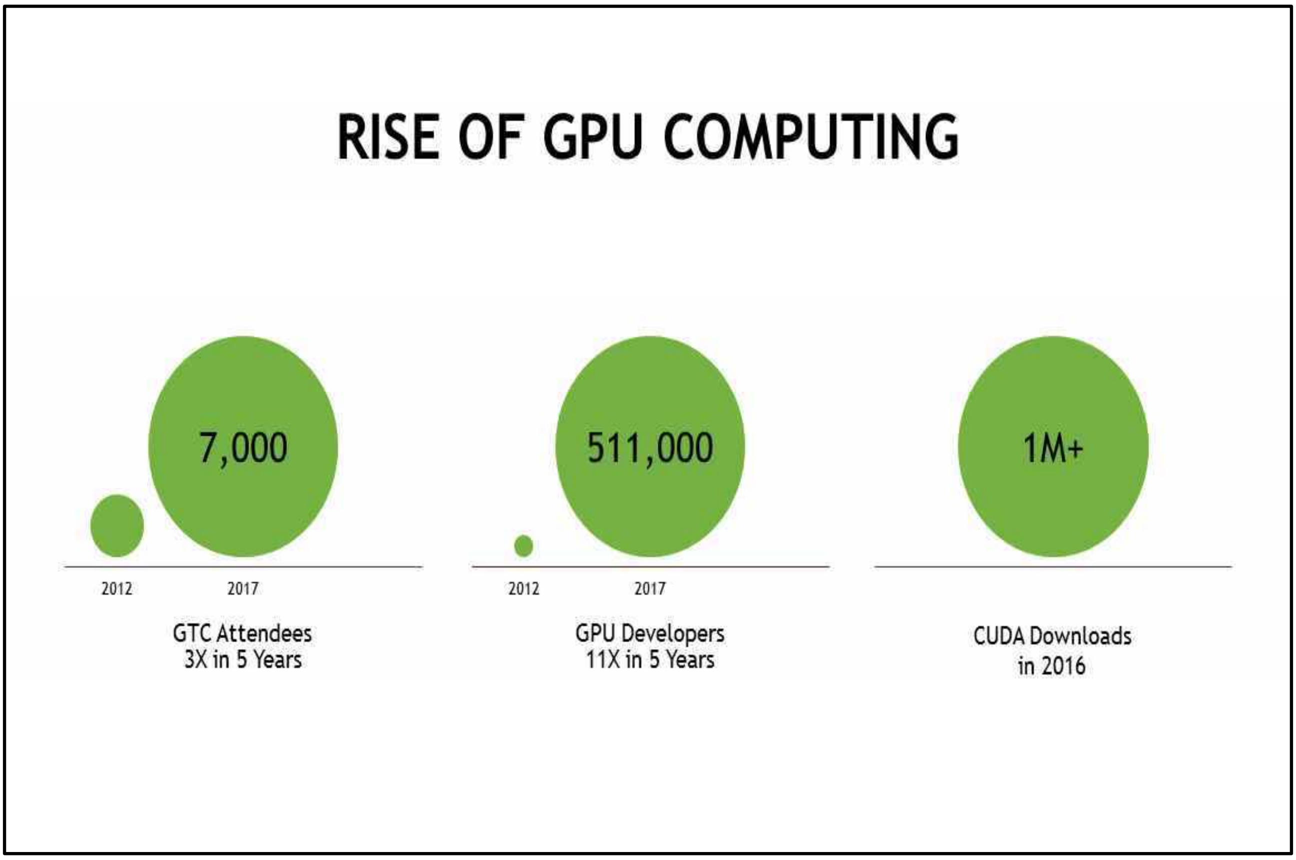

ERA of GPU Computing in AI/HPC
서완석상무, wseo@nvidia.com
NVIDIA

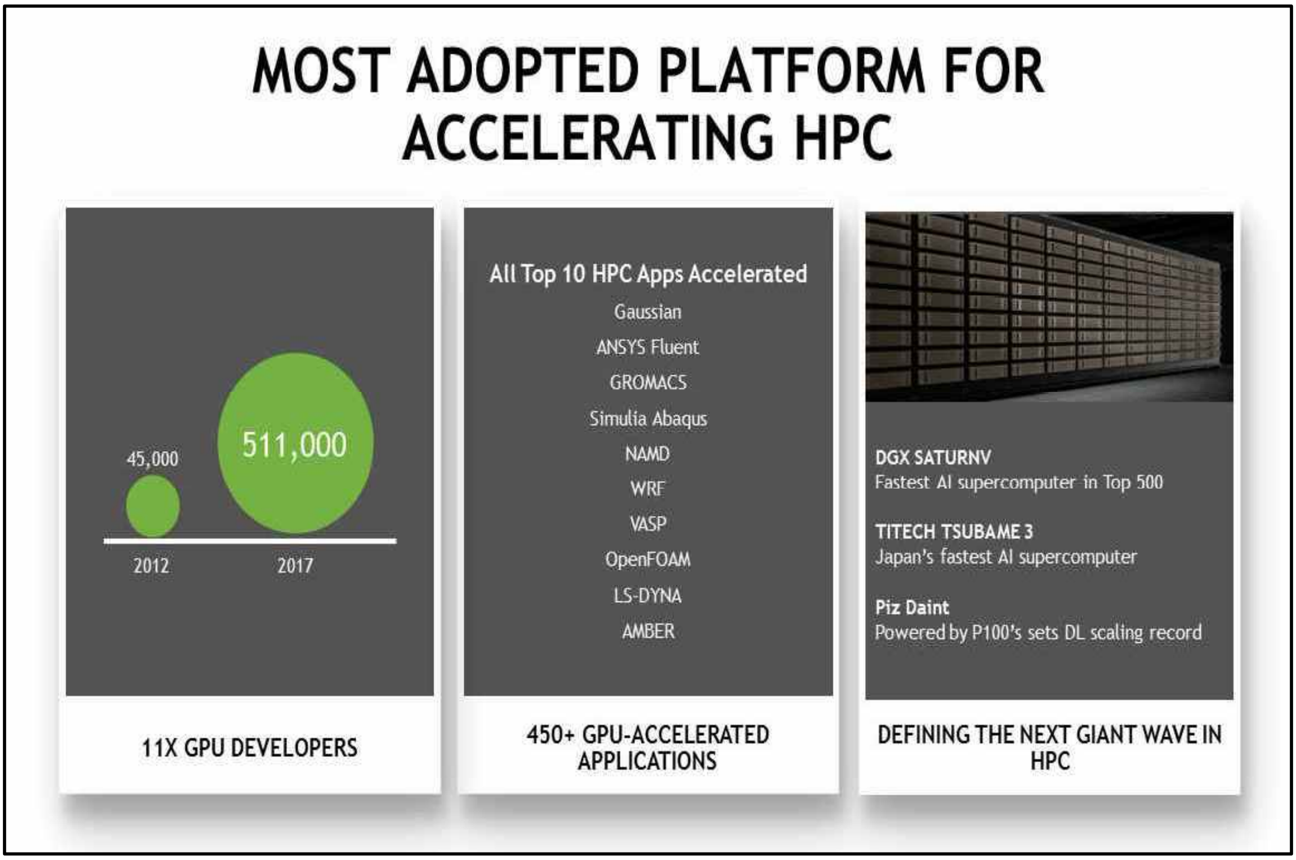
LIFE AFTER MOORE'S LAW



RISE OF GPU COMPUTING




RISE OF GPU COMPUTING
7,000
2012
2017
GTC Attendees
3X in 5 Years
511,000
2012
2017
GPU Developers
11X in 5 Years
1M+
CUDA Downloads
in 2016


MOST ADOPTED PLATFORM FOR ACCELERATING HPC
45,000
511,000
2012
2017
11X GPU DEVELOPERS
All Top 10 HPC Apps Accelerated
Gaussian
ANSYS Fluent
GROMACS
Simulia Abaqus
NAMD
WRF
VASP
OpenFOAM
LS-DYNA
AMBER
450+ GPU-ACCELERATED APPLICATIONS
DGX SATURNV
Fastest AI supercomputer in Top 500
TITECH TSUBAME 3
Japan's fastest AI supercomputer
Piz Daint
Powered by P100's sets DL scaling record
DEFINING THE NEXT GIANT WAVE IN HPC

SUPERCOMPUTERS DESIGNED FOR DATA SCIENCE AND COMPUTATION SCIENCE

THE GREEN 500

Tsubame 3

#1 on Green500, June 2017

TSUBAME 3.0 delivering 47.2 Petaflops at half-precision set for AI Computing using 2,160 P100 SMX

Tokyo Tech's Tsubame 3 will be AI/HPC hybrid

20 February 2017 | By Peter Judge — DatacenterDynamics, The Business of Data Centers

JAPAN KEEPS ACCELERATING WITH TSUBAME 3.0 AI SUPERCOMPUTER

February 17, 2017 Timothy Prickett Morgan — THENEXTPLATFORM

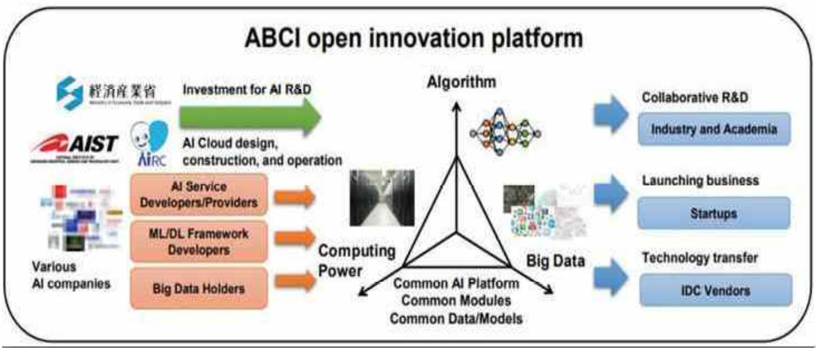
Next-Generation TSUBAME Will Be Petascale Supercomputer for AI

Michael Feldman | February 16, 2017 00:04 CET — TOP500

"NVIDIA's broad AI ecosystem will enable Tokyo Tech to begin training TSUBAME3.0 immediately to help us more quickly solve some of the world's once unsolvable problems."

- Satoshi Matsuoka, Prof Computer Science, TiTech & Project lead Tsubame 3

ABCI(AI BRIDGING CLOUD INFRASTRUCTURE)

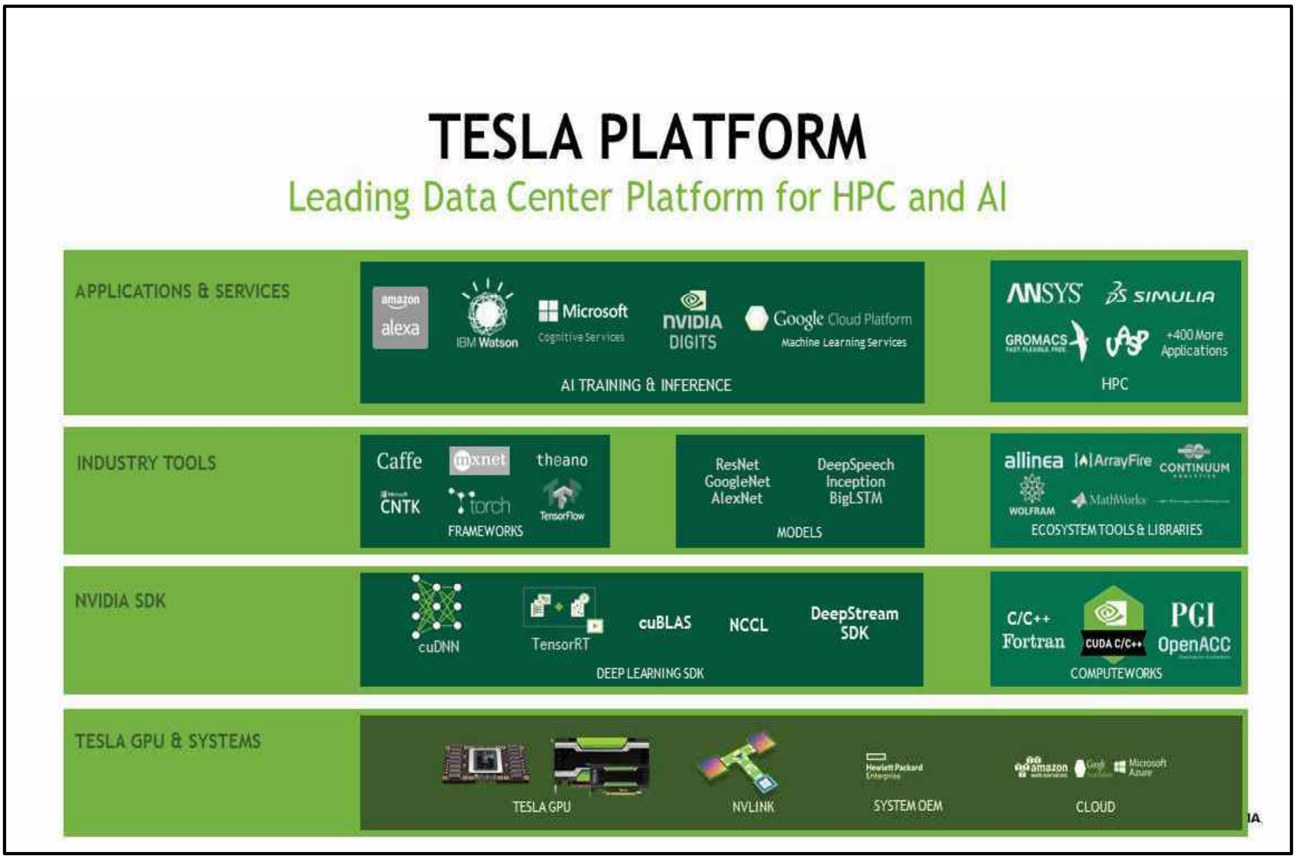


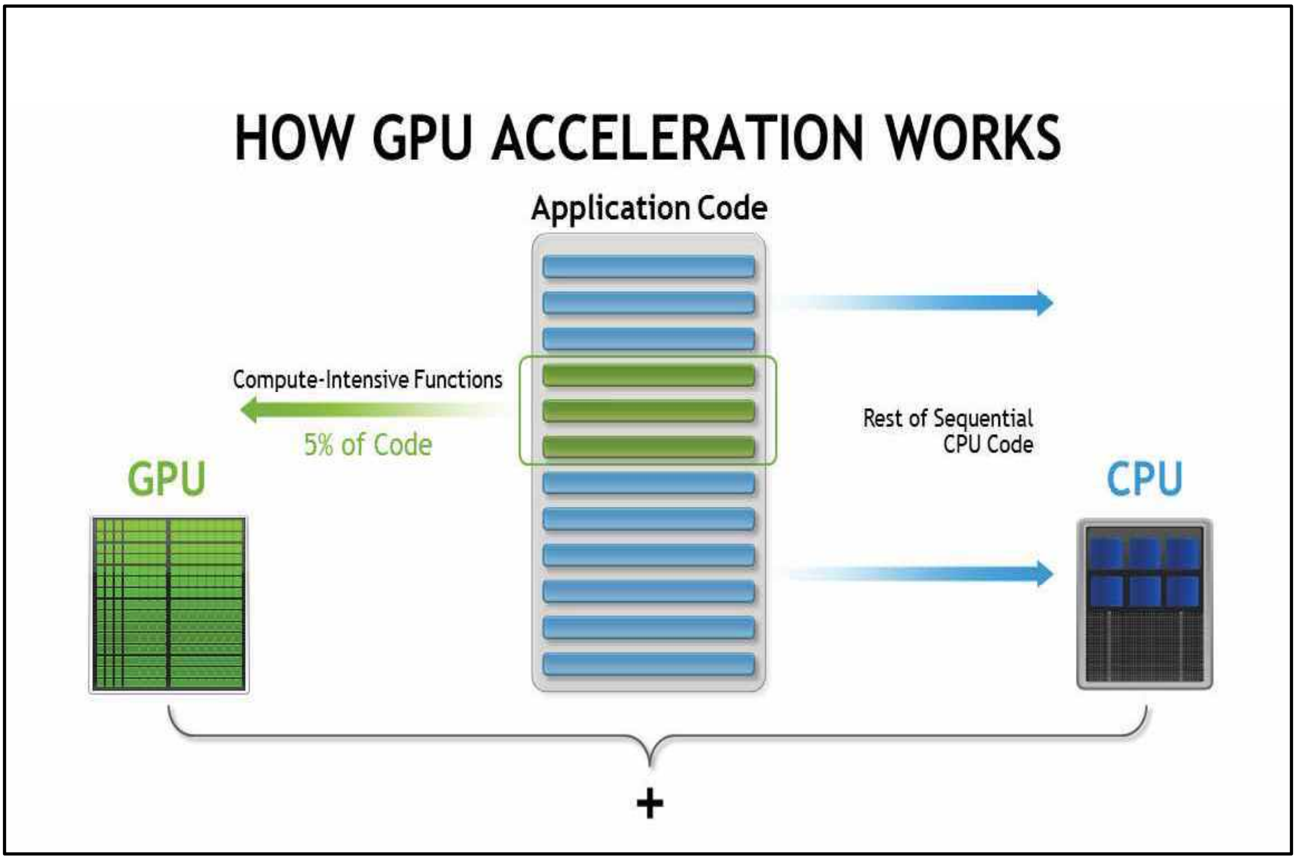
National Institute of Advanced Industrial Science and Technology (AIST) in Japan will offer cloud access to compute and storage capacity for artificial intelligence and data analytics workloads

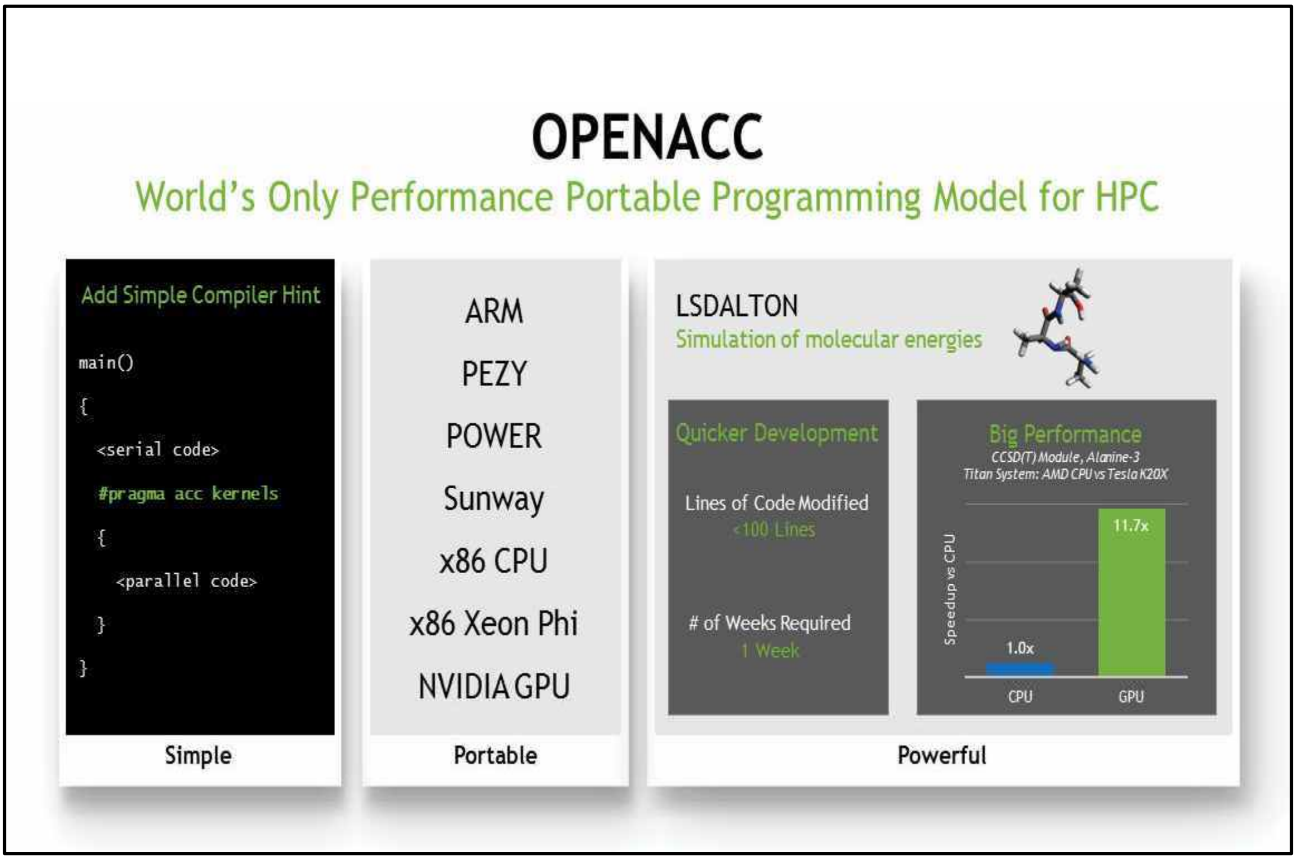
4352x Tesla V100

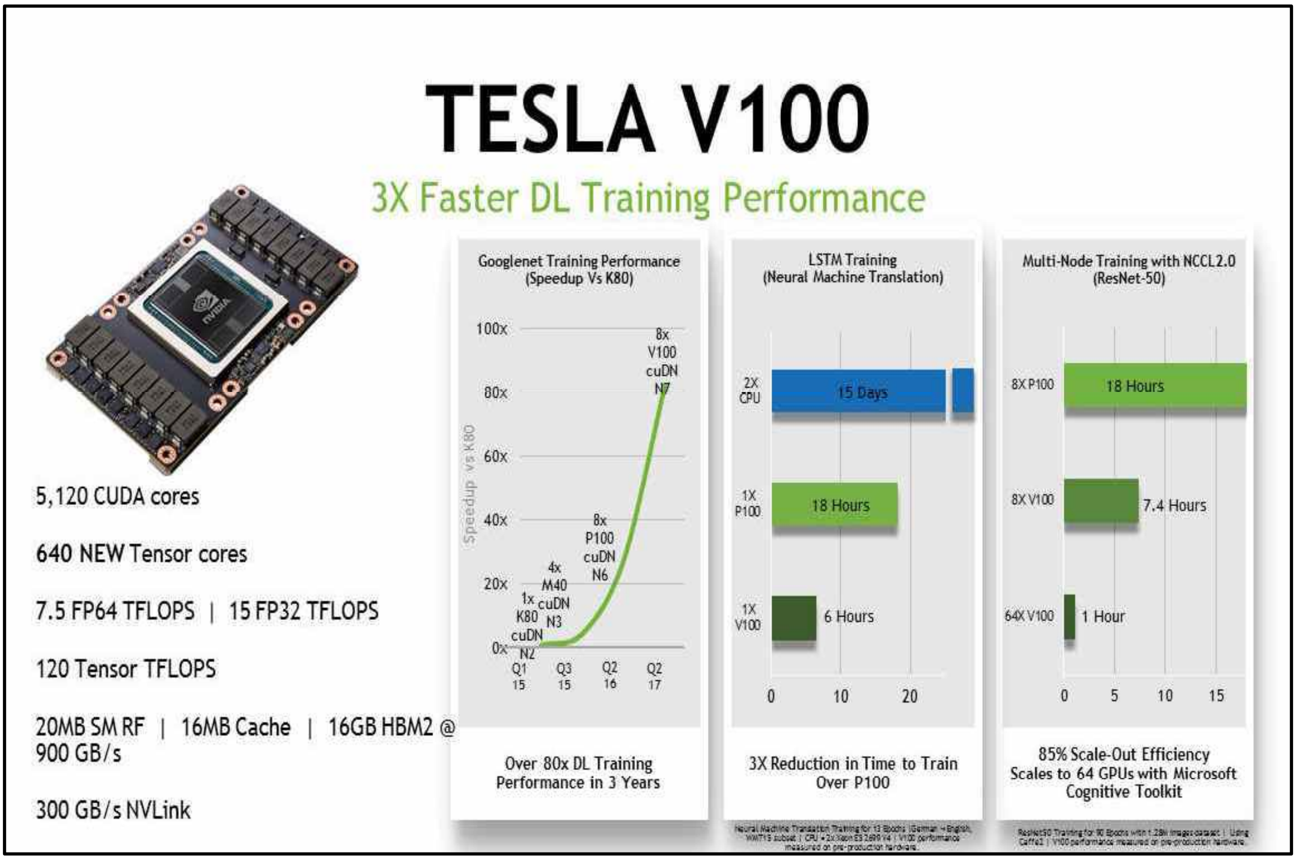
550PFLOSP with half precision (TensorCore + CPU)

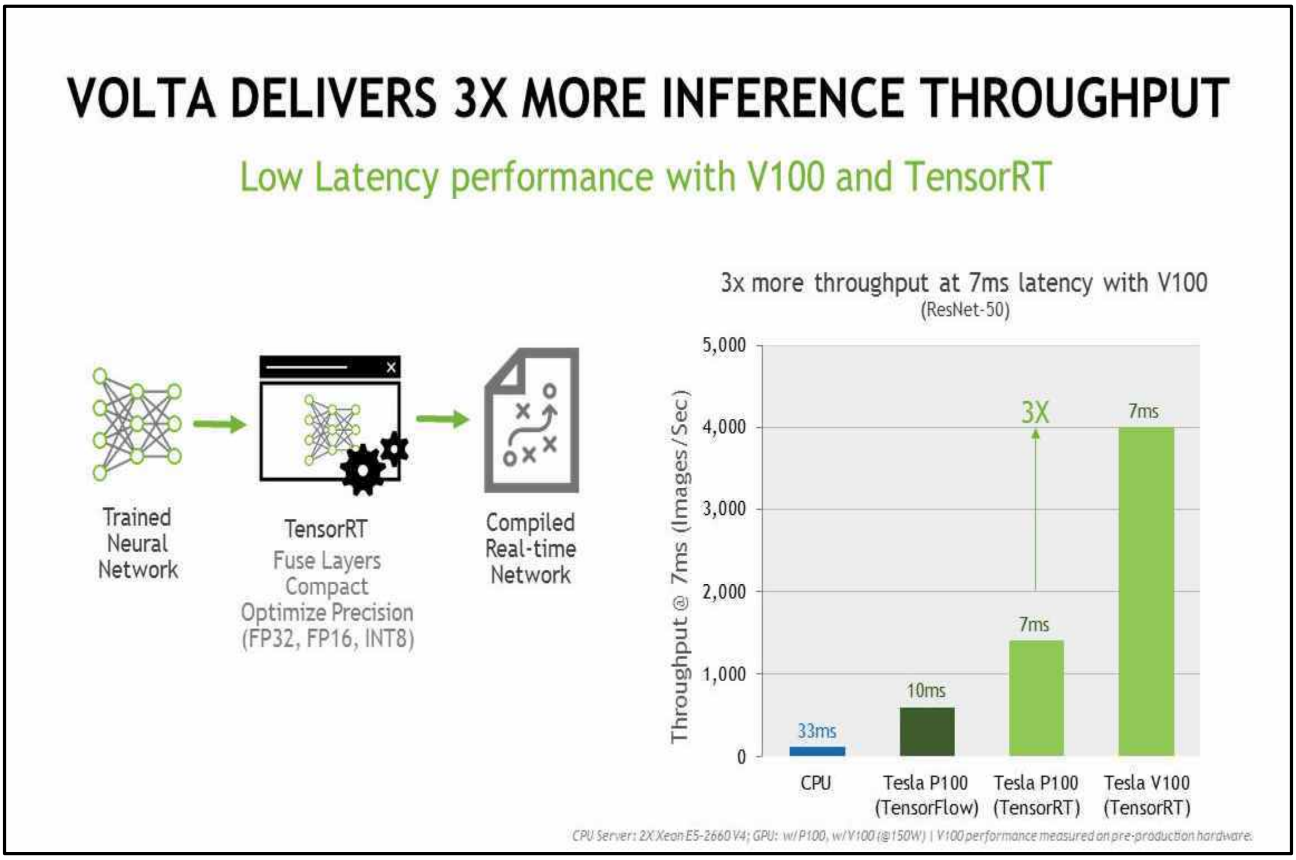
37PFLOPS with Double precision

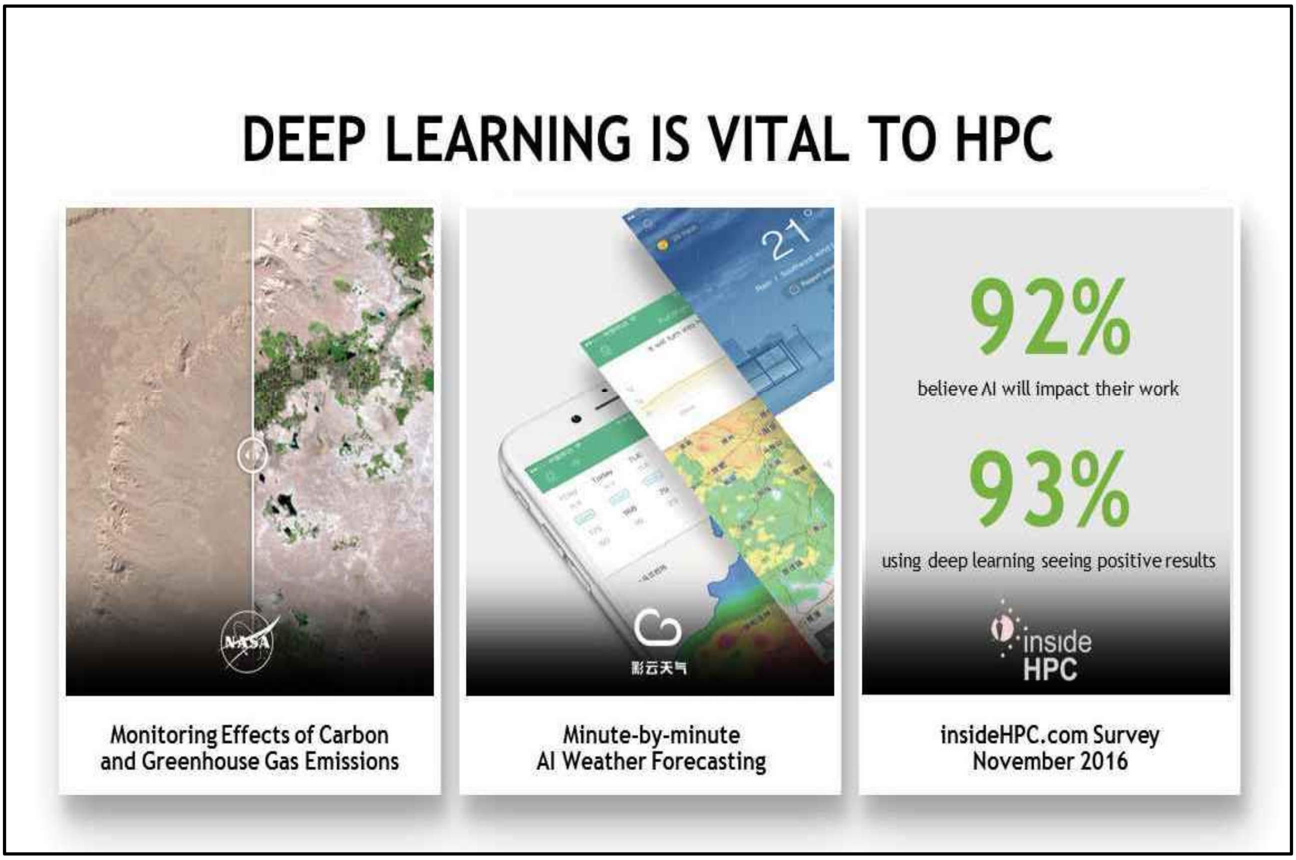

TESLA PLATFORM
Leading Data Center Platform for HPC and AI
APPLICATIONS & SERVICES
amazon alexa
IBM Watson
Microsoft Cognitive Services
NVIDIA DIGITS
Google Cloud Platform Machine Learning Services
AI TRAINING & INFERENCE
ANSYS
SIMULIA
GROMACS
VASP
+400 More Applications
HPC
INDUSTRY TOOLS
Caffe
mxnet
theano
CNTK
torch
TensorFlow
FRAMEWORKS
ResNet GoogleNet AlexNet
DeepSpeech Inception BigLSTM
MODELS
allinea
ArrayFire
CONTINUUM
WOLFRAM
MathWorks
ECOSYSTEM TOOLS & LIBRARIES
NVIDIA SDK
cuDNN
TensorRT
cuBLAS
NCCL
DeepStream SDK
DEEP LEARNING SDK
C/C++
Fortran
CUDA C/C++
PGI
OpenACC
COMPUTEWORKS
TESLA GPU & SYSTEMS
TESLA GPU
NVLINK
Hewlett Packard Enterprise
SYSTEM OEM
amazon
Microsoft Azure
CLOUD

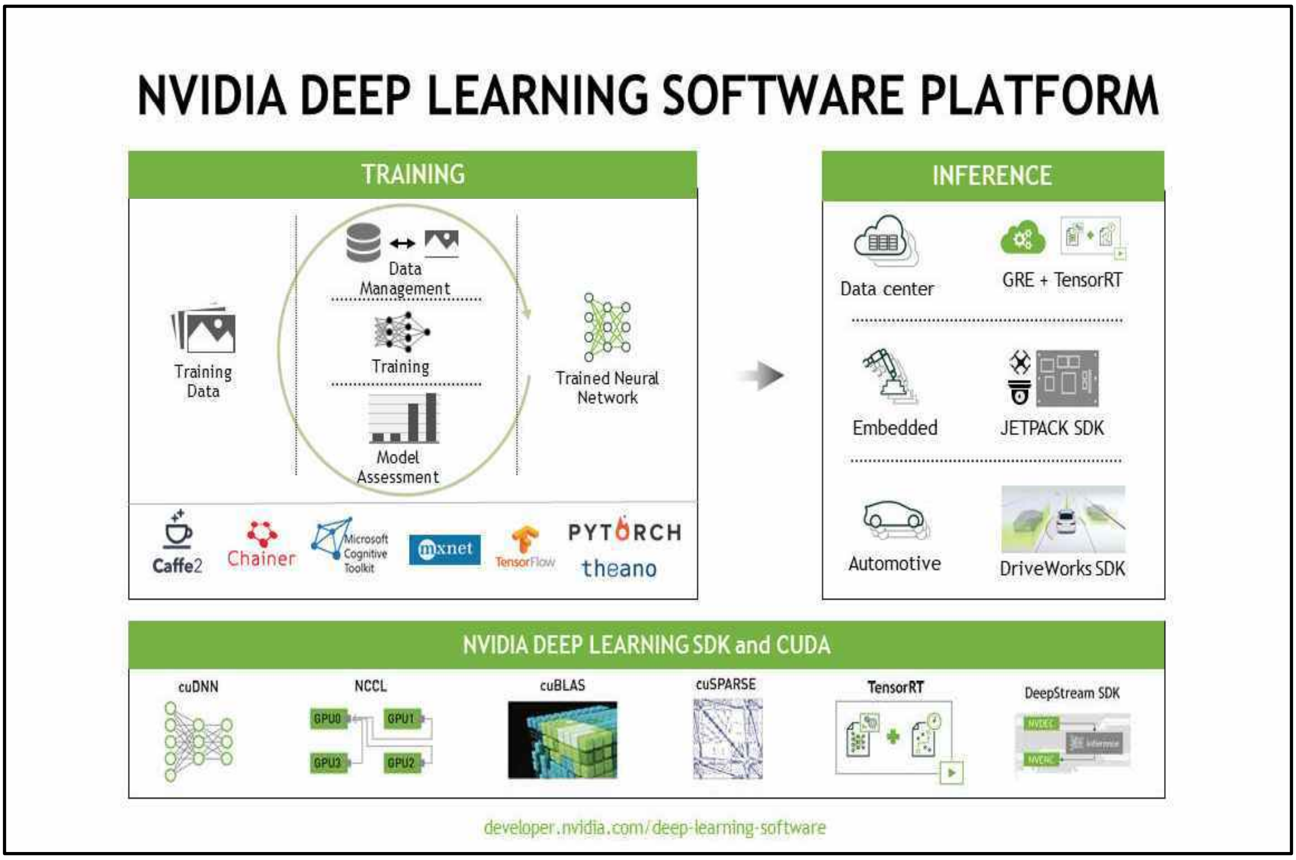

HOW GPU ACCELERATION WORKS
Application Code
Compute-Intensive Functions
5% of Code
Rest of Sequential CPU Code
GPU
CPU
+

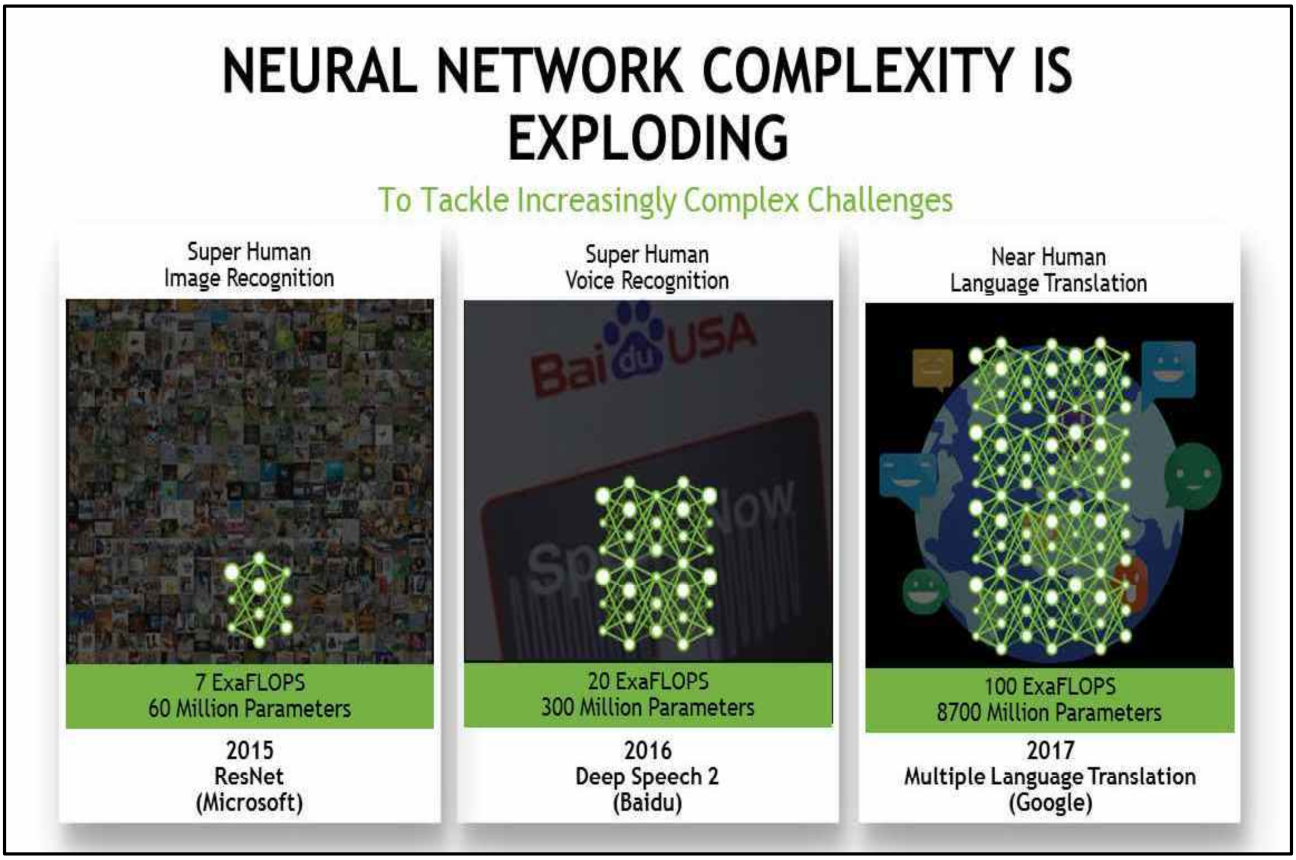

OPENACC
World's Only Performance Portable Programming Model for HPC
Add Simple Compiler Hint
main()
{
<serial code>
#pragma acc kernels
{
<parallel code>
}
}
Simple
ARM
PEZY
POWER
Sunway
x86 CPU
x86 Xeon Phi
NVIDIA GPU
Portable
LSDALTON
Simulation of molecular energies
Quicker Development
Lines of Code Modified
<100 Lines
# of Weeks Required
1 Week
Big Performance
CCSD(T) Module, Alanine-3
Titan System: AMD CPU vs Tesla K20X
Speedup vs CPU
1.0x
11.7x
CPU
GPU
Powerful

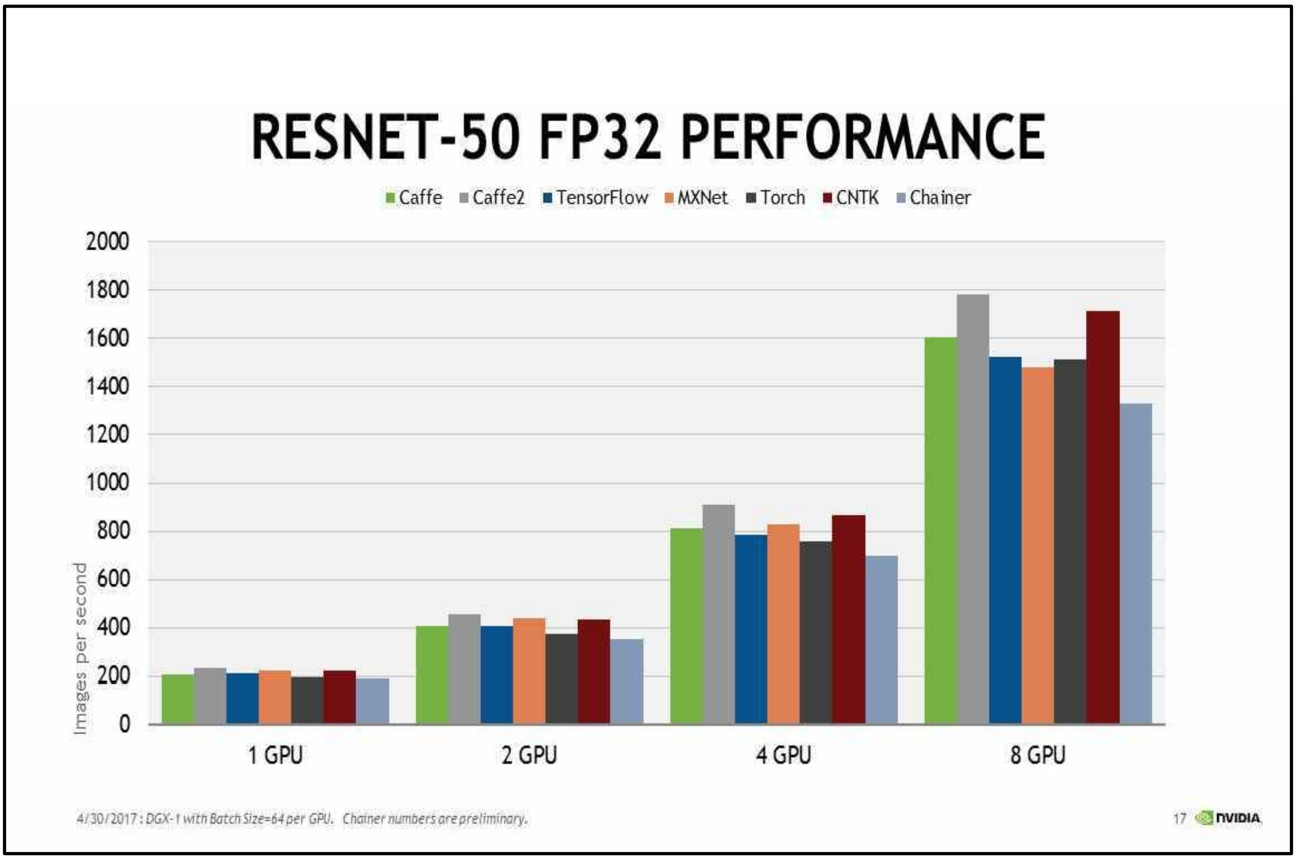

TESLA V100
3X Faster DL Training Performance
5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
120 Tensor TFLOPS
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink
Googlenet Training Performance (Speedup Vs K80)
Speedup vs K80
100x
80x
60x
40x
20x
0x
1x K80 cuDNN2
4x M40 cuDNN3
8x P100 cuDNN6
8x V100 cuDNN7
Q1 15
Q3 15
Q2 16
Q2 17
Over 80x DL Training Performance in 3 Years
LSTM Training (Neural Machine Translation)
2X CPU
15 Days
1X P100
18 Hours
1X V100
6 Hours
0
10
20
3X Reduction in Time to Train Over P100
Multi-Node Training with NCCL2.0 (ResNet-50)
8X P100
18 Hours
8X V100
7.4 Hours
64X V100
1 Hour
0
5
10
15
85% Scale-Out Efficiency
Scales to 64 GPUs with Microsoft Cognitive Toolkit

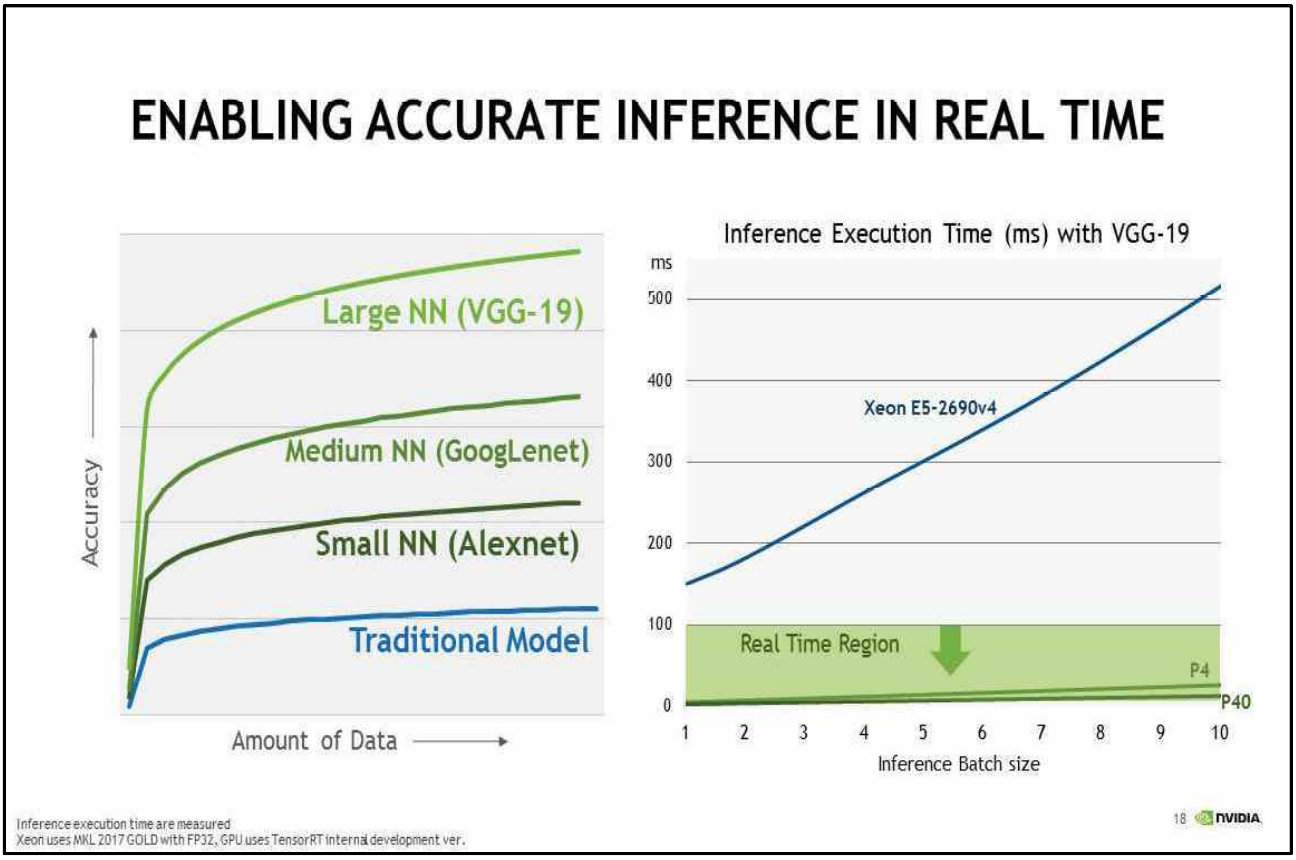

VOLTA DELIVERS 3X MORE INFERENCE THROUGHPUT
Low Latency performance with V100 and TensorRT
Trained Neural Network
TensorRT
Fuse Layers
Compact
Optimize Precision
(FP32, FP16, INT8)
Compiled Real-time Network
3x more throughput at 7ms latency with V100
(ResNet-50)
Throughput @ 7ms (Images/Sec)
5,000
4,000
3,000
2,000
1,000
0
3X
33ms
10ms
7ms
7ms
CPU
Tesla P100 (TensorFlow)
Tesla P100 (TensorRT)
Tesla V100 (TensorRT)
CPU Server: 2X Xeon E5-2660 V4; GPU: w/ P100, w/ V100 (@150W) | V100 performance measured on pre-production hardware.


DEEP LEARNING IS VITAL TO HPC
NASA
Monitoring Effects of Carbon and Greenhouse Gas Emissions
彩云天气
Minute-by-minute AI Weather Forecasting
92%
believe AI will impact their work
93%
using deep learning seeing positive results
inside HPC
insideHPC.com Survey November 2016


NVIDIA DEEP LEARNING SOFTWARE PLATFORM
TRAINING
Training Data
Data Management
Training
Model Assessment
Trained Neural Network
Caffe2
Chainer
Microsoft Cognitive Toolkit
mxnet
TensorFlow
PYTORCH
theano
INFERENCE
Data center
GRE + TensorRT
Embedded
JETPACK SDK
Automotive
DriveWorks SDK
NVIDIA DEEP LEARNING SDK and CUDA
cuDNN
NCCL
GPU0
GPU1
GPU3
GPU2
cuBLAS
cuSPARSE
TensorRT
DeepStream SDK
developer.nvidia.com/deep-learning-software

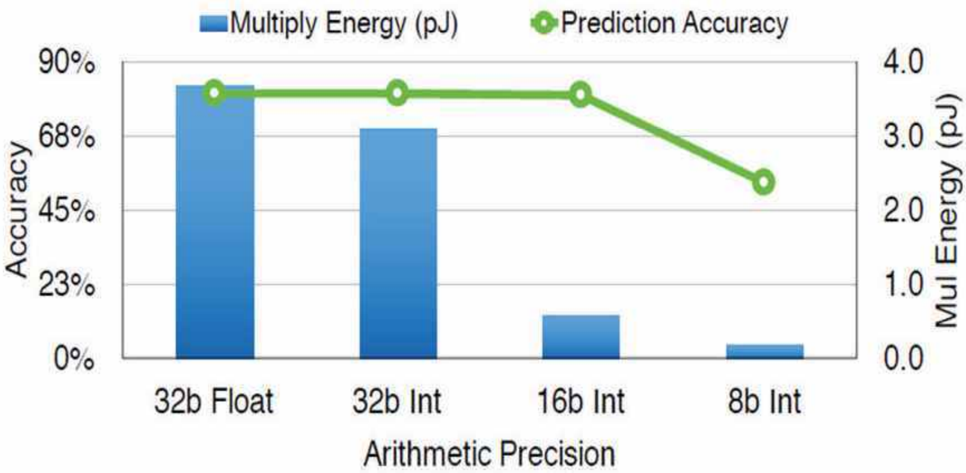

NEURAL NETWORK COMPLEXITY IS EXPLODING
To Tackle Increasingly Complex Challenges
Super Human Image Recognition
7 ExaFLOPS
60 Million Parameters
2015
ResNet
(Microsoft)
Super Human Voice Recognition
Baidu USA
20 ExaFLOPS
300 Million Parameters
2016
Deep Speech 2
(Baidu)
Near Human Language Translation
100 ExaFLOPS
8700 Million Parameters
2017
Multiple Language Translation
(Google)


RESNET-50 FP32 PERFORMANCE
Caffe
Caffe2
TensorFlow
MXNet
Torch
CNTK
Chainer
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Images per second
1 GPU
2 GPU
4 GPU
8 GPU
4/30/2017 : DGX-1 with Batch Size=64 per GPU. Chainer numbers are preliminary.
17
NVIDIA


ENABLING ACCURATE INFERENCE IN REAL TIME
Large NN (VGG-19)
Medium NN (GoogLenet)
Small NN (Alexnet)
Traditional Model
Accuracy
Amount of Data
Inference Execution Time (ms) with VGG-19
ms
500
400
300
200
100
0
Xeon E5-2690v4
Real Time Region
P4
P40
1 2 3 4 5 6 7 8 9 10
Inference Batch size
Inference execution time are measured
Xeon uses MKL 2017 GOLD with FP32, GPU uses TensorRT internal development ver.
18
NVIDIA

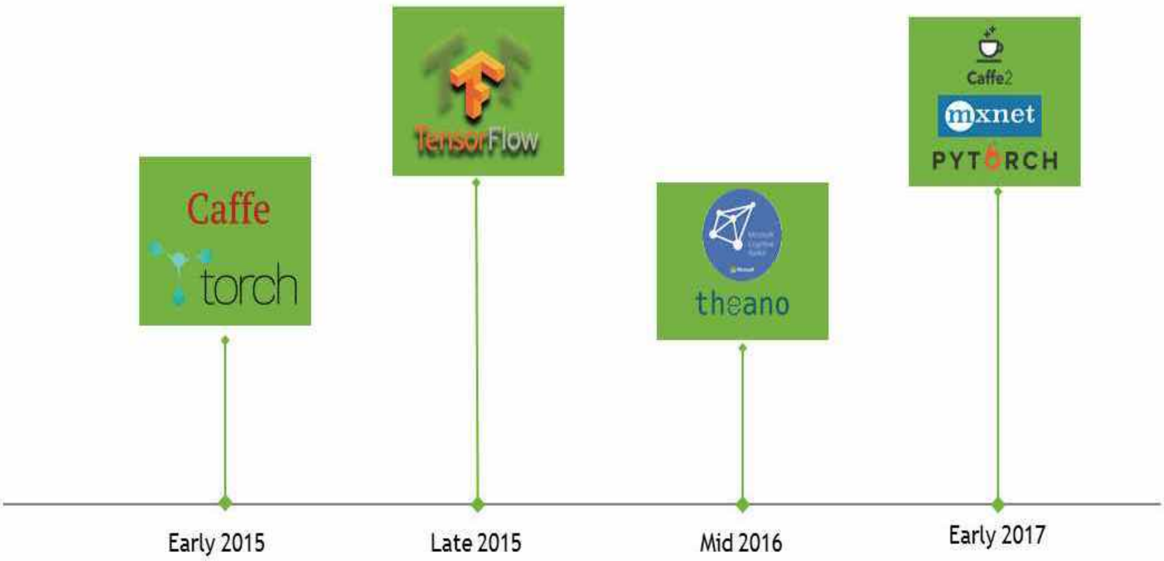
REDUCED PRECISION FOR INFERENCE

16 bits is a good spot

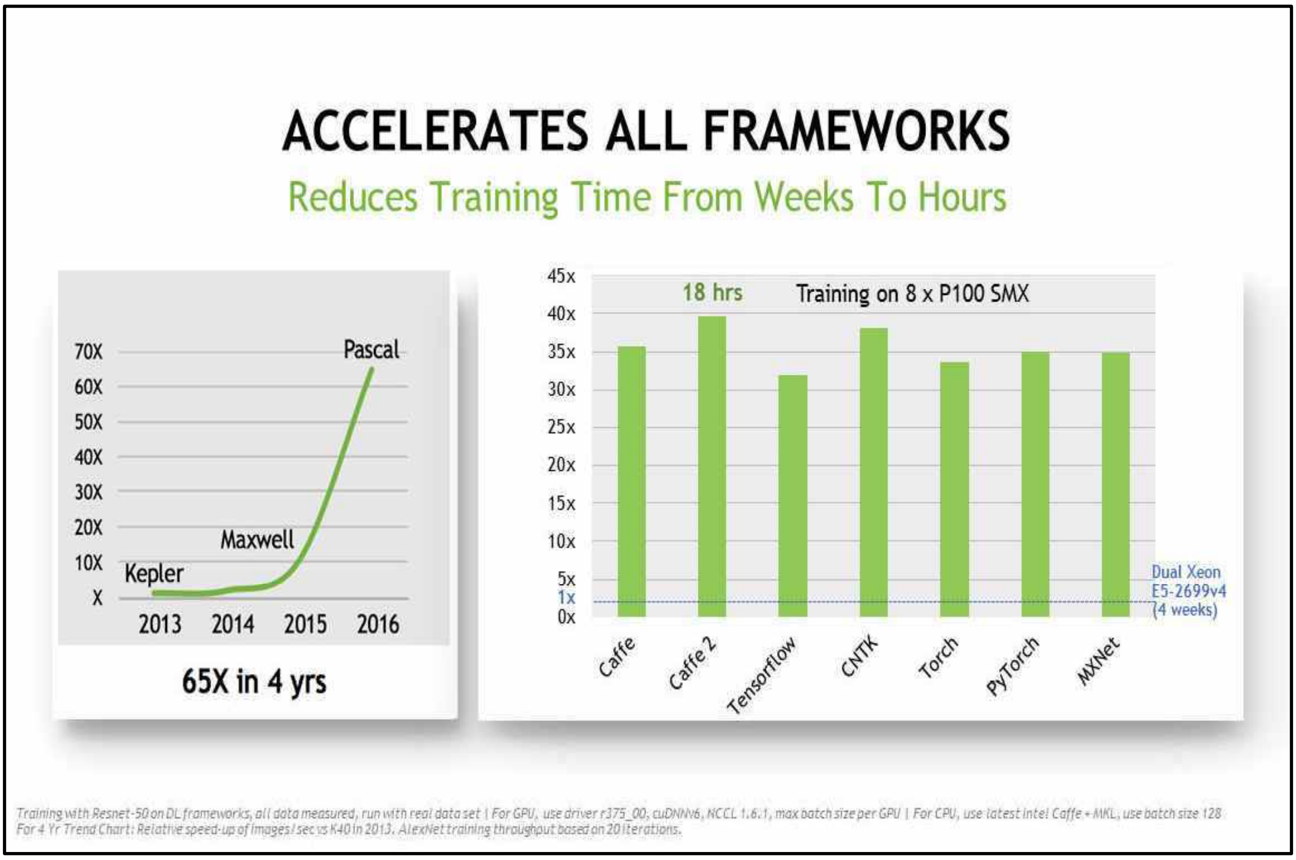


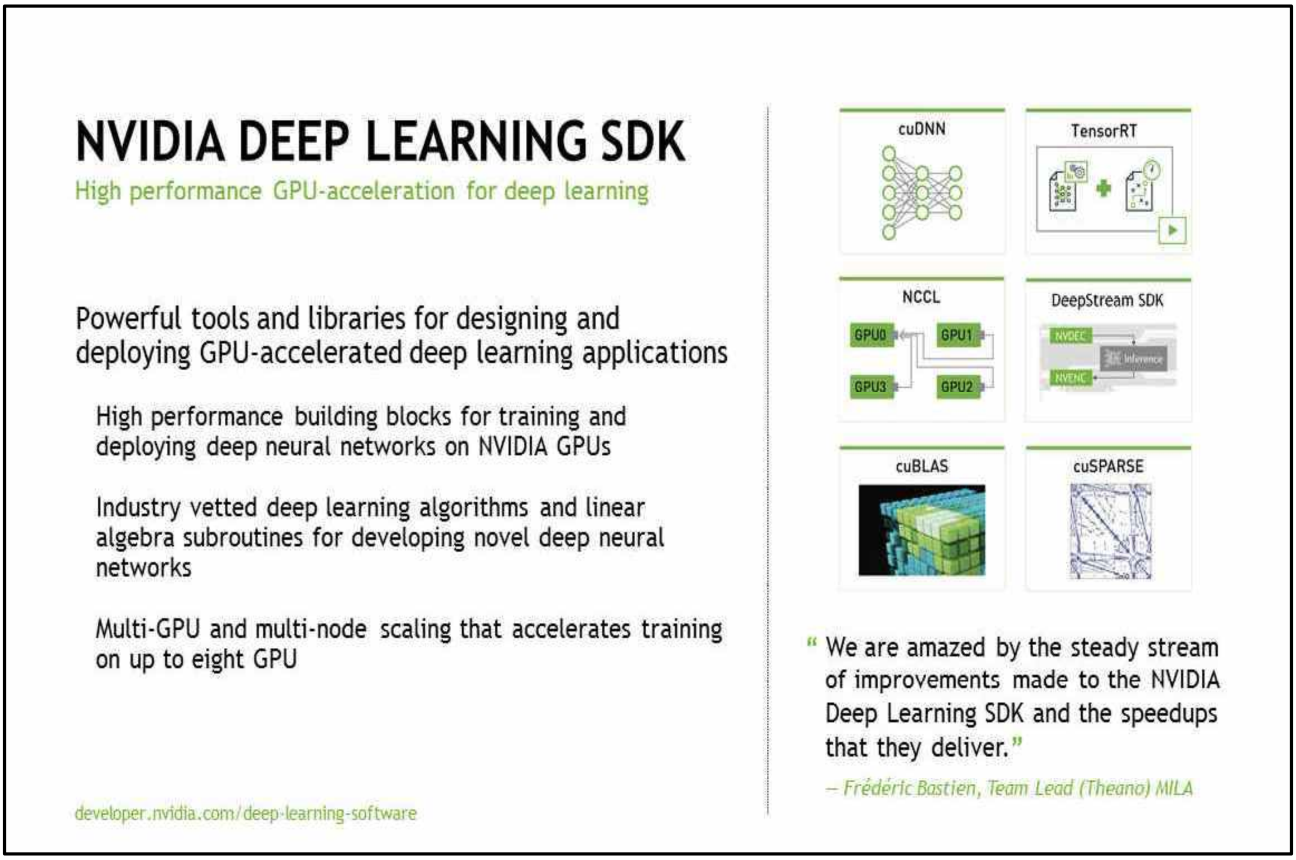
19 NVIDIA

NVIDIA PARTICIPATION IN FRAMEWORKS



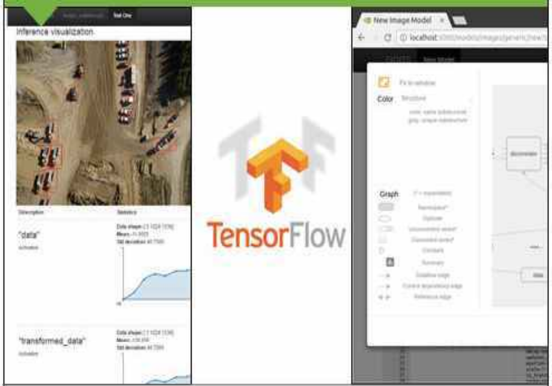
20 NVIDIA


ACCELERATES ALL FRAMEWORKS
Reduces Training Time From Weeks To Hours
70X
60X
50X
40X
30X
20X
10X
X
Pascal
Maxwell
Kepler
2013
2014
2015
2016
65X in 4 yrs
45x
40x
35x
30x
25x
20x
15x
10x
5x
1x
0x
18 hrs
Training on 8 x P100 SMX
Dual Xeon
E5-2699v4
(4 weeks)
Caffe
Caffe 2
Tensorflow
CNTK
Torch
PyTorch
MXNet
Training with Resnet-50 on DL frameworks, all data measured, run with real data set | For GPU, use driver r375_00, cuDNNv6, NCCL 1.6.1, max batch size per GPU | For CPU, use latest Intel Caffe + MKL, use batch size 128
For 4 Yr Trend Chart: Relative speed-up of images/sec vs K40 in 2013. AlexNet training throughput based on 20 iterations.


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 and deploying deep neural networks on NVIDIA GPUs
Industry vetted deep learning algorithms and linear algebra subroutines for developing novel deep neural networks
Multi-GPU and multi-node scaling that accelerates training on up to eight GPU
developer.nvidia.com/deep-learning-software
cuDNN
TensorRT
NCCL
GPU0
GPU1
GPU3
GPU2
DeepStream SDK
cuBLAS
cuSPARSE
“ 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

WHAT'S NEW IN DIGITS 6?

TENSORFLOW SUPPORT



Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS

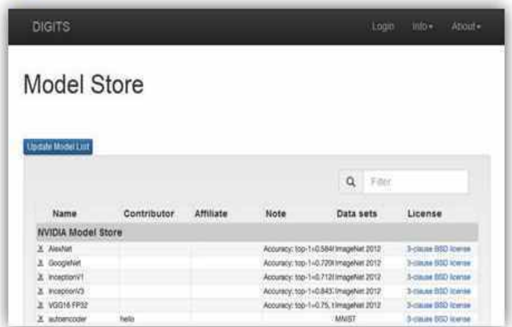
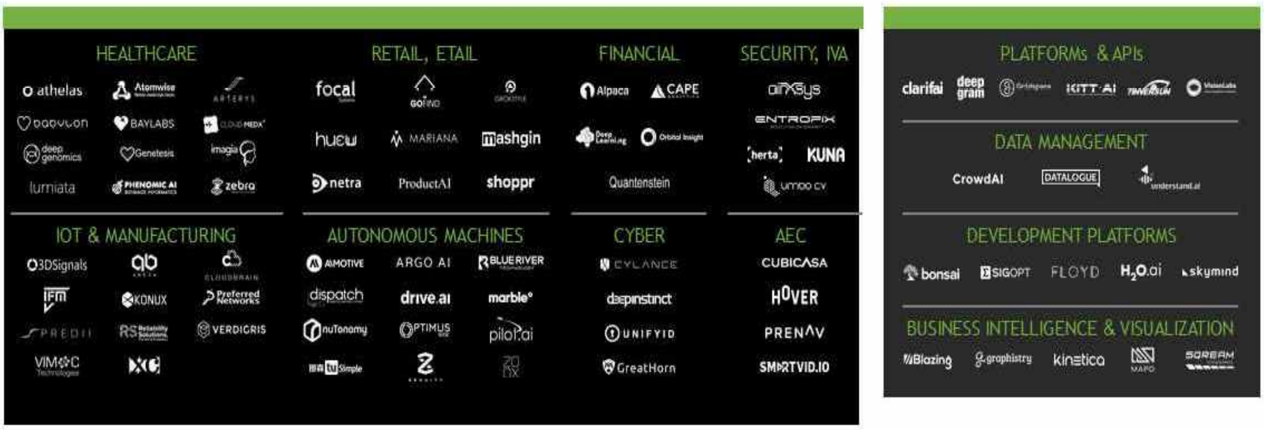


Image Classification: VGG-16, ResNet50
Object Detection: DetectNet

DIGITS 6 Release Candidate available now on Docker Hub for testing and feedback
General availability in September

NVIDIA INCEPTION — 1,300 DEEP LEARNING STARTUPS



NVIDIA
DEEP LEARNING APPLIANCE
DGX

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube

From 8 days on TITAN X to 8 hours

400 servers in a box

PERSONAL DGX

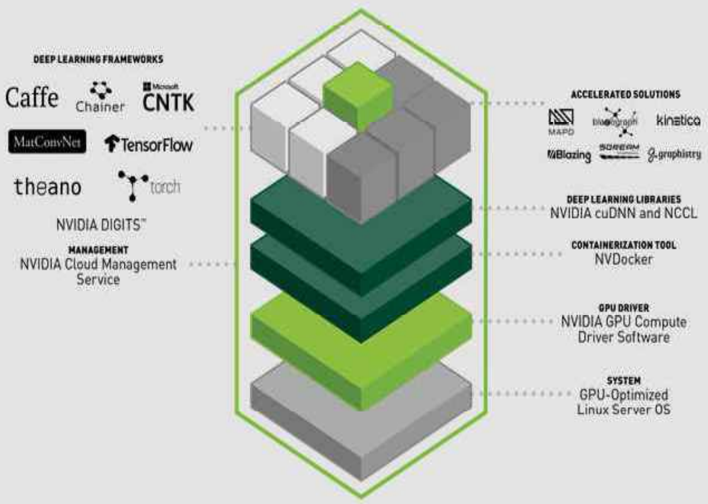
480 Tensor TFLOPS | 4x Tesla V100 16GB

NVLink Fully Connected | 3x DisplayPort

1500W | Water Cooled

DGX STACK

Fully integrated 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

DLI LABS MAY 2017

Industry	Lab Title	Level	Framework	Available Self-Paced Online	Available Onsite in Instructor-Led Workshops
Deep Learning Fundamentals	Applications of Deep Learning with Caffe, Theano and Torch	Beginner	Caffe, Theano, Torch	•	•
	Image Classification with DIGITS	Beginner	Caffe	•	•
	Object Detection with DIGITS	Beginner	Caffe	•	•
	Neural Network Deployment with DIGITS and TensorRT	Intermediate	Caffe	•	•
	Signal Processing with NVIDIA DIGITS	Beginner	Caffe	•	•
	Linear Classification with Tensorflow	Beginner	TensorFlow	•	•
	Image Segmentation with TensorFlow	Beginner	TensorFlow	•	•
	Deep Learning Workflows with TensorFlow, MXNet and NVIDIA-Docker	Beginner	TensorFlow	•	•
	Image Classification with Caffe2	Beginner	Caffe2		•
	Image Classification with Microsoft Cognitive Toolkit	Beginner	Cognitive Toolkit		•
Autonomous Vehicles	Introduction and Integration with DriveWorks on Drive PX2	Beginner	Caffe		•
	From Trained Neural Network Model to Deployment for Inference	Beginner	Caffe		•
Healthcare	Medical Image Segmentation with DIGITS	Beginner	Caffe	•	•
	Medical Image Analysis with R and MXNet	Intermediate	MXNet	•	•
	Image Classification with TensorFlow: Radiomics – 1p19q Chromosone Status Classification	Intermediate	TensorFlow	•	•
	Modeling Time Series Data with Recurrent Neural Networks in Keras	Intermediate	Theano	•	•
	Deep Learning for Genomics using DragoNN with Keras and Theano	Advanced	Theano	•	•
Robotics	Image Classification and Object Detection with NVIDIA Jetson TX2	Beginner	Caffe		•
Media & Entertainment	Deep Learning for Image and Video Captioning	Intermediate	TensorFlow		•
	Photo Editing with Generative Adversarial Networks with Tensorflow and DIGITS	Intermediate	Caffe, TensorFlow		•
Game Development	Deep Reinforcement Learning Agents on Atari 2600 Games	Beginner	TensorFlow		•

30 NVIDIA


DEEP LEARNING DAY
AGENDA
REGISTER NOW
CONTACT US
NVIDIA DEEP LEARNING DAY 2017
SEOUL / Oct 31 - Nov 2
지금 등록하기
https://www.nvidia.com/ko-kr/deep-learning-day/


감사합니다
NVIDIA