TOWARDS ACCELERATED DEEP LEARNING IN HPC AND HYPERSCALE ARCHITECTURES

Environnement logiciel pour l'apprentissage profond dans un contexte HPC

TERATECH Juin 2017

Gunter Roth, François Courteille

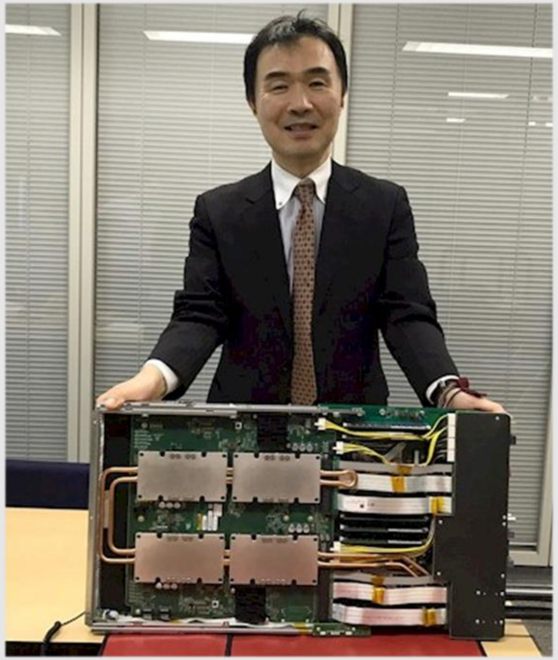
NVIDIA

SUPERCOMPUTERS DESIGNED FOR AI SUPERCOMPUTING

Tsubame 3

#1 Green500 System

Powered by 2160 P100s



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

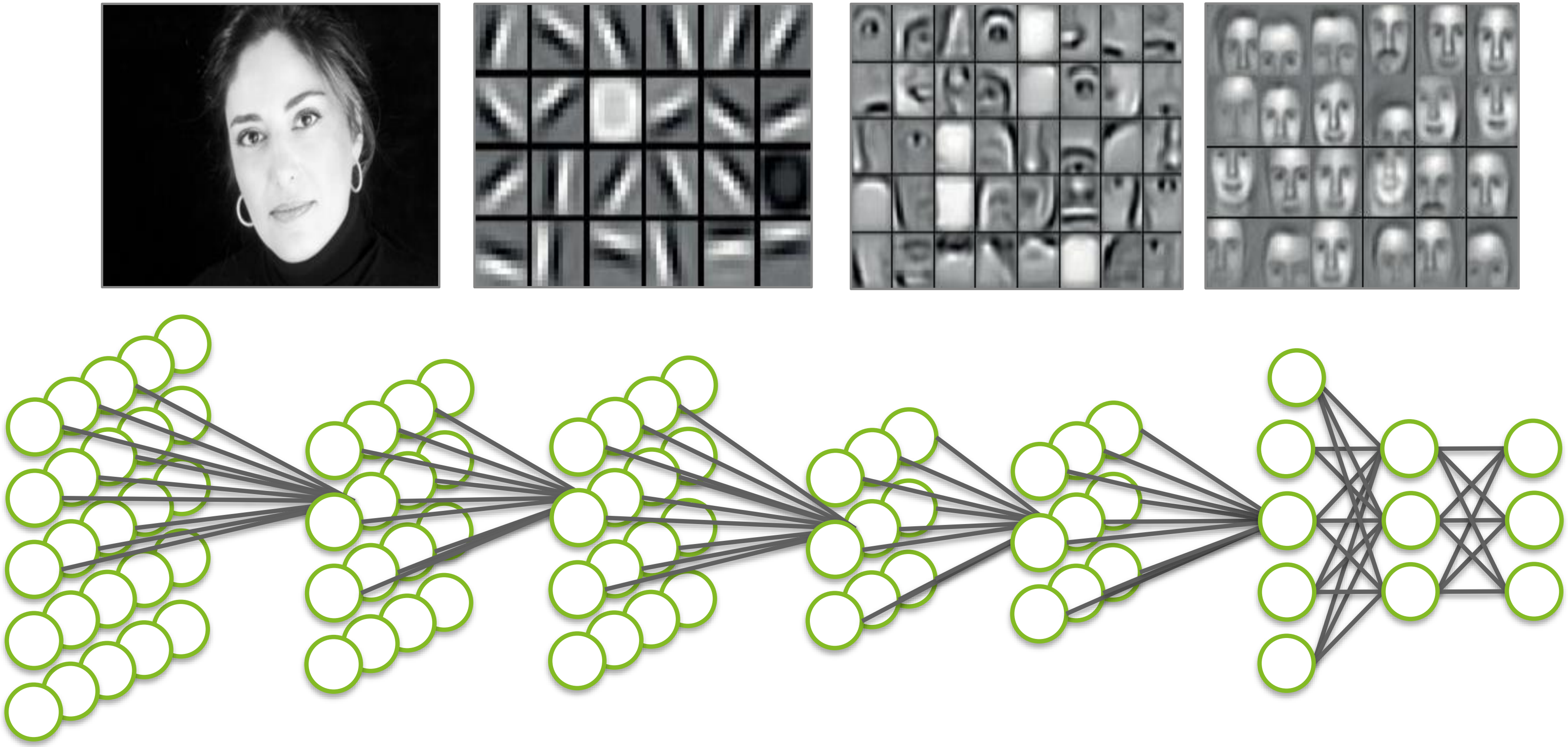
Next-Generation TSUBAME Will Be Petascale Supercomputer for AI

Michael Feldman | February 18, 2017 00:04 CET — TOP500 The List.

"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

WHAT IS DEEP LEARNING?



Typical Network

Task objective
e.g. identify face
Training data
10-100M images
Network architecture
10 layers
1B parameters
Learning algorithm
~30 exaflops
~30 GPU days

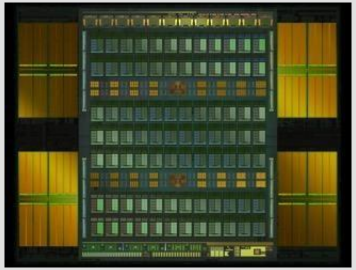
Image classification

Training AlexNet [~60 Millions parameters] requires ~27,000 flops/input data byte

Training VGG [~138 Millions parameters] requires ~150,000 flops/input data byte

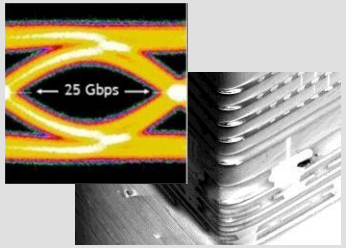
INTRODUCING TESLA V100

Volta Architecture



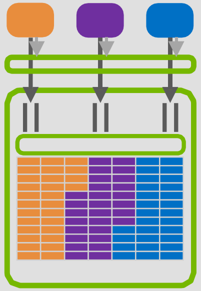
Most Productive GPU

Improved NVLink & HBM2



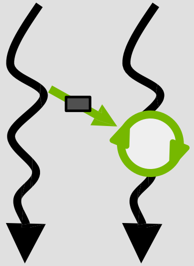
Efficient Bandwidth

Volta MPS



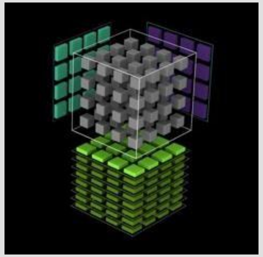
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



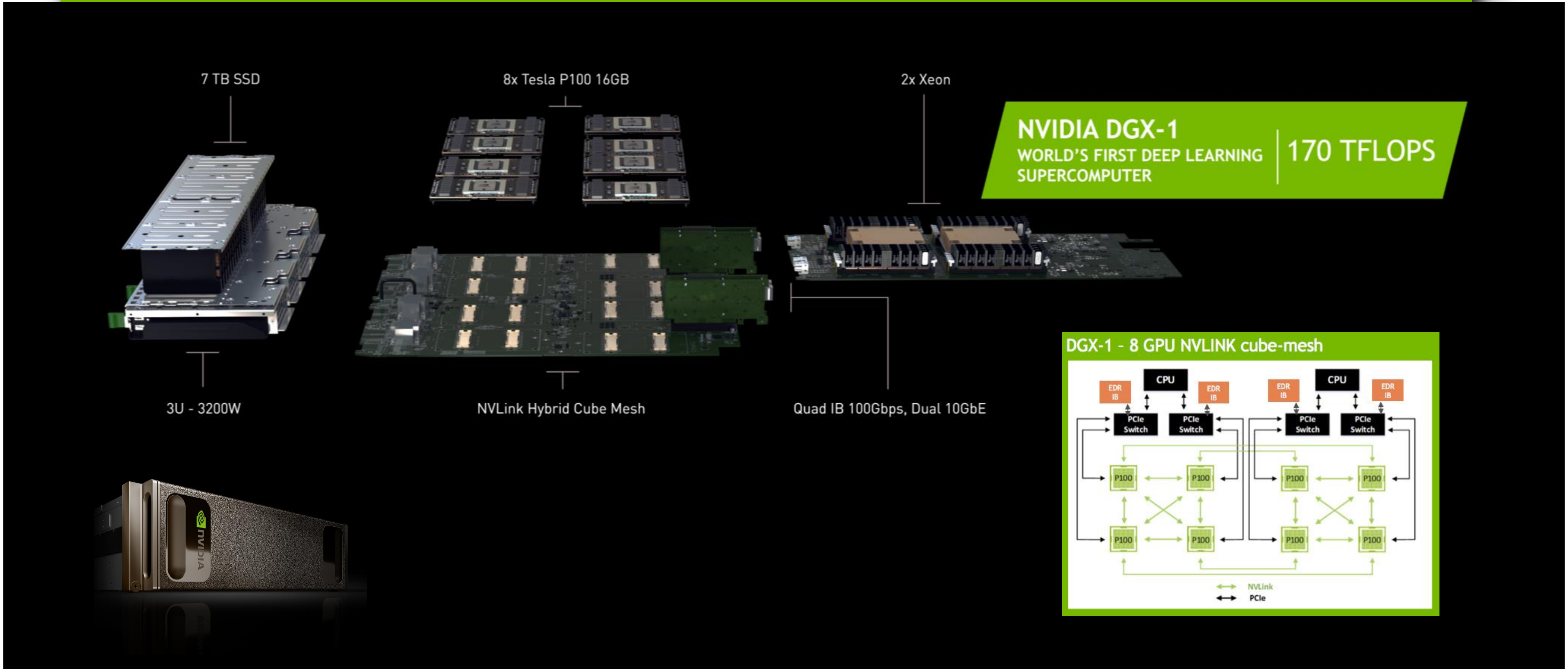
120 Programmable TFLOPS Deep Learning

The Fastest and Most Productive GPU for Deep Learning and HPC

GPU PERFORMANCE COMPARISON

	P100	V100	Ratio
Training acceleration	10 TOPS	120 TOPS	12x
Inference acceleration	21 TFLOPS	120 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

NVIDIA DGX-1 DEEP LEARNING SYSTEM



NVIDIA DGX SATURNV

124 node Cluster

124 NVIDIA DGX-1 Nodes – 992 P100 GPUs

- 8x NVIDIA Tesla P100 SXM GPUs – NVLINK CubeMesh
- 2x Intel Xeon 20 core GPUs
- 512TB DDR4 System Memory
- SSD – 7 TB scratch + 0.5 TB OS

Mellanox 36 port EDR L1 and L2 switches

- 4 ports per system
- Partial Fat tree topology

Ubuntu 14.04, CUDA 8, OpenMPI 1.10.3

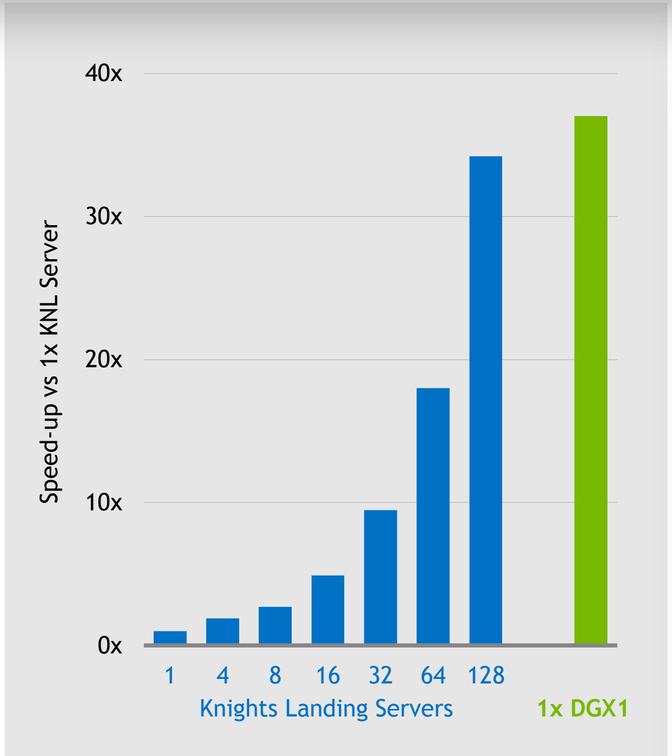
- NVIDIA GPU BLAS + Intel MKL (NVIDIA GPU HPL)

Deep Learning applied research

- Many users, frameworks, algorithms, networks, new approaches
- Embedded, robotic, auto, hyperscale, HPC

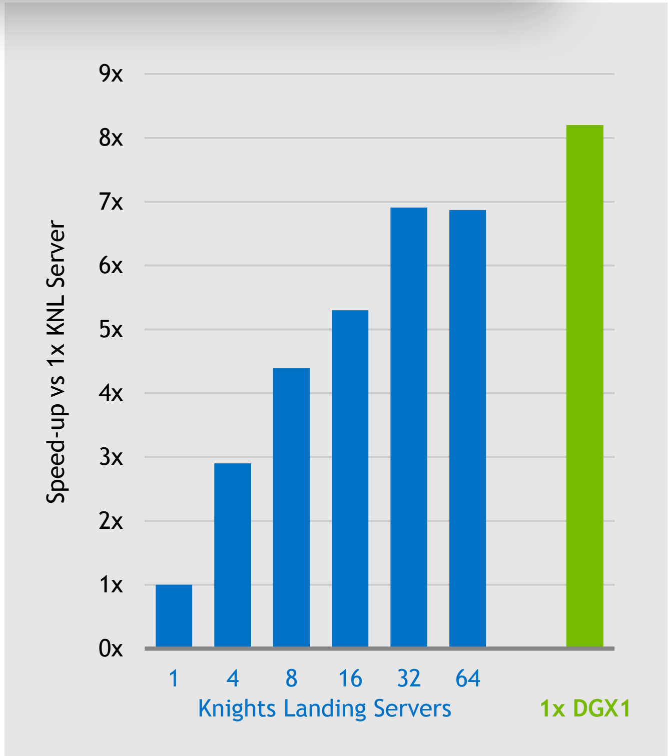
ONE ARCHITECTURE BUILT FOR BOTH DATA SCIENCE & COMPUTATIONAL SCIENCE

GPU-Accelerated Server



AlexNet Training

DGX-1 Faster than 128 Knights Landing Servers



GTC-P: Plasma Turbulence

DGX-1 Faster than 64 Knights Landing Servers

Based on AlexNet Batch size 256, weak scaling up to 32 KNL servers, 64 & 128 estimated based on ideal scaling, Xeon Phi 7250 Nodes

GTC-P, Grid Size A, Systems: NVIDIA DGX-1, 8xP100, Intel KNL 7250 68 core Flat-Quadrant mode, Omnipath

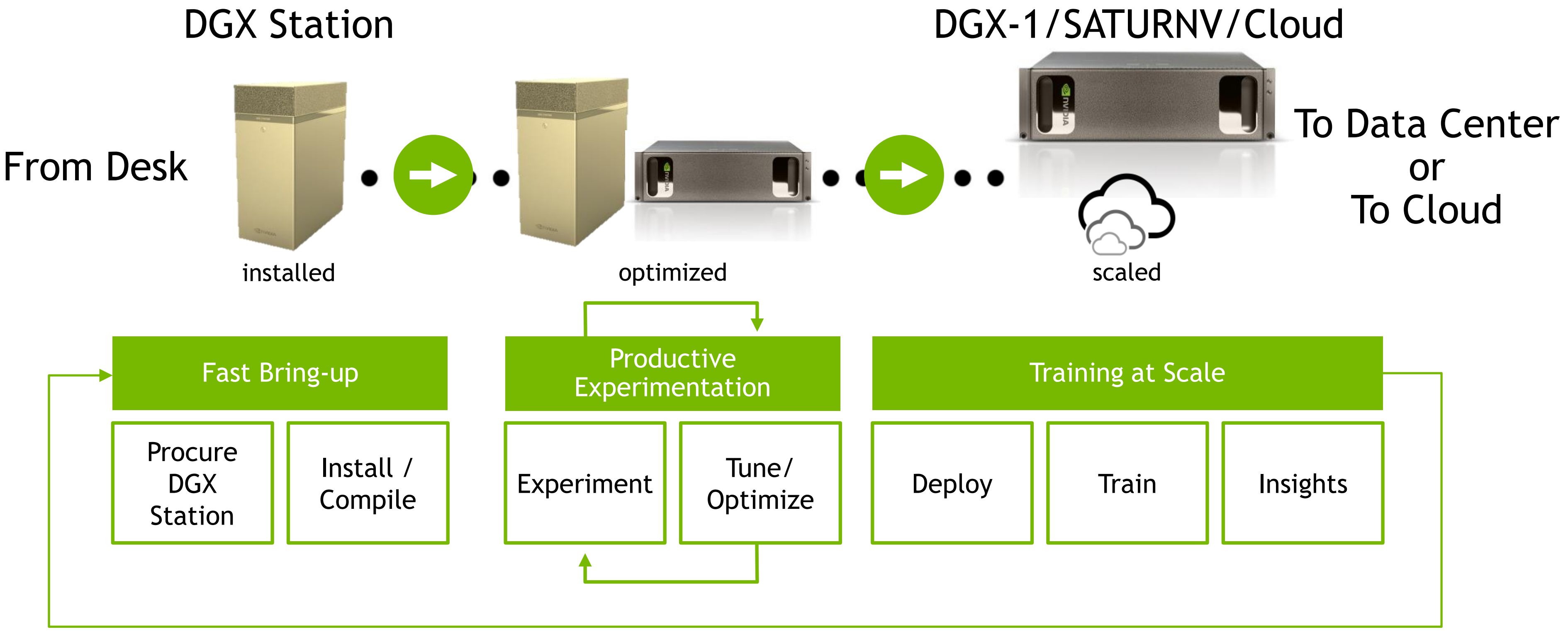
GREEN500 ISC17

Top 13 Systems (measured), 50% Efficiency Improvement, 2.5x Comp.

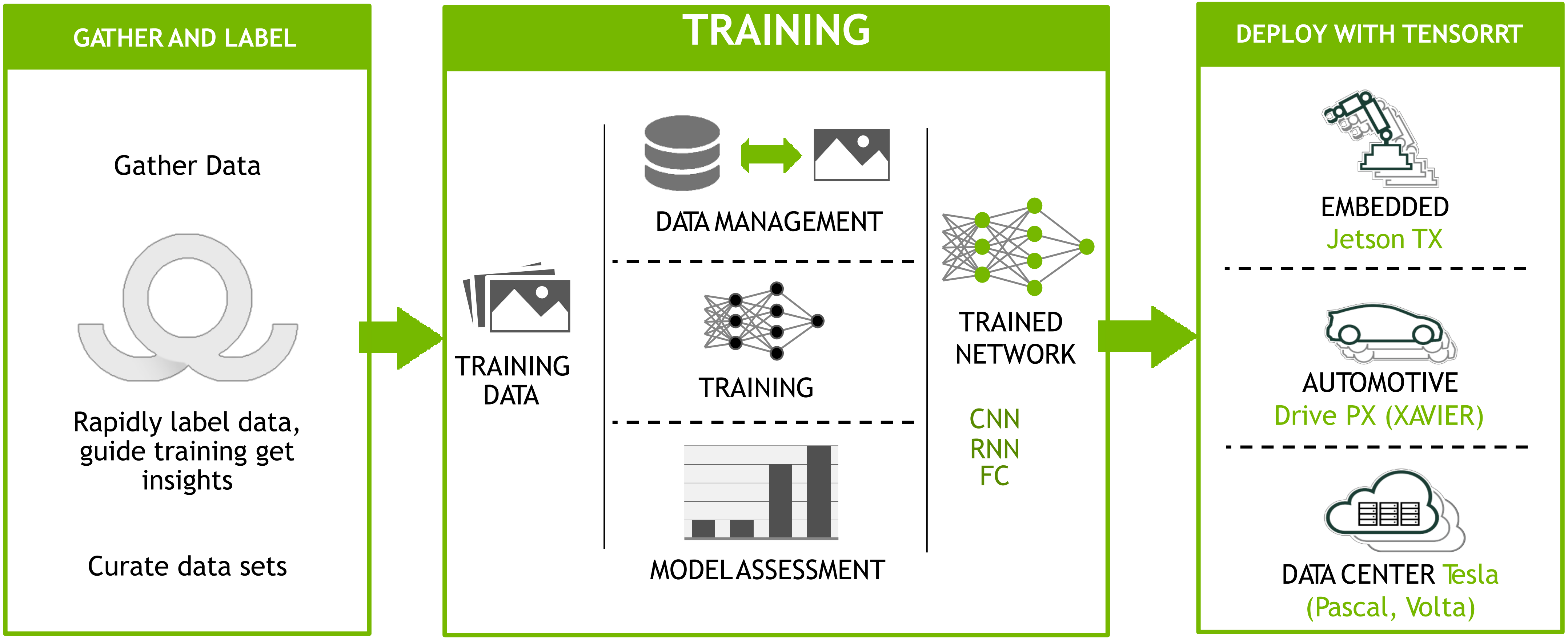
ISC17				
Rank	System	Site	Accelerator	GF/W
1	TSUBAME3.0	GSIC Center, Tokyo Institute of Technology	NVIDIA Tesla P100	14.1
2	kukai	Yahoo Japan Corporation	NVIDIA Tesla P100	14.0
3	AIST AI Cloud	National Institute of Advanced Industrial Science and Technology	NVIDIA Tesla P100	12.7
4	RAIDEN GPU subsystem	Center for Advanced Intelligence Project, RIKEN	NVIDIA Tesla P100	10.6
5	Piz Daint	Swiss National Supercomputing Centre (CSCS)	NVIDIA Tesla P100	10.4
6	Wilkes-2	University of Cambridge	NVIDIA Tesla P100	10.2
7	RCF2	National Institute for Environmental Studies	NVIDIA Tesla P100	9.8
8	DGX Saturn V	NVIDIA Corporation	NVIDIA Tesla P100	9.5
9	Reedbush-H	Information Technology Center, The University of Tokyo	NVIDIA Tesla P100	8.6
10	JADE	University of Oxford	NVIDIA Tesla P100	8.4

DL FROM DEVELOPMENT TO PRODUCTION

Accelerated Deep Learning Value with DGX Solutions



NVIDIA DEEP LEARNING SOFTWARE PLATFORM



ACCELERATED DEEP LEARNING TRAINING STACK

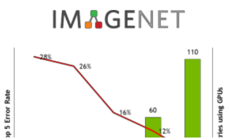
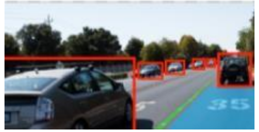


Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Network description, Workflow, Hyper-parameter Sweep, Experiment, Data and Job Management

DL SW Libraries: Tensor/Graph Execution Engines (AKA Frameworks)

Architecture Specific Optimization Layer

ACCELERATED DEEP LEARNING TRAINING STACK

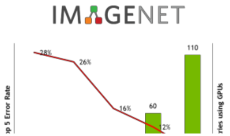
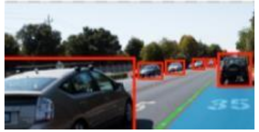


Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

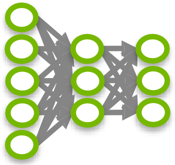
Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Network description, Workflow, Hyper-parameter Sweep, Experiment, Data and Job Management

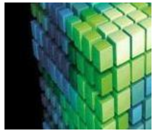
DL SW Libraries: Tensor/Graph Execution Engines (AKA Frameworks)



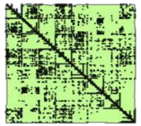
cuDNN

DEEP LEARNING

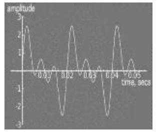
cuBLAS



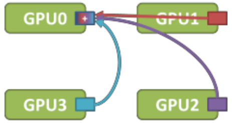
cuSPARSE



cuFFT



MATH LIBRARIES



MULTI-GPU

ACCELERATED DEEP LEARNING TRAINING STACK

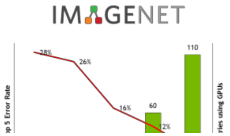
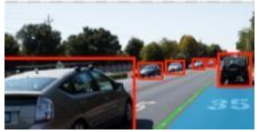


Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

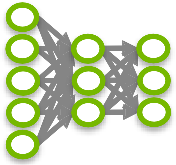
Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Network description, Workflow, Hyper-parameter Sweep,
Experiment, Data and Job Management

TensorFlow PYTORCH torch Caffe2 Caffe mxnet theano Microsoft CNTK Chainer DL4J Deeplearning4j KERAS MATLAB MINERVA OpenDeep Purine Pylearn2

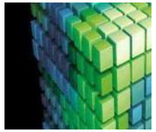
DEEP LEARNING FRAMEWORKS



cuDNN

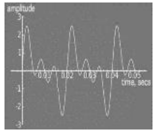
DEEP LEARNING

cuBLAS

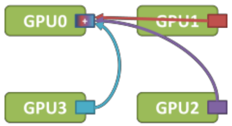


cuSPARSE

cuFFT



MATH LIBRARIES



MULTI-GPU

ACCELERATED DEEP LEARNING TRAINING STACK

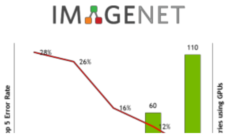
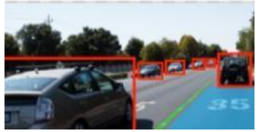


Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

Sentiment Analysis

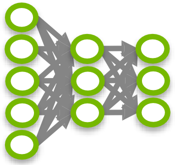
NATURAL LANGUAGE PROCESSING

Productivity Layer/Rapid experimentation: DIGITS, NVIDIA GPU Cloud

UI / JOB MANAGEMENT / DATASET VERSIONING/ VISUALIZATION

TensorFlow · PYTORCH · torch · Caffe2 · Caffe · mxnet · theano · Microsoft CNTK · Chainer · DL4J Deeplearning4j · K KERAS · MATLAB · OpenDeep · Purine · MINERVA · Pylearn2

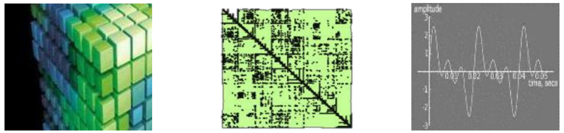
DEEP LEARNING FRAMEWORKS



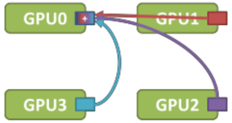
cuDNN

DEEP LEARNING

cuBLAS · cuSPARSE · cuFFT



MATH LIBRARIES



MULTI-GPU

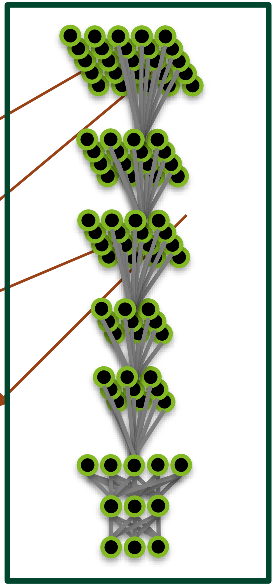
CUDNN LIBRARY OVERVIEW

Stateless, Layer API that is easy to integrate into training frameworks

cudnnConv()
cudnnActivation()

cudnnConv()
cudnnActivation()

:



- Forward and backward paths for many common layer types
- Forward and backward convolution routines
- LSTM, GRU, and Persistent RNNs
- Arbitrary dimension ordering/striding/sub-regions for 4d tensors
- Tensor transformation functions (NCHW, CHWN, NHWC)
- Context-based API allows for easy multithreading

OPTIMIZING FOR GPUS

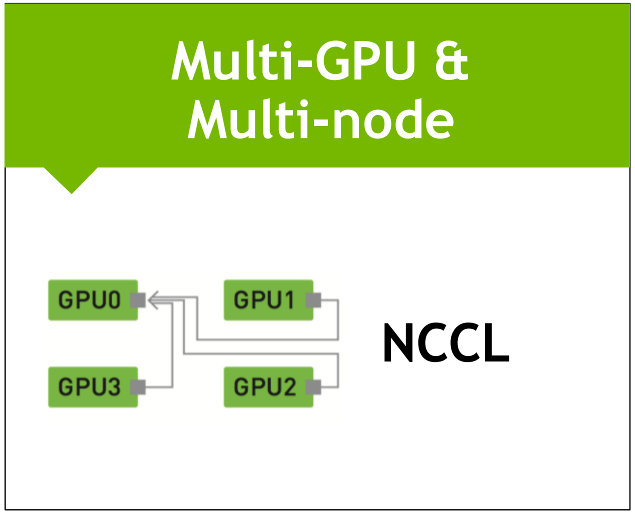
NCCL – NVIDIA Collective Communication Library

Optimized to achieve high bandwidth over PCIe and NVLink

Supports arbitrary number of GPUs installed in a single

Can be used in either single- or multi-process (e.g., MPI) applications.

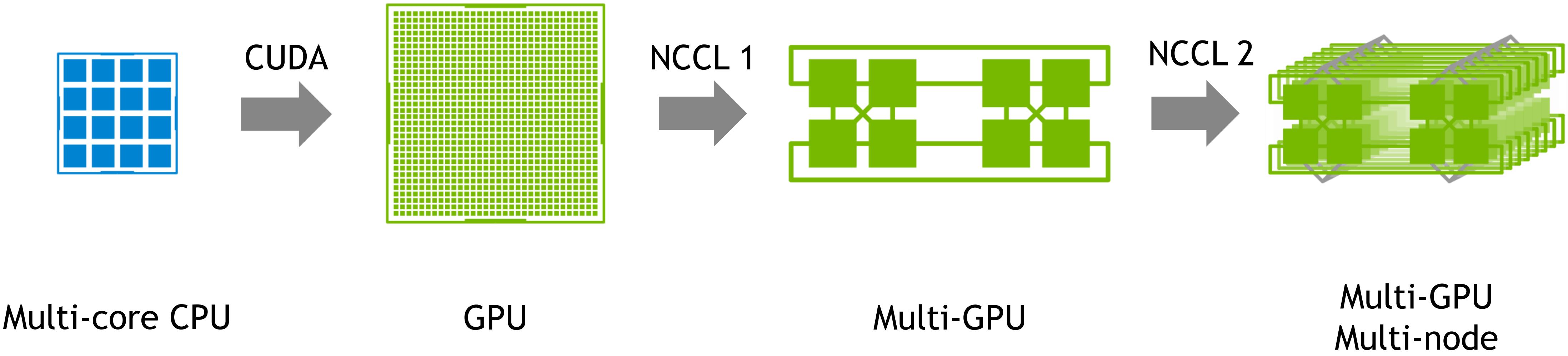
NCCL functions: all-reduce, all-gather, reduce-scatter, reduce, broadcast



DEEP LEARNING ON GPUS

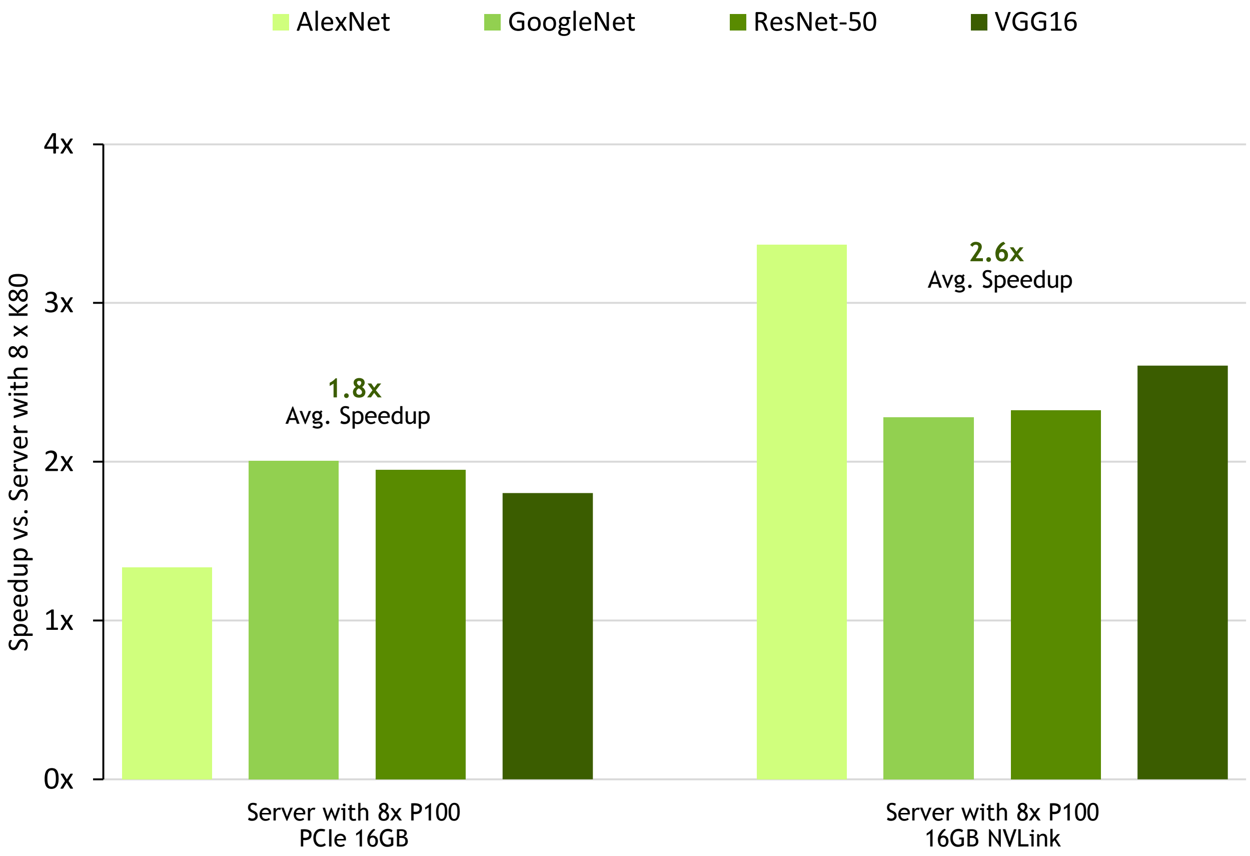
Making DL training times shorter

Deeper neural networks, larger data sets … training is a very, very long operation !



CAFFE Deep Learning Framework

Training on 8x P100 GPU Server vs 8 x K80 GPU Server



GPU Servers: Single Xeon E5-2690 v4@2.6GHz with GPUs configs as shown
Ubuntu 14.04.5, CUDA 8.0.42, cuDNN 6.0.5; NCCL 1.6.1, data set: ImageNet
batch sizes: AlexNet (128), GoogleNet (256), ResNet-50 (64), VGG-16 (32)

CAFFE

Deep Learning

A popular, GPU-accelerated Deep Learning framework developed at UC Berkeley

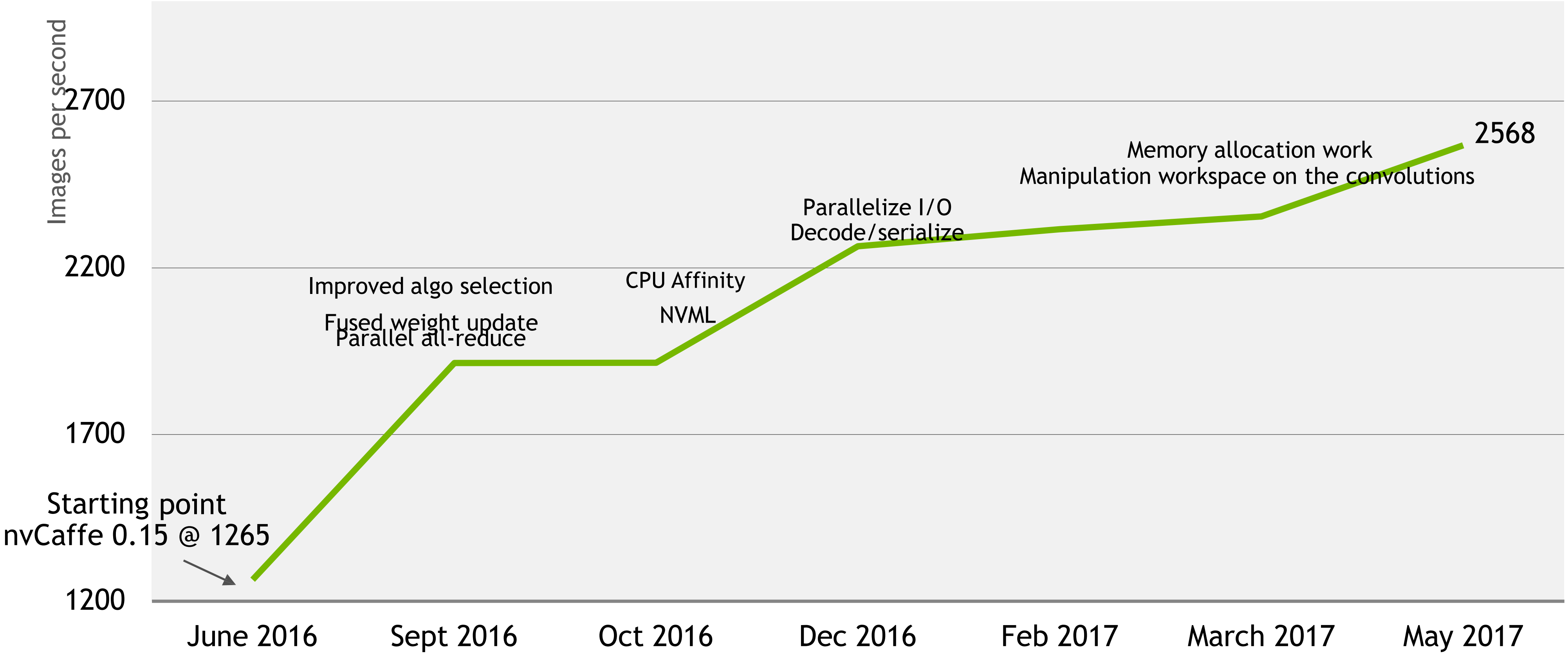
VERSION
1.0

ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU

More Information
http://caffe.berkeleyvision.org/

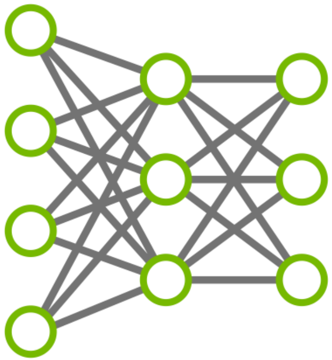
NVCAFFE V0.16 TRAINING ALEXNET



Single P100 GPU, Batch Size=128

NVIDIA TensorRT

Optimizations



TRAINED
NEURAL NETWORK

- Fuse network layers
- Eliminate concatenation layers
- Kernel specialization
- Auto-tuning for target platform
- Tuned for given batch size

OPTIMIZED
INFERENCE
RUNTIME

developer.nvidia.com/tensorrt

NVIDIA TensorRT

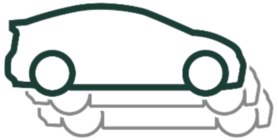
High-performance Inference for Production

DATA CENTER

Tesla P4

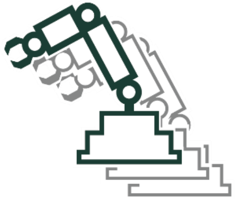
Tesla P40

AUTOMOTIVE



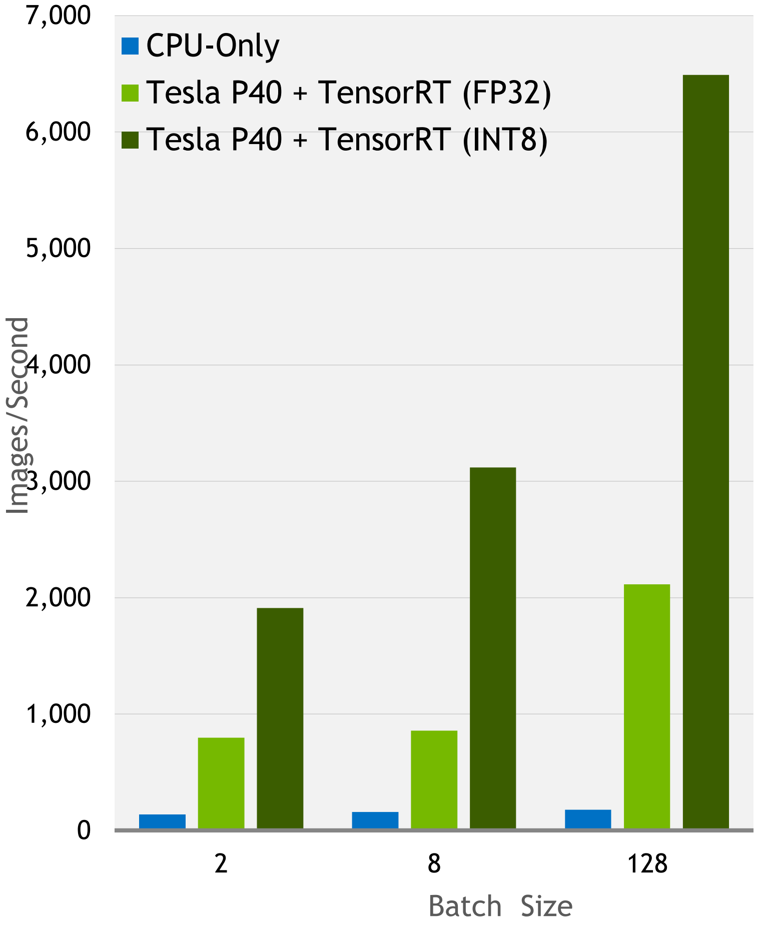
Drive PX2

EMBEDDED



Jetson TX1

Up to 36x More Image/sec

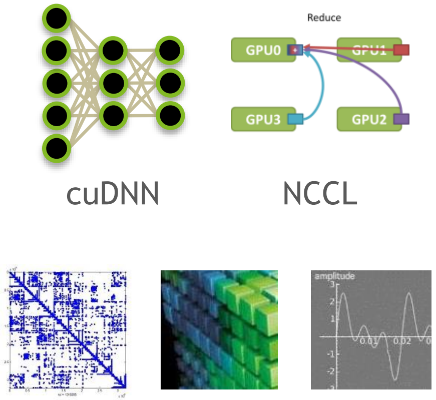


GoogLenet, CPU-only vs Tesla P40 + TensorRT
CPU: 1 socket E4 2690 v4 @2.6 GHz, HT-on
GPU: 2 socket E5-2698 v3 @2.3 GHz, HT off, 1 P40 card in the box

developer.nvidia.com/tensorrt

A TRUE DL APPLIANCE

Accelerated Deep Learning



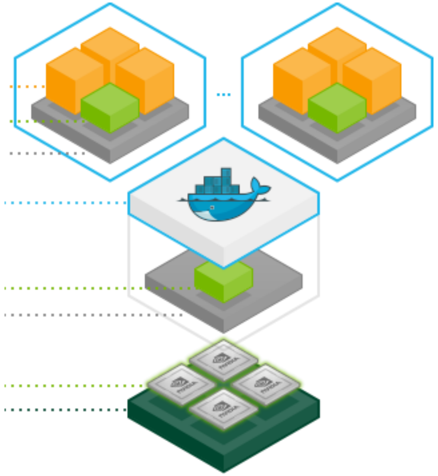
cuDNN NCCL

cuSPARSE cuBLAS cuFFT

Container Based Applications

Digits

DL Frameworks



NVIDIA Cloud Management

AI Researchers ⟷ Enterprise Data Scientists

INTELLIGENT HPC

DL Driving Future HPC Breakthroughs

- Trained networks as solvers
- Super-resolution of coarse simulations
- Low- and mixed-precision
- Simulation for training, network in production

From calendar time to real time?

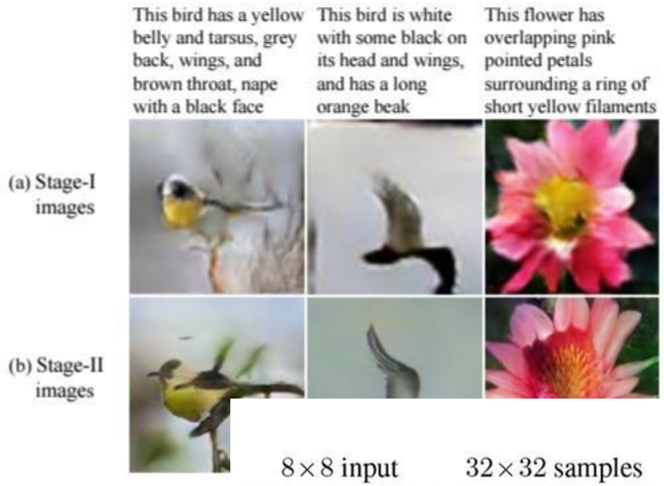
Pre-processing → Simulation → Post-processing

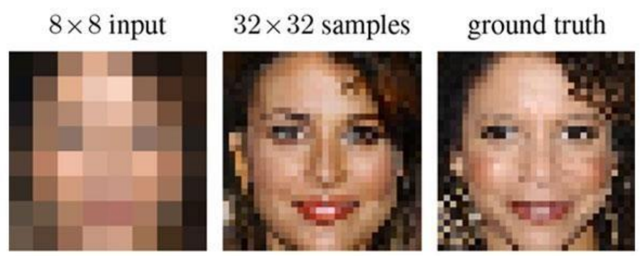
Pre-processing:
- Select/classify/augment/ distribute input data
- Control job parameters

Post-processing:
- Analyze/reduce/augment output data
- Act on output data

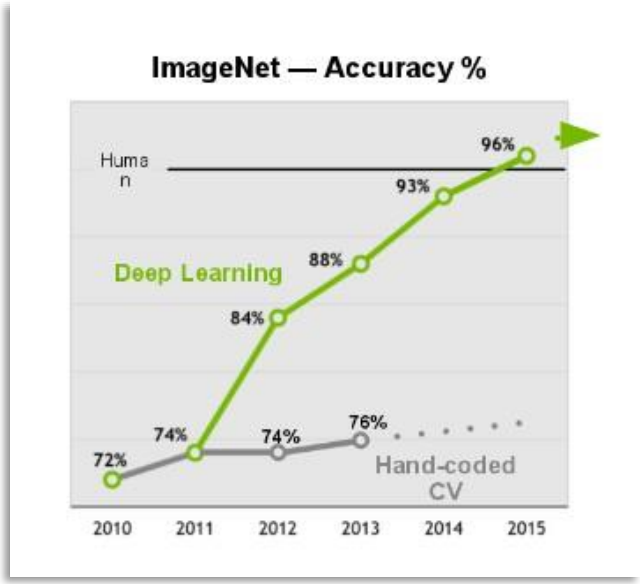
WHY THE EXCITEMENT?

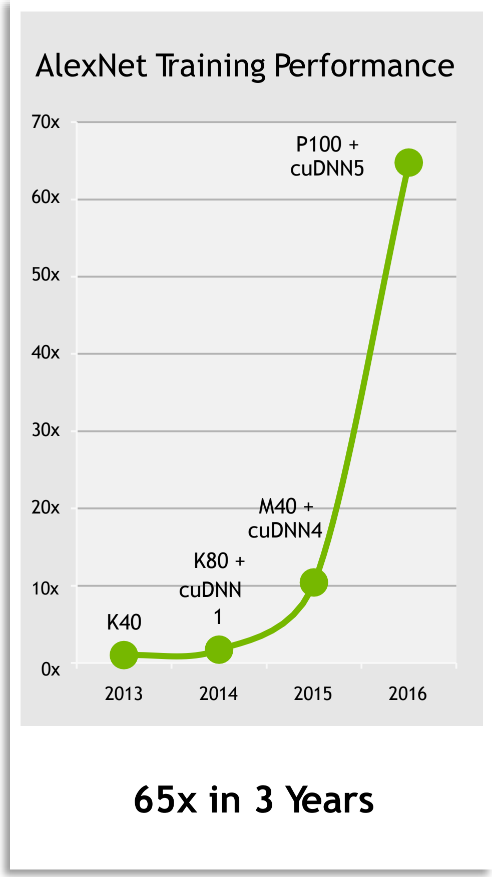
GPUs as Enablers of Breakthrough Results





Dahl et al. 2017





We can generate photorealistic images from textual descriptions and super-enhance blurry photos!

Achieve super-human accuracy in classification

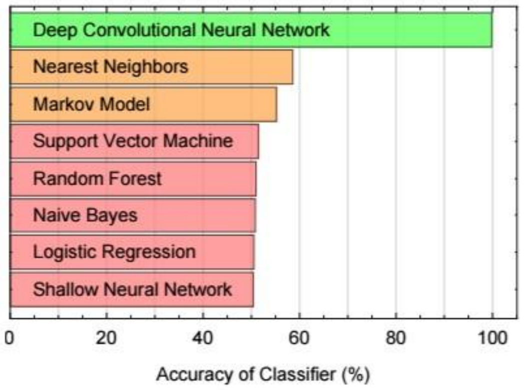
And we are getting faster fast

Paper: H.Zhang et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, arXiv:1612.03242

NVIDIA

DL FOR SIGNAL PROCESSING

Looking for Gravitational Waves



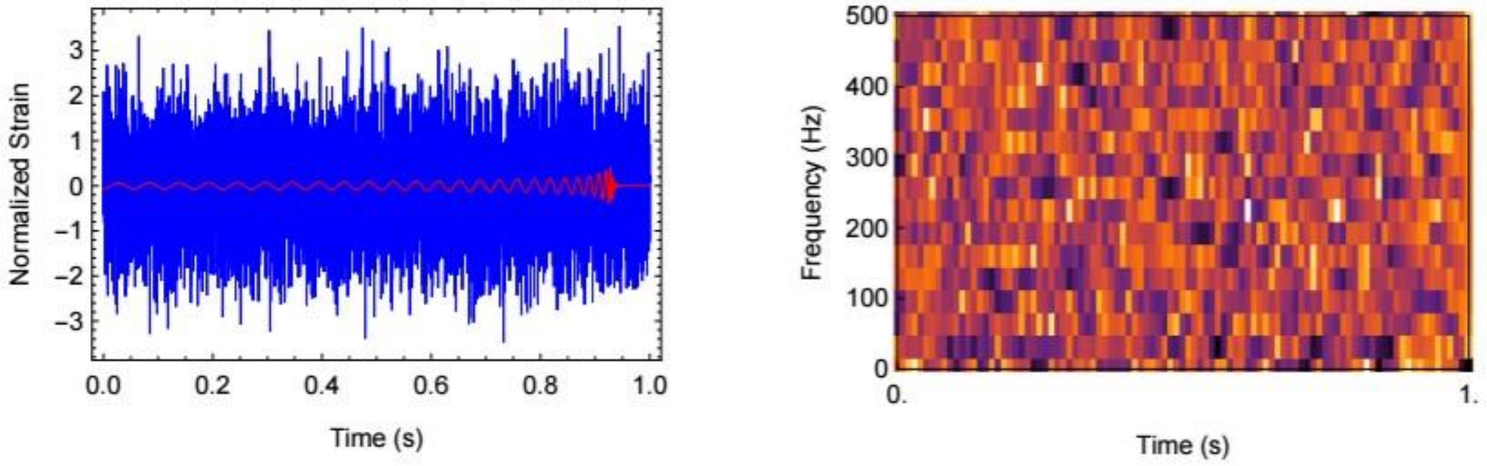


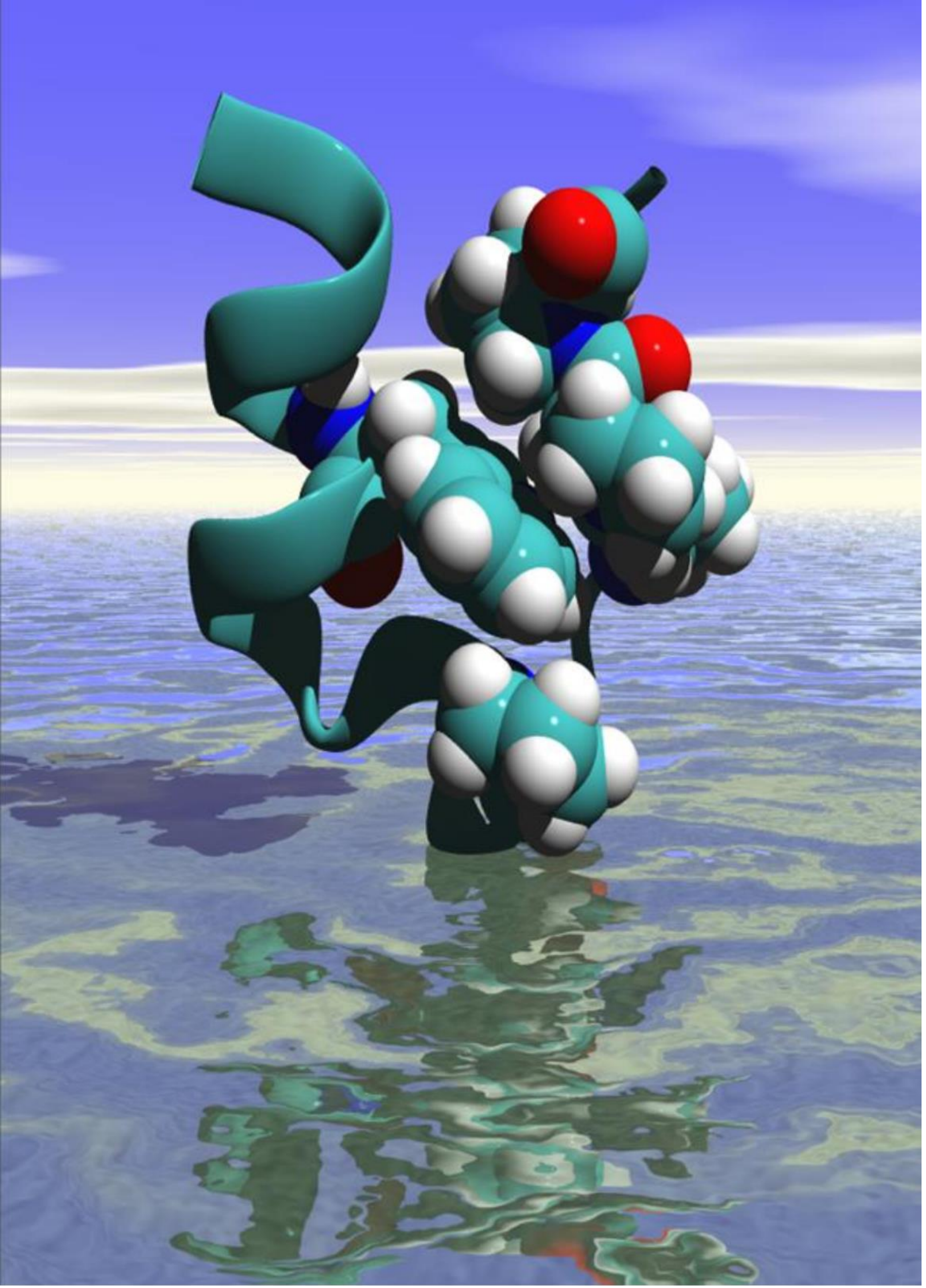
FIG. 2. Left panel: The blue curve is a sample of an input to our DNN algorithm. It contains a BBH GW signal (red) which was whitened with aLIGO's PSD design sensitivity (see Figure 3) and superimposed in noisy data with SNR = 0.5. Right panel: The corresponding spectrogram showing that the BBH GW signal on the left is not visible and thus cannot be detected by any algorithm trained for image recognition. Nevertheless, our DNN detects the presence of this signal from the time-series data, and reconstructs the source's parameters with excellent accuracy.

Classifier: Detect Presence of GWs

↓

Regression: Parameter Estimation (i.e., masses of the two black holes)

From: D.George, E.A.Huerta. Deep Neural Networks to Enable Real-time Multimessenger Astrophysics, arXiv:1701.00008 [astro-ph.IM]



AI Quantum Breakthrough

Background
Developing a new drug costs $2.5B and takes 10-15 years. Quantum chemistry (QC) simulations are important to accurately screen millions of potential drugs to a few most promising drug candidates.

Challenge
QC simulation is computationally expensive so researchers use approximations, compromising on accuracy. To screen 10M drug candidates, it takes 5 years to compute on CPUs.

Solution
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at 1-10/millionths of the cost of current computational methods.
Essentially the DL model is trained to learn Hamiltonian of the Schrodinger equation.

Impact
Faster, more accurate screening at far lower cost

UF UNIVERSITY of FLORIDA

NVIDIA

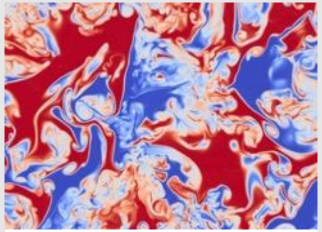
THE HOPE AND PROMISE OF DL IN HPC

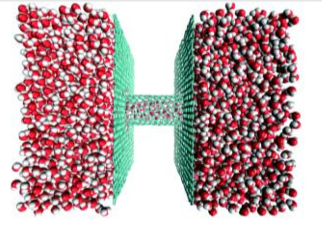
"As the results clearly show, the ANI method is a potential game-changer for molecular simulation. Even the current version, ANI-1, is more accurate vs. the reference DFT level of theory in the provided test cases than DFTB, and PM6, two of the most widely used semi-empirical QM methods. Besides being accurate, a single point energy, and eventually forces, can be calculated as many as six orders of magnitude faster than through DFT."

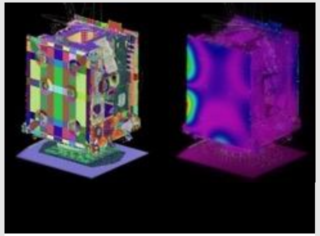
— J.S. Smith et al., ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost.
Chem. Sci., 2017

AI SUPERCOMPUTING IS THE NEW COMPUTING MODEL

Extending The Reach of HPC By Combining Computational & Data Science


Turbulent Flow

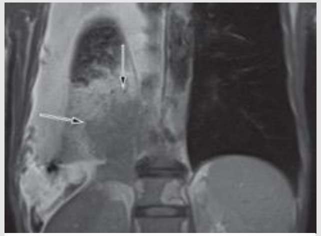

Molecular Dynamics


Structural Analysis


N-body Simulation

COMPUTATIONAL SCIENCE

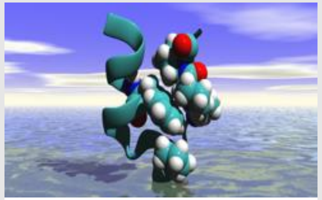

"What's happening?"


"Is there cancer?"


"Next move?"


"What does she mean?"

DATA SCIENCE


Drug Discovery


Clean Energy


Understanding Universe


Monitoring Climate Change

COMPUTATIONAL & DATA SCIENCE

MORE DEEP LEARNING RESOURCES

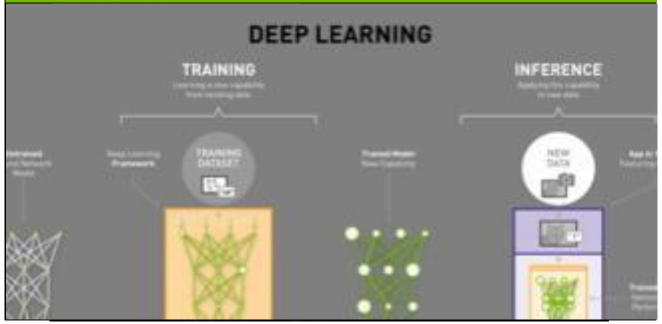
VISIT THE DEEP LEARNING WEBPAGE

http://www.nvidia.com/object/deep-learning.html

RESOURCES

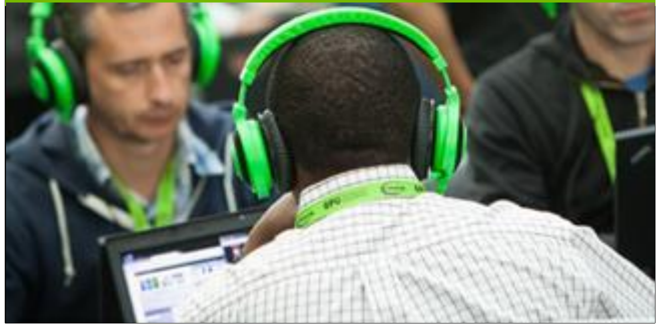
For Executives, Developers and Data Scientists

INTRO MATERIALS



CASE STUDIES

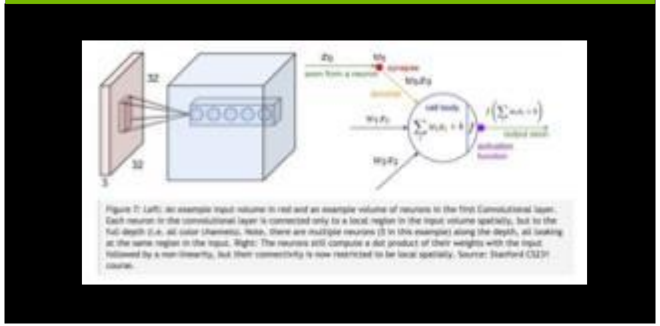
SELF-PACED LABS



ON-SITE WORKSHOPS

PARTNER COURSES

TECHNICAL BLOGS



NVIDIA DEEP LEARNING INSTITUTE

Hands-on Training for Data Scientists and Software Engineers

Training organizations and individuals to solve challenging problems using Deep Learning

On-site workshops and online courses presented by certified experts

Covering complete workflows for proven application use cases

Self-driving cars, recommendation engines, medical image classification, intelligent video analytics and more

www.nvidia.com/dli

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

QUESTIONS?