TEN YEARS OF GPU COMPUTING

2006
CUDA Launched

2008
World’s First GPU Top500 System

2010
Fermi: World’s First HPC GPU

2012
AlexNet beats expert code by huge margin using GPUs

2014
Stanford Builds AI Machine using GPUs

2016
World’s First 3-D Mapping of Human Genome

2016
Google Outperforms Humans in ImageNet

2016
GPU-Trained AI Machine Beats World Champion in Go

2012
Oak Ridge Deploys World’s Fastest Supercomputer w/ GPUs

2010
World’s First Atomic Model of HIV Capsid

2012
Discovered How H1N1 Mutates to Resist Drugs

2016
World’s First HPC GPU

2016
AlexNet beats expert code by huge margin using GPUs
A NEW ERA OF COMPUTING

“
It’s clear we’re moving from a mobile first to an AI-first world”

Sundar Pichai, Google CEO

1995: PC INTERNET
WinTel, Yahoo!
1 billion PC users

2005: MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

2015: AI & IOT
Deep Learning, GPU
100s of billions of devices
TOUCHING OUR LIVES

Bringing grandmother closer to family by bridging language barrier

Predicting sick baby’s vitals like heart rate, blood pressure, survival rate

Enabling the blind to “see” their surrounding, read emotions on faces
FUELING ALL INDUSTRIES

- Increasing public safety with smart video surveillance at airports & malls
- Providing intelligent services in hotels, banks and stores
- Separating weeds as it harvests, reduces chemical usage by 90%
WHAT DOES HPC HAVE TO DO WITH AI?
EVERYTHING!!!
TESLA PLATFORM
Leading Data Center Platform for HPC and AI

APPLICATIONS & SERVICES

INDUSTRY TOOLS

NVIDIA SDK

TESLA GPU & SYSTEMS
NVIDIA POWERS WORLD’S LEADING DATA CENTERS FOR HPC AND AI
NVIDIA
ONE ARCHITECTURE FOR ALL PRODUCTS

GPU Computing

Computer Graphics

Artificial Intelligence
U.S. TO BUILD TWO FLAGSHIP SUPERCOMPUTERS
Pre-Exascale Systems Powered by the Tesla Platform

Summit & Sierra Supercomputers

100-300 PFLOPS Peak
IBM POWER9 CPU + NVIDIA Volta GPU
NVLink High Speed Interconnect
40 TFLOPS per Node, >3,400 Nodes
2017
70% of top HPC APPS ACCELERATED

Intersect360 Survey of Top Apps

- 9 of top 10 Apps Accelerated
- 35 of top 50 Apps Accelerated

Top 25 Apps in Survey

- GROMACS
- SIMULIA Abaqus
- NAMD
- AMBER
- ANSYS Mechanical
- Exelis IDL
- MSC NASTRAN
- ANSYS Fluent
- WRF
- VASP
- OpenFOAM
- CHARMM
- Quantum Espresso
- LAMMPS
- NWChem
- LS-DYNA
- Schrodinger
- Gaussian
- GAMESS
- ANSYS CFX
- Star-CD
- CCSM
- COMSOL
- Star-CCM+
- BLAST

Intersect360, Nov 2015
“HPC Application Support for GPU Computing”
INTRODUCING TESLA P100
New GPU Architecture to Enable the World’s Fastest Compute Node

Pascal Architecture
Highest Compute Performance

NVLink
GPU Interconnect for Maximum Scalability

CoWoS HBM2
Unifying Compute & Memory in Single Package

Page Migration Engine
Simple Parallel Programming with Virtually Unlimited Memory Space
CUDA 8 - WHAT’S NEW

**P100 Support**
New Pascal Architecture
Stacked Memory
NVLINK
FP16 math

**Unified Memory**
Large Datasets
Demand Paging
New Tuning APIs
Standard C/C++ Allocators

**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
GIANT LEAPS IN EVERYTHING
HIGHEST ABSOLUTE PERFORMANCE DELIVERED

NVLink for Max Scalability, More than 45x Faster with 8x P100
PASCAL ARCHITECTURE
### TESLA P100 ACCELERATOR

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compute</strong></td>
<td>5.3 TF DP · 10.6 TF SP · 21.2 TF HP</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>HBM2: 720 GB/s · 16 GB</td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td>NVLink (up to 8 way) + PCIe Gen3</td>
</tr>
<tr>
<td><strong>Programmability</strong></td>
<td>Page Migration Engine</td>
</tr>
<tr>
<td></td>
<td>Unified Memory</td>
</tr>
<tr>
<td><strong>Availability</strong></td>
<td>DGX-1: Order Now</td>
</tr>
<tr>
<td></td>
<td>Cray, Dell, HP, IBM: Q1 2017</td>
</tr>
</tbody>
</table>
## GPU PERFORMANCE COMPARISON

<table>
<thead>
<tr>
<th></th>
<th>P100</th>
<th>M40</th>
<th>K40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double Precision TFlop/s</td>
<td>5.3</td>
<td>0.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Single Precision TFlop/s</td>
<td>10.6</td>
<td>7.0</td>
<td>4.3</td>
</tr>
<tr>
<td>Half Precision T flop/s</td>
<td>21.2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Memory Bandwidth (GB/s)</td>
<td>720</td>
<td>288</td>
<td>288</td>
</tr>
<tr>
<td>Memory Size</td>
<td>16GB</td>
<td>12GB, 24GB</td>
<td>12GB</td>
</tr>
</tbody>
</table>
### IEEE 754 Floating Point on GP100

**3 sizes, 3 speeds, all fast**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Half precision</th>
<th>Single precision</th>
<th>Double precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout</td>
<td>s5.10</td>
<td>s8.23</td>
<td>s11.52</td>
</tr>
<tr>
<td>Issue rate</td>
<td>pair every clock</td>
<td>1 every clock</td>
<td>1 every 2 clocks</td>
</tr>
<tr>
<td>Subnormal support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomic Addition</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**New** indicates features that are new to GP100.
HALF-PRECISION FLOATING POINT (FP16)

• 16 bits
  
  \[
  \begin{array}{cccccccc}
  s & e & x & p & . & f & r & a & c
  \end{array}
  \]
  
  • 1 sign bit, 5 exponent bits, 10 fraction bits

• \(2^{40}\) Dynamic range
  
  • Normalized values: 1024 values for each power of 2, from \(2^{-14}\) to \(2^{15}\)
  
  • Subnormals at full speed: 1024 values from \(2^{-24}\) to \(2^{-15}\)

• Special values
  
  • +- Infinity, Not-a-number

USE CASES

Deep Learning Training
Radio Astronomy
Sensor Data
Image Processing
END-TO-END PRODUCT FAMILY

**HYPERSONE HPC**
- Training - Tesla P100
- Inference - Tesla P40 & P4
- Deep learning training & inference

**STRONG-SCALE HPC**
- Tesla P100 with NVLink
- HPC and DL data centers with workloads scaling to multiple GPUs

**MIXED-APPS HPC**
- Tesla P100 with PCI-E
- HPC data centers with mix of CPU and GPU workloads

**FULLY INTEGRATED DL SUPERCOMPUTER**
- DGX-1
- Fully integrated deep learning solution
Two fully connected quads, connected at corners

160GB/s per GPU bidirectional to Peers

Load/store access to Peer Memory

Full atomics to Peer GPUs

High speed copy engines for bulk data copy

PCIe to/from CPU
UNIFIED MEMORY
PAGE MIGRATION ENGINE
Support Virtual Memory Demand Paging

49-bit Virtual Addresses

Sufficient to cover 48-bit CPU address + all GPU memory

GPU page faulting capability

Can handle thousands of simultaneous page faults

Up to 2 MB page size

Better TLB coverage of GPU memory
KEPLER/MAXWELL UNIFIED MEMORY

CUDA 6+

Kepler GPU  CPU

Unified Memory

Allocate Up To GPU Memory Size

Simpler Programming & Memory Model

Single allocation, single pointer, accessible anywhere
Eliminate need for *explicit copy*
Greatly simplifies code porting

Performance Through Data Locality

Migrate data to accessing processor
 Guarantee global coherency
Still allows explicit hand tuning

Allocate Up To GPU Memory Size
PASCAL UNIFIED MEMORY
Large datasets, simple programming, High Performance

CUDA 8

Enable Large Data Models
Oversubscribe GPU memory
Allocate up to system memory size

Tune Unified Memory Performance
Usage hints via cudaMemAdvise API
Explicit prefetching API

Simpler Data Access
CPU/GPU Data coherence
Unified memory atomic operations

Allocate Beyond GPU Memory Size

Unified Memory

Pascal GPU

CPU
CUDA 8
CUDA 8 - WHAT'S NEW

**P100 Support**
- New Pascal Architecture
- Stacked Memory
- NVLINK
- FP16 math

**Unified Memory**
- Large Datasets
- Demand Paging
- New Tuning APIs
- Standard C/C++ Allocators

**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
ENHANCED PROFILING
The longest running kernel is not always the most critical optimization target.
DEPENDENCY ANALYSIS

Visual Profiler

Unguided Analysis

Generating critical path

Dependency Analysis

Functions on critical path

The following table shows metrics collected from a dependency analysis of the program execution. The data is summarized per function type. Use the "Dependency Analysis" menu on the main toolbar to visualize analysis results on the timeline.

<table>
<thead>
<tr>
<th>Function Name</th>
<th>Time on Critical Path (%)</th>
<th>Time on Critical Path</th>
<th>Waiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>cudaMalloc</td>
<td>32.72 %</td>
<td>127.302 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>jacobi_kernelfloat const <em>, float, int, float</em>)</td>
<td>20.61 %</td>
<td>80.248 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>copy_kernel(float*, float const *, int, int)</td>
<td>17.46 %</td>
<td>68.004 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>&lt;Other&gt;</td>
<td>12.61 %</td>
<td>49.113 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaMemcpy</td>
<td>10.75 %</td>
<td>41.844 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaMemcpyDtoH</td>
<td>5.18 %</td>
<td>20.131 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaSetupArgument</td>
<td>0.14 %</td>
<td>534.684 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaFree</td>
<td>0.11 %</td>
<td>424.883 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>[CUDA memory free]</td>
<td>0.10 %</td>
<td>400.25 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaMemcpyHtoD</td>
<td>0.09 %</td>
<td>336.781 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaGetDeviceProperties</td>
<td>0.08 %</td>
<td>319.677 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaLaunch</td>
<td>0.05 %</td>
<td>192.598 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaConfigureCall</td>
<td>0.05 %</td>
<td>186.452 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaDeviceTotalMem_v2</td>
<td>0.05 %</td>
<td>182.833 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaDeviceGetName</td>
<td>0.00 %</td>
<td>15.022 ms</td>
<td>0 ns</td>
</tr>
<tr>
<td>cudaSetDevice</td>
<td>0.00 %</td>
<td>12.393 ms</td>
<td>0 ns</td>
</tr>
</tbody>
</table>
APIs, GPU activities not in critical path are greyed out
MORE CUDA 8 PROFILER FEATURES

Unified Memory Profiling

CPU Profiling

OpenACC Profiling

NVLink Topology and Bandwidth profiling
OPENACC
World’s Only Performance Portable Programming Model for HPC

Add Simple Compiler Hint

```c
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
SINGLE OPENACC CODE RUNS ON ALL CPU & GPU PLATFORMS

CloverLeaf- Hydrodynamics Mini-Application

<table>
<thead>
<tr>
<th>Platform</th>
<th>Speedup vs Single Haswell Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicore Haswell</td>
<td>9x</td>
</tr>
<tr>
<td>Multicore Broadwell</td>
<td>10x</td>
</tr>
<tr>
<td>Multicore POWER8</td>
<td>11x</td>
</tr>
<tr>
<td>Xeon Phi Knights Landing</td>
<td>25x 15x</td>
</tr>
<tr>
<td>1x Tesla P100</td>
<td>52x</td>
</tr>
<tr>
<td>2x Tesla P100</td>
<td>93x</td>
</tr>
</tbody>
</table>

(PGI OpenACC in early development stage)
LSDalton
Quantum Chemistry
12X speedup in 1 week

Numeca
CFD
10X faster kernels
2X faster app

PowerGrid
Medical Imaging
40 days to 2 hours

INCOMP3D
CFD
3X speedup

NekCEM
Computational Electromagnetics
2.5X speedup
60% less energy

COSMO
Climate Weather
40X speedup
3X energy efficiency

CloverLeaf
CFD
4X speedup
Single CPU/GPU code

MAESTRO CASTRO
Astrophysics
4.4X speedup
4 weeks effort
# OPENACC FOR EVERYONE

New PGI Community Edition Now Available

<table>
<thead>
<tr>
<th>Programming Models</th>
<th>Platforms</th>
<th>Updates</th>
<th>Support</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenACC, CUDA Fortran, OpenMP, C/C++/Fortran Compilers and Tools</td>
<td>x86, OpenPOWER, NVIDIA GPU</td>
<td>1-2 times a year</td>
<td>User Forums</td>
<td>Annual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6-9 times a year</td>
<td>PGI Support</td>
<td>Perpetual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6-9 times a year</td>
<td>PGI Enterprise Services</td>
<td>Volume/Site</td>
</tr>
</tbody>
</table>

**FREE**
PERFORMANCE
MOLECULAR DYNAMICS
AMBER Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

AMBER
Molecular Dynamics

Suite of programs to simulate molecular dynamics on biomolecule

VERSION
16.3

ACCELERATED FEATURES
PMEMD Explicit Solvent & GB; Explicit & Implicit Solvent, REMD, aMD

SCALABILITY
Multi-GPU and Single-Node

More Information
http://ambermd.org/gpus

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: GB-Myoglobin
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.
HOOMD-Blue Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: microsphere

More Information
http://codeblue.umich.edu/hoomd-blue/index.html
LAMMPS Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

LAMMPS
Classical molecular dynamics package

VERSION
2016

ACCELERATED FEATURES
Lennard-Jones, Gay-Berne, Tersoff, many more potentials

SCALABILITY
Multi-GPU and Multi-Node

More Information

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: EAM
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.
GROMACS Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

CPU Server: Dual Xeon E5-2690 v4 @ 2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: Water 3M

GROMACS
Molecular Dynamics
Simulation of biochemical molecules with complicated bond interactions

VERSION
5.1.2

ACCELERATED FEATURES
PME, Explicit & Implicit Solvent

SCALABILITY
Multi-GPU and Multi-Node
Scales to 4xP100

More Information
http://www.gromacs.org
NAMD Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: STMV

NAMD
Geoscience (Oil & Gas)

Designed for high-performance simulation of large molecular systems

VERSION
2.11

ACCELERATED FEATURES
Full electrostatics with PME and most simulation features

SCALABILITY
Up to 100M atom capable, multi-GPU, Scale Scales to 2xP100

More Information
http://www.ks.uiuc.edu/Research/namd/
MATERIALS SCIENCE
VASP Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

# of CPU Only Servers

6 CPU Servers
13 CPU Servers
18 CPU Servers
14 CPU Servers
19 CPU Servers

2xP100
4xP100
8xP100
2xP100
4xP100
8xP100

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset:B.hR105
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

VASP
Material Science (Quantum Chemistry)

Package for performing ab-initio quantum-mechanical molecular dynamics (MD) simulations

VERSION
5.4.1

ACCELERATED FEATURES
RMM-DIIS, Blocked Davidson, K-points and exact-exchange

SCALABILITY
Multi-GPU and Multi-Node

More Information
BENCHMARKS
Linpack Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: HPL.dat

Linpack Benchmark
Measures floating point computing power
VERSION
2.1
ACCELERATED FEATURES
All
SCALABILITY
Multi-GPU and Multi-Node

More Information
https://www.top500.org/project/linpack/
HPCG Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

HPCG Benchmark

Exercises computational and data access patterns that closely match a broad set of important HPC applications

VERSION
3.0

ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

More Information
http://www.hpcg-benchmark.org/index.html

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ P100s PCIe (12GB or 16GB)
CUDA Version: CUDA 8.0.42, Dataset: 256x256x256 local size
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.
DEEP LEARNING
CAFFE Deep Learning Framework
Training on 8x P100 GPU Server vs 8x K80 GPU Server

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/

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)
TensorFlow Deep Learning Framework

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

- AlexNet
- GoogleNet
- ResNet-50
- ResNet-152
- VGG16

**ACCELERATED FEATURES**
Full framework accelerated

**SCALABILITY**
Multi-GPU and multi-node

More Information
https://www.tensorflow.org/

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), ResNet-152 (32), VGG-16 (32)
Torch Deep Learning Framework
Training on 8x P100 GPU Server vs 8x 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), InceptionV3 (64), ResNet-50 (64), VGG16-16 (32)
CNTK Deep Learning Framework
Training on 8x P100 GPU Server vs 8 x K80 GPU Server

- 1.0
- 2.0
- 3.0
- 4.0

Speedup vs. Server with 8 x K80

- AlexNet
- ResNet-50

2.7x
Avg. Speedup

1.6x
Avg. Speedup

Server with 8x P100 PCIe 16GB
Server with 8x P100 16GB NVLink

CNTK
Deep Learning Training

A free, easy-to-use, open-source, commercial-grade toolkit that trains deep learning algorithms to learn like the human brain.

VERSION 1.0

ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU and multi-node

More Information
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
- DL4J
- Keras
- Microsoft CNTK
- mxnet
- Purine
- TensorFlow
- theano

NVIDIA DEEP LEARNING SDK
- cuDNN
- TensorRT
- DeepStream SDK
- cuBLAS
- cuSPARSE
- NCCL

developer.nvidia.com/deep-learning-software
NVIDIA DEEP LEARNING SOFTWARE PLATFORM

DIGITS

- Data Management
- Training
- Model Assessment

TensorRT

- Embedded
- Automotive
- Data center

NVIDIA DEEP LEARNING SDK

developer.nvidia.com/deep-learning-software
NVIDIA DIGITS
Interactive Deep Learning GPU Training System

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

Schedule, monitor, and manage neural network training jobs

Analyze accuracy and loss in real time

Track datasets, results, and trained neural networks

Scale training jobs across multiple GPUs automatically

developer.nvidia.com/digits
NVIDIA cuDNN
Accelerating Deep Learning

High performance building blocks for deep learning frameworks

Drop-in acceleration for widely used deep learning frameworks such as Caffe, CNTK, Tensorflow, Theano, Torch and others

Accelerates industry vetted deep learning algorithms, such as convolutions, LSTM, fully connected, and pooling layers

Fast deep learning training performance tuned for NVIDIA GPUs

developer.nvidia.com/cudnn
INTRODUCING NVIDIA TensorRT
High Performance Inference Engine

User Experience: Instant Response
45x Faster with Pascal + TensorRT

1x CPU (14 cores) 260 ms
P4 11 ms
P40 6 ms

Faster, more responsive AI-powered services such as voice recognition, speech translation
Efficient inference on images, video, & other data in hyperscale production data centers

Based on VGG-19 from Intel Caffe Github: https://github.com/intel/caffe/tree/master/models/mkl2017_vgg_19
CPU: Intel Caffe, batch size = 4, Intel E5-2690v4, using Intel MKL 2017
GPU: Caffe, batch size = 4, using TensorRT internal version
NVIDIA DGX-1
WORLD’S FIRST DEEP LEARNING SUPERCOMPUTER

170 TFLOPS
8x Tesla P100 16GB
NVLink Hybrid Cube Mesh
Optimized Deep Learning Software
Dual Xeon
7 TB SSD Deep Learning Cache
Dual 10GbE, Quad IB 100Gb
3RU - 3200W
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
ACCELERATE EVERY FRAMEWORK

ACADEMIA

<table>
<thead>
<tr>
<th>Framework</th>
<th>Institution(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAFFE</td>
<td>Berkeley University of California</td>
</tr>
<tr>
<td>TORCH</td>
<td>NYU</td>
</tr>
<tr>
<td>THEANO</td>
<td>Université de Montréal</td>
</tr>
<tr>
<td>MATCONVNET</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>MOCHA.JL</td>
<td>MIT, Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>PURINE</td>
<td>NUS, National University of Singapore</td>
</tr>
<tr>
<td>MINERVA</td>
<td>NYU, University of Washington, Carnegie Mellon University, Microsoft</td>
</tr>
<tr>
<td>MXNET*</td>
<td></td>
</tr>
</tbody>
</table>

START-UPS

<table>
<thead>
<tr>
<th>Framework</th>
<th>Institution(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TENSORFLOW</td>
<td>Google</td>
</tr>
<tr>
<td>TORCH</td>
<td>Facebook</td>
</tr>
<tr>
<td>CNTK</td>
<td>Microsoft</td>
</tr>
<tr>
<td>CHAINER</td>
<td>Preferred Networks</td>
</tr>
<tr>
<td>DL4J</td>
<td>SKYMIND</td>
</tr>
<tr>
<td>KERAS</td>
<td>SCHULTS LABORATORIES</td>
</tr>
<tr>
<td>OPENDEEP</td>
<td>VITRUVIAN</td>
</tr>
</tbody>
</table>

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai
BENEFITS FOR AI RESEARCHERS

Fastest DL Supercomputer

Design Big Networks

Reduce Training Times

DL SDK Ongoing Updates

cuDNN

NCCL

cuSPARSE

cuBLAS

cuFFT

Fastest DL Supercomputer

Design Big Networks

Reduce Training Times

DL SDK Ongoing Updates
BENEFITS FOR INDUSTRY DATA SCIENTISTS

Purpose Built DL Supercomputer

Accelerates All Major Frameworks

Simple Appliance, Software Updates

NVIDIA Expert Support

NVIDIA GPU PLATFORM

TORCH
NYU facebook

CAFFE
Berkeley

THEANO
Université de Montréal

MATCONVNET
University of Oxford

TENSORFLOW
Google

WATSON
IBM

CNTK
Microsoft
GPU TECHNOLOGY CONFERENCE
May 8 - 11, 2017 | Silicon Valley | #GTC17
www.gputechconf.com

CONNECT
Connect with technology experts from NVIDIA and other leading organizations

LEARN
Gain insight and valuable hands-on training through hundreds of sessions and research posters

DISCOVER
See how GPUs are creating amazing breakthroughs in important fields such as deep learning and AI

INNOVATE
Hear about disruptive innovations from startups

REGISTER EARLY: SAVE UP TO $240 AT WWW.GPUTECHCONF.COM

Don’t miss the world’s most important event for GPU developers
May 8 - 11, 2017 in Silicon Valley