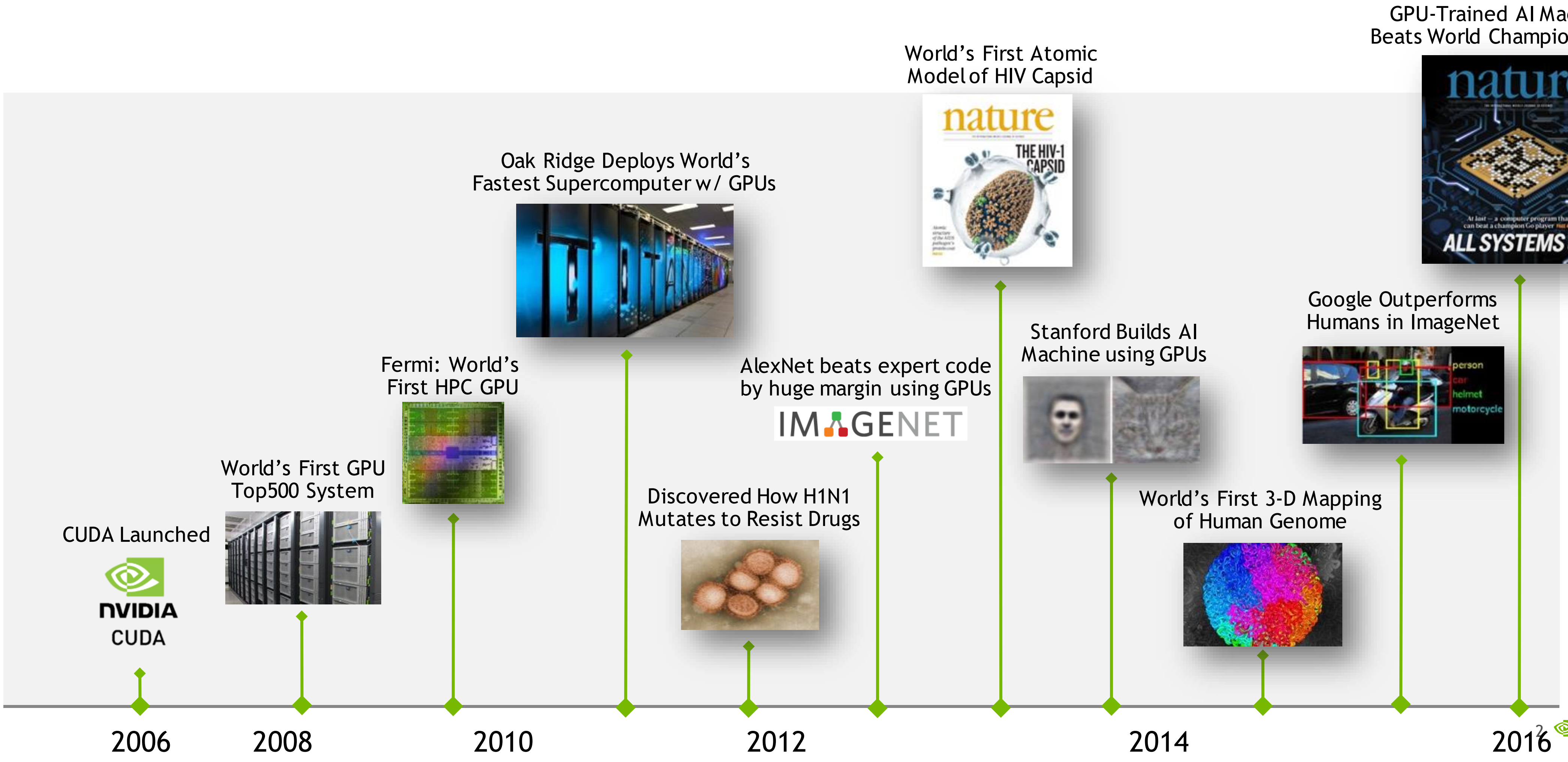
NVIDIA IN HPC AND AI

April 2017, OSC SUG

TEN YEARS OF GPU COMPUTING



A NEW ERA OF COMPUTING

" It's clear we're moving from a mobile first to an AI-first world "

Sundar Pichai, Google CEO

PC INTERNET
WinTel, Yahoo!
1 billion PC users

MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

AI & IOT
Deep Learning, GPU
100s of billions of devices

1995 **2005** **2015**

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!!!

https://code.facebook.com
Code
Android
iOS
Web
Backend
Hardware
Engineering at Facebook
AIR FLOW
Facebook to open-source AI hardware design

TESLA PLATFORM

Leading Data Center Platform for HPC and AI

APPLICATIONS & SERVICES

AI TRAINING & INFERENCE
- amazon alexa
- IBM Watson
- Microsoft Cognitive Services
- NVIDIA DIGITS
- Google Cloud Platform Machine Learning Services

HPC
- ANSYS
- SIMULIA
- GROMACS
- VASP
- +400 More Applications

INDUSTRY TOOLS

FRAMEWORKS
- Caffe
- mxnet
- theano
- CNTK
- torch
- TensorFlow

MODELS
- ResNet
- GoogleNet
- AlexNet
- DeepSpeech
- Inception
- BigLSTM

ECOSYSTEM TOOLS & LIBRARIES
- allinea
- ArrayFire
- CONTINUUM
- WOLFRAM
- MathWorks

NVIDIA SDK

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

COMPUTEWORKS
- C/C++
- Fortran
- CUDA C/C++
- PGI OpenACC

TESLA GPU & SYSTEMS

- TESLA GPU
- NVLINK
- SYSTEM OEM: DELL, Hewlett Packard Enterprise, IBM
- CLOUD: amazon webservices, Google Cloud Platform, Microsoft Azure

NVIDIA POWERS WORLD'S LEADING DATA CENTERS FOR HPC AND AI

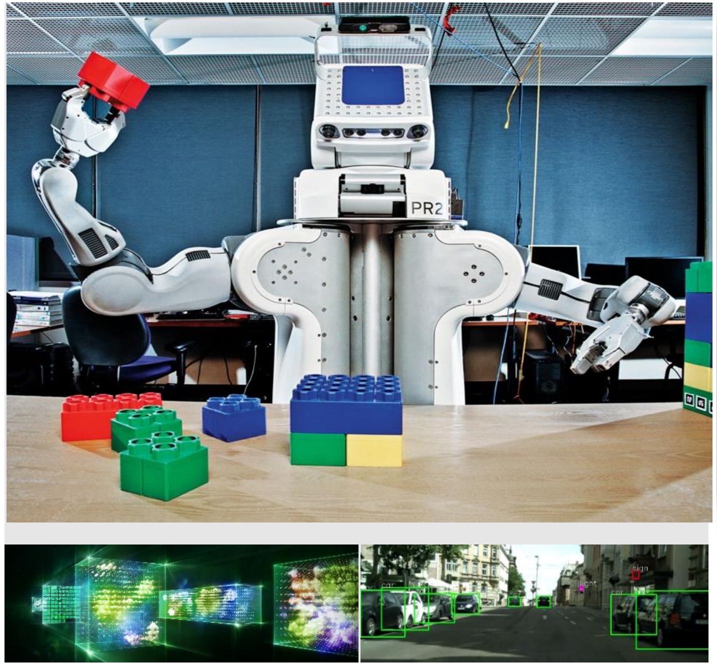
amazon web services™ facebook Google IBM

Lawrence Livermore National Laboratory Microsoft OAK RIDGE National Laboratory twitter

NVIDIA
ONE ARCHITECTURE FOR ALL PRODUCTS

GPU Computing

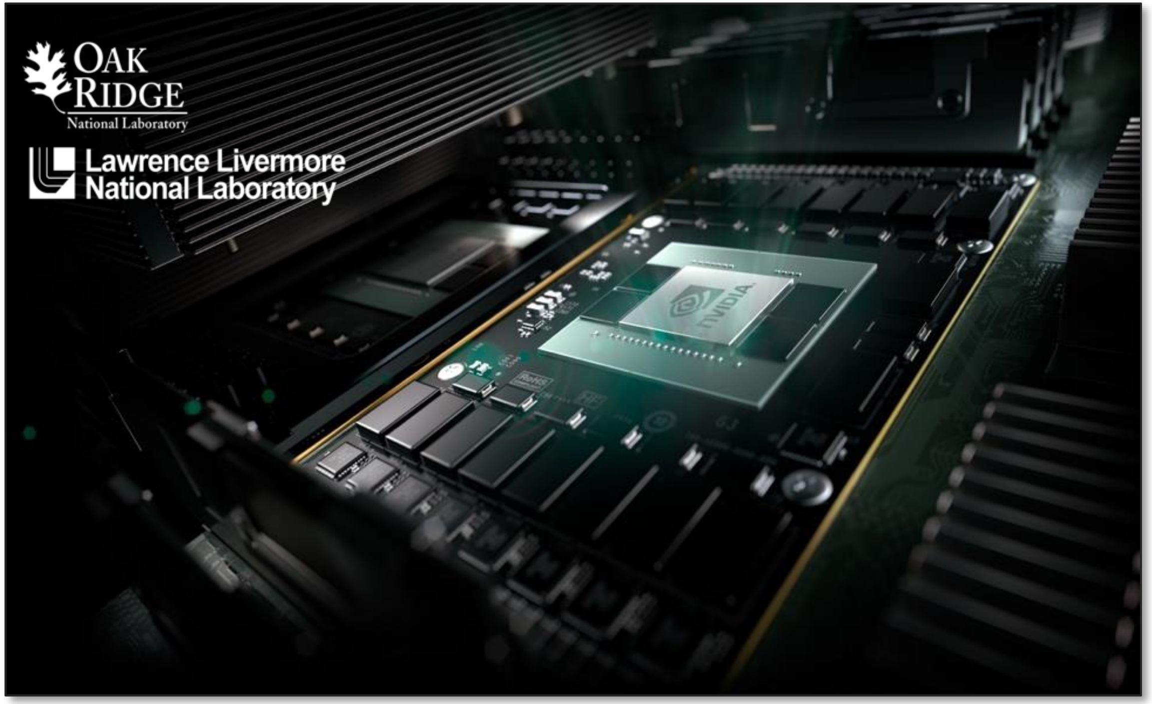
Computer Graphics



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

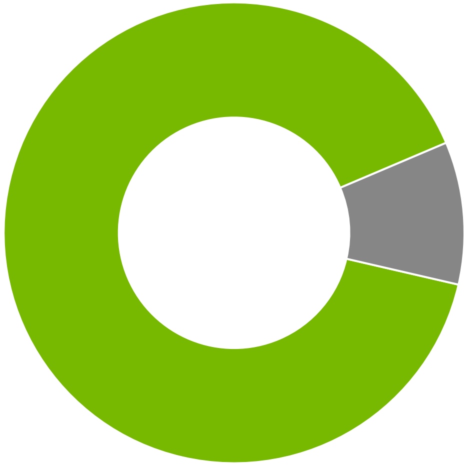
NVLink High Speed Interconnect

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

Intersect360, Nov 2015
"HPC Application Support for GPU Computing"

TOP 25 APPS IN SURVEY

GROMACS
SIMULIA Abaqus
NAMD
AMBER
ANSYS Mechanical
Exelis IDL
MSC NASTRAN

LAMMPS
NWChem
LS-DYNA
Schrodinger

Gaussian
GAMESS

ANSYS Fluent
WRF
VASP
OpenFOAM
CHARMM
Quantum Espresso

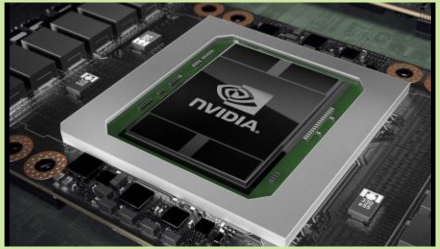
ANSYS CFX
Star-CD
CCSM
COMSOL
Star-CCM+
BLAST

= All popular functions accelerated
= Some popular functions accelerated
= In development
= Not supported

INTRODUCING TESLA P100

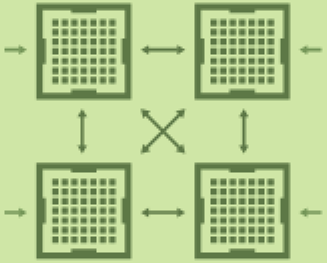
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



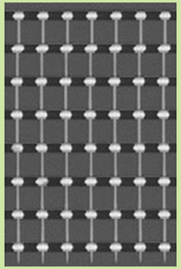
Highest Compute Performance

NVLink



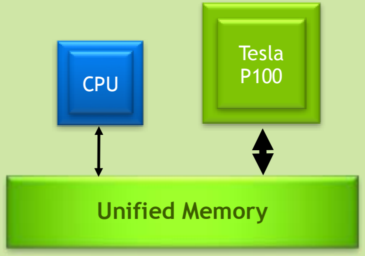
GPU Interconnect for Maximum Scalability

CoWoS HBM2

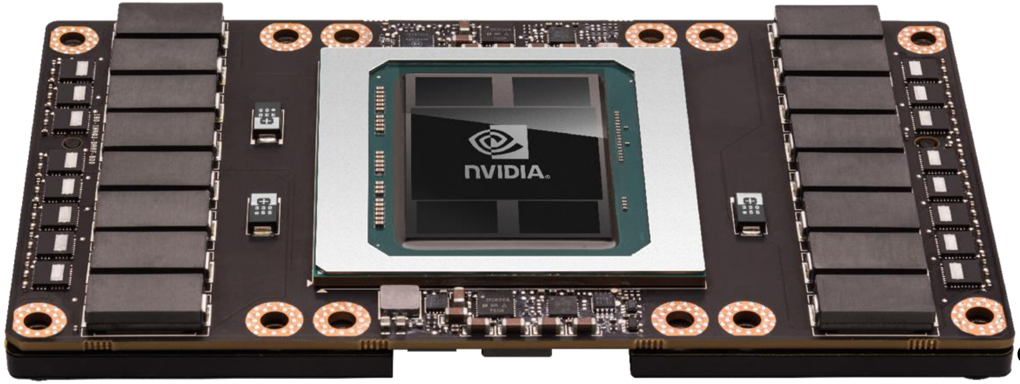


Unifying Compute & Memory in Single Package

Page Migration Engine



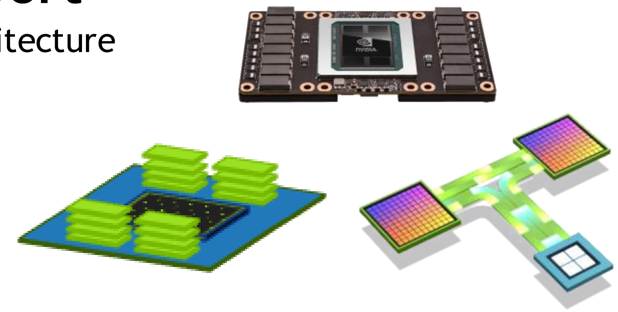
Simple Parallel Programming with Virtually Unlimited Memory Space



CUDA 8 - WHAT'S NEW

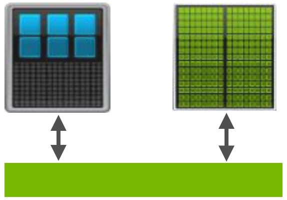
P100 Support

New Pascal Architecture
Stacked Memory
NVLINK
FP16 math



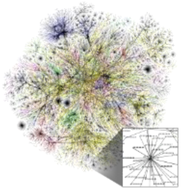
Unified Memory

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



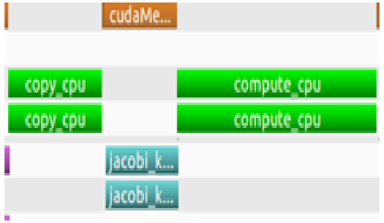
Libraries

New nvGRAPH library
cuBLAS improvements for Deep Learning



Developer Tools

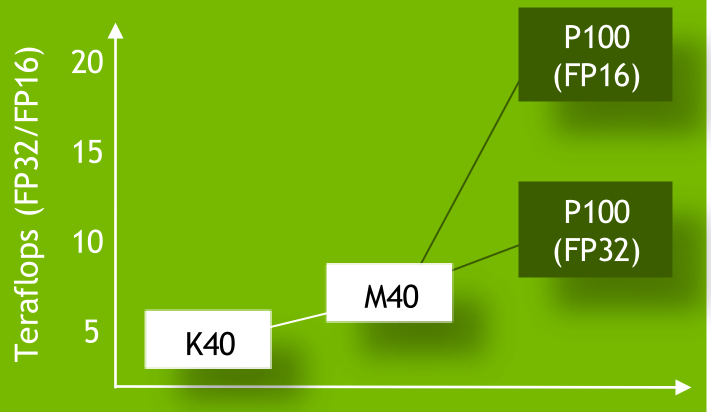
Critical Path Analysis
2x faster compile time
OpenACC profiling
Debug CUDA Apps on display GPU



GIANT LEAPS IN EVERYTHING

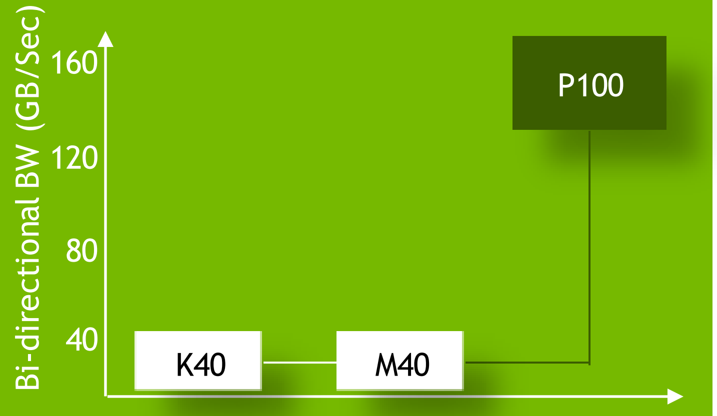
PASCAL ARCHITECTURE

21 Teraflops of FP16 for Deep Learning



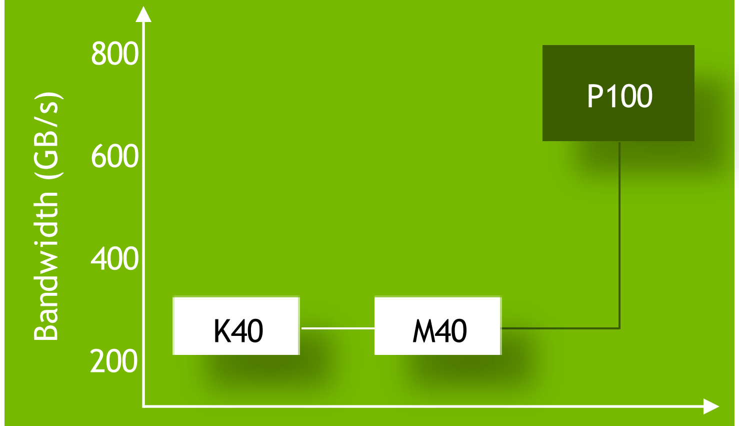
NVLINK

5x GPU-GPU Bandwidth



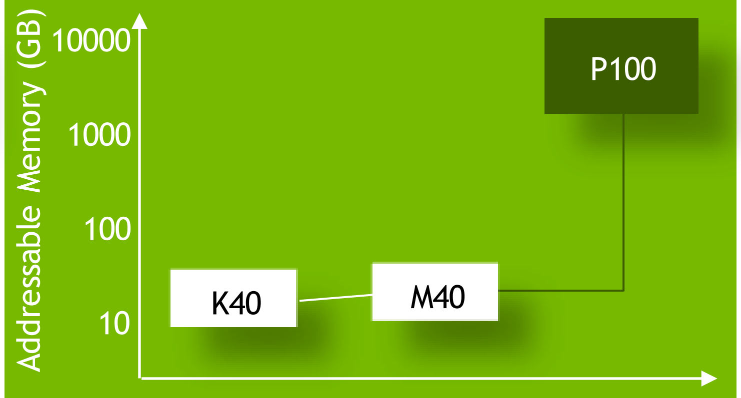
CoWoS HBM2 Stacked Mem

3x Higher for Massive Data Workloads



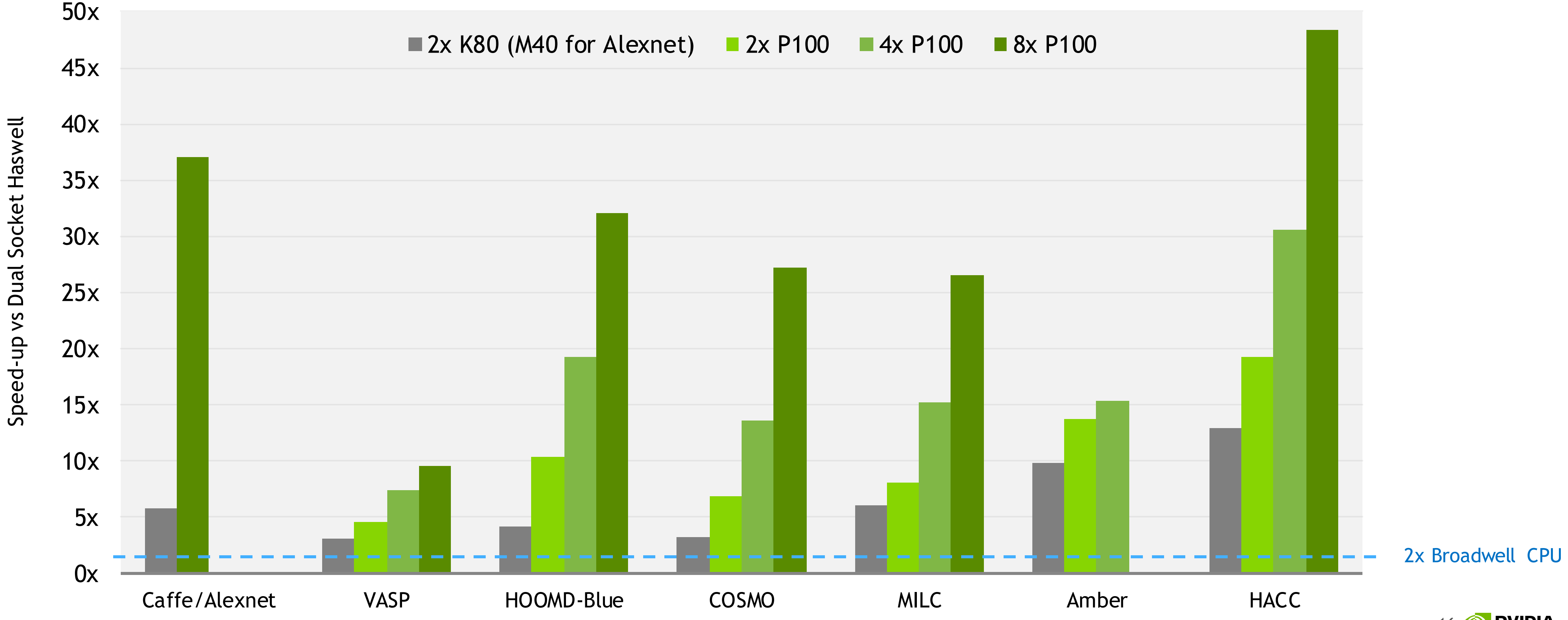
PAGE MIGRATION ENGINE

Virtually Unlimited Memory Space



HIGHEST ABSOLUTE PERFORMANCE DELIVERED

NVLink for Max Scalability, More than 45x Faster with 8x P100



PASCAL ARCHITECTURE

TESLA P100 ACCELERATOR

	
Compute	5.3 TF DP · 10.6 TF SP · 21.2 TF HP
Memory	HBM2: 720 GB/s · 16 GB
Interconnect	NVLink (up to 8 way) + PCIe Gen3
Programmability	Page Migration Engine Unified Memory
Availability	DGX-1: Order Now Cray, Dell, HP, IBM: Q1 2017

GPU PERFORMANCE COMPARISON

	P100	M40	K40
Double Precision TFlop/s	5.3	0.2	1.4
Single Precision TFlop/s	10.6	7.0	4.3
Half Precision Tflop/s	21.2	NA	NA
Memory Bandwidth (GB/s)	720	288	288
Memory Size	16GB	12GB, 24GB	12GB

IEEE 754 FLOATING POINT ON GP100

3 sizes, 3 speeds, all fast

Feature	New Half precision	Single precision	Double precision
Layout	s5.10	s8.23	s11.52
Issue rate	pair every clock	1 every clock	1 every 2 clocks
Subnormal support	Yes	Yes	Yes
Atomic Addition	Yes	Yes	New Yes

HALF-PRECISION FLOATING POINT (FP16)

- 16 bits

s	e	x	p			.	f	r	a	c						

  - 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

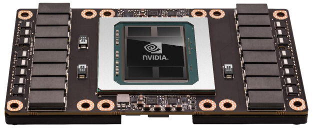
Radio Astronomy

Sensor Data

Image Processing

END-TO-END PRODUCT FAMILY

HYPERSCALE HPC

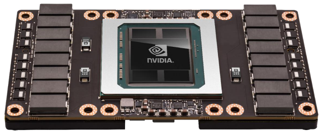


Training - Tesla P100

Inference - Tesla P40 & P4

Deep learning training & inference

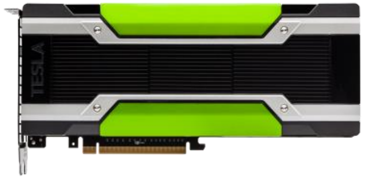
STRONG-SCALE HPC



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

NVLINK

NVLINK - GPU CLUSTER

Two fully connected quads,
connected at corners

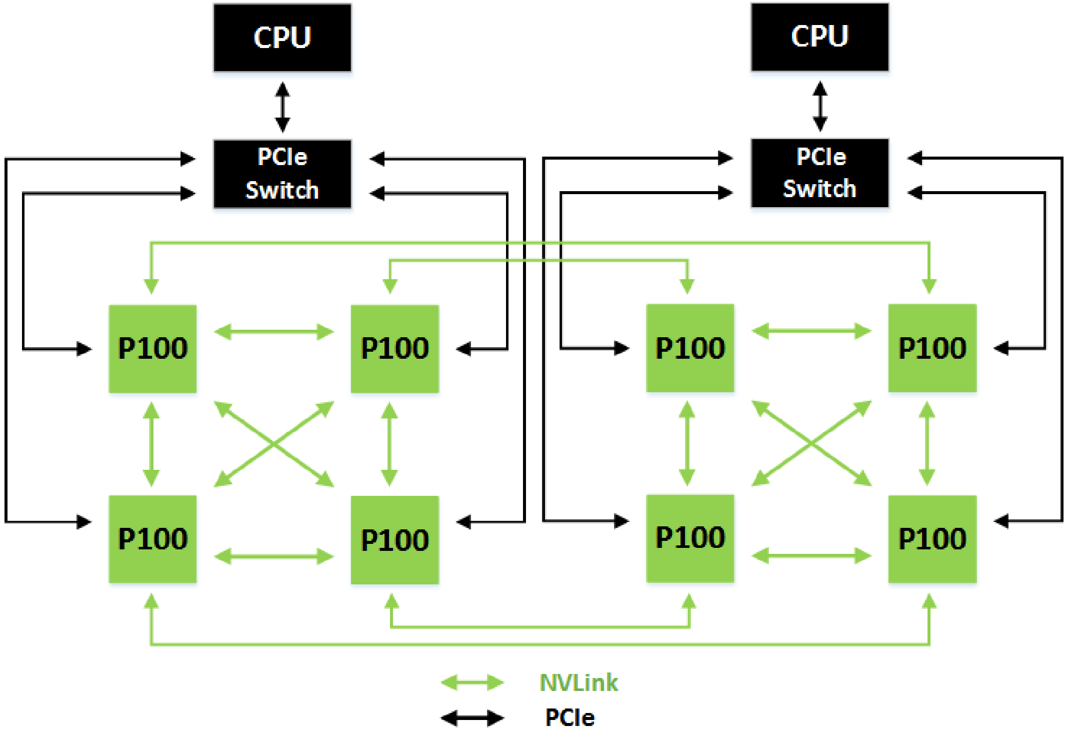
160GB/s per GPU bidirectional to Peers

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

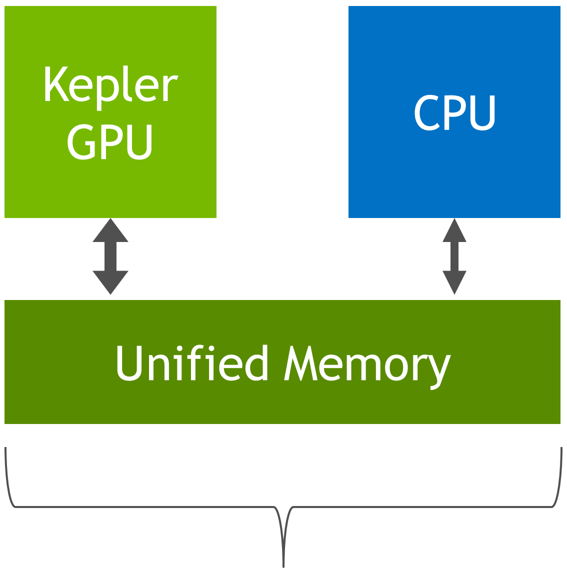
- Can handle thousands of simultaneous page faults

Up to 2 MB page size

- Better TLB coverage of GPU memory

KEPLER/MAXWELL UNIFIED MEMORY

CUDA 6+



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

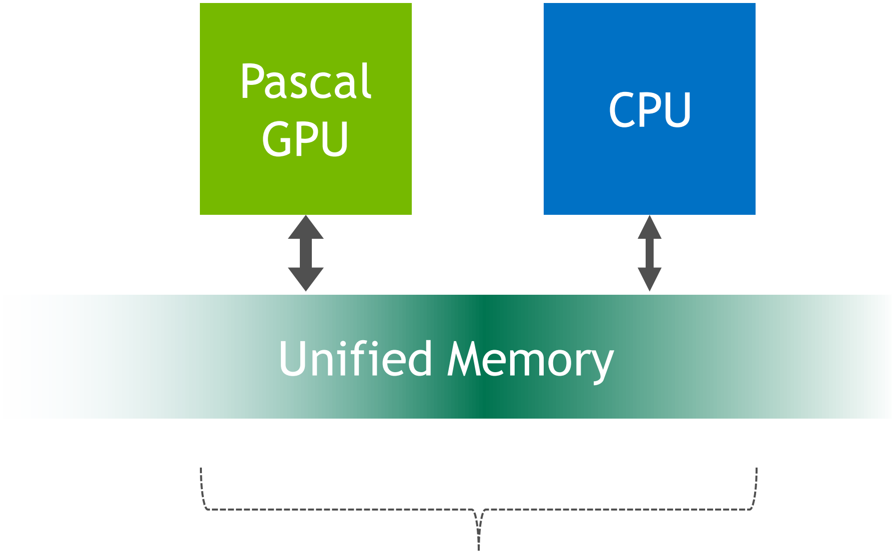
Performance Through Data Locality

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

PASCAL UNIFIED MEMORY

Large datasets, simple programming, High Performance

CUDA 8



Allocate Beyond
GPU Memory Size

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

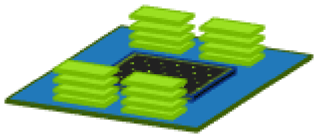
- CPU/GPU Data coherence
- Unified memory atomic operations

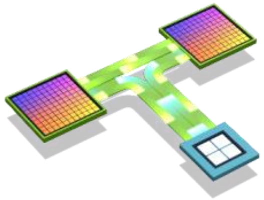
CUDA 8

CUDA 8 - WHAT'S NEW

P100 Support

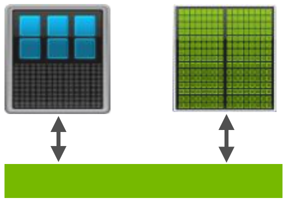
New Pascal Architecture
Stacked Memory
NVLINK
FP16 math





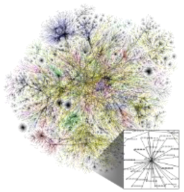
Unified Memory

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



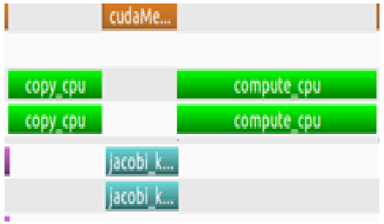
Libraries

New nvGRAPH library
cuBLAS improvements for Deep Learning



Developer Tools

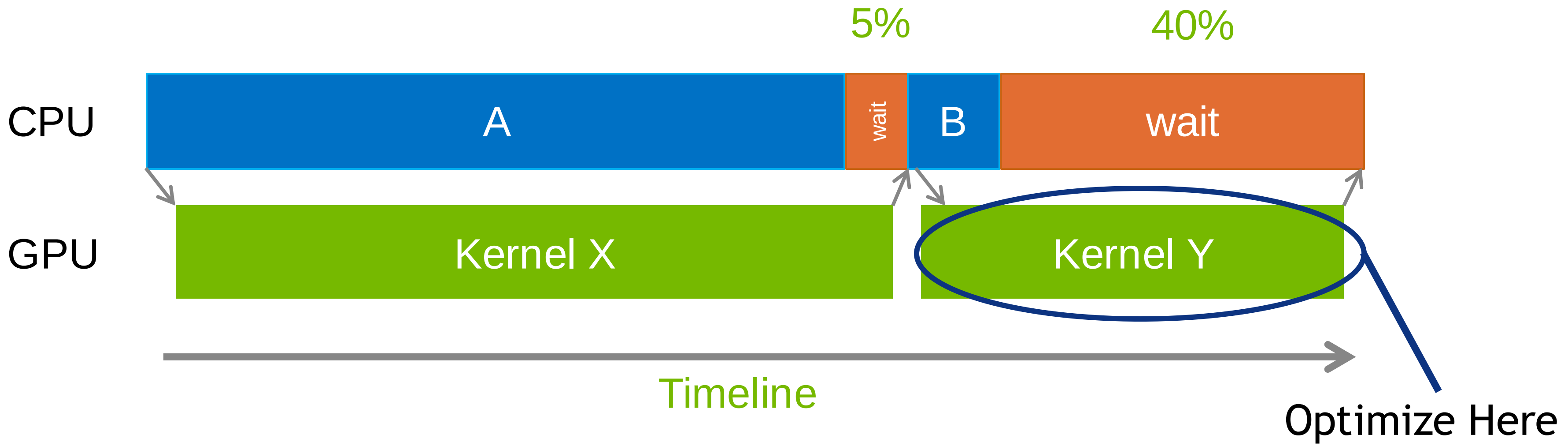
Critical Path Analysis
2x faster compile time
OpenACC profiling
Debug CUDA Apps on display GPU



ENHANCED PROFILING

DEPENDENCY ANALYSIS

Easily Find the Critical Kernel To Optimize



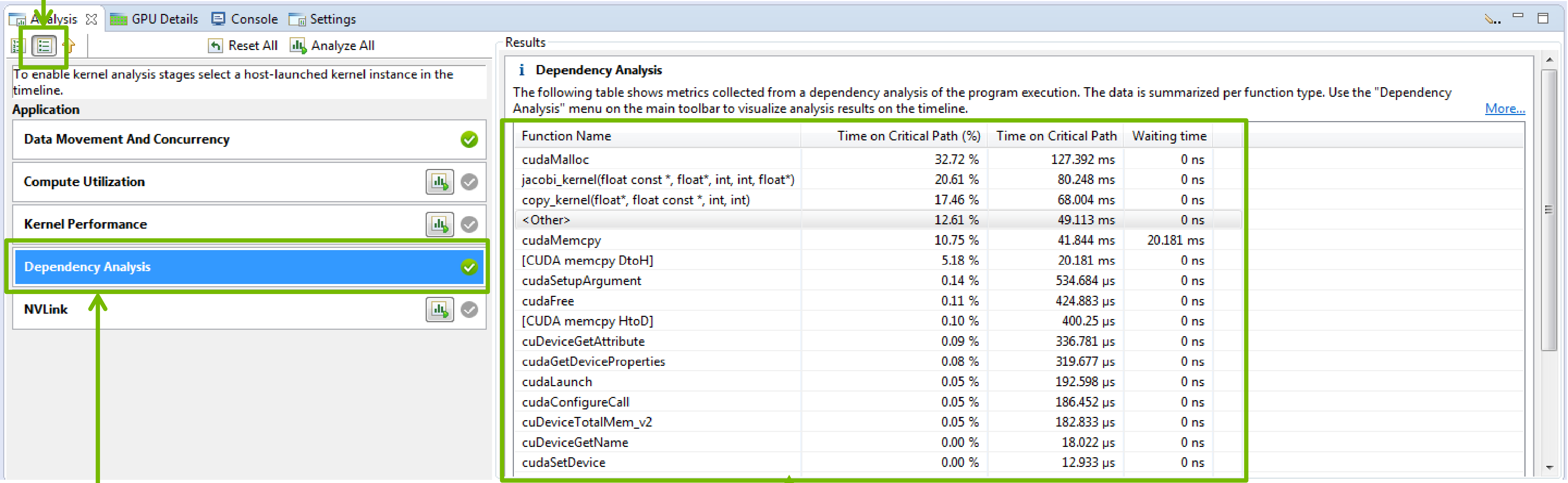
The longest running kernel is not always the most critical optimization target

DEPENDENCY ANALYSIS

Visual Profiler

Unguided Analysis

Generating critical path



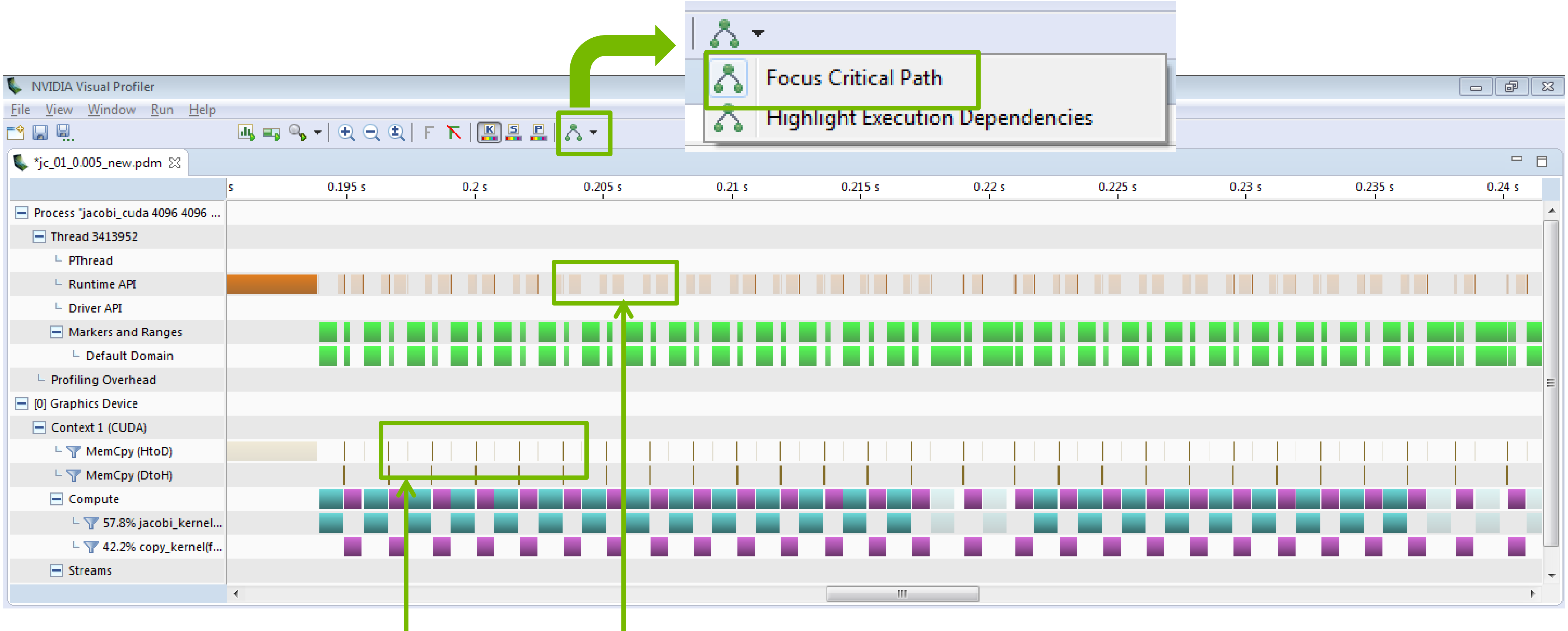
Function Name	Time on Critical Path (%)	Time on Critical Path	Waiting time
cudaMalloc	32.72 %	127.392 ms	0 ns
jacobi_kernel(float const *, float*, int, int, float*)	20.61 %	80.248 ms	0 ns
copy_kernel(float*, float const *, int, int)	17.46 %	68.004 ms	0 ns
<Other>	12.61 %	49.113 ms	0 ns
cudaMemcpy	10.75 %	41.844 ms	20.181 ms
[CUDA memcpy DtoH]	5.18 %	20.181 ms	0 ns
cudaSetupArgument	0.14 %	534.684 µs	0 ns
cudaFree	0.11 %	424.883 µs	0 ns
[CUDA memcpy HtoD]	0.10 %	400.25 µs	0 ns
cuDeviceGetAttribute	0.09 %	336.781 µs	0 ns
cudaGetDeviceProperties	0.08 %	319.677 µs	0 ns
cudaLaunch	0.05 %	192.598 µs	0 ns
cudaConfigureCall	0.05 %	186.452 µs	0 ns
cuDeviceTotalMem_v2	0.05 %	182.833 µs	0 ns
cuDeviceGetName	0.00 %	18.022 µs	0 ns
cudaSetDevice	0.00 %	12.933 µs	0 ns

Dependency Analysis

Functions on critical path

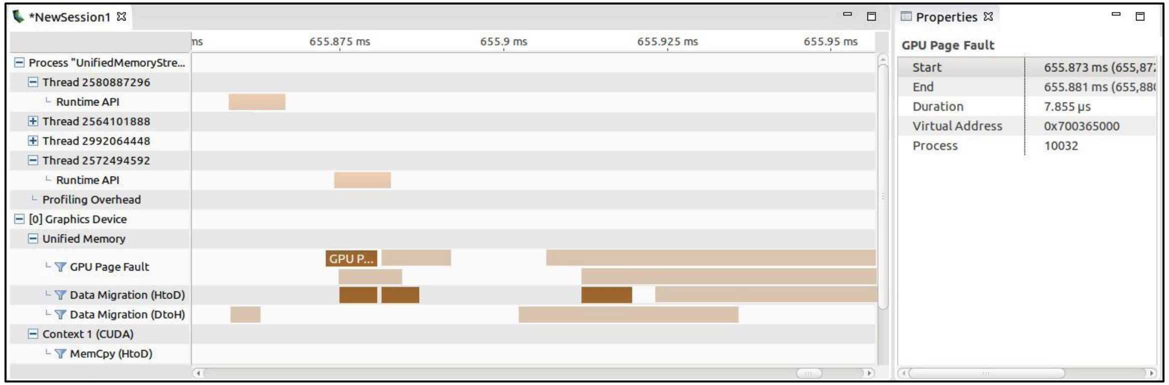
DEPENDENCY ANALYSIS

Visual Profiler

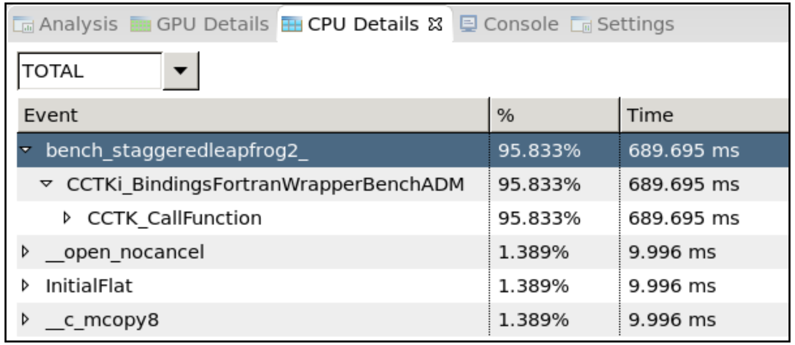


APIs, GPU activities not in critical path are greyed out

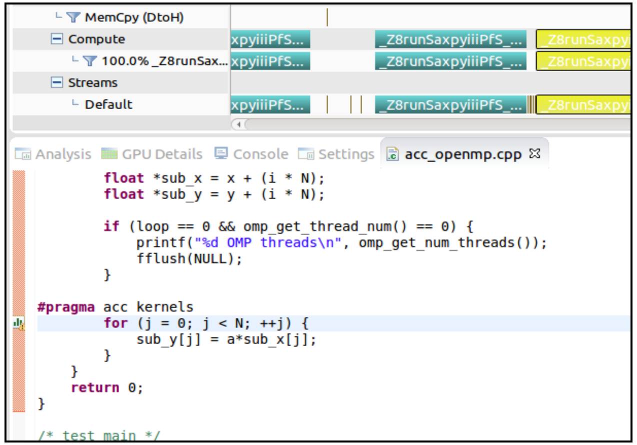
MORE CUDA 8 PROFILER FEATURES



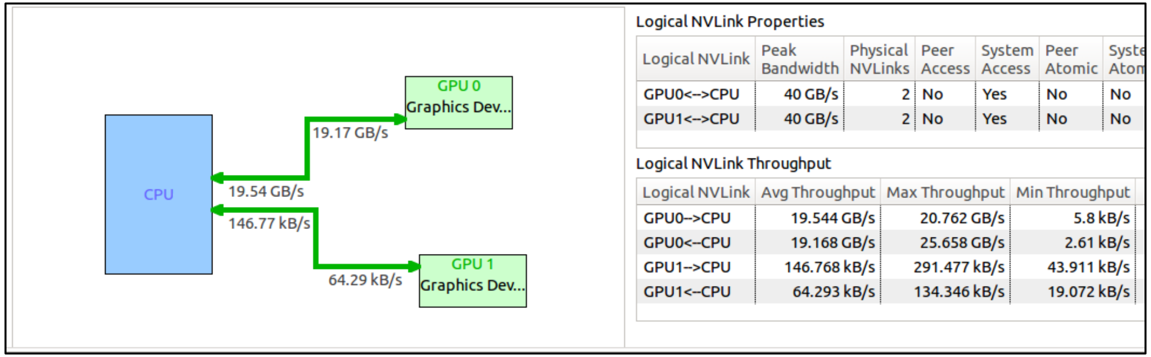
Unified Memory Profiling



CPU Profiling



OpenACC Profiling



NVLink Topology and Bandwidth profiling

OPENACC

World’s Only Performance Portable Programming Model for HPC

Add Simple Compiler Hint

```
main()
{
  <serial code>
  #pragma acc kernels
  {
    <parallel code>
  }
}
```

Simple

ARM

PEZY

POWER

Sunway

x86 CPU

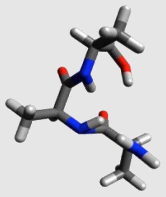
x86 Xeon Phi

NVIDIA GPU

Portable

LSDALTON

Simulation of molecular energies



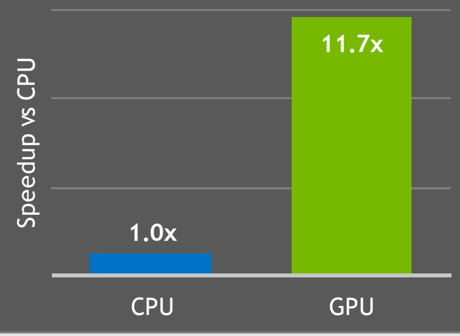
Quicker Development

Lines of Code Modified
<100 Lines

# of Weeks Required
1 Week

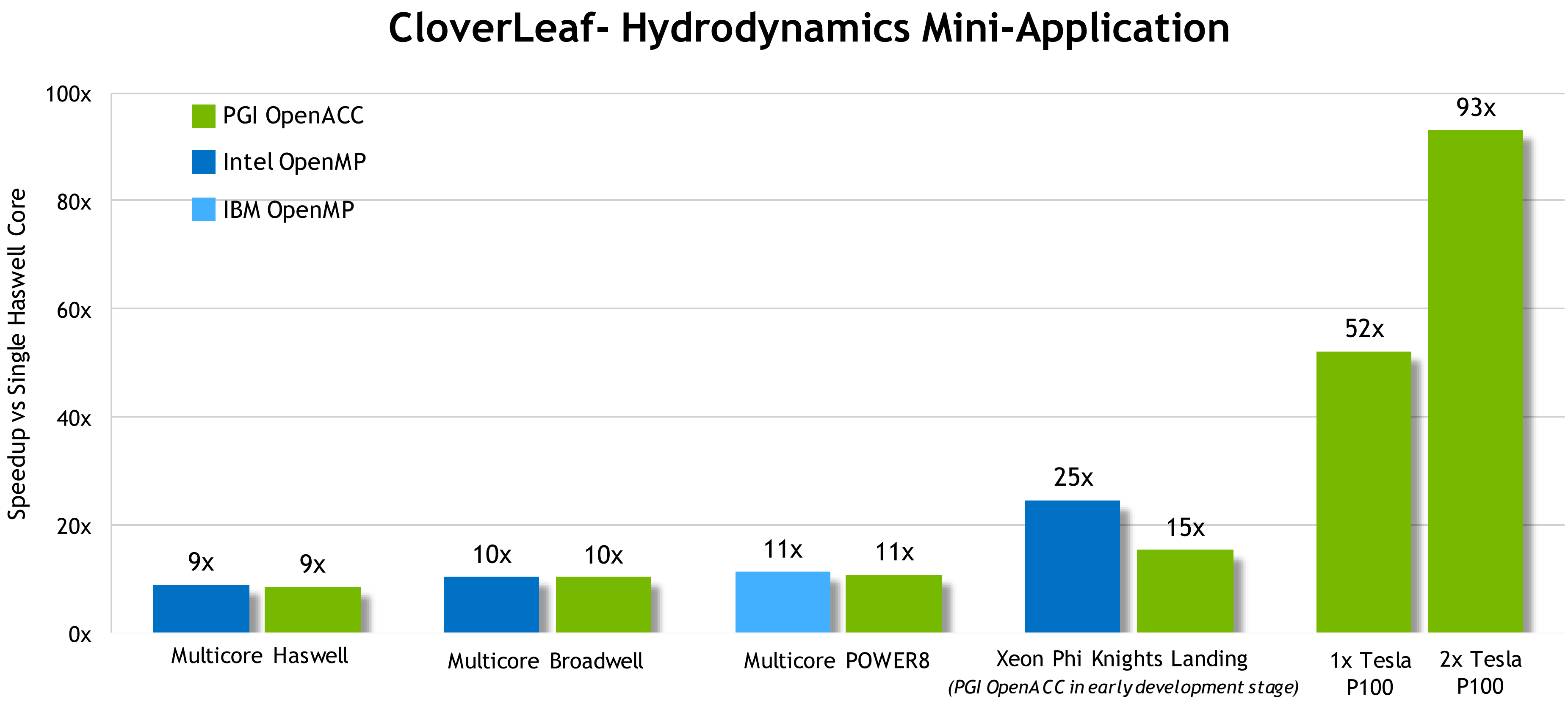
Big Performance

CCSD(T) Module, Alanine-3
Titan System: AMD CPU vs Tesla K20X

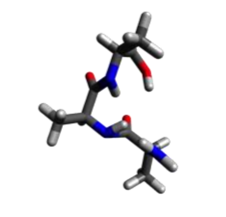


Powerful

SINGLE OPENACC CODE RUNS ON ALL CPU & GPU PLATFORMS



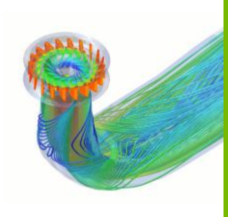
CloverLeaf bm32 dataset



LSDalton

Quantum Chemistry

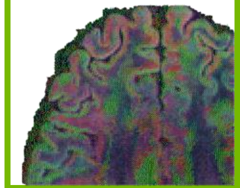
12X speedup
in 1 week



Numeca

CFD

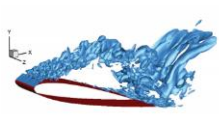
10X faster kernels
2X faster app



PowerGrid

Medical Imaging

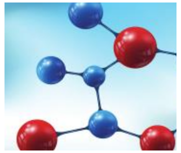
40 days to
2 hours



INCOMP3D

CFD

3X speedup



NekCEM

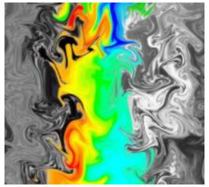
Computational Electromagnetics

2.5X speedup
60% less energy

COSMO

Climate Weather

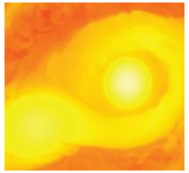
40X speedup
3X energy efficiency



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

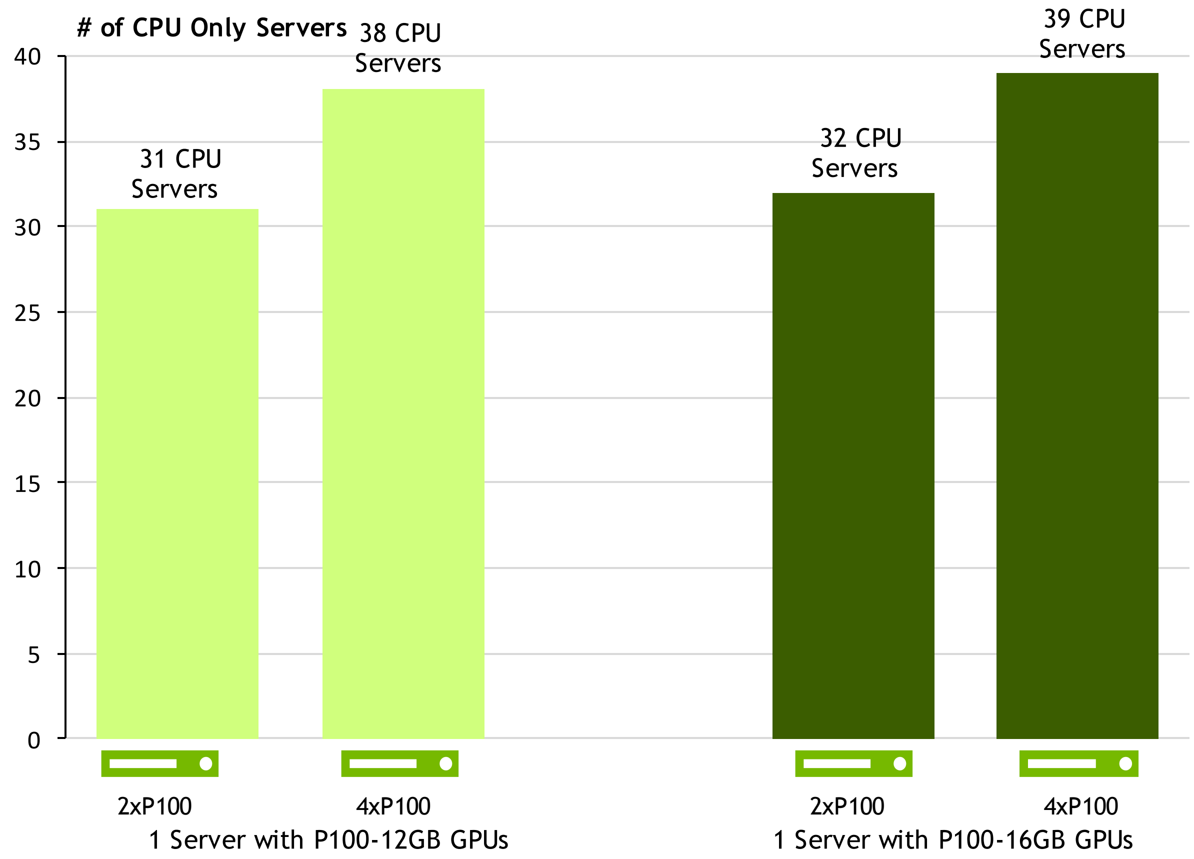
	FREE PGI Community EDITION	PGI Professional EDITION	PGI Enterprise EDITION
PROGRAMMING MODELS OpenACC, CUDA Fortran, OpenMP, C/C++/Fortran Compilers and Tools	✓	✓	✓
PLATFORMS x86, OpenPOWER, NVIDIA GPU	✓	✓	✓
UPDATES	1-2 times a year	6-9 times a year	6-9 times a year
SUPPORT	User Forums	PGI Support	PGI Enterprise Services
LICENSE	Annual	Perpetual	Volume/Site

PERFORMANCE

MOLECULAR DYNAMICS

AMBER Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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.

AMBER

Molecular Dynamics

Suite of programs to simulate molecular dynamics on biomolecule

VERSION
16.3

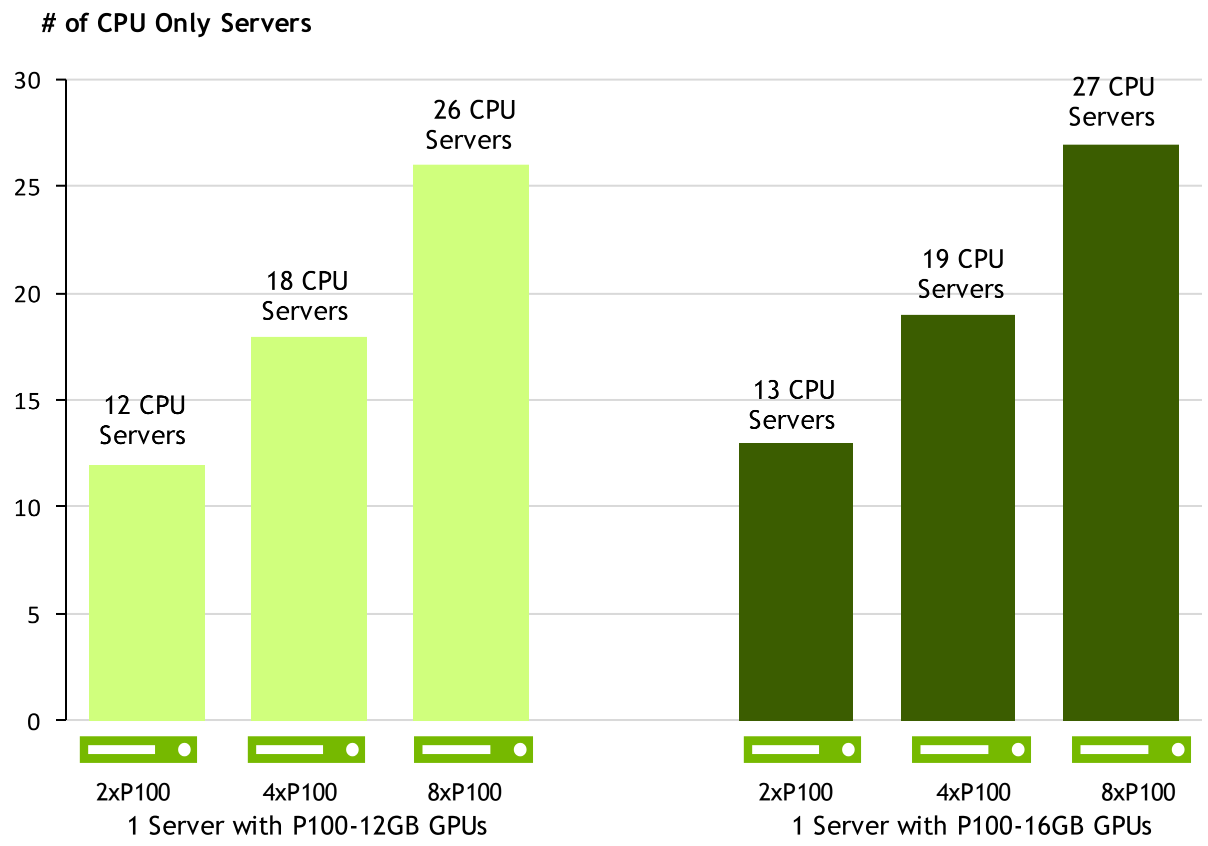
ACCELERATED FEATURES
PMEMD Explicit Solvent & GB; Explicit & Implicit Solvent, REMD, aMD

SCALABILITY
Multi-GPU and Single-Node

More Information
http://ambermd.org/gpus

HOOMD-Blue Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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
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

Molecular Dynamics

Particle dynamics package written grounds up for GPUs

VERSION
1.3.3

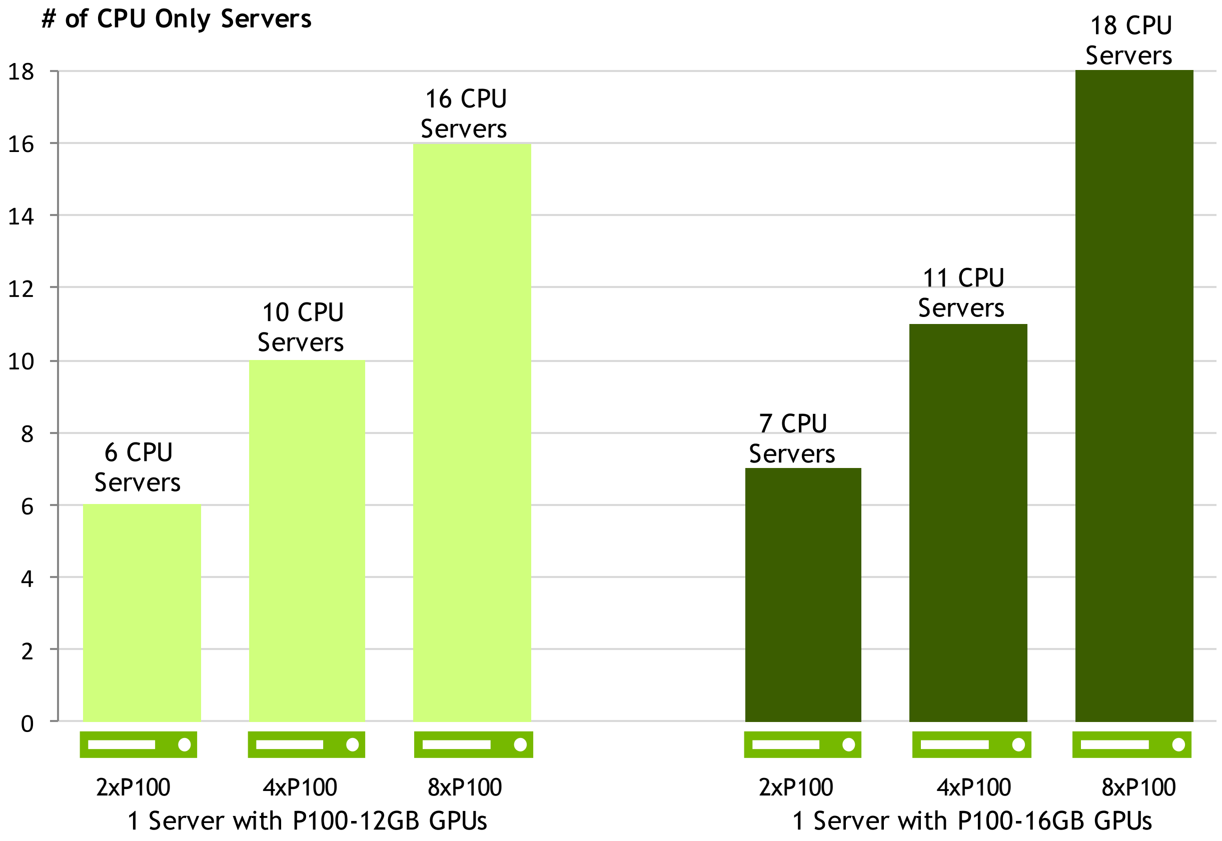
ACCELERATED FEATURES
CPU & GPU versions available

SCALABILITY
Multi-GPU and Multi-Node

More Information
http://codeblue.umich.edu/hoomd-blue/index.html

LAMMPS Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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.

LAMMPS

Molecular Dynamics

Classical molecular dynamics package

VERSION
2016

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

SCALABILITY
Multi-GPU and Multi-Node

More Information
http://lammps.sandia.gov/index.html

GROMACS Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers

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
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

Molecular Dynamics

Simulation of biochemical molecules with complicated bond interactions

VERSION
5.1.2

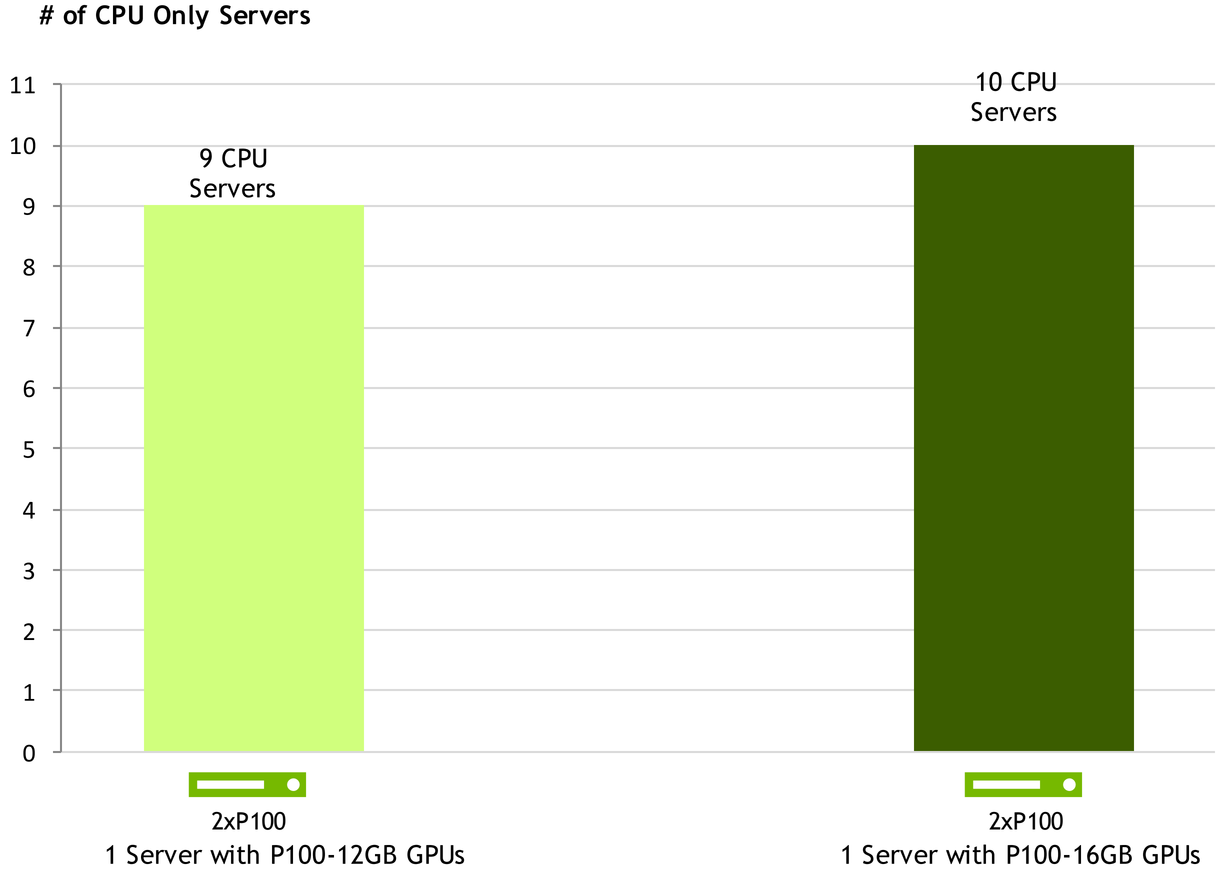
ACCELERATED FEATURES
PME, Explicit & Implicit Solvent

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

More Information
http://www.gromacs.org l

NAMD Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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
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.

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

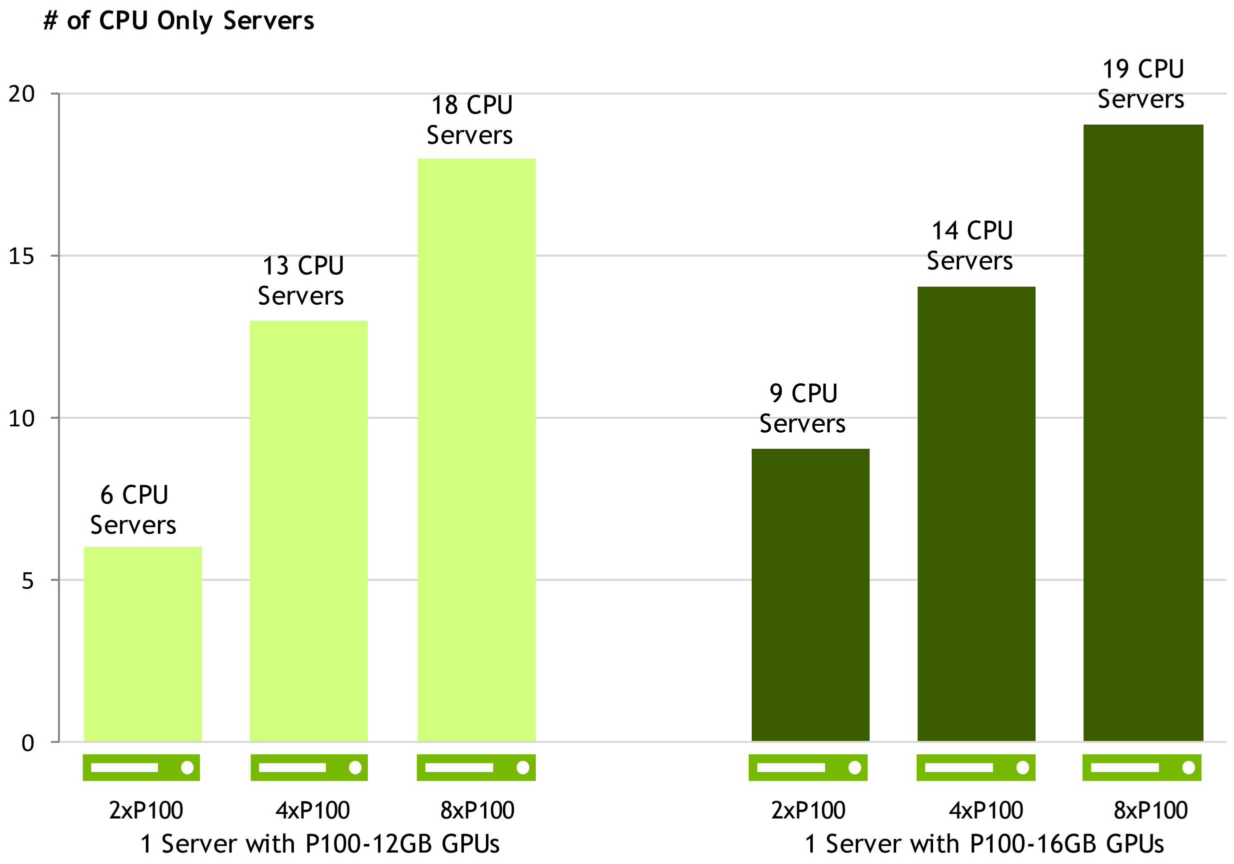
SCALABILITY
Up to 100M atom capable, multi-GPU, Scale
Scales to 2xP100

More Information
http://www.ks.uiuc.edu/Research/namd/

MATERIALS SCIENCE

VASP Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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

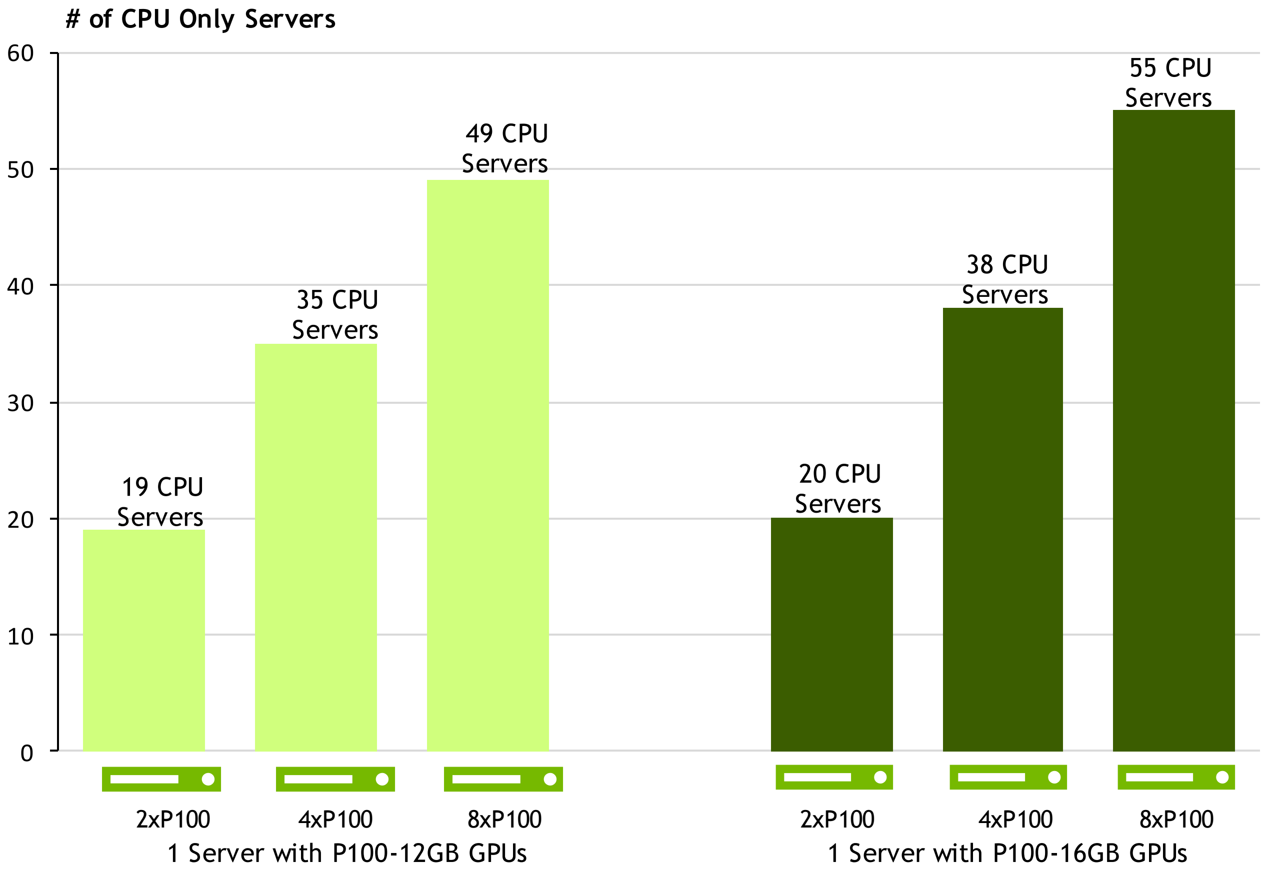
SCALABILITY
Multi-GPU and Multi-Node

More Information
http://www.vasp.at/index.php/news/44-administrative/115-new-release-vasp-5-4-1-with-gpu-support

BENCHMARKS

Linpack Performance Equivalency

Single GPU Server vs Multiple CPU-Only Servers



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
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.

Linpack
Benchmark

Measures floating point computing power

VERSION
2.1

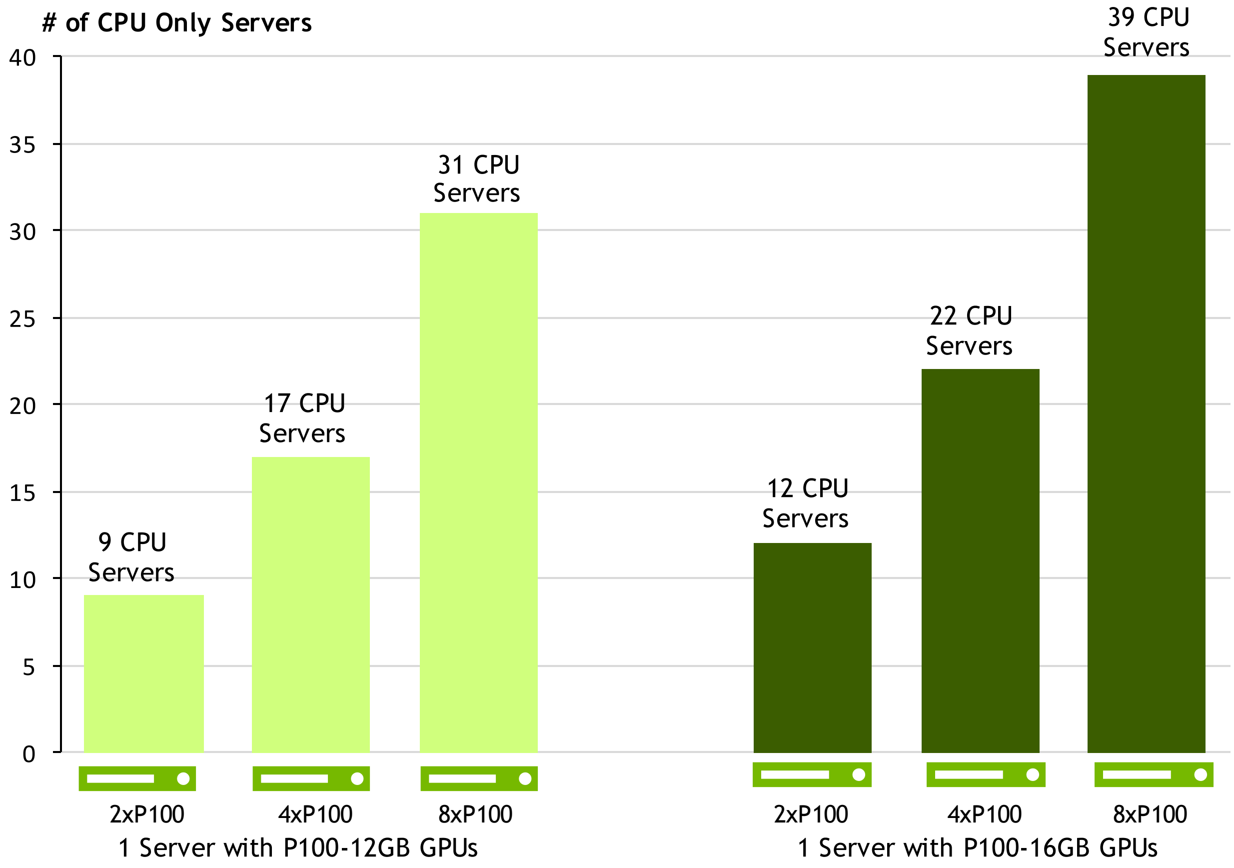
ACCELERATED FEATURES
All

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



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.

HPCG

Benchmark

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

VERSION
3.0

ACCELERATED FEATURES
All

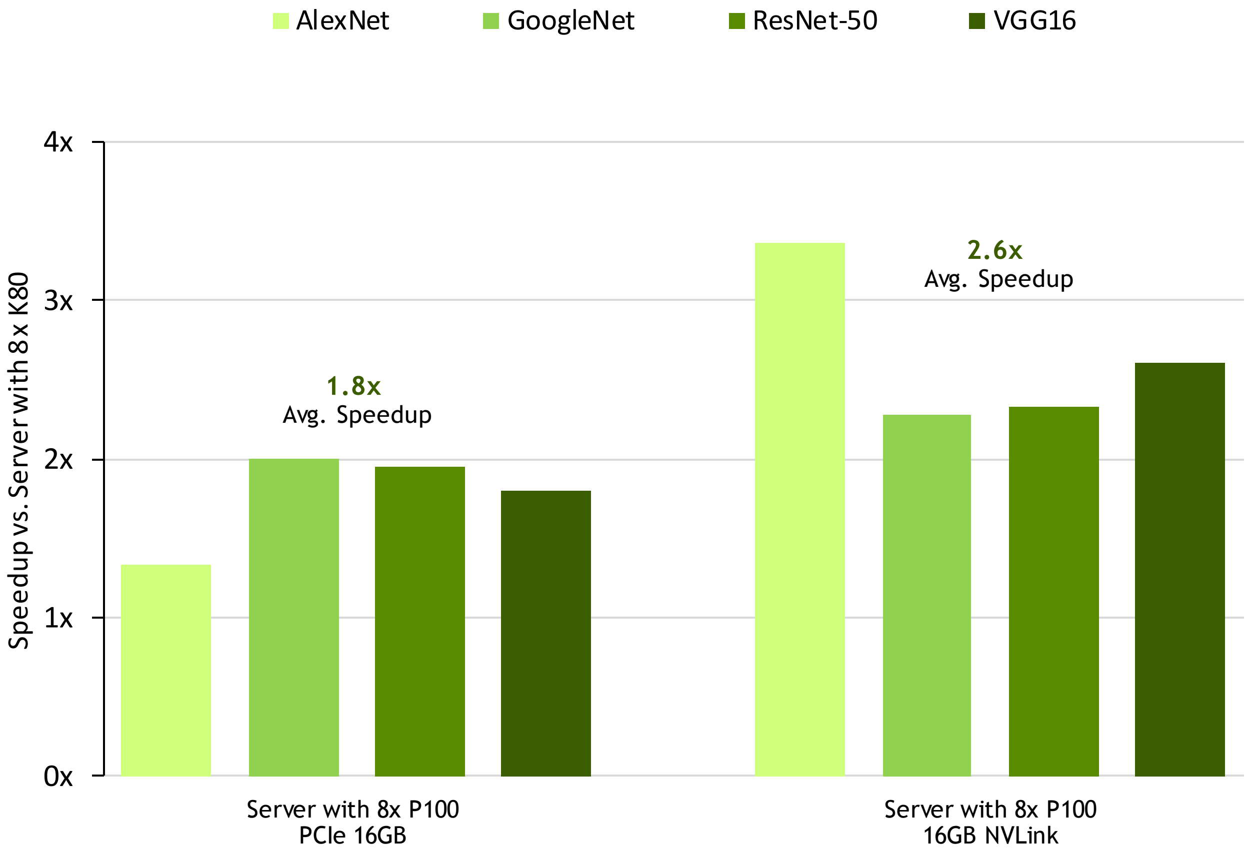
SCALABILITY
Multi-GPU and Multi-Node

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

DEEP LEARNING

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

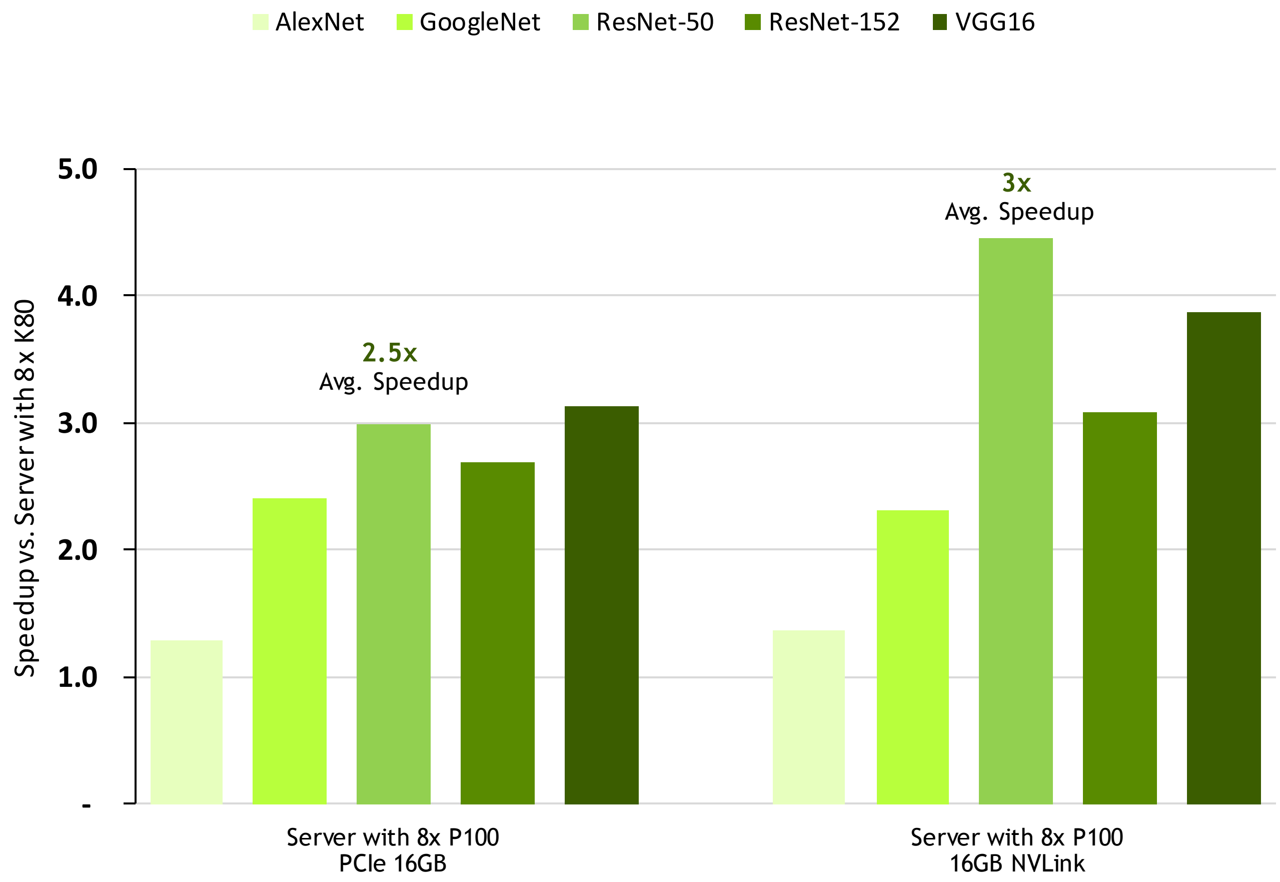
ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU

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

TensorFlow 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), ResNet-152 (32), VGG-16 (32)

TensorFlow

Deep Learning Training

An open-source software library for numerical computation using data flow graphs.

VERSION
1.0

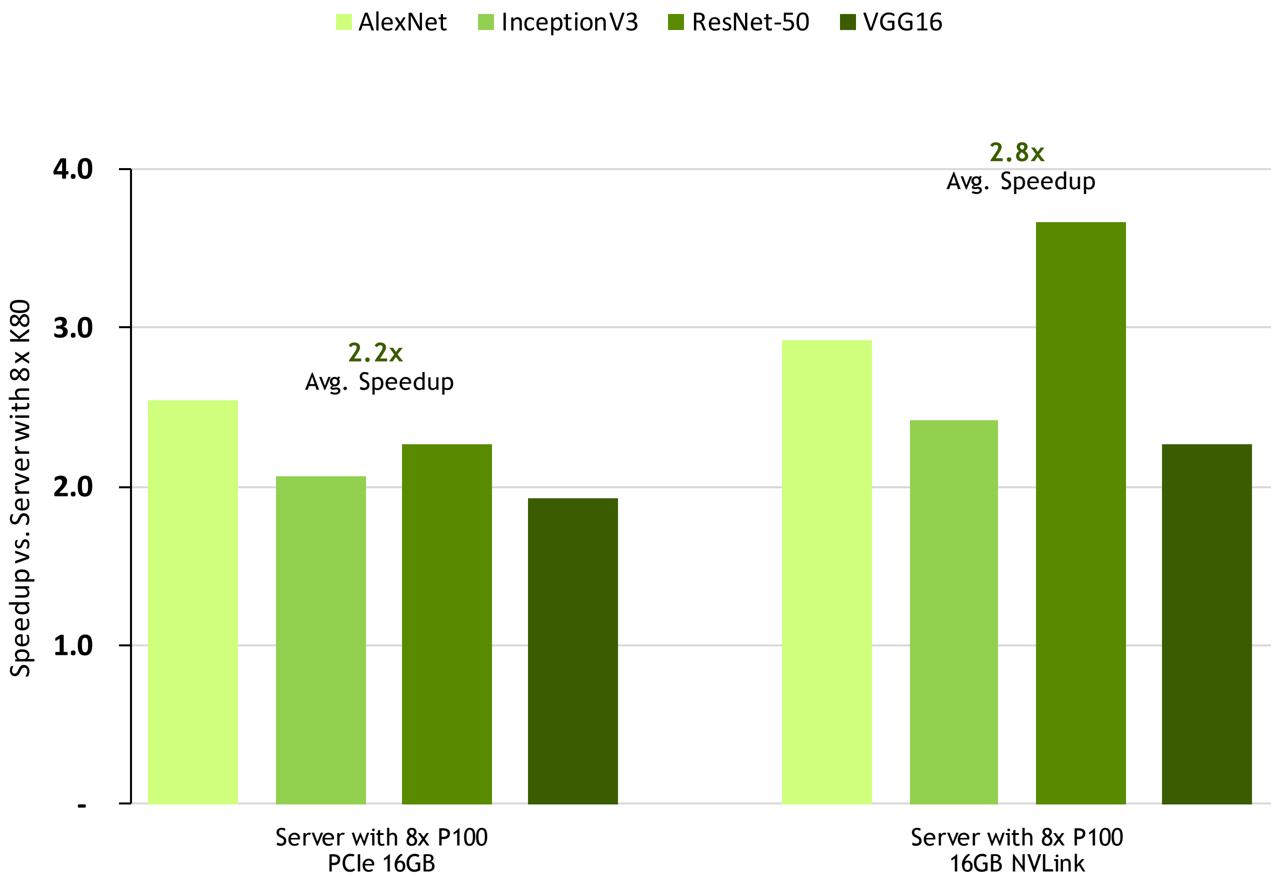
ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU and multi-node

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

Torch 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), InceptionV3 (64), ResNet-50 (64), VGG-16 (32)

Torch

Deep Learning Training

A scientific computing framework with wide support for machine learning algorithms that puts GPUs first.

VERSION
7.0

ACCELERATED FEATURES
Full framework accelerated

SCALABILITY
Multi-GPU

More Information
https://www.torch.ch

CNTK 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), ResNet-50 (64)

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
www.microsoft.com/en-us/research/product/cognitive-toolkit/

DEEP LEARNING SOFTWARE

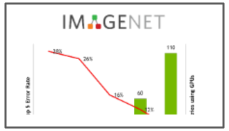
POWERING THE DEEP LEARNING ECOSYSTEM

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

KERAS

MatConvNet

Microsoft CNTK

MINERVA

mxnet

OpenDeep

Purine

Pylearn2

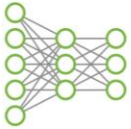
TensorFlow

torch

theano

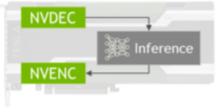
NVIDIA DEEP LEARNING SDK

cuDNN

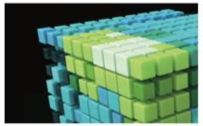


TensorRT

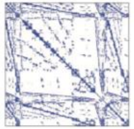
DeepStream SDK



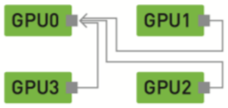
cuBLAS



cuSPARSE

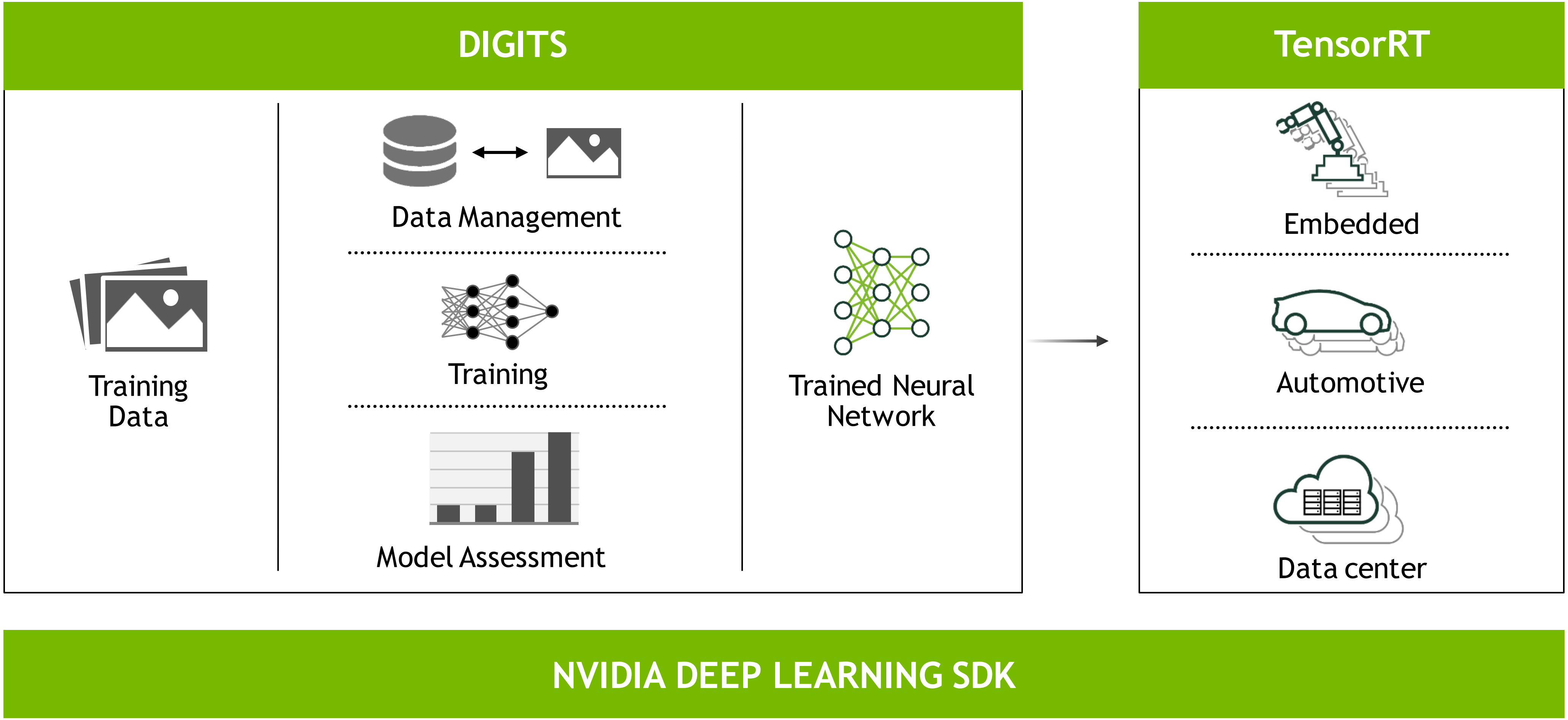


NCCL



developer.nvidia.com/deep-learning-software

NVIDIA DEEP LEARNING SOFTWARE PLATFORM



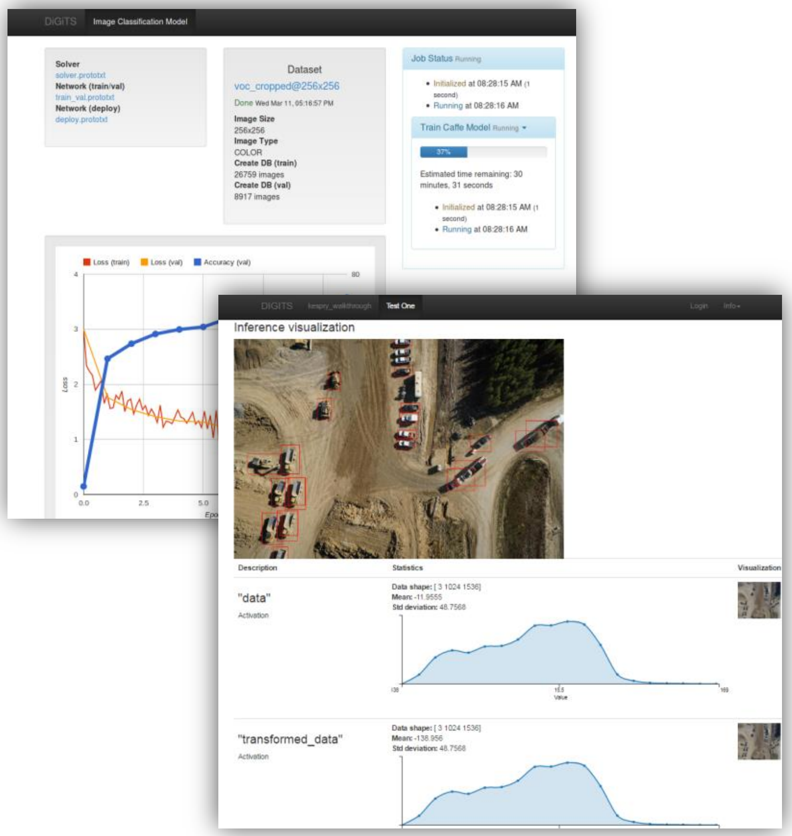
developer.nvidia.com/deep-learning-software

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

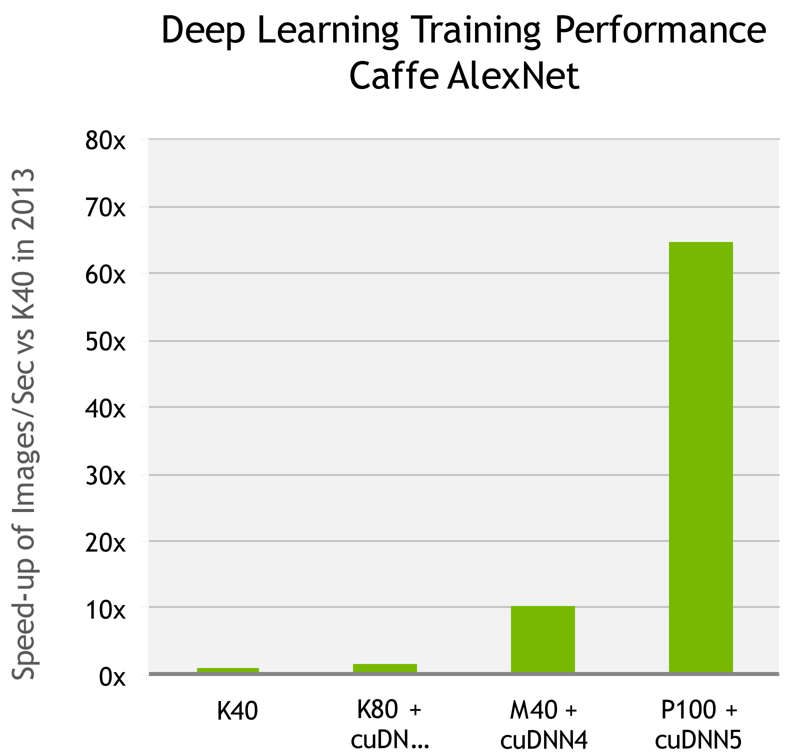
High performance building blocks for deep learning frameworks

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



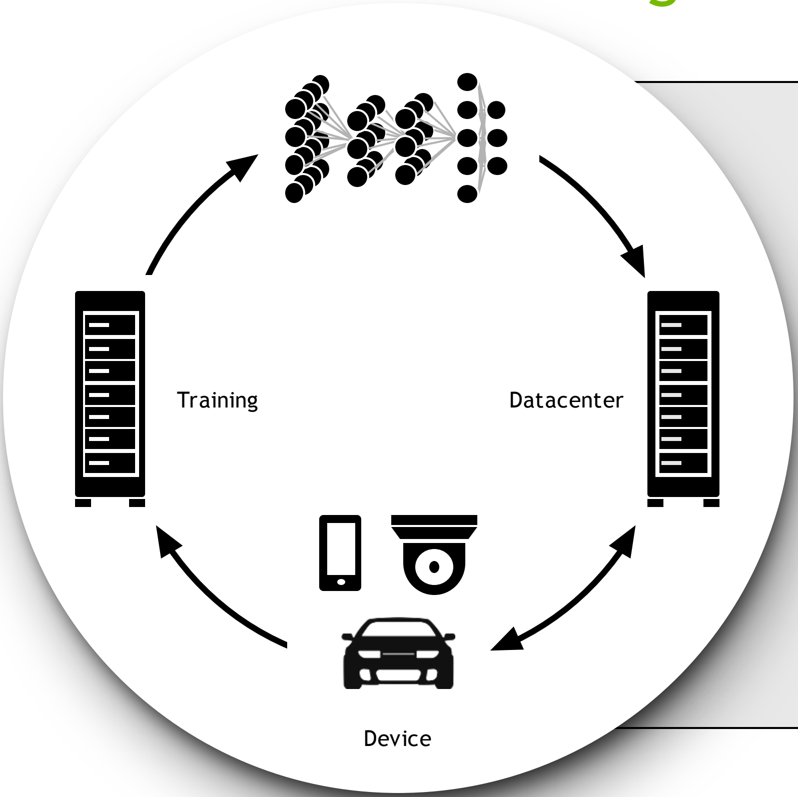
AlexNet training throughput on CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04
M40 bar: 8x M40 GPUs in a node, P100: 8x P100 NVLink-enabled

"NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time."

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley

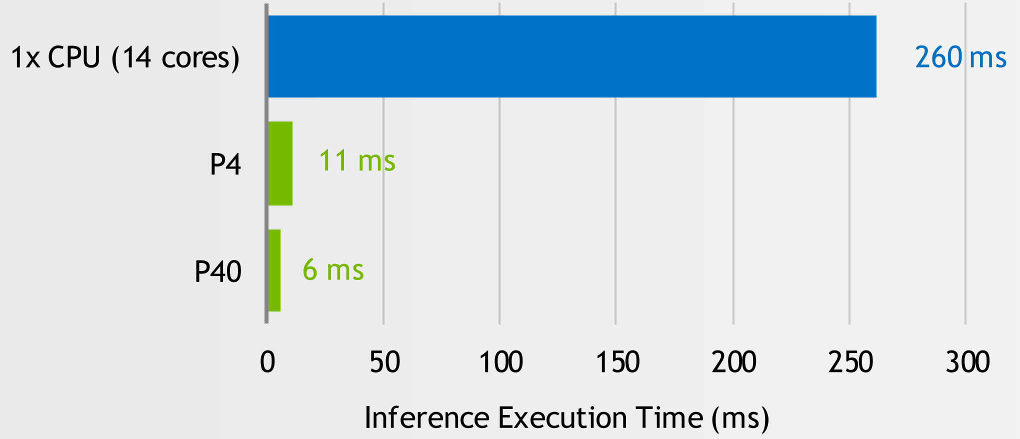
INTRODUCING NVIDIA TensorRT

High Performance Inference Engine



User Experience: Instant Response

45x Faster with Pascal + TensorRT



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 IntelCaffe Github: https://github.com/intel/caffe/tree/master/models/mkl2017_vgg_19
CPU: IntelCaffe, 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

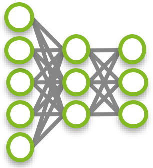
Dual 10GbE, Quad IB 100Gb

3RU – 3200W

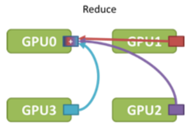
NVIDIA DGX-1 SOFTWARE STACK

Optimized for Deep Learning Performance

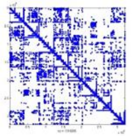
Accelerated Deep Learning



cuDNN

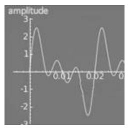


NCCL



cuSPARSE

cuBLAS



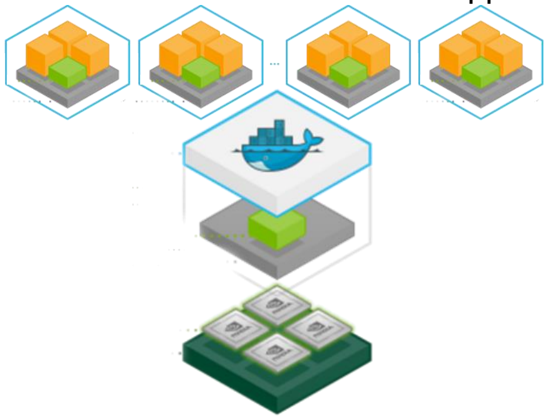
cuFFT

Container Based Applications

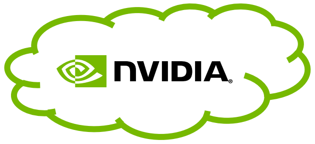
Digits

DL Frameworks

GPU Apps



NVIDIA Cloud Management

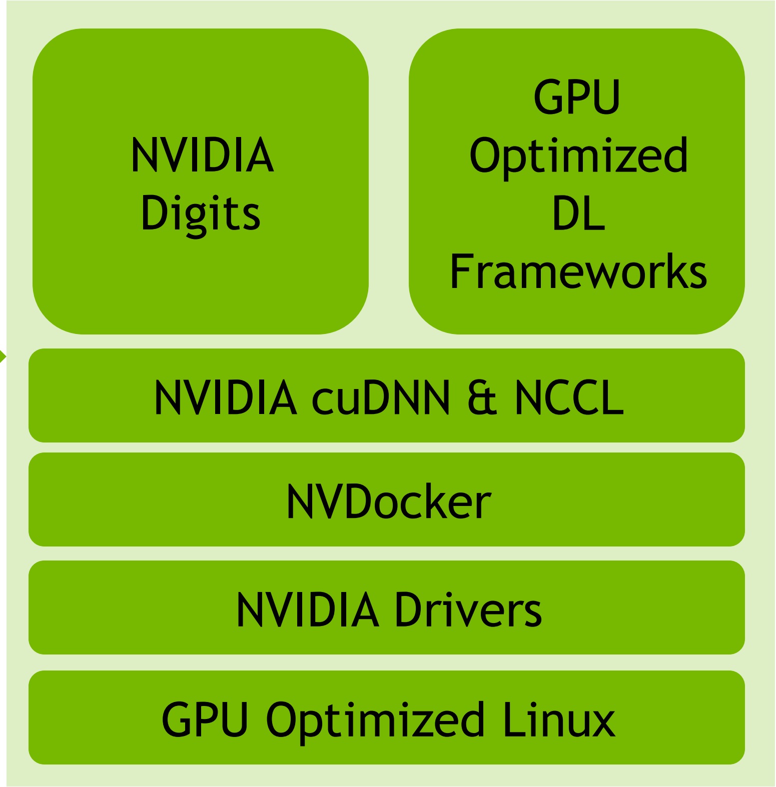


NVIDIA DGX-1 SOFTWARE STACK

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



NVIDIA DGX-1

ACCELERATE EVERY FRAMEWORK

ACADEMIA

CAFFE	TORCH
Berkeley UNIVERSITY OF CALIFORNIA	NYU

THEANO	MATCONVNET
Université de Montréal	UNIVERSITY OF OXFORD

MOCHA.JL	PURINE
Massachusetts Institute of Technology	NUS National University of Singapore

MINERVA	MXNET*
NYU, Microsoft	UNIVERSITY of WASHINGTON, Carnegie Mellon University

TENSORFLOW	TORCH	CNTK
Google	facebook	Microsoft

START-UPS

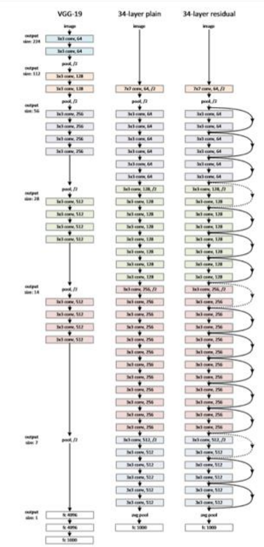
CHAINER	DL4J	KERAS	OPENDEEP
Preferred Networks	SKYMIND	SCHULTS LABORATORIES	VITRUVIAN

NVIDIA DGX-1

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai

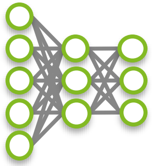
BENEFITS FOR AI RESEARCHERS

Fastest
DL Supercomputer

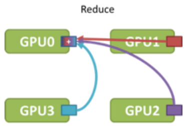


Design
Big Networks

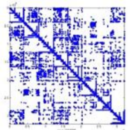
Reduce
Training Times



cuDNN

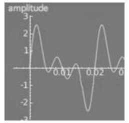


NCCL



cuSPARSE

cuBLAS

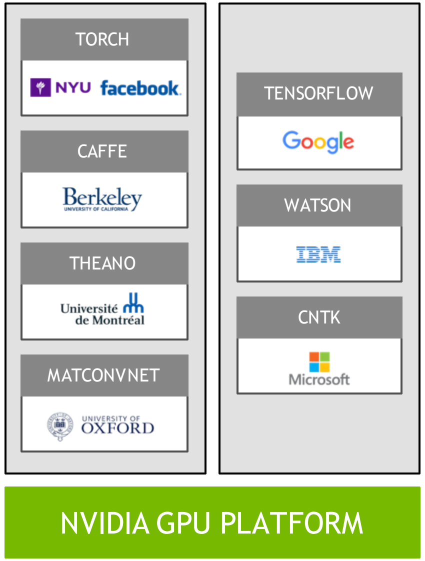


cuFFT

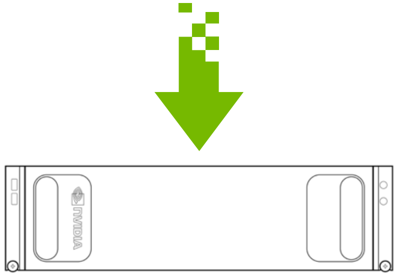
DL SDK
Ongoing Updates

BENEFITS FOR INDUSTRY DATA SCIENTISTS

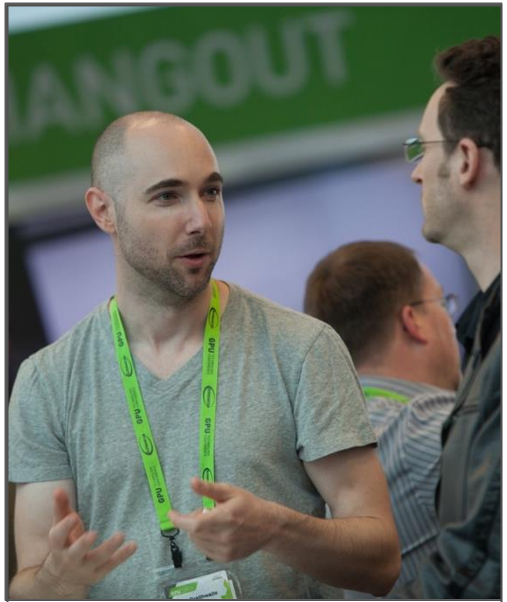
Purpose Built DL Supercomputer



Accelerates All Major Frameworks



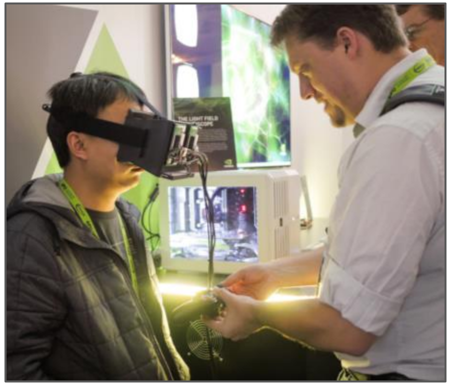
Simple Appliance, Software Updates



NVIDIA Expert Support

GPU TECHNOLOGY CONFERENCE

May 8 - 11, 2017 | Silicon Valley | #GTC17
www.gputechconf.com

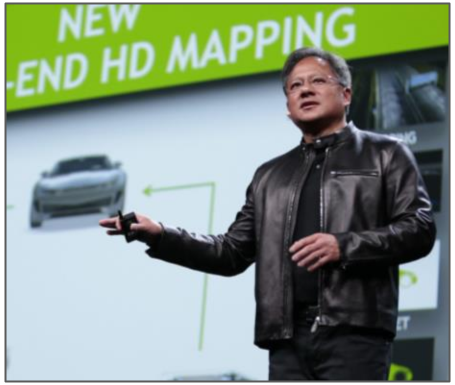


CONNECT

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

NVIDIA