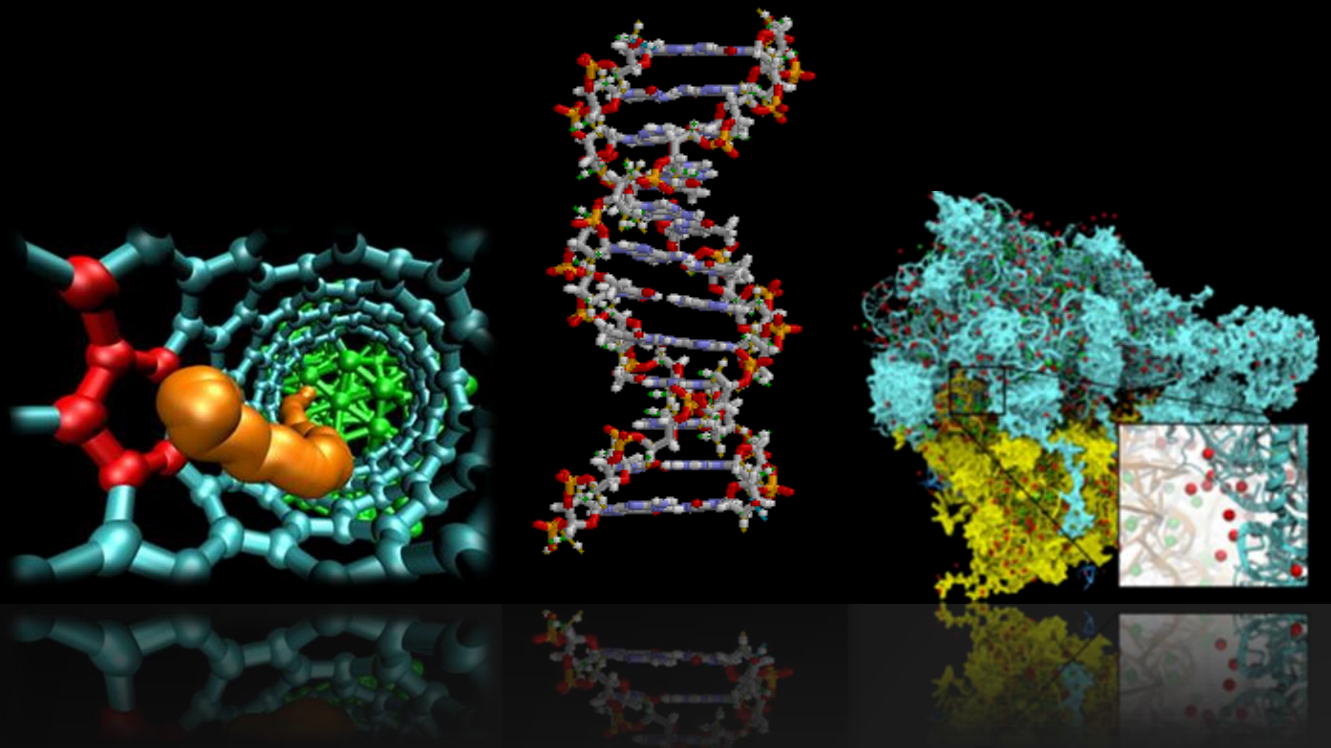
NVIDIA Computational Chemistry & Biology



Mark Berger
Senior Alliance Manager
Life and Material Sciences
mberger@nvidia.com

Updated: Aug. 4, 2015

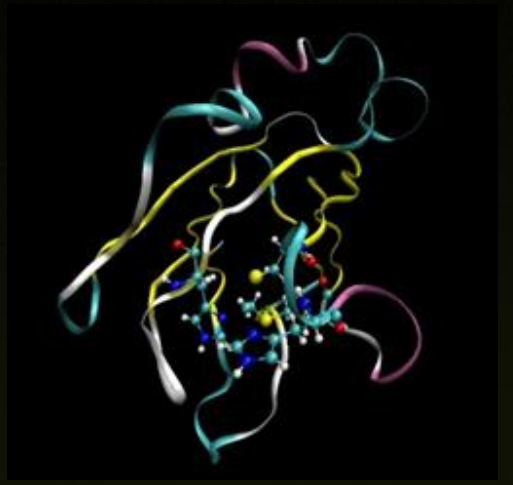
Four Cool Things

- Molecular Dynamics and Quantum Update
- A bit about NVIDIA future technologies
- Deep Learning
- OpenACC

Overview of Life & Material Accelerated Apps

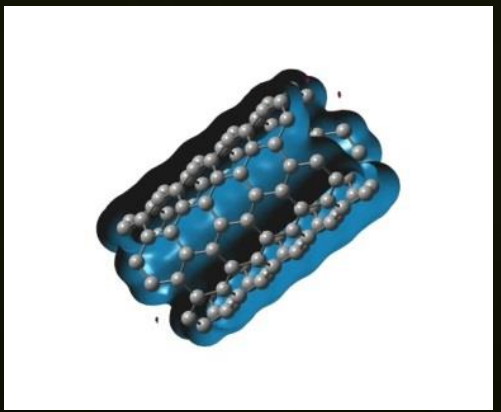
MD: All key codes are GPU-accelerated

- ACEMD*, AMBER (PMEMD)*, BAND, CHARMM, DESMOND, ESPResso, Folding@Home, GPUgrid.net, GROMACS, HALMD, HOOMD-Blue*, LAMMPS, Lattice Microbes*, mdcore, NAMD, OpenMM, SOP-GPU*
- Great multi-GPU performance!
- Focus: on dense (up to 16) GPU nodes & large # of GPU nodes



QC: All key codes are ported or optimizing:

- GPU-accelerated and available today:
  - ABINIT, ACES III, ADF, BigDFT, CP2K, GAMESS, Quantum Espresso/PWscf, MOLCAS, MOPAC2012, NWChem, OCTOPUS*, QUICK, Q-Chem, TeraChem*
- Active GPU acceleration projects:
  - CASTEP, CPMD, GAMESS, Gaussian, NWChem, ONETEP, Quantum Supercharger Library*, VASP & more
- Focus: on using GPU-accelerated math libraries, OpenACC directives



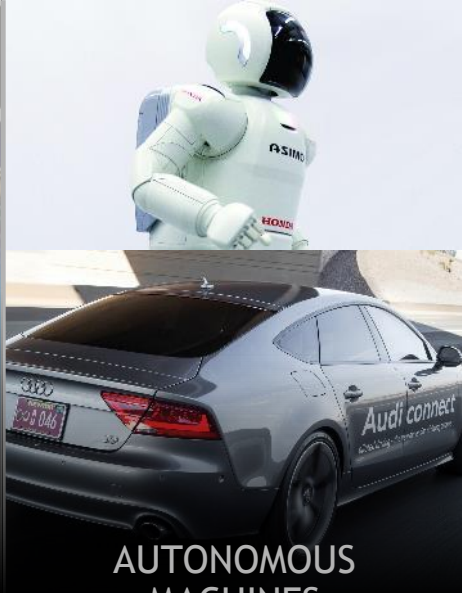
green* = application where all the workload is on GPU

GAMING

DESIGN

ENTERPRISE
VIRTUALIZATION

HPC & CLOUD
SERVICE PROVIDERS



AUTONOMOUS
MACHINES

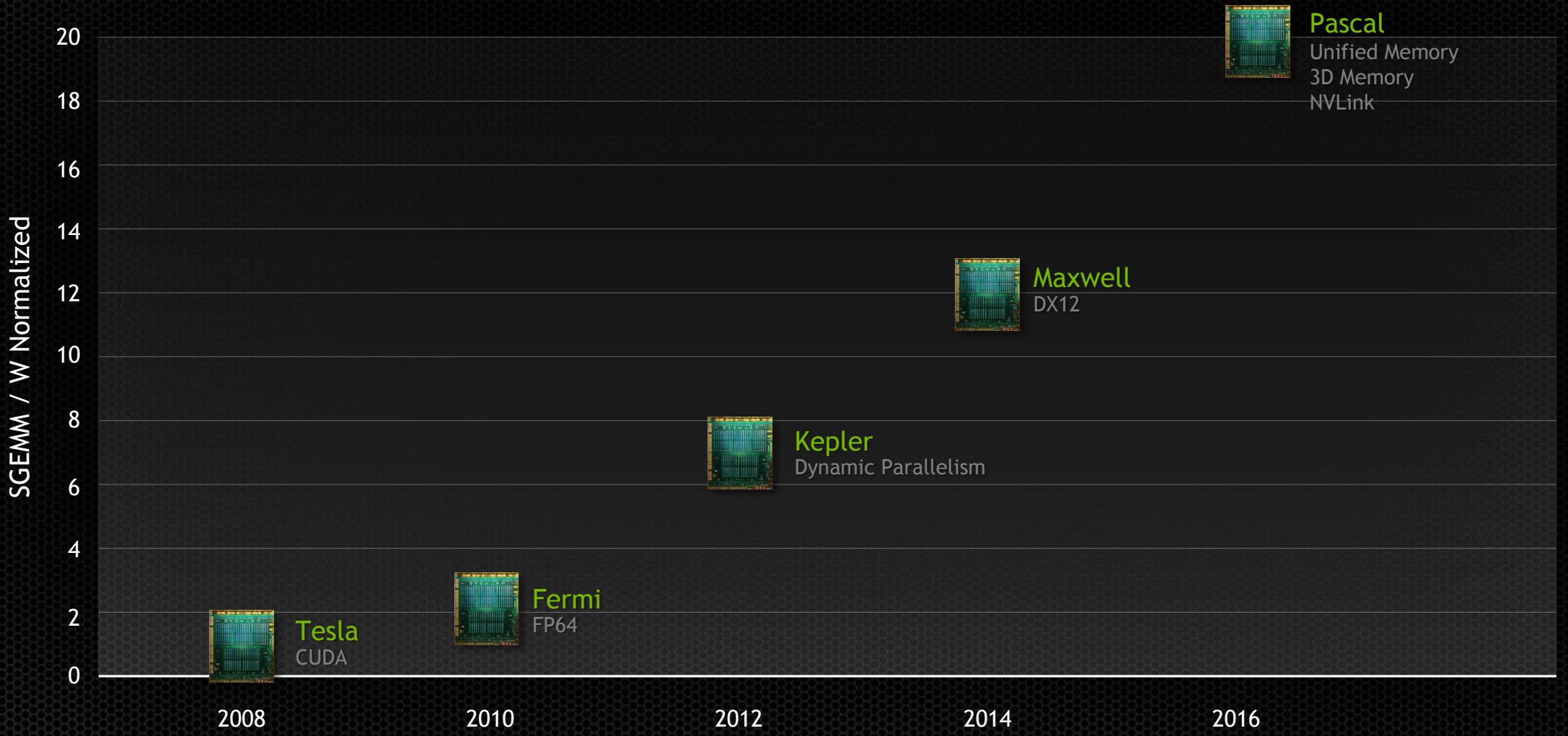
PC

DATA
CENTER

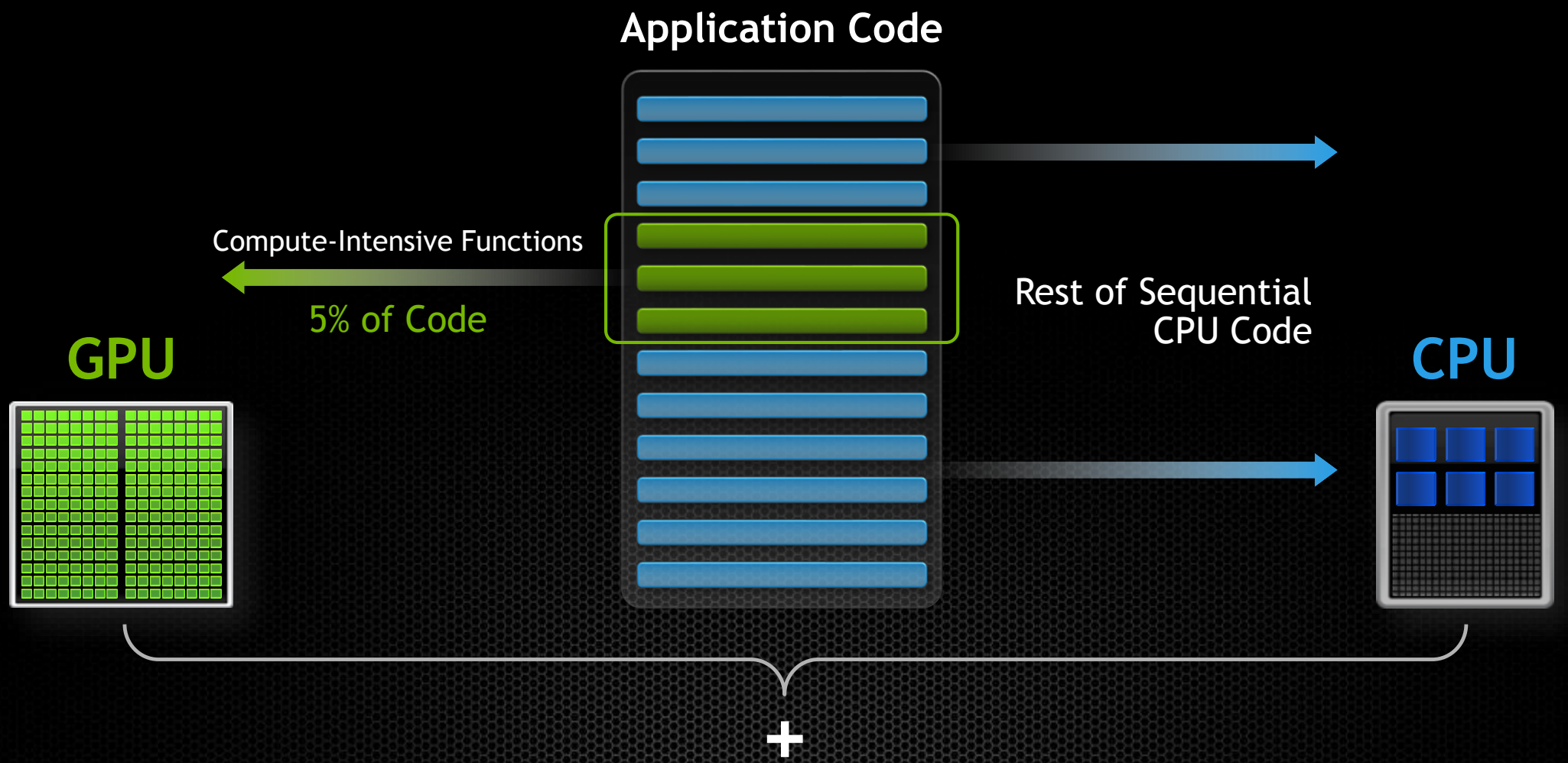
MOBILE

The World Leader in Visual Computing

Strong CUDA GPU Roadmap

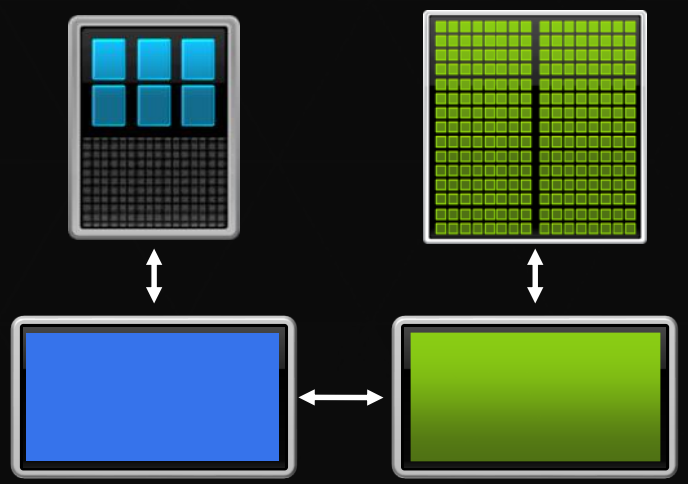


How GPU Acceleration Works

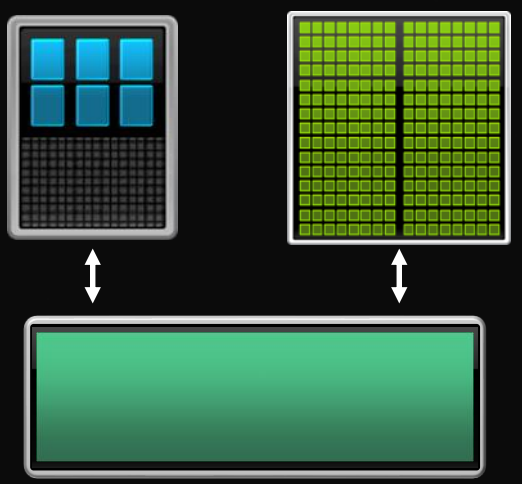


Unified Memory: Simpler & Faster with NVLink

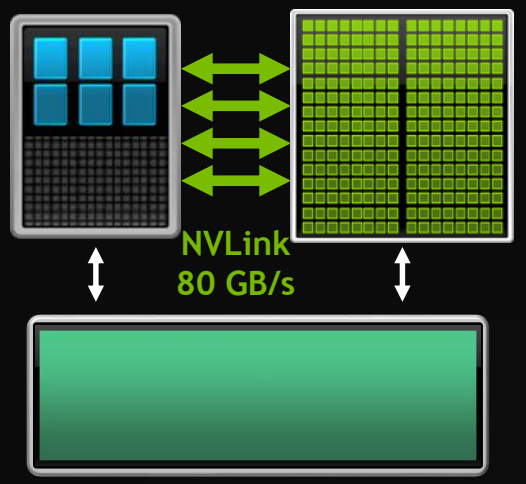
Traditional Developer View



Developer View With Unified Memory



Developer View With Pascal & NVLink



Share Data Structures at
CPU Memory Speeds, not PCIe speeds
Oversubscribe GPU Memory

Tesla Platform

- *Tesla GPU – K40 & K80*
- *NVLink*
- *IBM Power*
- *ARM*

Tesla Accelerated Computing Platform

Data Center Infrastructure

System Solutions

CRAY
DELL
IBM
hp
amazon web services EC2

Communication

Mellanox TECHNOLOGIES
MVAPICH
Open MPI

Infrastructure Management

Adaptive COMPUTING
Bright Computing
IBM PLATFORM COMPUTING

GPU Accelerators
GPU Boost
…

Interconnect
GPU Direct
NVLink
…

System Management
NVML / NVIDIA-SMI
…

Development

Programming Languages

C/C++
Fortran
OpenACC
python

Development Tools

allinea DDT
PGI
VAMPIR

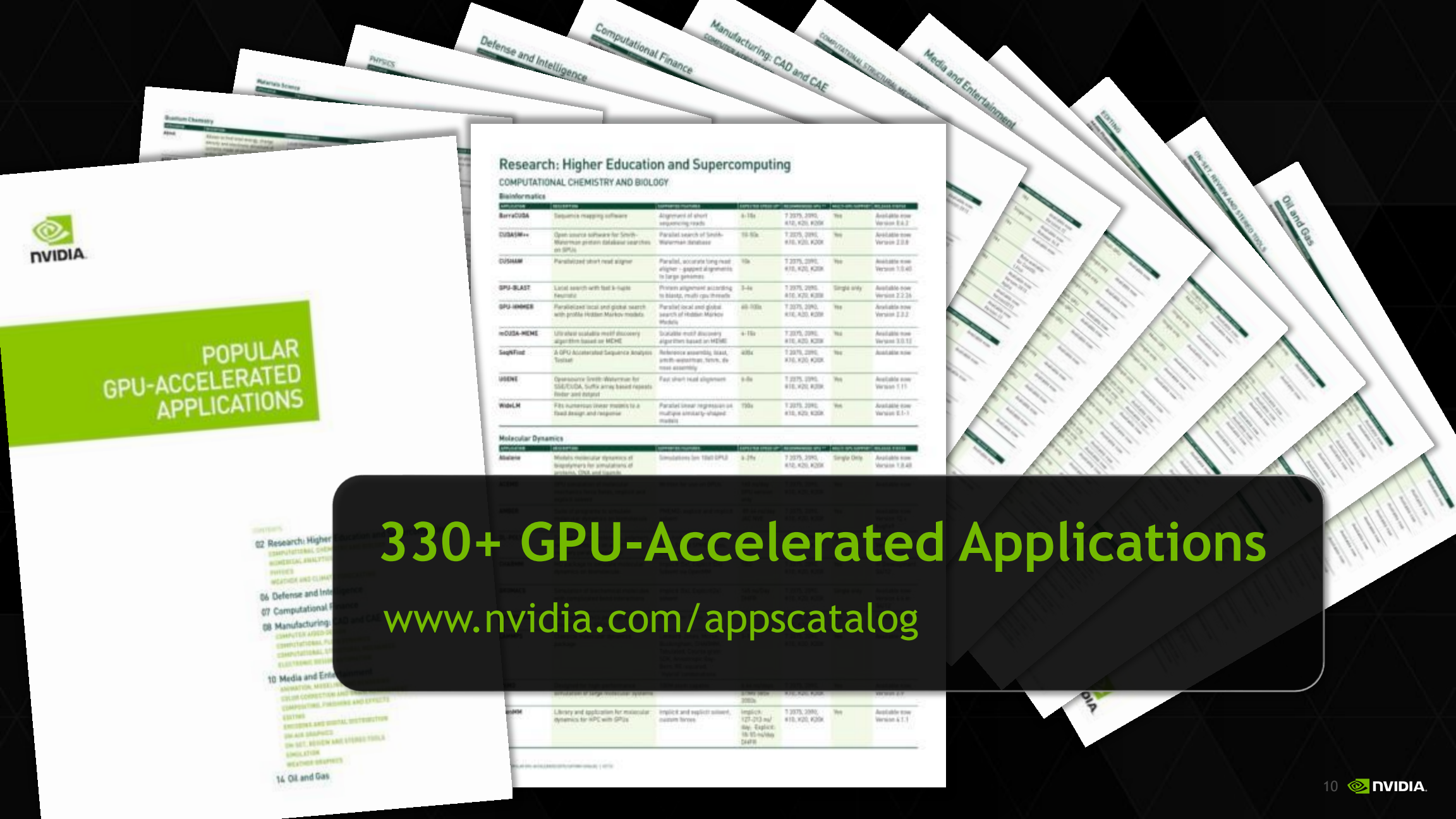
Software Solutions

Kitware
MATLAB
ROGUE WAVE SOFTWARE

Compiler Solutions
LLVM
…

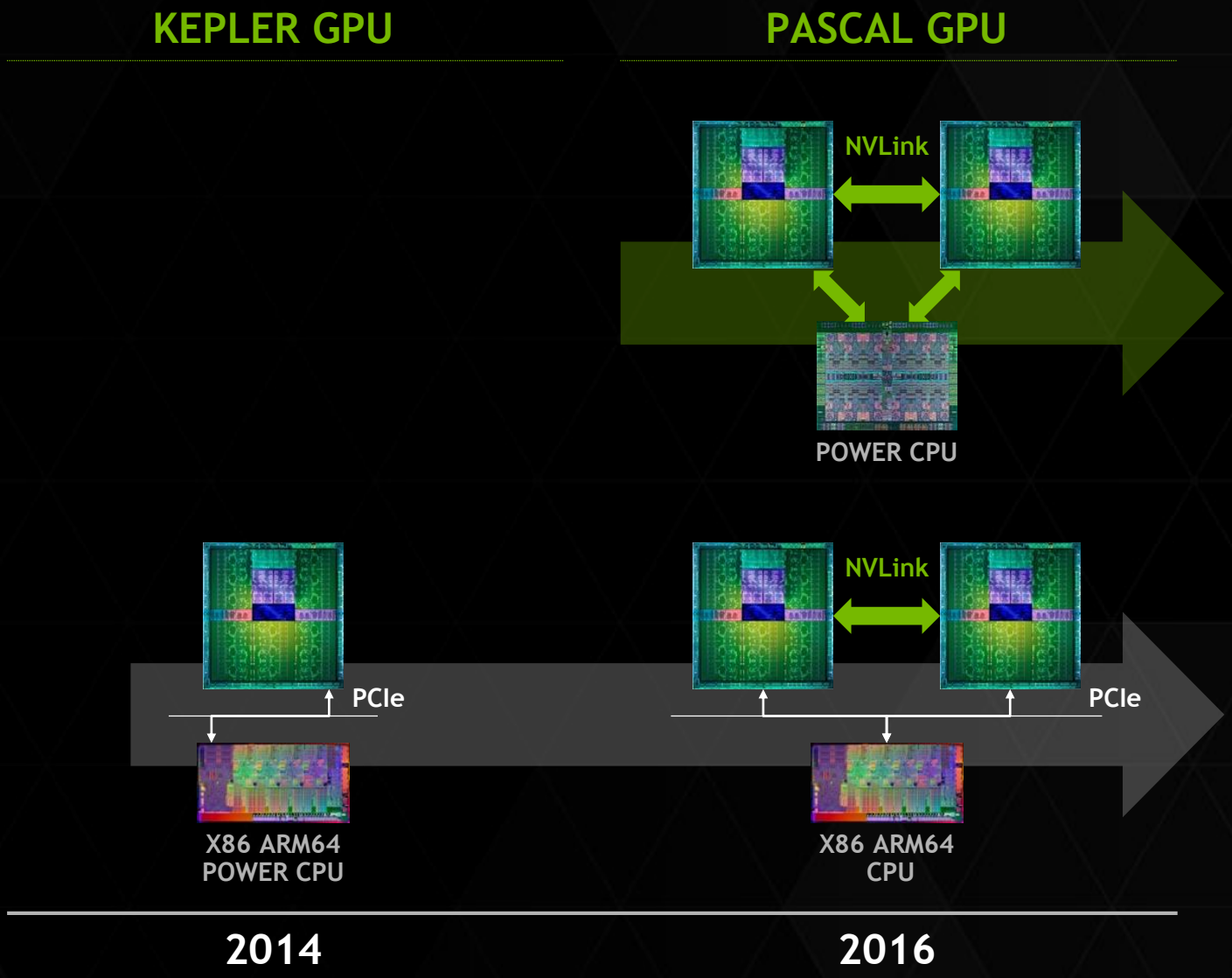
Profile and Debug
CUDA Debugging API
…

Libraries
cuBLAS
…


NVIDIA
POPULAR GPU-ACCELERATED APPLICATIONS
Research: Higher Education and Supercomputing
COMPUTATIONAL CHEMISTRY AND BIOLOGY
Bioinformatics
Molecular Dynamics
330+ GPU-Accelerated Applications
www.nvidia.com/appscatalog

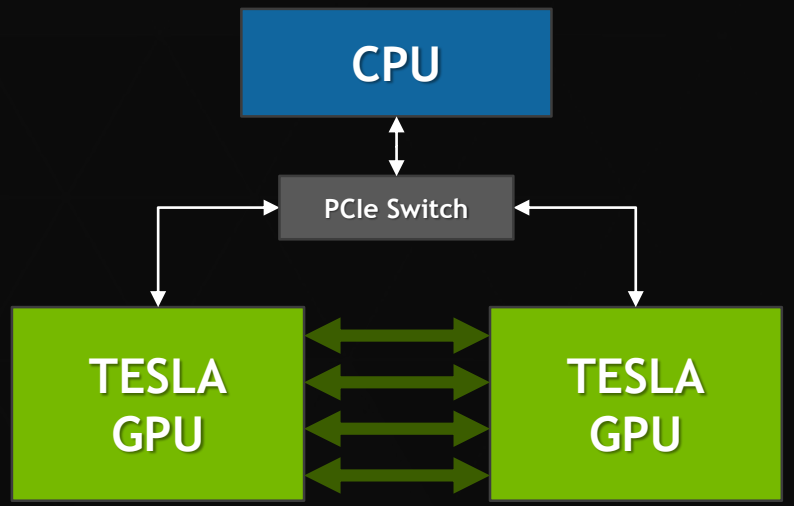
NVLink

High-speed GPU Interconnect



NVLink Unleashes Multi-GPU Performance

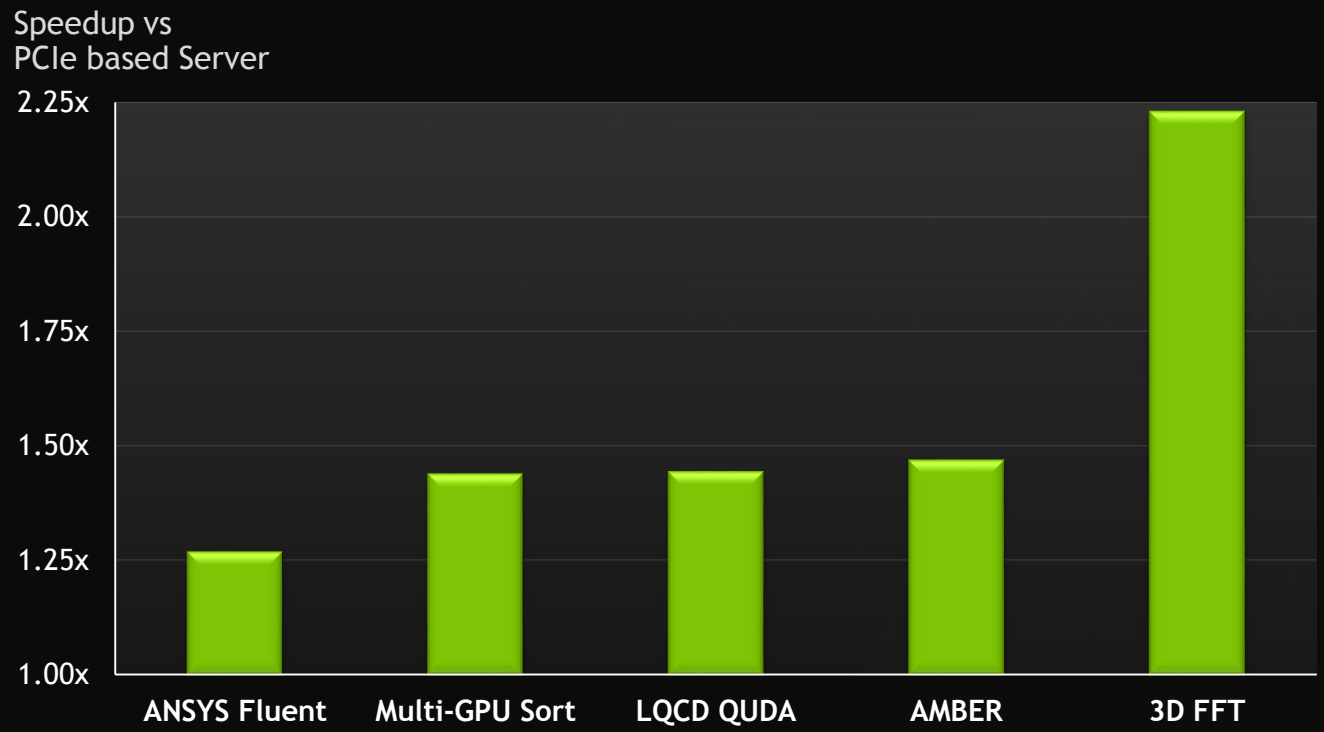
GPUs Interconnected with NVLink



5x Faster than
PCIe Gen3 x16

Over 2x Application Performance Speedup

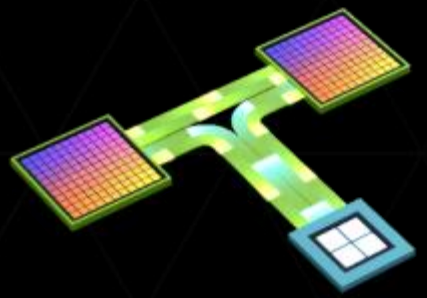
When Next-Gen GPUs Connect via NVLink Versus PCIe



To learn more: http://www.nvidia.com/object/nvlink.html

3D FFT, ANSYS: 2 GPU configuration, All other apps comparing 4 GPU configuration
AMBER Cellulose (256x128x128), FFT problem size (256^3)

Major Data Center OEMs Support NVLink



Bull atos technologies

CRAY

DELL

hp

IBM

QCT™ Quanta CLOUD TECHNOLOGY

TYAN

US to Build Two Flagship Supercomputers

Powered by the Tesla Platform

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

IBM

NVIDIA

100-300 PFLOPS Peak

10x in Scientific App Performance

IBM POWER9 CPU + NVIDIA Volta GPU

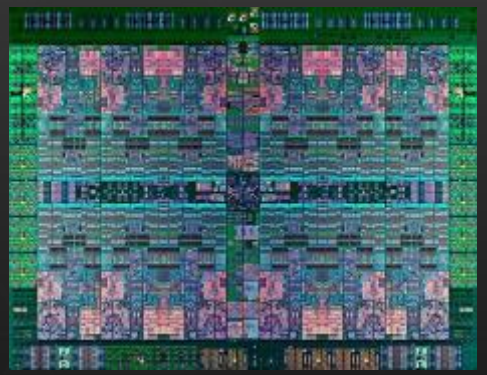
NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

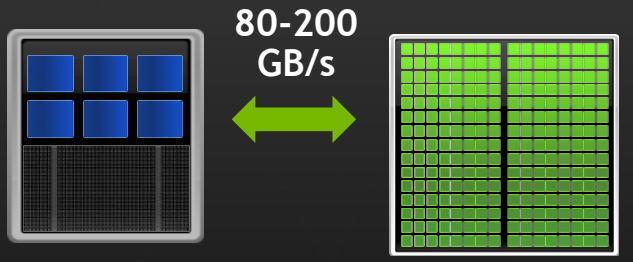
2017

Major Step Forward on the Path to Exascale

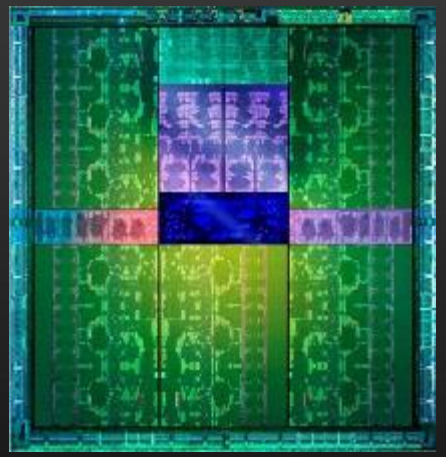
Accelerated Computing
5x Higher Energy Efficiency



IBM POWER CPU
Most Powerful Serial Processor



NVIDIA NVLink
Fastest CPU-GPU Interconnect



NVIDIA Volta GPU
Most Powerful Parallel Processor

GPUS MAKE MACHINE LEARNING ACCESSIBLE

Deep learning with COTS HPC systems

A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro

ICML 2013

"Now You Can Build Google's $1M Artificial Brain on the Cheap"

WIRED

GOOGLE DATACENTER

1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
$5,000,000

STANFORD AI LAB

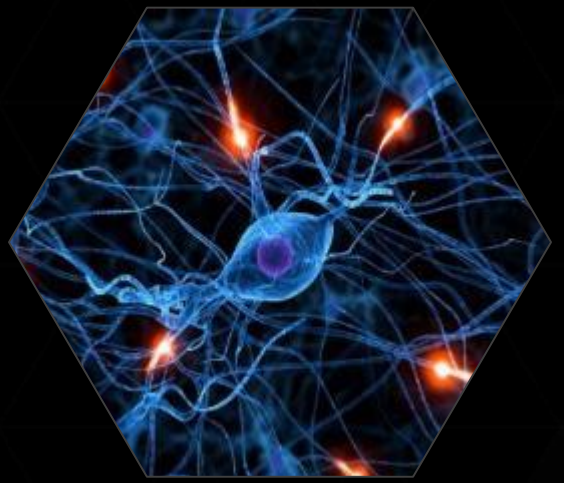
3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts
$33,000

3 DRIVERS FOR DEEP LEARNING

More Data

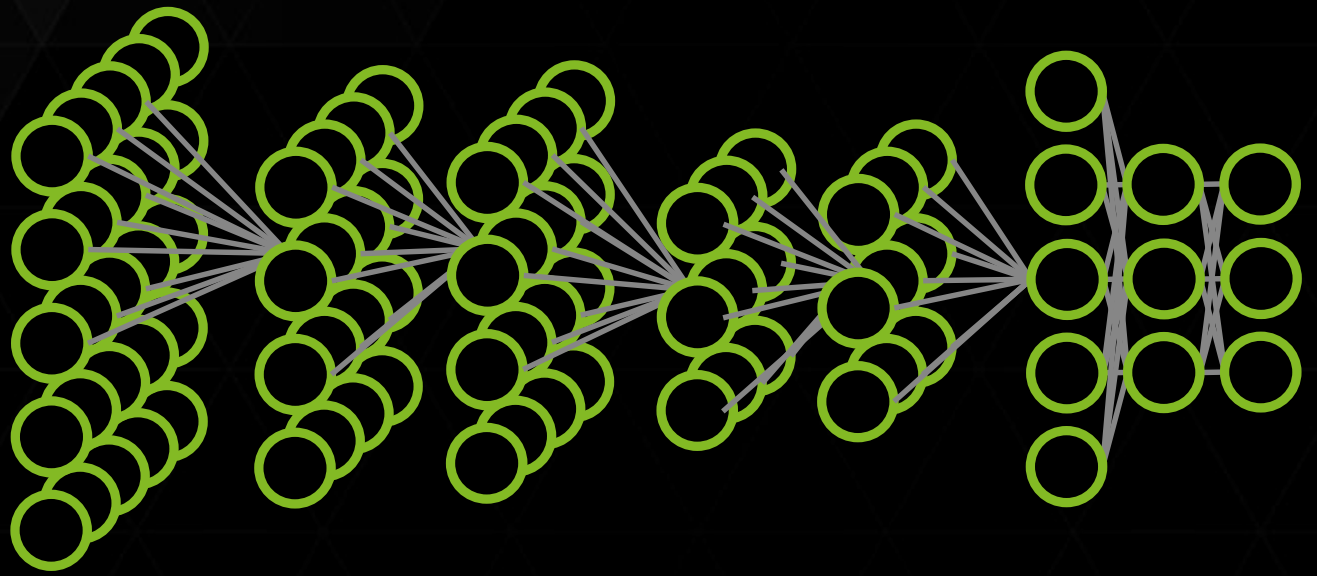
Better Models

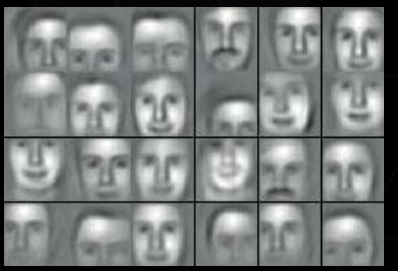


Powerful GPU Accelerators

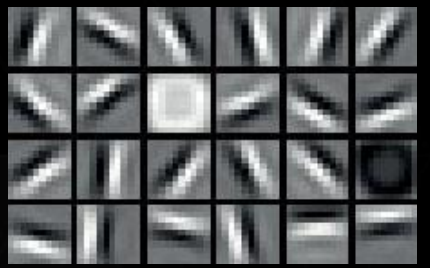
WHAT IS DEEP LEARNING?

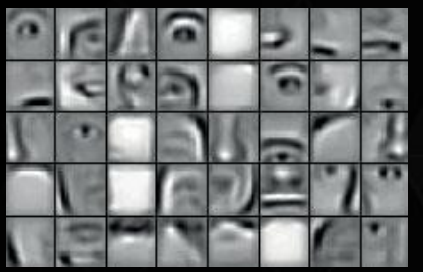
Input



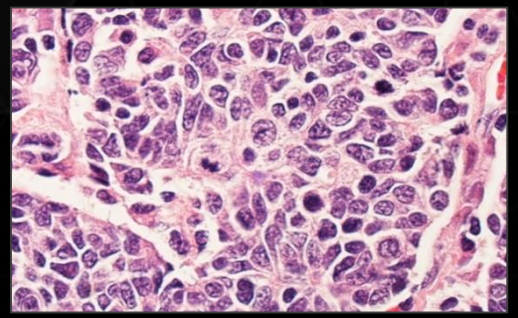


Result





DEEP LEARNING REVOLUTIONIZING MEDICAL RESEARCH

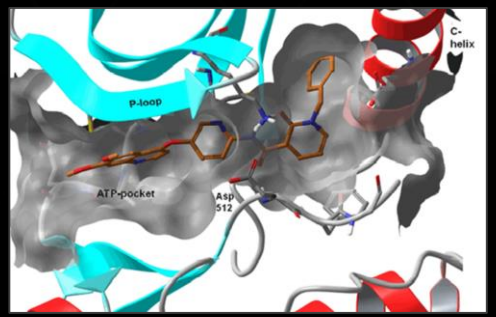


Detecting Mitosis in Breast Cancer Cells

— IDSIA

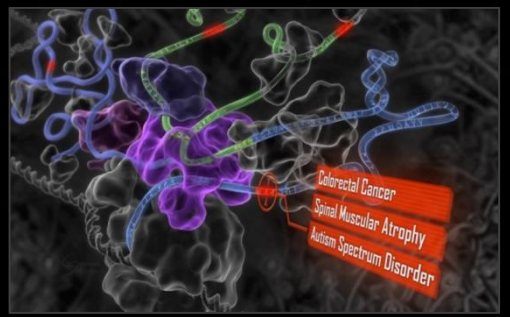
Molecular Activity Prediction for Drug Discovery

— Merck



Predicting the Toxicity of New Drugs

— Johannes Kepler University



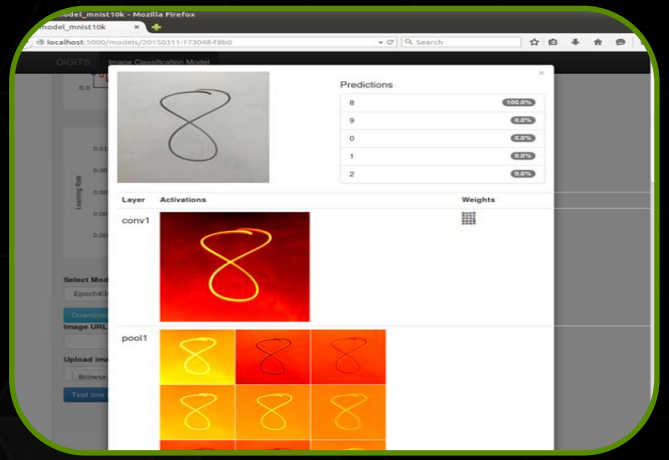
Understanding Gene Mutation to Prevent Disease

— University of Toronto

Deep Learning Performance Doubles

for Data Scientists and Researchers

Train Models up to 2x Faster with Automatic Multi-GPU Scaling



DIGITS 2

2x Faster Single GPU Training Support for Larger Models

cuDNN 3

2x Larger Datasets Instruction-level Profiling

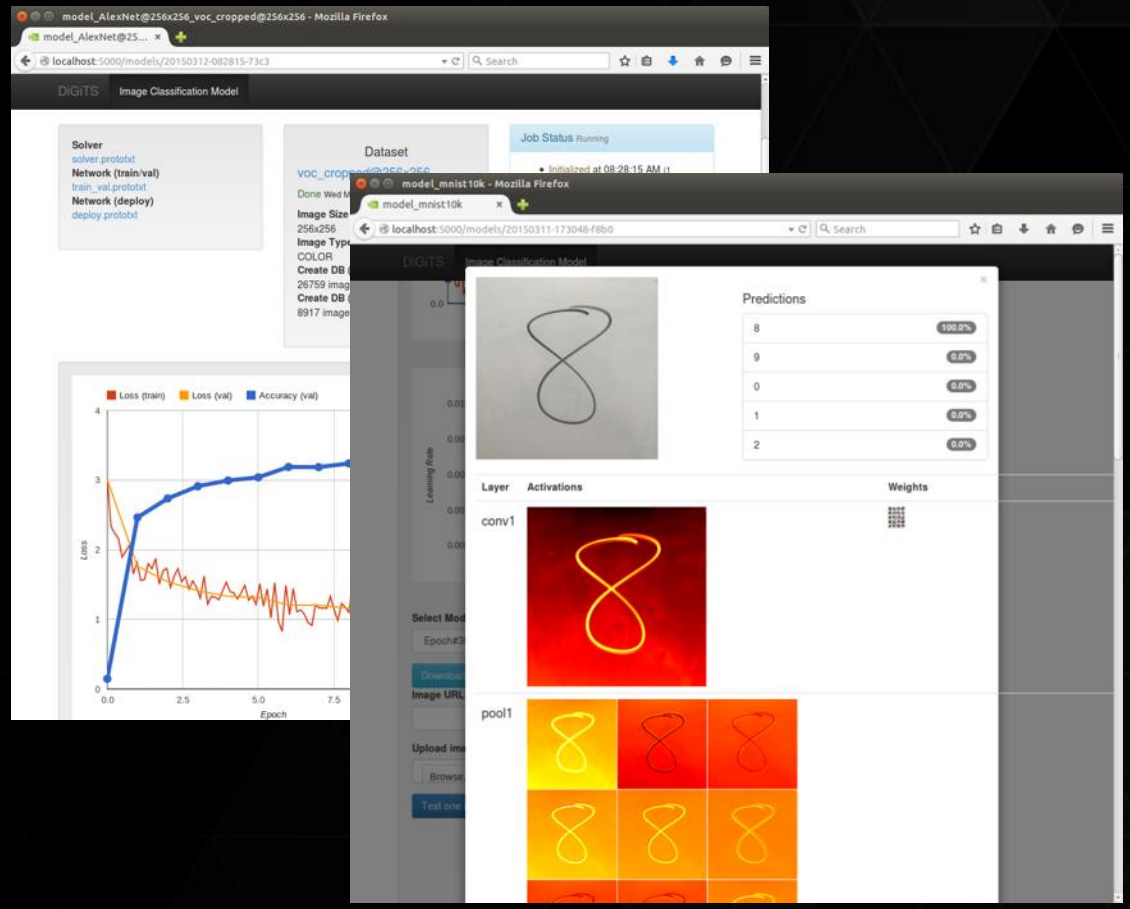
CUDA 7.5

NVIDIA® DIGITS™

Interactive Deep Learning GPU Training System

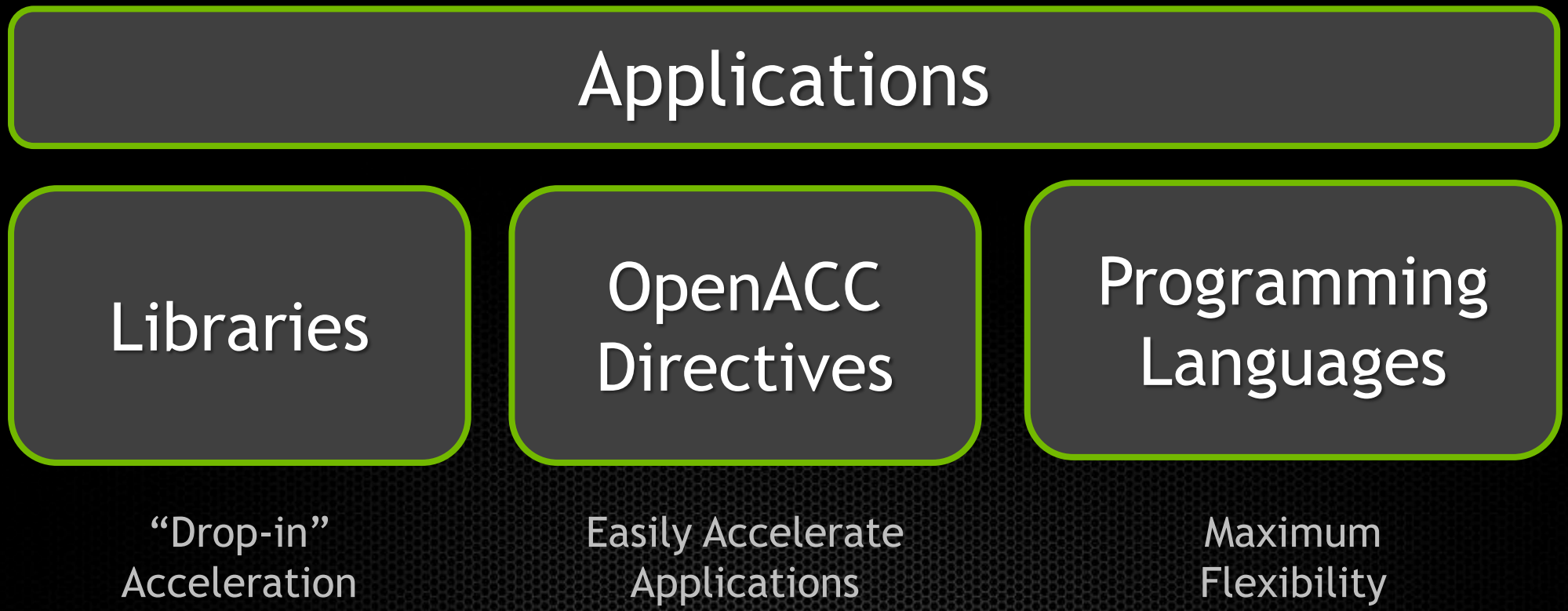
Data Scientists & Researchers:

- Quickly design the best deep neural network (DNN) for your data
- Visually monitor DNN training quality in real-time
- Manage training of many DNNs in parallel on multi-GPU systems



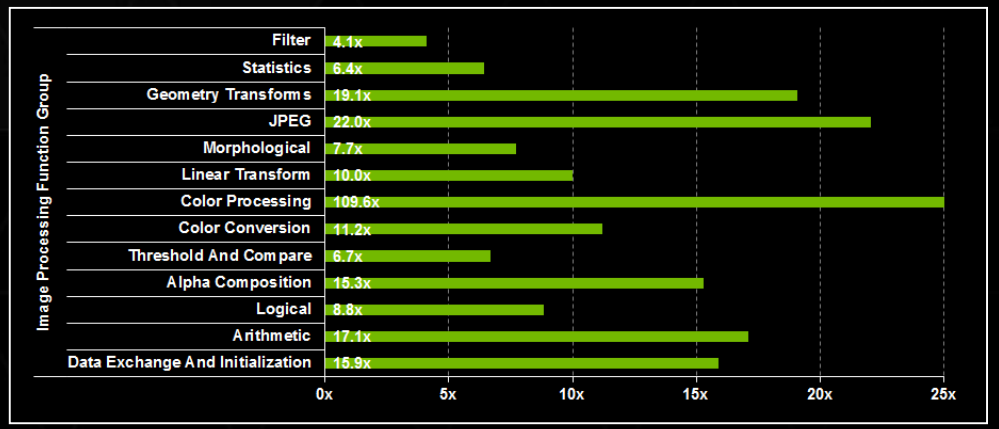
developer.nvidia.com/digits

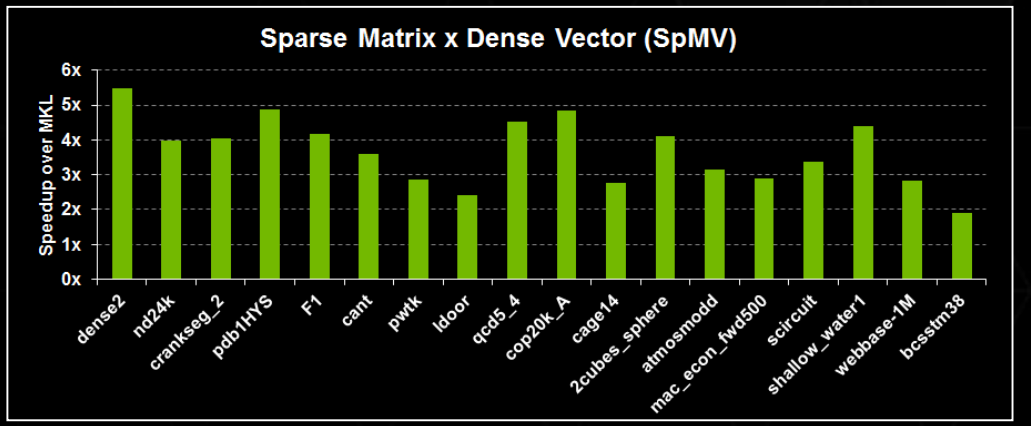
3 Ways to Accelerate Applications

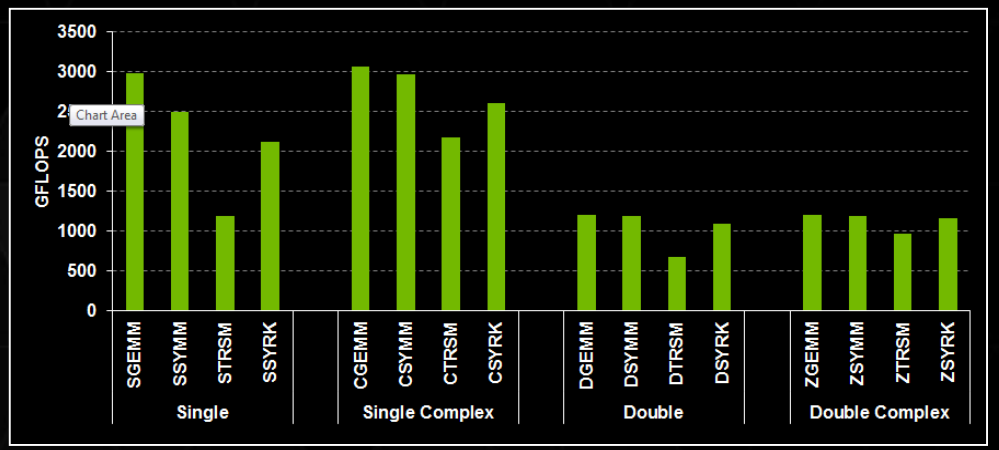


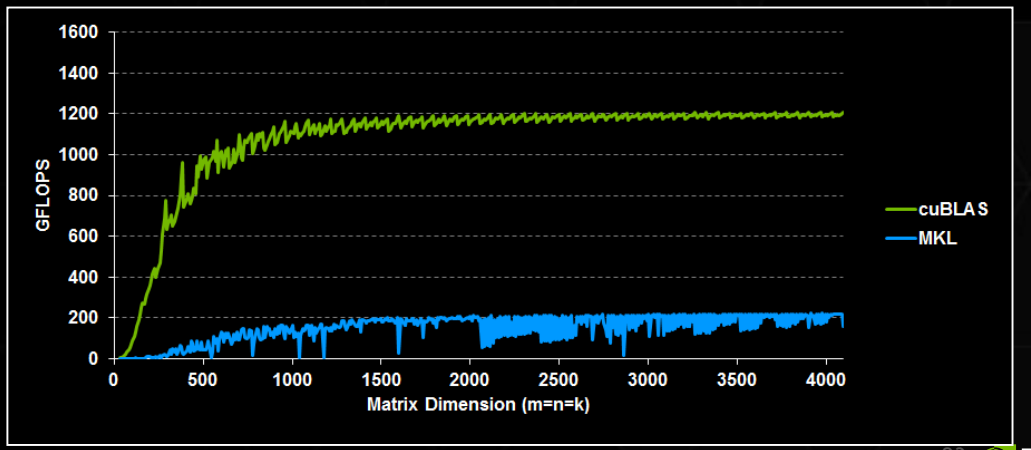
5X-10X SPEEDUP USING NVIDIA LIBRARIES

BLAS | LAPACK | SPARSE | FFT | Math | Deep Learning | Graphs | Image & Signal Processing









NVIDIA

OpenACC

Simple | Powerful | Portable

Fueling the Next Wave of Scientific Discoveries in HPC

```
main()
{
  <serial code>
  #pragma acc kernels
  //automatically runs on GPU
  {
    <parallel code>
  }
}
```

University of Illinois
PowerGrid- MRI Reconstruction

70x Speed-Up
2 Days of Effort

RIKEN Japan
NICAM- Climate Modeling

7-8x Speed-Up
5% of Code Modified

8000+

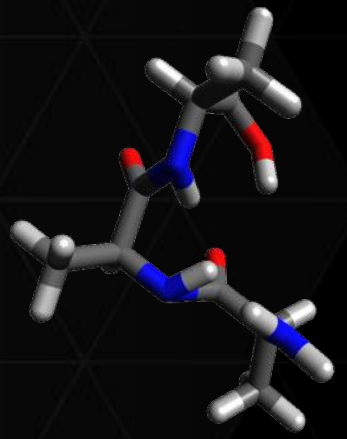
Developers

using OpenACC

http://www.cray.com/sites/default/files/resources/OpenACC_213462.12_OpenACC_Cosmo_CS_FNL.pdf
http://www.hpcwire.com/off-the-wire/first-round-of-2015-hackathons-gets-underway
http://on-demand.gputechconf.com/gtc/2015/presentation/S5297-Hisashi-Yashiro.pdf
http://www.openacc.org/content/experiences-porting-molecular-dynamics-code-gpus-cray-xk7

LS-DALTON

LARGE-SCALE APPLICATION FOR CALCULATING HIGH-ACCURACY MOLECULAR ENERGIES



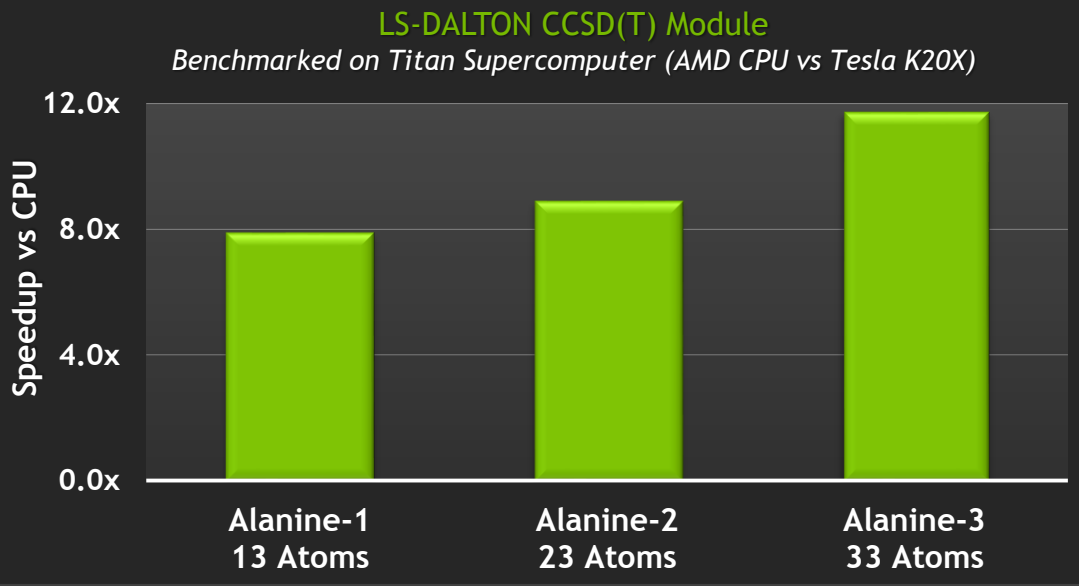
"OpenACC makes GPU computing approachable for domain scientists. Initial OpenACC implementation required only minor effort, and more importantly, no modifications of our existing CPU implementation."

Janus Juul Eriksen, PhD Fellow
qLEAP Center for Theoretical Chemistry, Aarhus University

Minimal Effort

Lines of Code Modified	# of Weeks Required	# of Codes to Maintain
<100 Lines	1 Week	1 Source

Big Performance

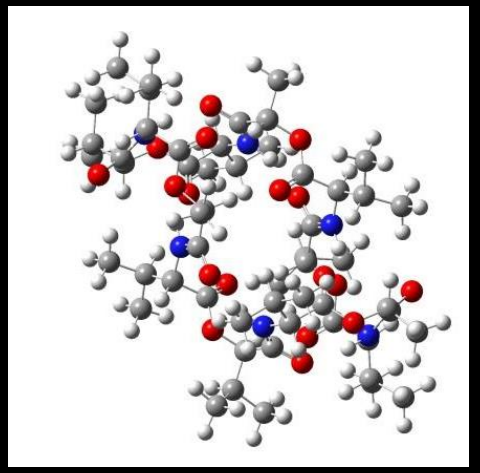


Select Slides from “Enabling Gaussian 09 on GPGPUs” at GTC March 2014

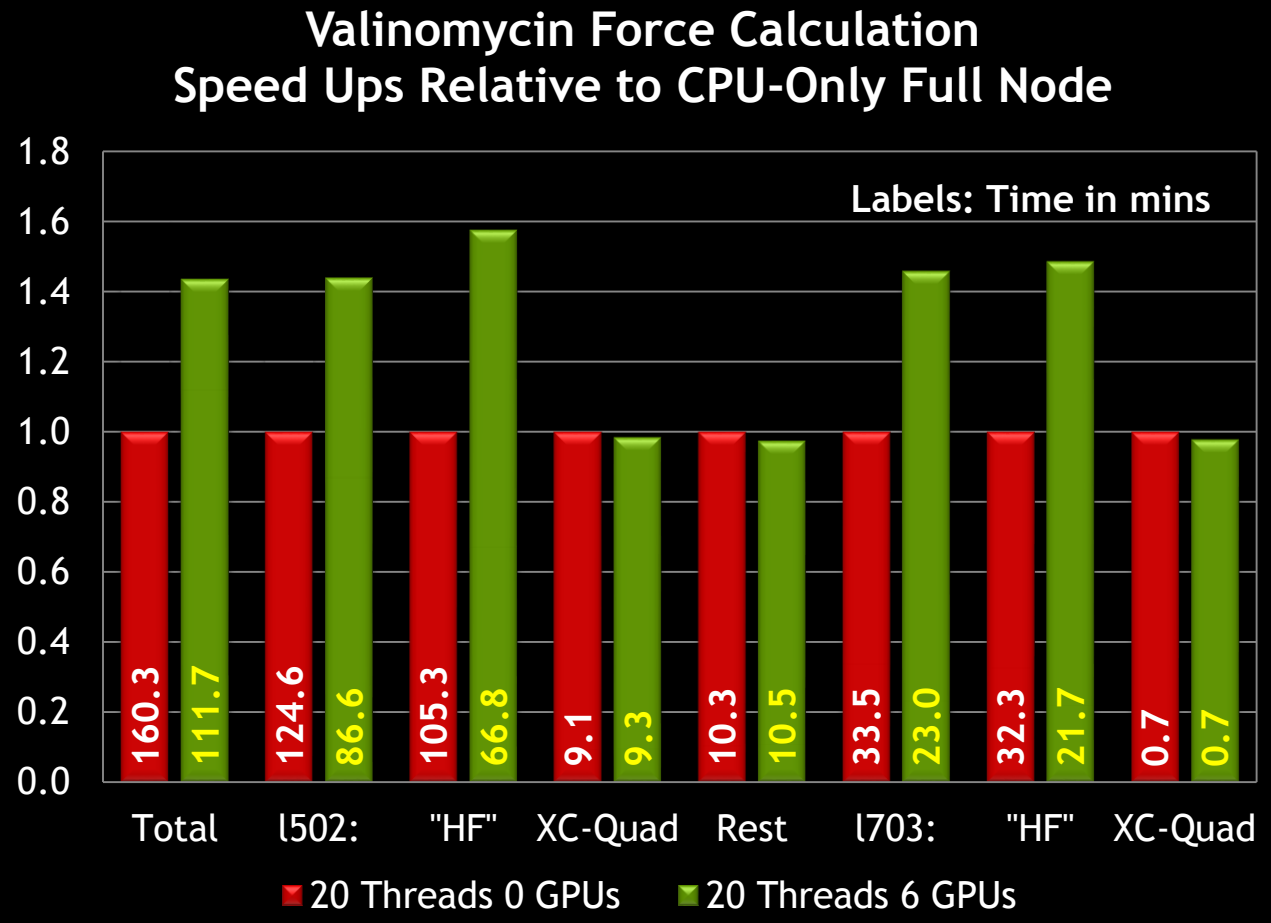
- In 2011 Gaussian, Inc., NVIDIA Corp. and PGI started a long-term project to enable all the performance critical paths of Gaussian on GPGPUs.
  - Ultimate goal is to show significant performance improvement by using accelerators in conjunction with CPUs
  - Initial efforts are directed towards creating an infrastructure that will leverage the current CPU code base and at the same time minimize the additional maintenance effort associated with running on GPUs.
- Current status of this work for Direct Hartree-Fock and triples-correction calculations as applied in for example Coupled Cluster calculations that uses mostly the directives based OpenACC framework.

- Slides & Audio: http://on-demand.gputechconf.com/gtc/2014/video/S4613-enabling-gaussian-09-gpgpus.mp4

EARLY PERFORMANCE RESULTS (DIRECT SCF)

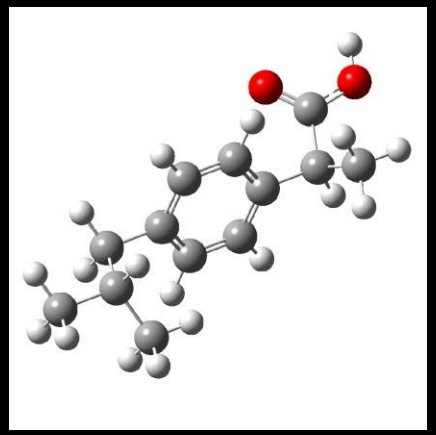


Method	rB3LYP
No. of Atoms	168
Basis Set	6-31G(3df,3p)
No. of Basis Funcs	3 642
No. of Cycles	17

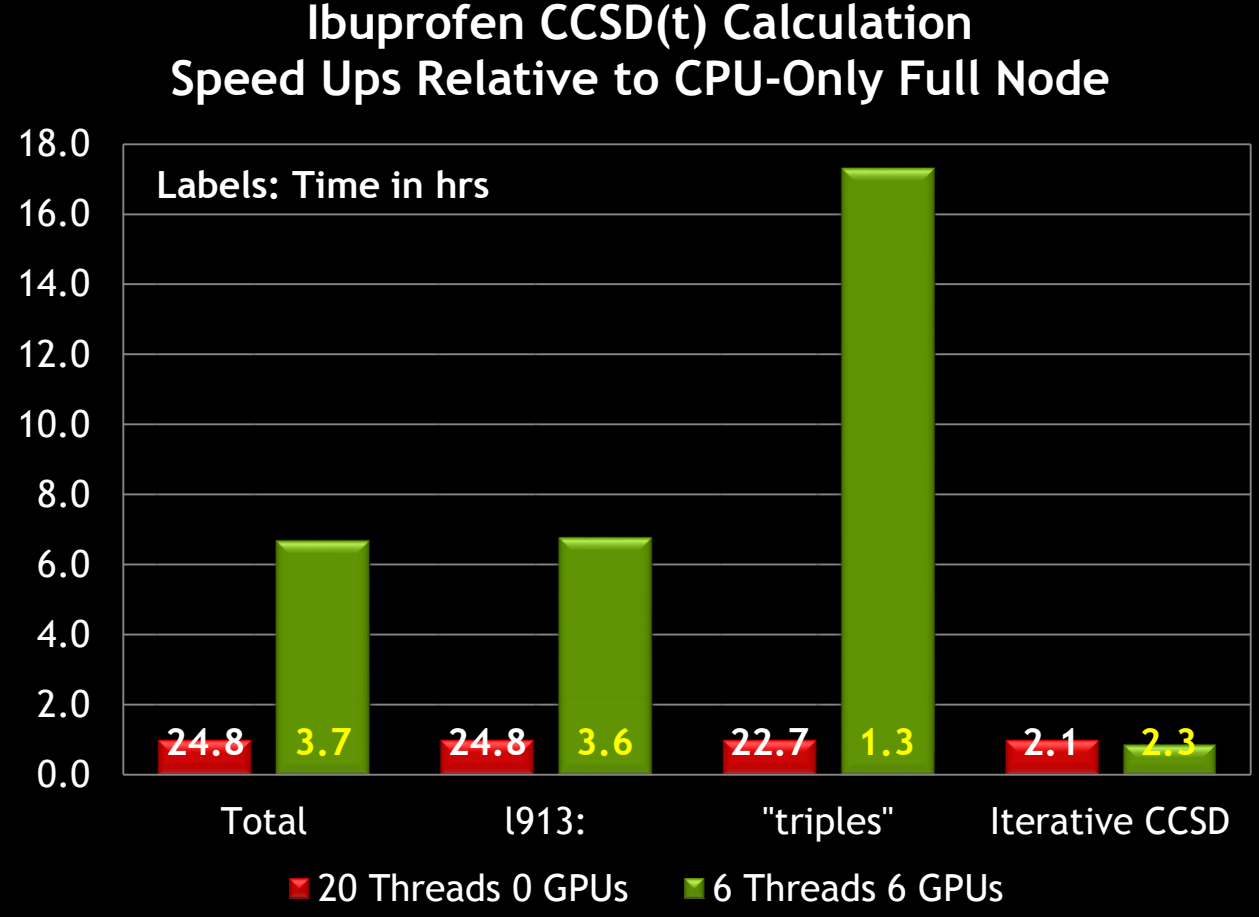


System: 2 Sockets E5-2690 V2 (2x 10 Cores @ 3.0 GHz); 128 GB RAM (DD3-1600); Used 108 GB
GPUs: 6 Tesla K40m (15 SMPs @ 875 MHz); 12 GB Global Memory

EARLY PERFORMANCE RESULTS (CCSD(T))



Method	CCSD(t)
No. of Atoms	33
Basis Set	6-31G(d,p)
No. of Basis Funcs	315
No. Occ Orbitals	41
No. Virt Orbitals	259
No. of Cycles	15
No. CCSD iters	16



System: 2 Sockets E5-2690 V2 (2x 10 Cores @ 3.0 GHz); 128 GB RAM (DD3-1600); Used 108 GB
GPUs: 6 Tesla K40m (15 SMPs @ 875 MHz); 12 GB Global Memory

Introducing the New OpenACC Toolkit

Free Toolkit Offers Simple & Powerful Path to Accelerated Computing

http://developer.nvidia.com/openacc

PGI Compiler
Free OpenACC compiler for academia

NVProf Profiler
Easily find where to add compiler directives

GPU Wizard
Identify which GPU libraries can jumpstart code

Code Samples
Learn from examples of real-world algorithms

Documentation
Quick start guide, Best practices, Forums

GPU VASP Collaboration

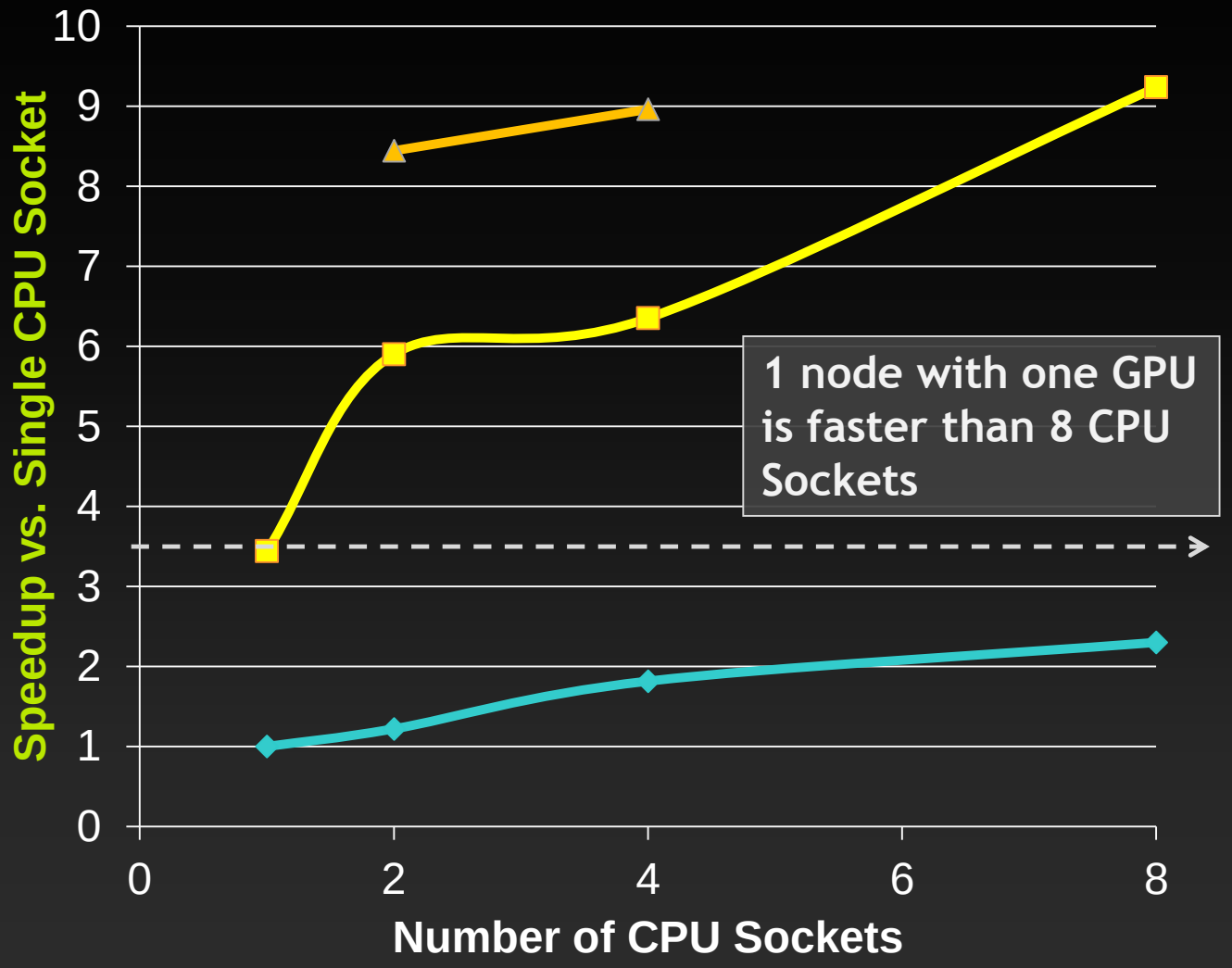
- **Collaborators**

- **2013-2014 Project Scope**

  Minimization algorithms to calculate electronic ground state
  - Blocked Davidson (ALGO = NORMAL & FAST)
  - RMM-DIIS (ALGO = VERYFAST & FAST)
  - K-Points

- **Earlier work**
  - *Speeding up plane-wave electronic-structure calculations using graphics-processing units*, Maintz, Eck, Dronskowski
  - *VASP on a GPU: application to exact-exchange calculations of the stability of elemental boron*, Hutchinson, Widom
  - *Accelerating VASP Electronic Structure Calculations Using Graphic Processing Units*, Hacene, Anciaux-Sedrakian, Rozanska, Klahr, Guignon, Fleurat-Lessard

VASP GPU Performance: Results Nial-MD (blocked Davidson)



CPU only (8 cores/CPU)

1 GPU : 1 CPU ratio (1 core/GPU)

2 GPU : 1 CPU ratio (1 core/GPU)

- Measured on K40 and dual-socket Sandy Bridge (8 cores per socket @2.9GHz)
- Default K40 clocks (no GPU Boost)
- FDR Infiniband
- >1 core per GPU exceeds GPU memory

NVIDIA

Test Drive K80 GPUs!

Experience The Acceleration

www.nvidia.com/GPUTestDrive

- Run Computational Chemistry Apps on Tesla K80 GPUs today
- Try Preconfigured Apps:
  AMBER, NAMD, GROMACS, LAMMPS, Quantum Espresso, TeraChem
  Or Load Your Own
- Sign up for FREE GPU Test Drive on remotely hosted clusters