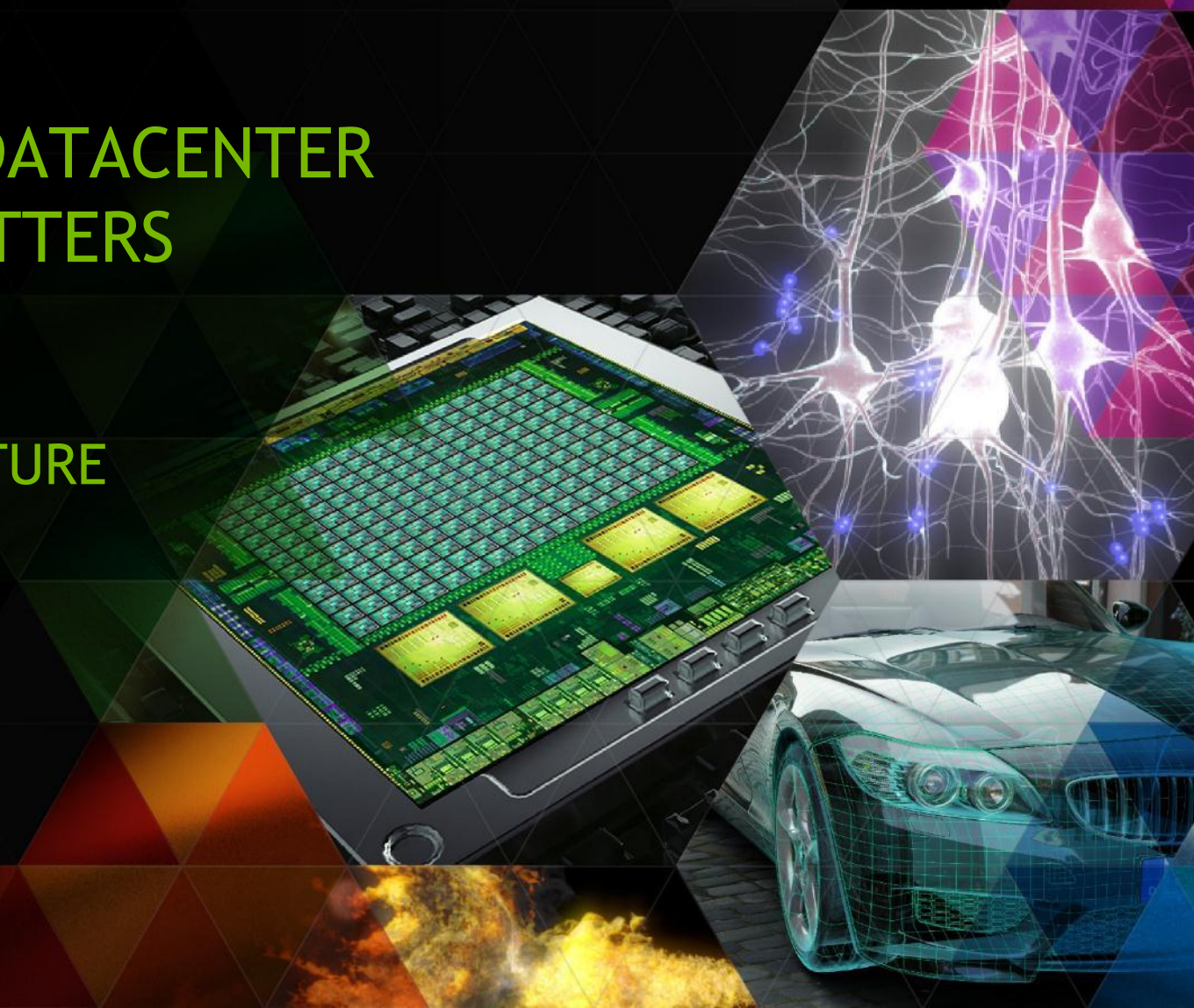
THE ACCELERATED DATACENTER
WHY TOPOLOGY MATTERS

MARC HAMILTON
VP, SOLUTION ARCHITECTURE
NVIDIA

HPCS AMSTERDAM
22 JULY 2015



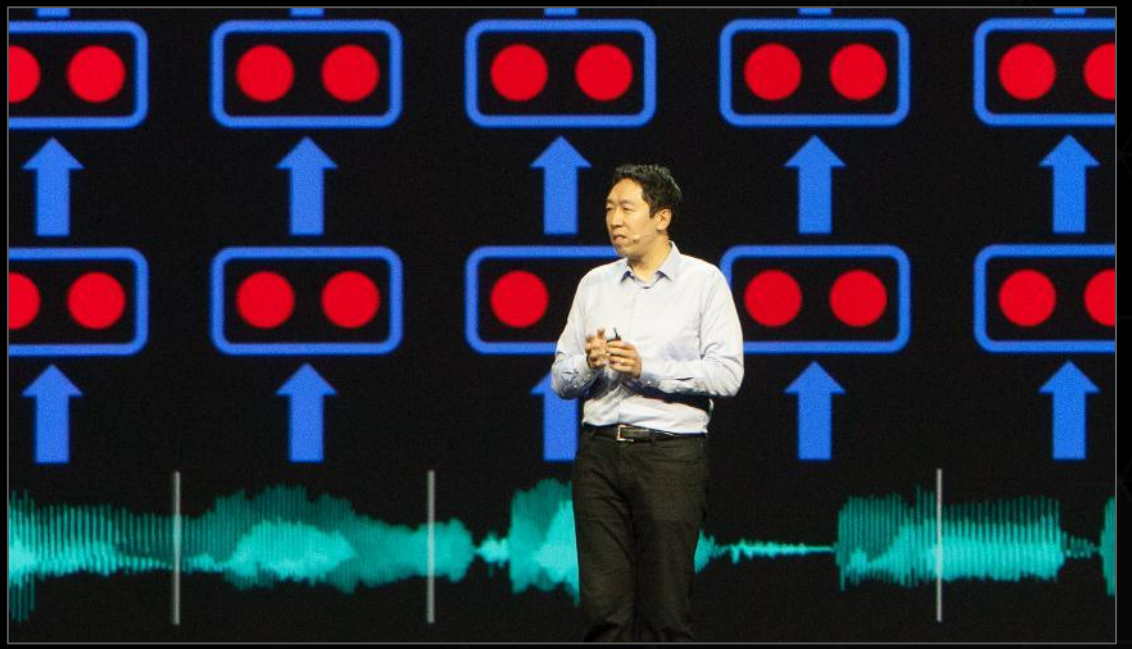
The Accelerated Data Center – View 1

"TSUBAME3.0 in 2016 will not only run HPC workloads, but also will have Big Data and Cloud Computing features"
Prof. Satoshi Matsuoka,
Tokyo Tech & ISC'16 Chair

The Accelerated Data Center – View 2

"We love GPU cards. We just use a lot of them."

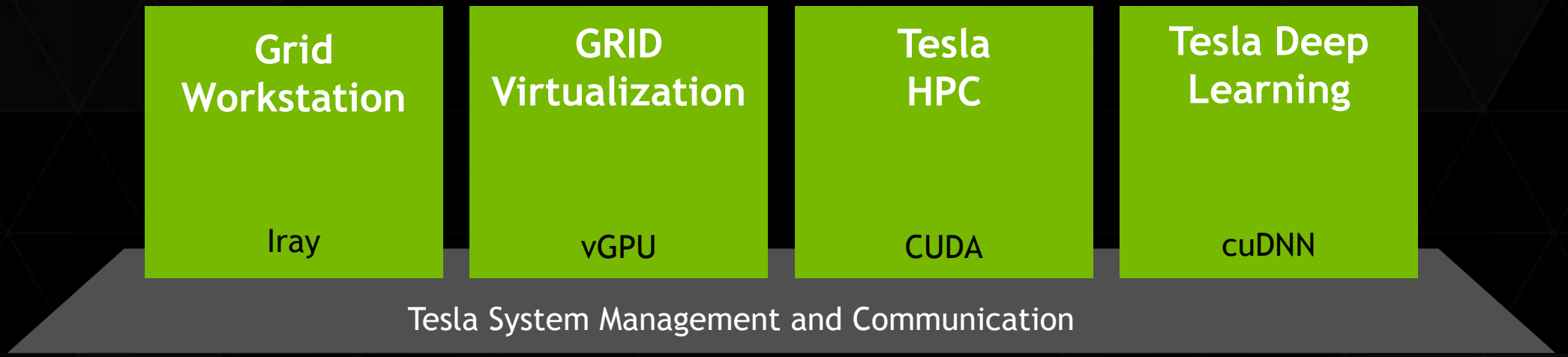
— Jeff Dean, Google



"In five years, we think 50% of queries will be speech or images."

— Andrew Ng, Baidu

The Industry's Most Versatile GPU-Accelerated Platform for the Datacenter



Flexible Pascal & NVLink Architecture for the Datacenter

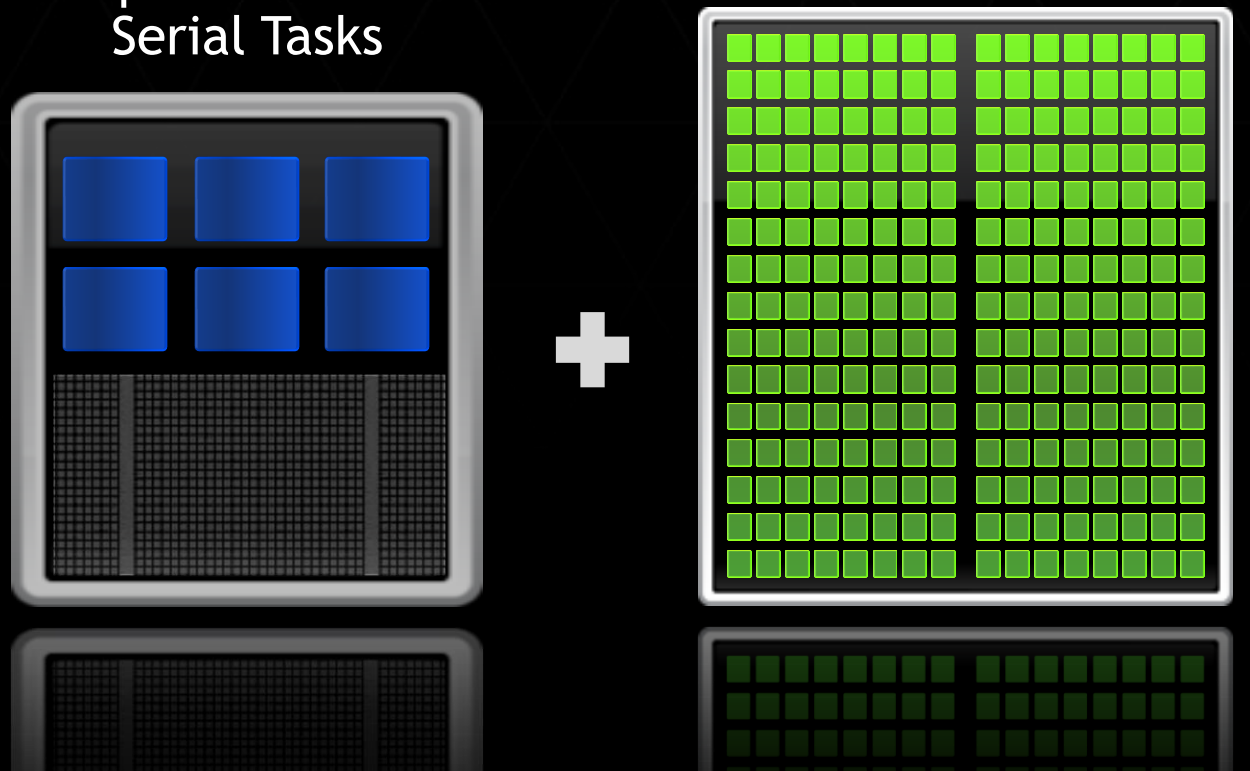
Why Topology Matters

CPU + GPU – Simple Topology?

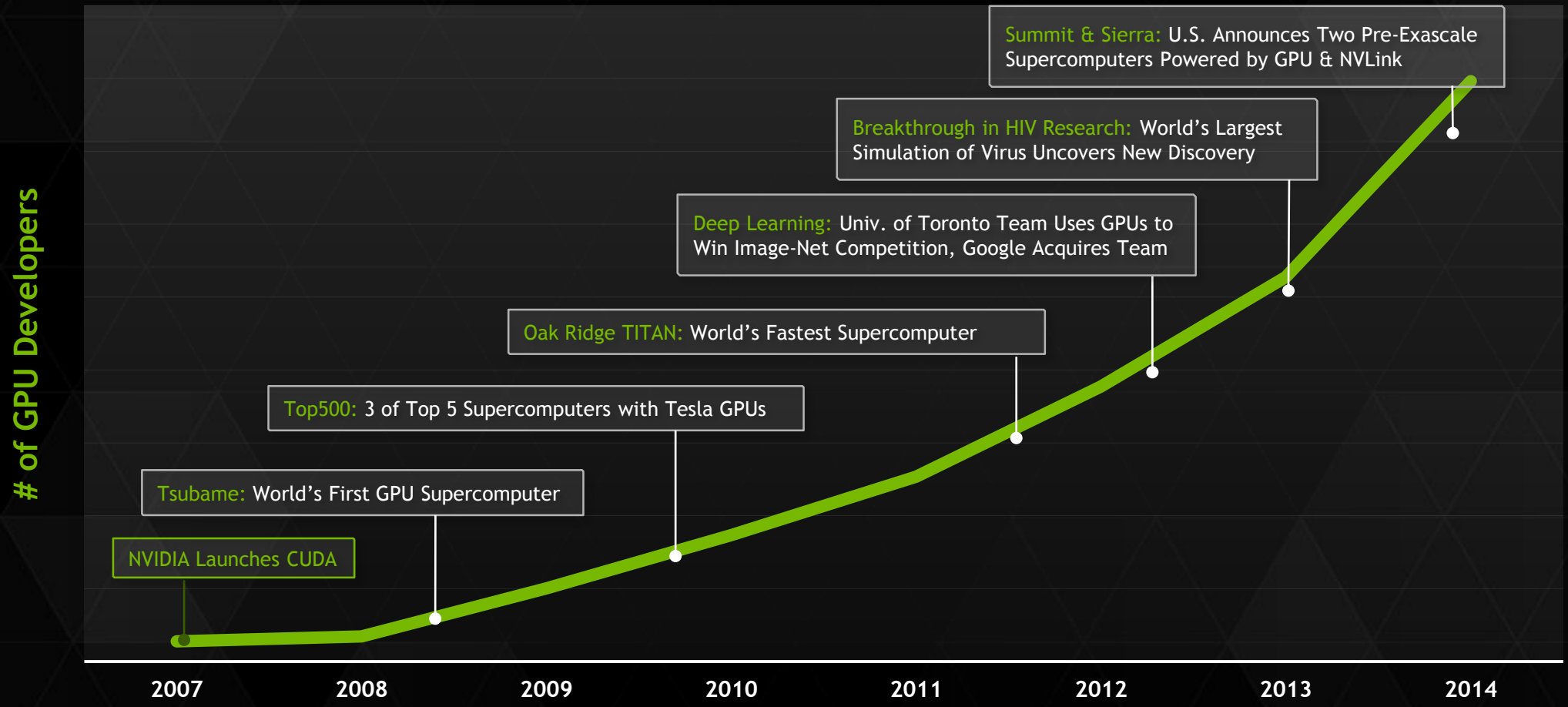
10x Performance & 5x Energy Efficiency

CPU
Optimized for
Serial Tasks

GPU Accelerator
Optimized for
Parallel Tasks



GPU Accelerator Redefined Parallel Computing in HPC



Tesla Accelerated Computing Platform

Data Center Infrastructure

System Solutions

Communication

Infrastructure Management

Development

Programming Languages

Development Tools

Software Solutions

GPU Accelerators	Interconnect	System Management	Compiler Solutions	Profile and Debug	Accelerated Libraries
GPU Boost	*GPU Direct NVLink*	*NVML*	*LLVM*	*CUPTI*	*cuBLAS*

Enterprise Services Support & Maintenance

Three Ways to Accelerate Your Application

"Drop-in" Acceleration

Annotate code with compiler hints

Modern language features (unified memory, for_each, lambda)

OPENACC Addresses Complementary Segment

Two Types of Developers in HPC

	 Computer Scientist	 Domain Scientist
Developer Goal	Maximum Performance, New Applications	More Science, Less Programming
Preferred Approach	CUDA C/C++, Fortran, Python, Java	OpenACC
NSF XSEDE *(Ratio of Users with Allocation in 2014)*	1	6

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

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+

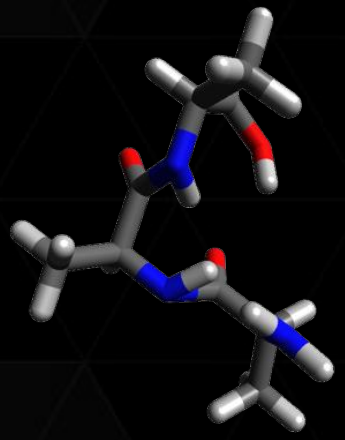
Developers

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



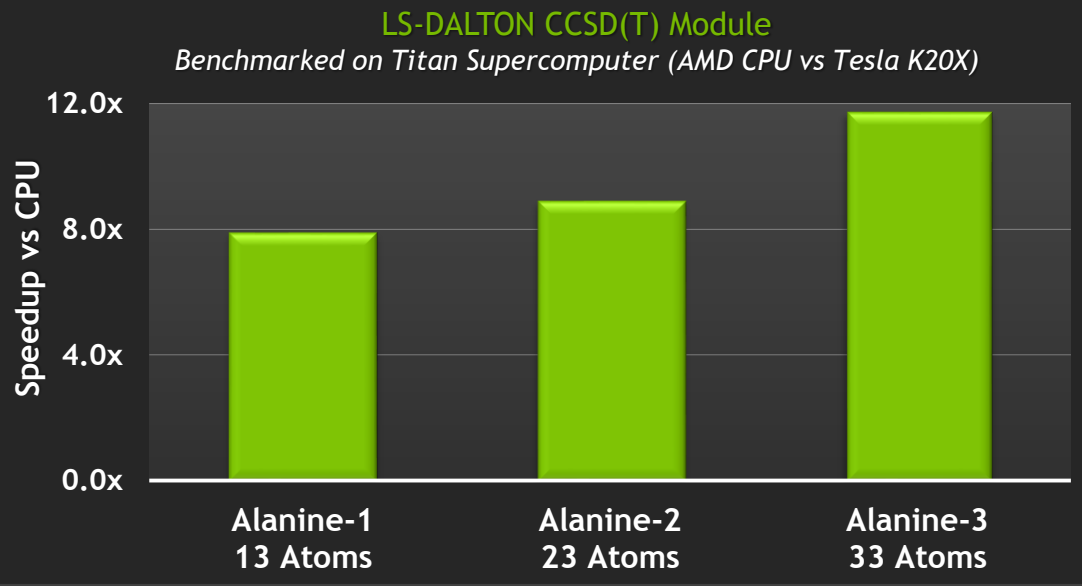
"*OpenACC makes GPU computing approachable for domain scientists. Initial OpenACC implementation required only minor effort, and more importantly, no modifications of our existing CPU implementation.*"

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

Minimal Effort

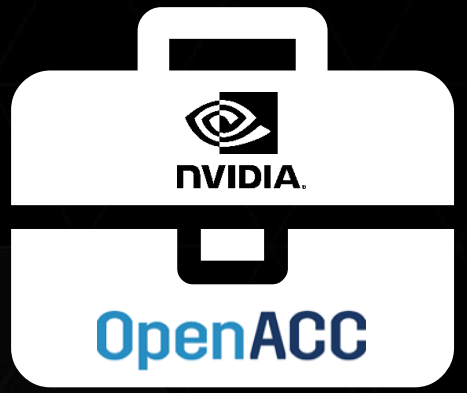
Lines of Code Modified	# of Weeks Required	# of Codes to Maintain
<100 Lines	1 Week	1 Source

Big Performance



Introducing the NVIDIA OpenACC Toolkit

Free Toolkit Offers Simple & Powerful Path to Accelerated Computing



PGI Compiler
Free OpenACC compiler for academia

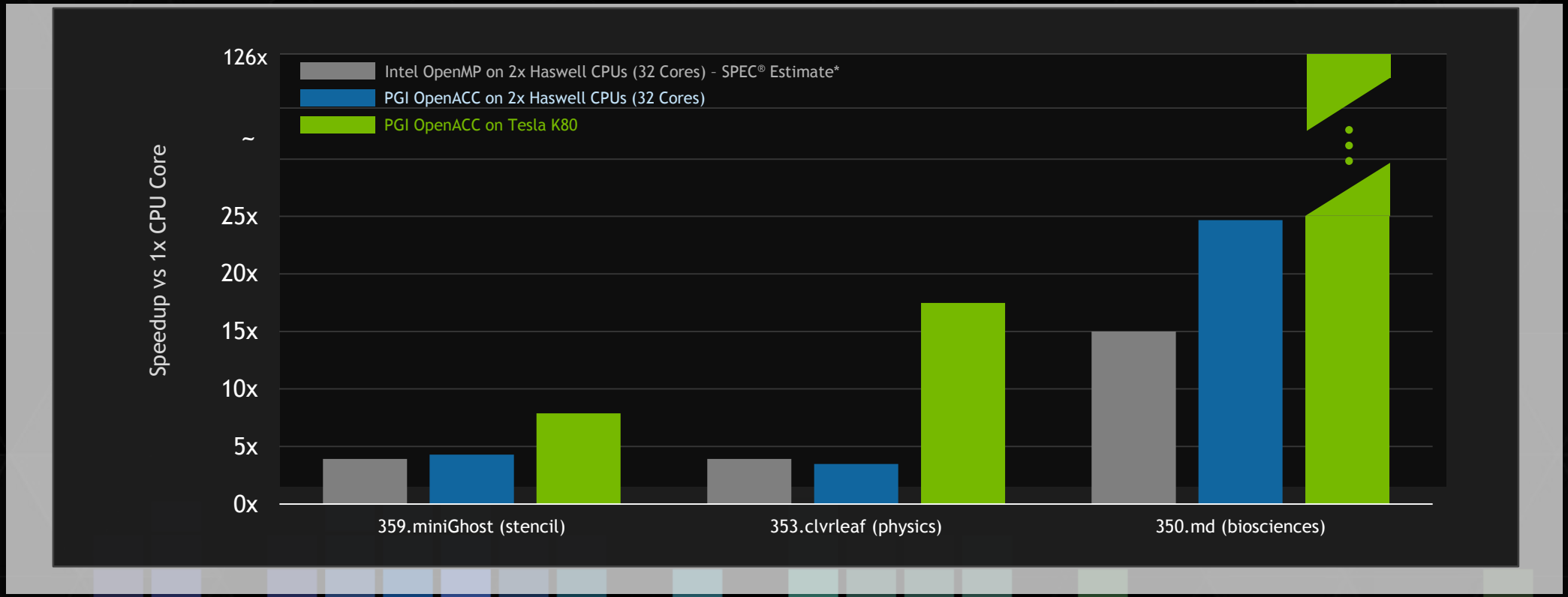
NVProf Profiler
Easily find where to add compiler directives

Code Samples
Learn from examples of real-world algorithms

Documentation
Quick start guide, Best practices, Forums

Download at http://www.nvidia.com/openacc

OpenACC Delivers Performance Portability



PGI®

Supermicro SYS-2028GR-TRT, Intel Xeon E5-2698 v3, 32 cores, NVIDIA Tesla K80, 256GB of System Memory
PGI 15.7 Beta OpenACC Multicore and K80 results from SPEC ACCEL™ measured June 2015.
* Intel 15.0.90 OpenMP results use Cloverleaf reference application and SPEC OMP®2012 using workloads from SPEC ACCEL.
SPEC® and the benchmark names SPEC ACCEL™ and SPEC OMP® are registered trademarks of the Standard Performance Evaluation Corporation.
More info: SPEC ACCEL www.spec.org/accel, OMP2012 www.spec.org/omp2012, CloverLeaf OpenMP uk-mac.github.io/CloverLeaf/, miniGhost ref 1.0.1 https://mantevo.org/download/

Vision: Mainstream Parallel Programming

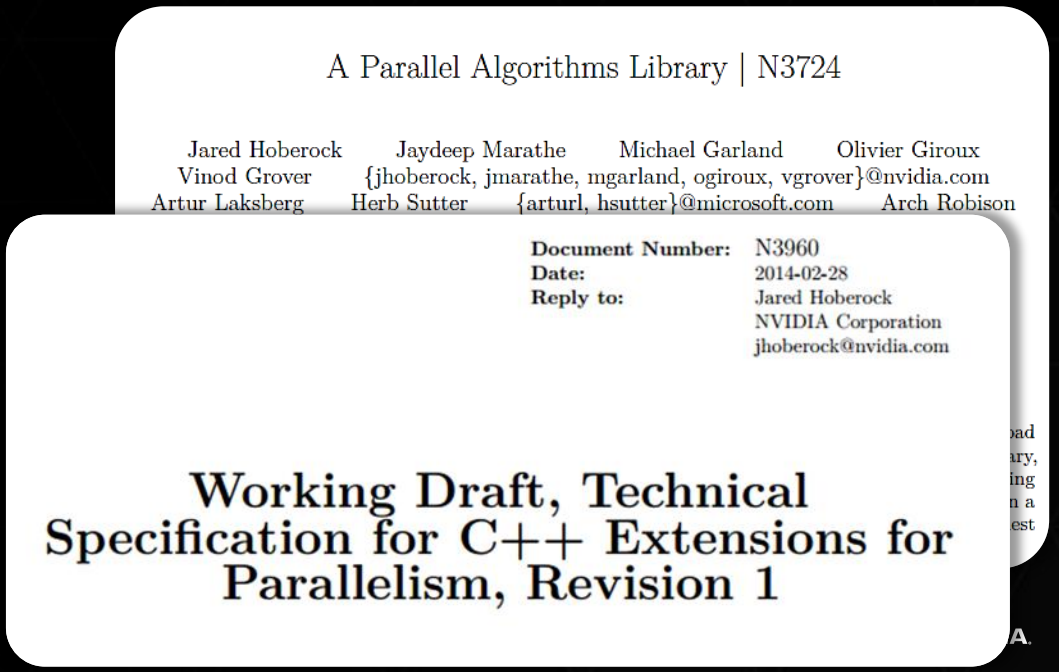
Enable more programmers to write portable parallel software in their language of choice

Embrace and evolve standards in key languages

CUDA continues to evolve as the target low-level platform for GPU acceleration

Portable, High-level Parallel Code TODAY

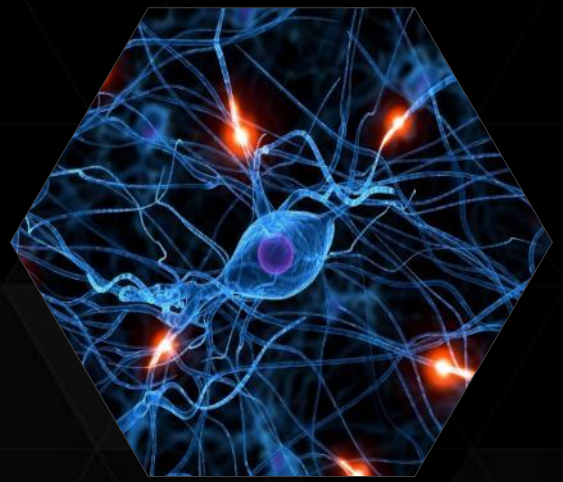
- Thrust library allows the same C++ code to target both:
  - NVIDIA GPUs
  - x86, ARM and POWER CPUs
- Thrust was the inspiration for a proposal to the ISO C++ Committee
- Committee voted unanimously to accept as official tech. specification working draft



N3960 Technical Specification Working Draft:
http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2014/n3960.pdf
Prototype:
https://github.com/n3554/n3554

Three Trends Driving HPC Advances

Algorithms



Large Datasets

Powerful Accelerators

Some Topology Options

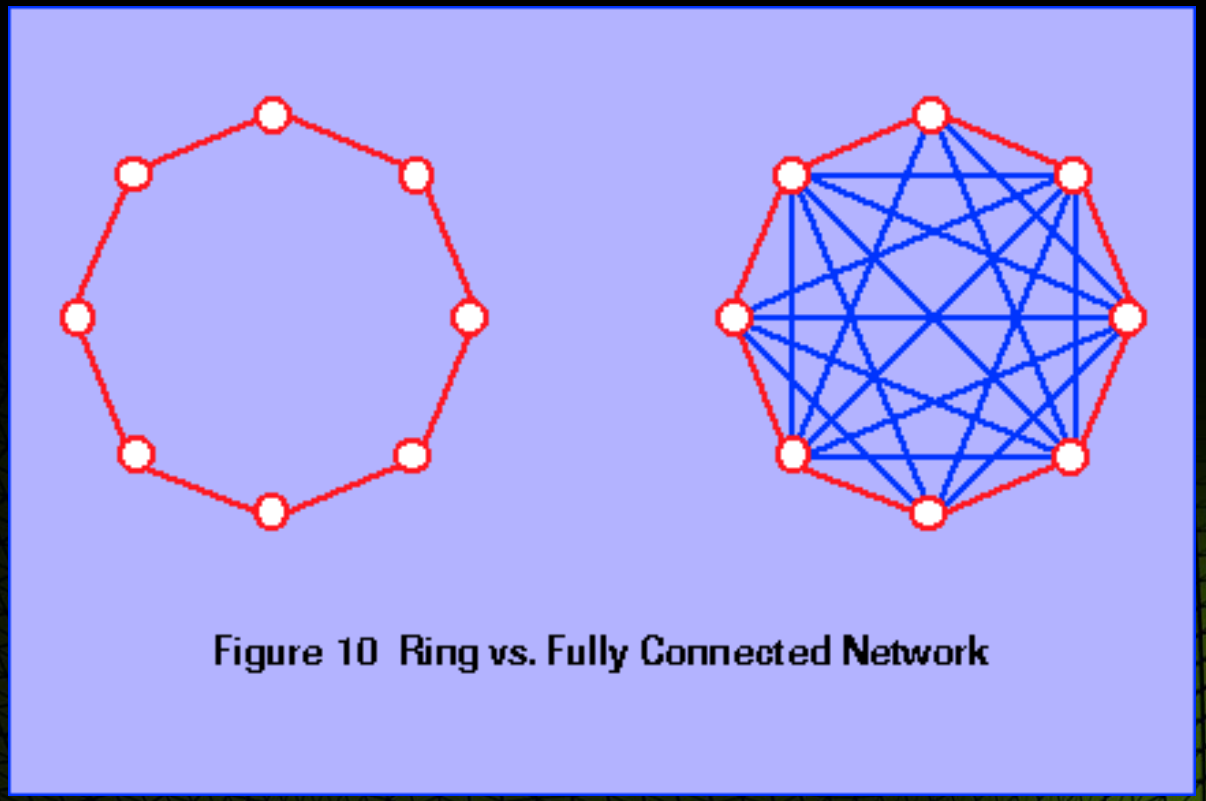
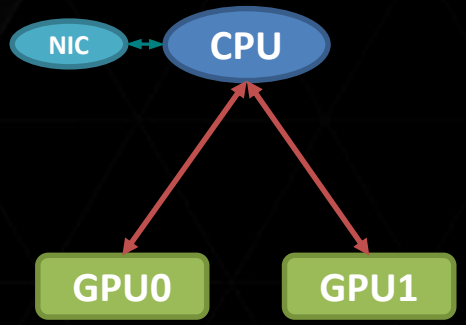
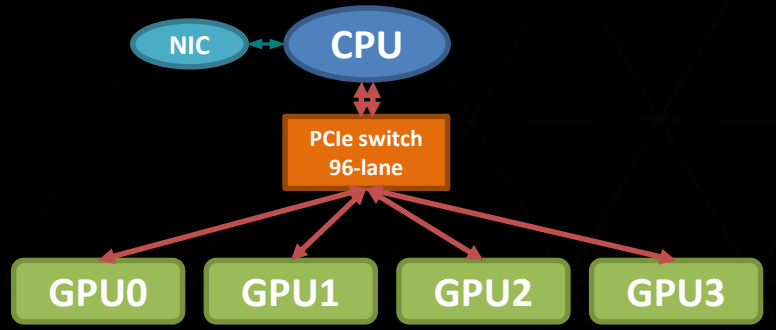


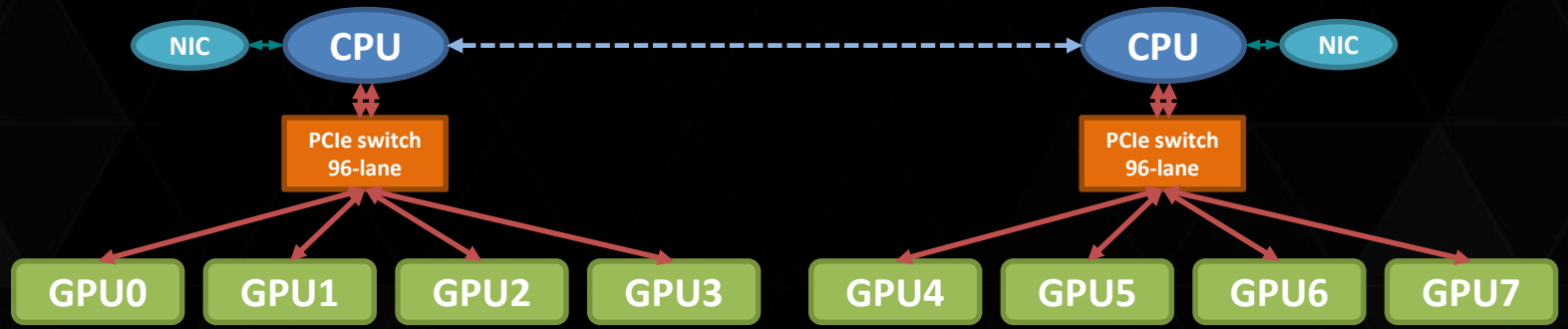
Figure 10 Ring vs. Fully Connected Network

Current PCIe Topologies For Single Nodes

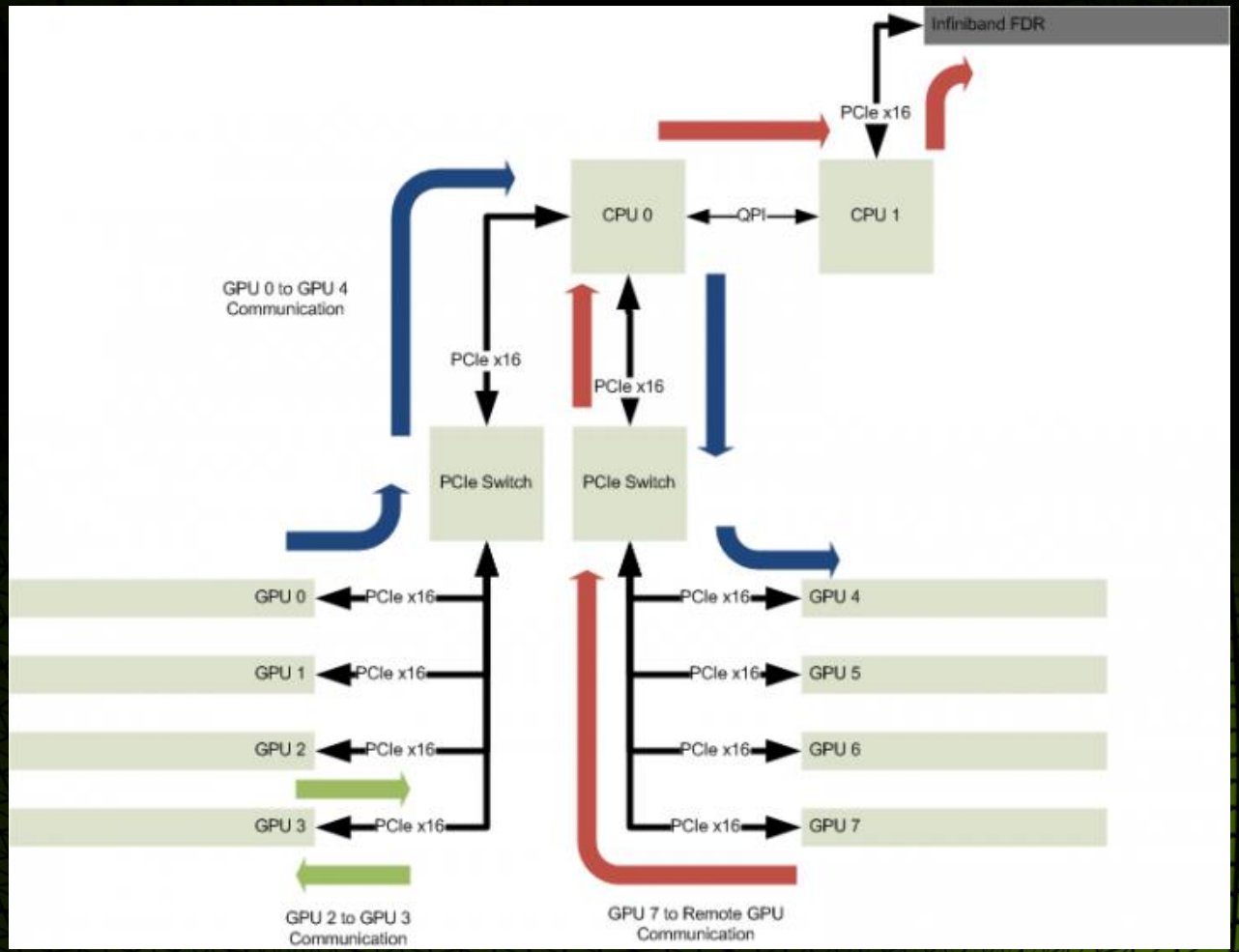
Optimizing for host-to-device bandwidth





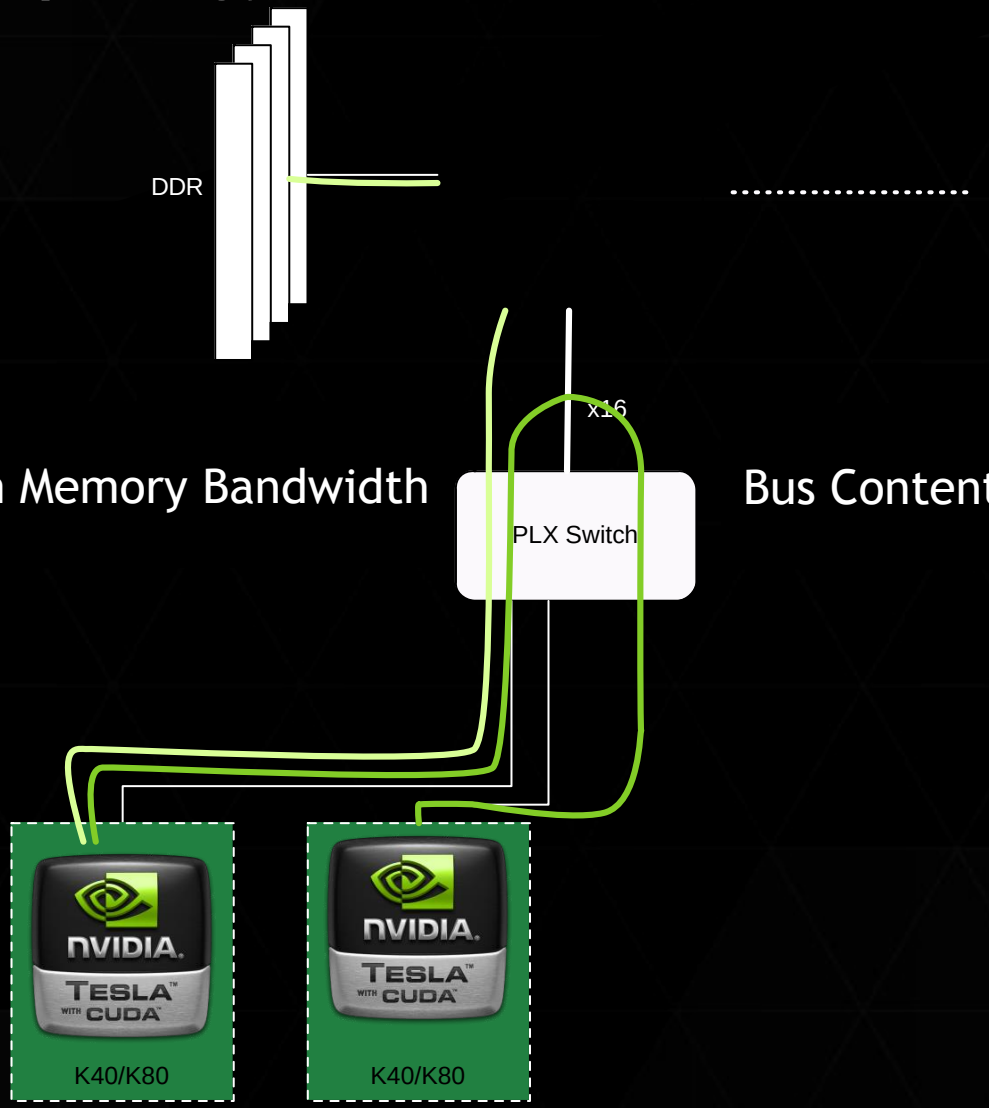


Some 8-GPU Topology Options



http://www.cirrascale.com/blog/index.php/exploring-the-pcie-bus-routes/

x86 GPU Topology – Bandwidth Limits Reached



Limited GPU <-> System Memory Bandwidth

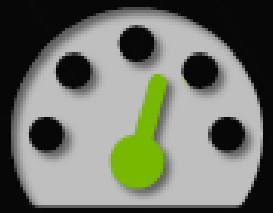
Bus Contention in multi-GPU system

PCIe Bandwidth Limits Reached

	8 GPU		4 GPU		
	PCIe		PCIe		Comment
Point-to-Point					
1D nearest-neighbors	8		12		GB/s to each neighbor
2D nearest-neighbors	9		12		GB/s to each neighbor
3D nearest-neighbors	6		NA		GB/s to each neighbor
Collectives					
Broadcast	7.6		11.3		GB/s received by each
Gather	12		12		GB/s received by master
All-Gather	12		12		GB/s received by each
Scatter	12		12		GB/s unique bytes sent
All-to-all	3.5		6.0		GB/s unique bytes sent
Reduce	7.6		11.3		GB/s unique sent by each
All-Reduce	4.6		8.0		GB/s unique sent by each

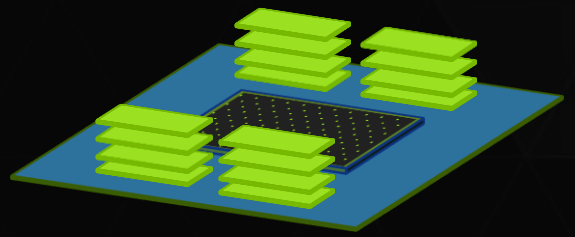
Pascal: Next Generation GPU

Peak Performance



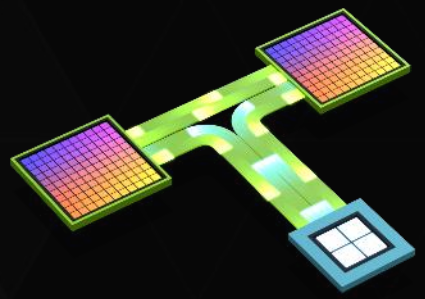
>3 TeraFLOPS

Stacked Memory



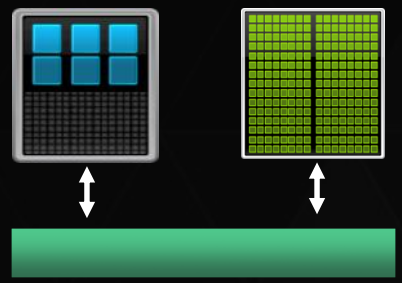
4x Higher Bandwidth (~1 TB/s)
Larger Capacity (16 GB)

NVLink High-Speed Interconnect



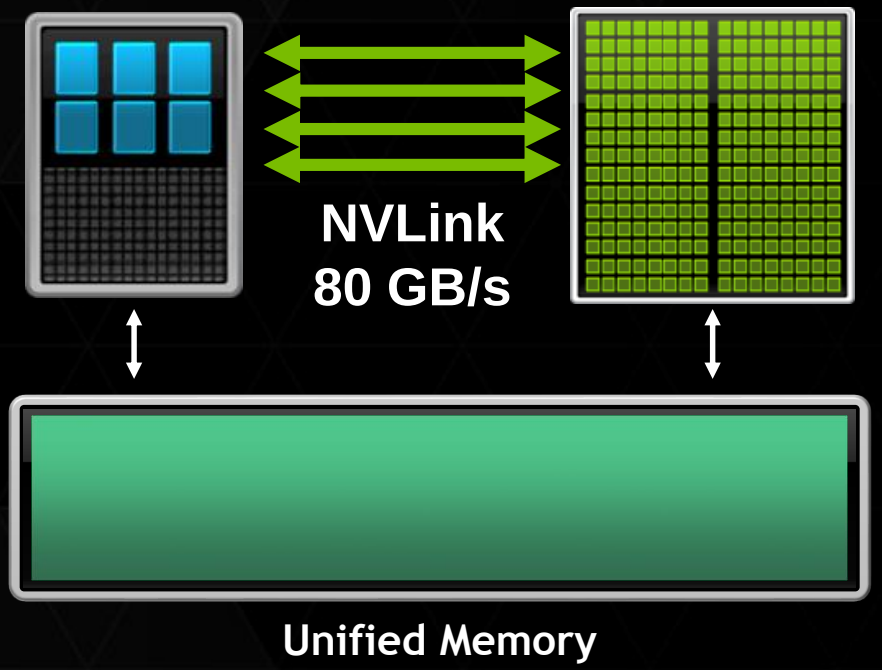
80 GB/sec
POWER CPU & GPU-to-GPU Interconnect

Unified Memory

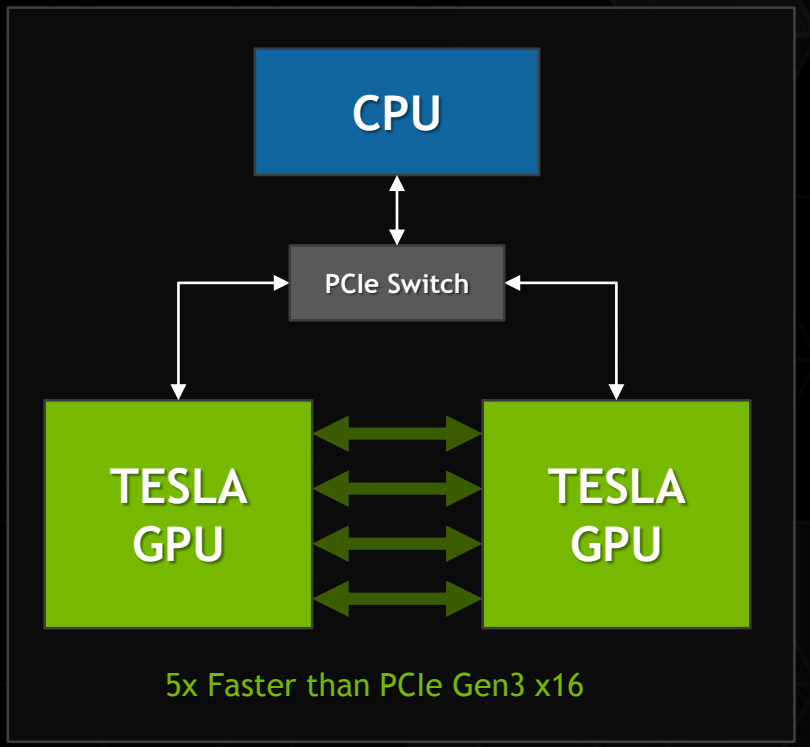


Single Memory Space
Lower Developer Effort

NVLink + Unified Memory: Simpler, Faster



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



Eliminate Multi-GPU
Scaling Bottlenecks

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

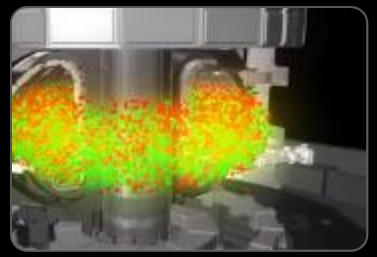
NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

2017

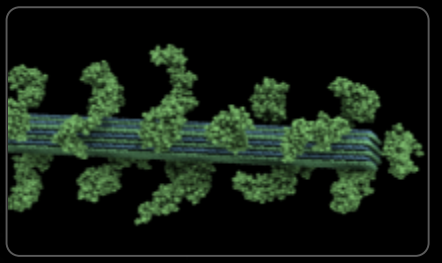
Major Step Forward on the Path to Exascale

CORAL: Built for Grand Scientific Challenges



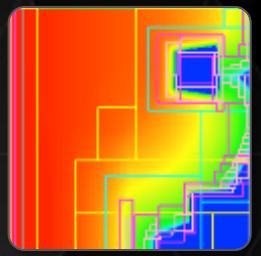
Fusion Energy
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

Climate Change
Study climate change adaptation and mitigation scenarios; realistically represent detailed features

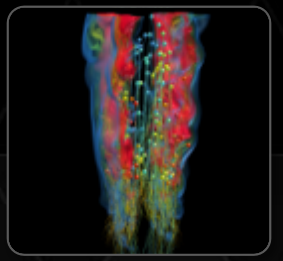


Biofuels
Search for renewable and more efficient energy sources

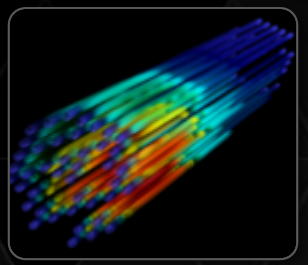
Astrophysics
Radiation transport – critical to astrophysics, laser fusion, atmospheric dynamics, and medical imaging



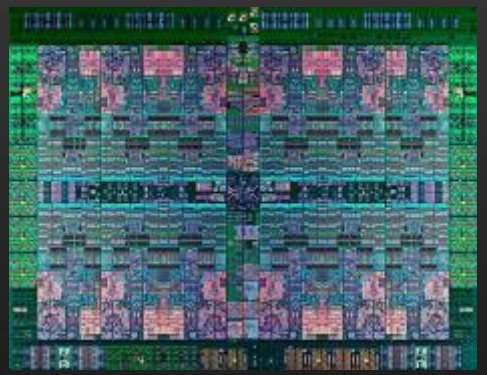
Combustion
Combustion simulations to enable the next gen diesel/bio-fuels to burn more efficiently



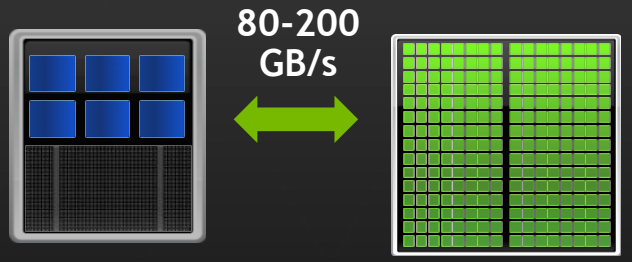
Nuclear Energy
Unprecedented high-fidelity radiation transport calculations for nuclear energy applications



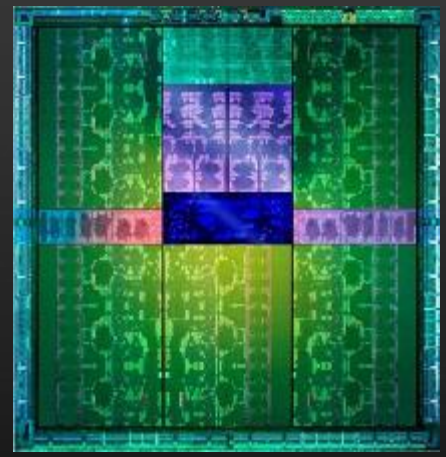
Accelerated Computing
5x Higher Energy Efficiency



IBM POWER CPU
Most Powerful Serial Processor



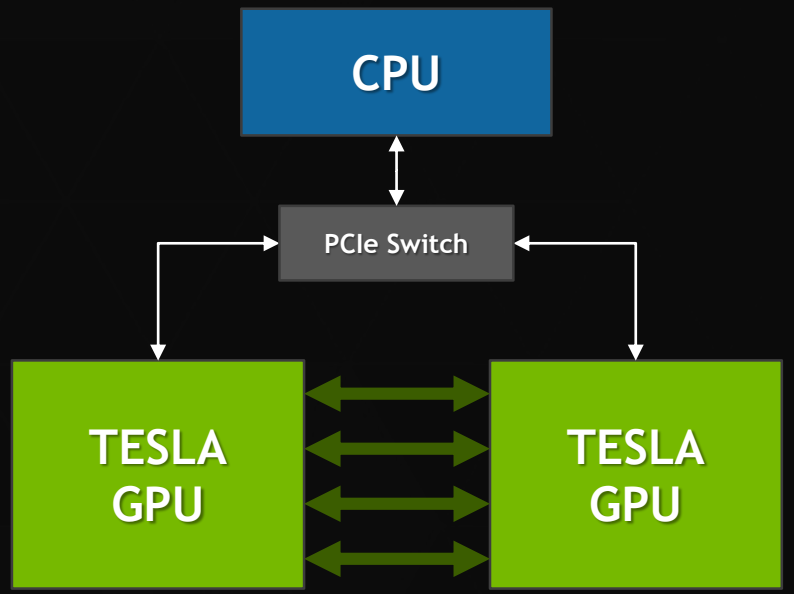
NVIDIA NVLink
Fastest CPU-GPU Interconnect



NVIDIA Volta GPU
Most Powerful Parallel Processor

NVLink Unleashes Multi-GPU Performance

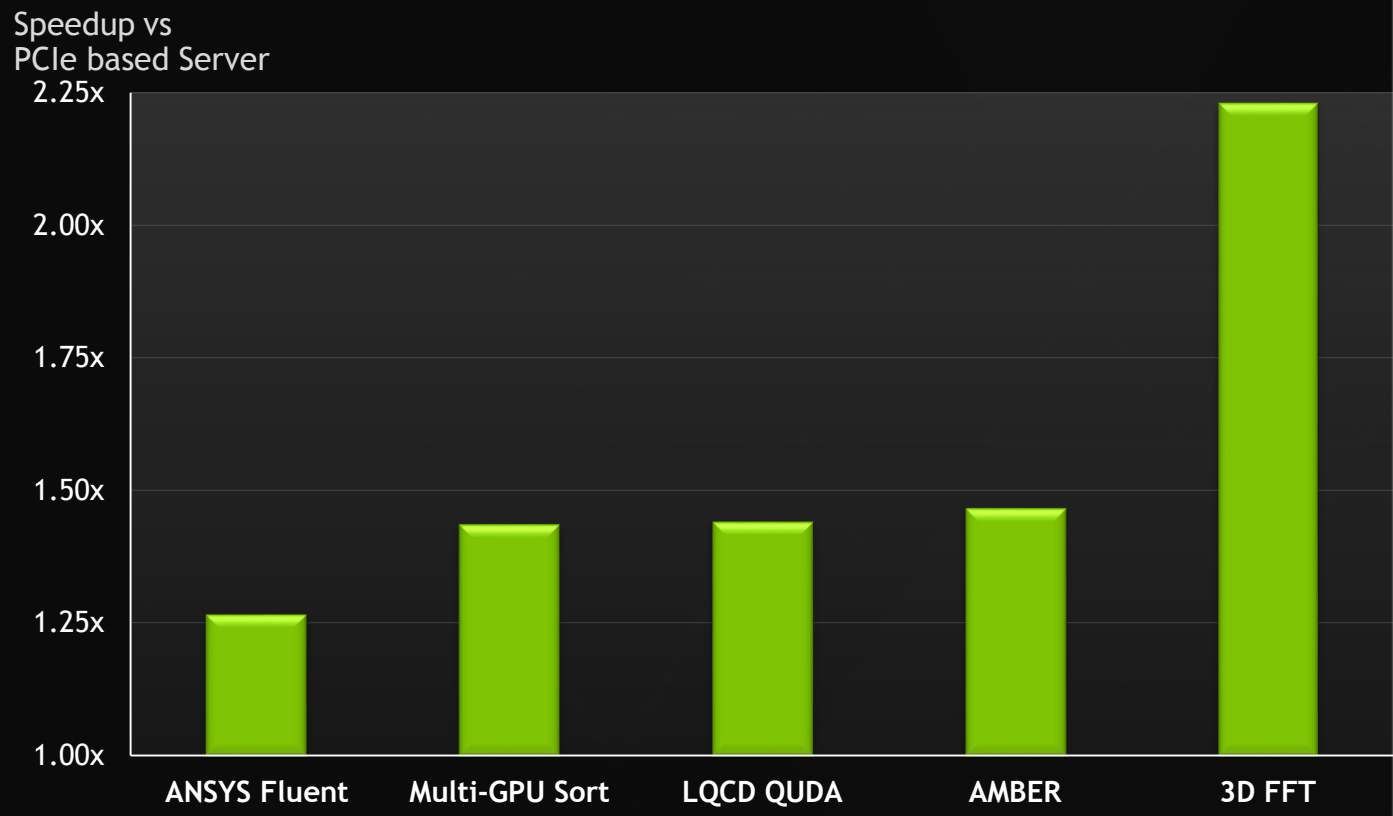
GPUs Interconnected with NVLink



5x Faster than
PCIe Gen3 x16

Over 2x Application Performance Speedup

When Next-Gen GPUs Connect via NVLink Versus PCIe



3D FFT, ANSYS: 2 GPU configuration [illegible] comparing 4 GPU connected
AMBER Cellulose [illegible] 3D FFT problem size 256^3

The Industry's Most Versatile GPU-Accelerated Platform for the Datacenter

Grid Workstation	GRID Virtualization	Tesla HPC	Tesla Deep Learning
Iray	vGPU	CUDA	cuDNN

Tesla System Management and Communication

Flexible Pascal & NVLink Architecture for the Datacenter