NVIDIA ECOSYSTEM

Piero Altoe,

ASTERICS-OBELICS Workshop, 14 December 2016, Rome

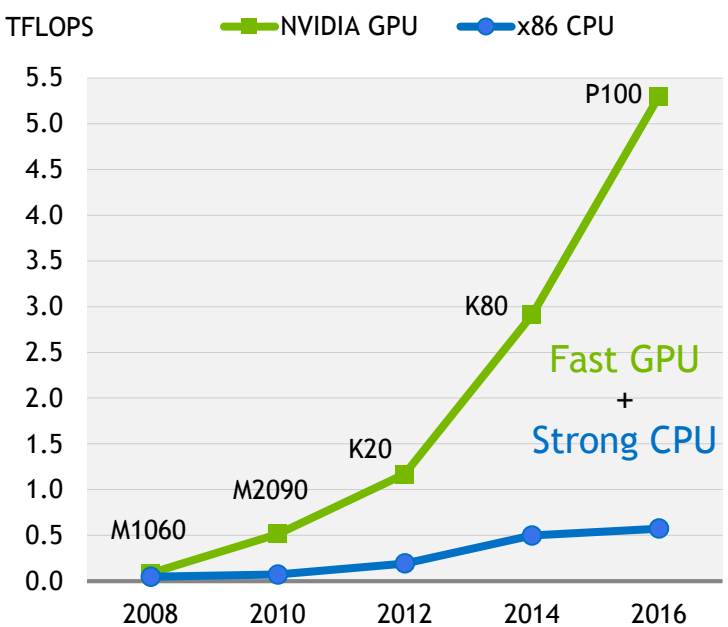
NVIDIA

NVIDIA ECOSYSTEM

TESLA ACCELERATED COMPUTING PLATFORM

Focused on Co-Design for Accelerated Data Center

Fast GPU Engineered for High Throughput



Productive Programming Model & Tools

NVIDIA CUDA

OpenACC
Directives for Accelerators

Expert Co-Design

APPLICATION

MIDDLEWARE

SYS SW

LARGE SYSTEMS

PROCESSOR

Accessibility

阿里云计算 Alibaba Cloud Computing

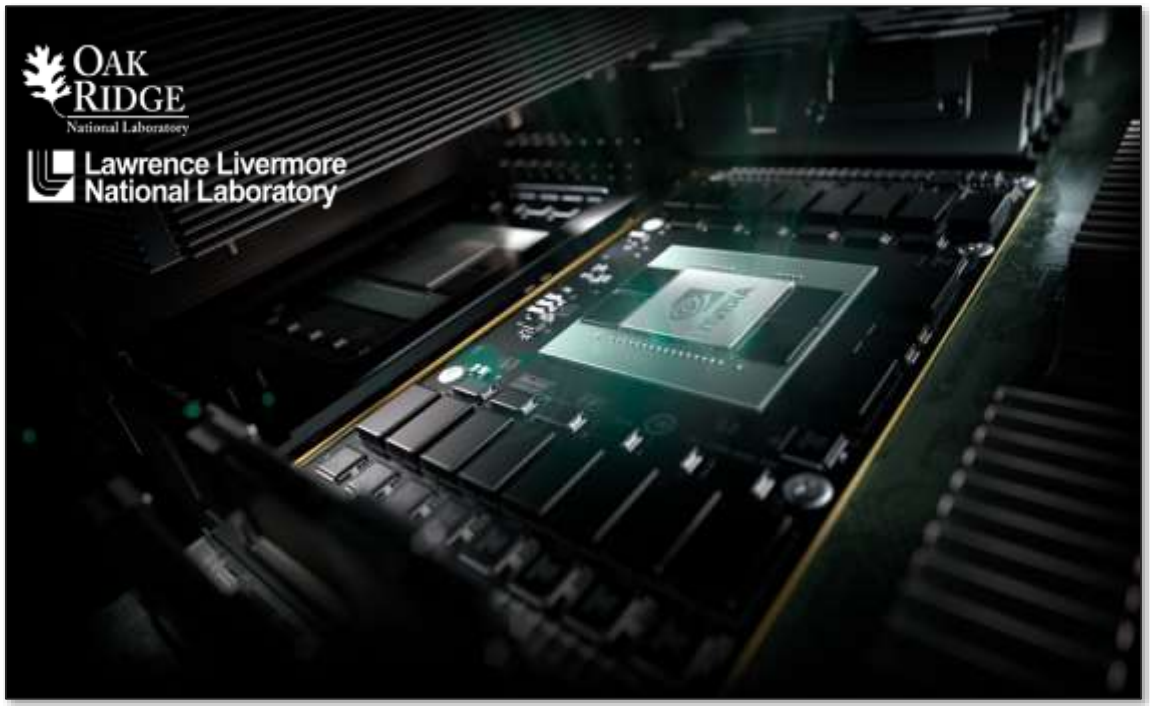
amazon web services

IBM SOFTLAYER

Microsoft Azure

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

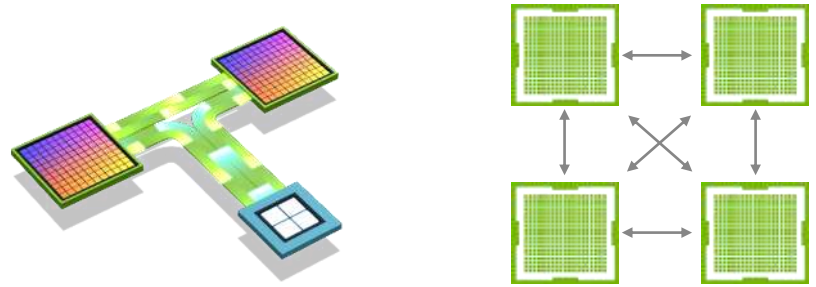
40 TFLOPS per Node, >3,400 Nodes

2017

TESLA PLATFORM LEADS IN EVERY WAY

PROCESSOR

INTERCONNECT



SOFTWARE

OpenACC
Directives for Accelerators

NVIDIA
CUDA

ECOSYSTEM

ParaView

torch

IBM

NAMD
Scalable Molecular Dynamics

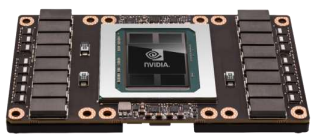
DMTK

TensorFlow

TESLA PLATFORM FOR STRONG SCALING HPC

END-TO-END PRODUCT FAMILY

HYPERSCALE HPC

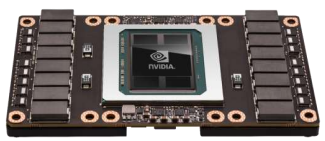


Training - Tesla P100

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

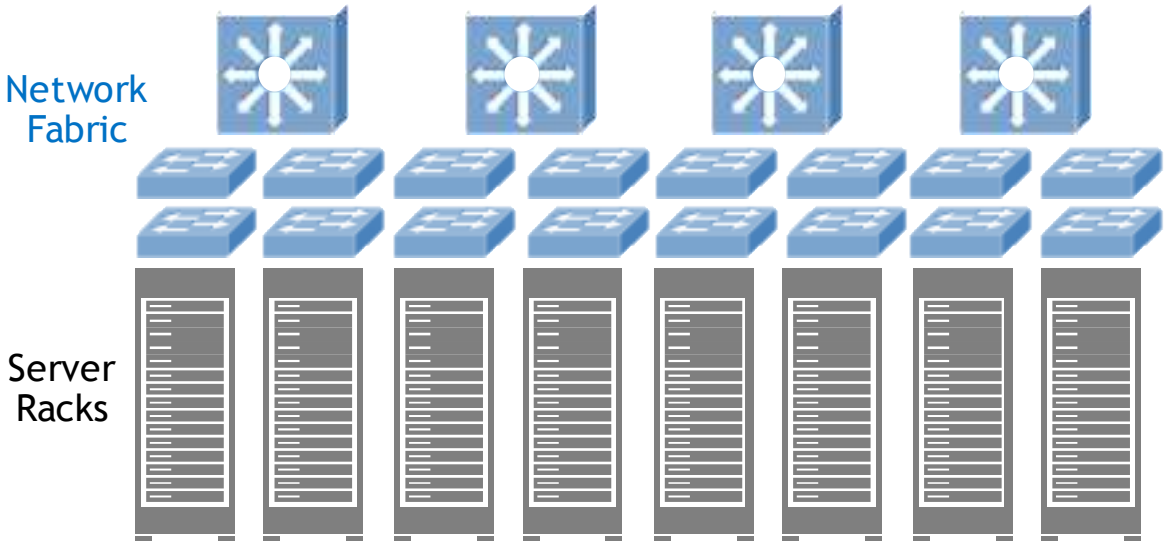
FULLY INTEGRATED DL SUPERCOMPUTER

DGX-1

Fully integrated deep learning solution

WEAK NODES

Lots of Nodes Interconnected with
Vast Network Overhead



STRONG NODES

Few Lightning-Fast Nodes with
Performance of Hundreds of Weak Nodes

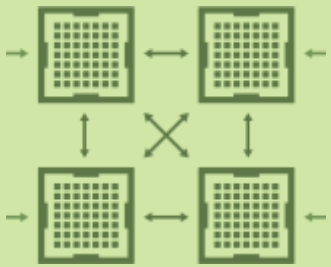
INTRODUCING TESLA P100

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

Pascal Architecture

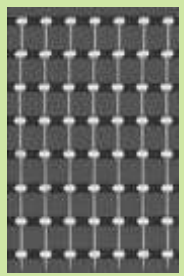
Highest Compute Performance

NVLink



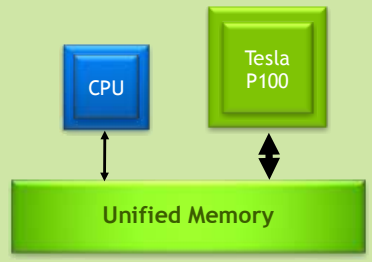
GPU Interconnect for Maximum Scalability

CoWoS HBM2



Unifying Compute & Memory in Single Package

Page Migration Engine

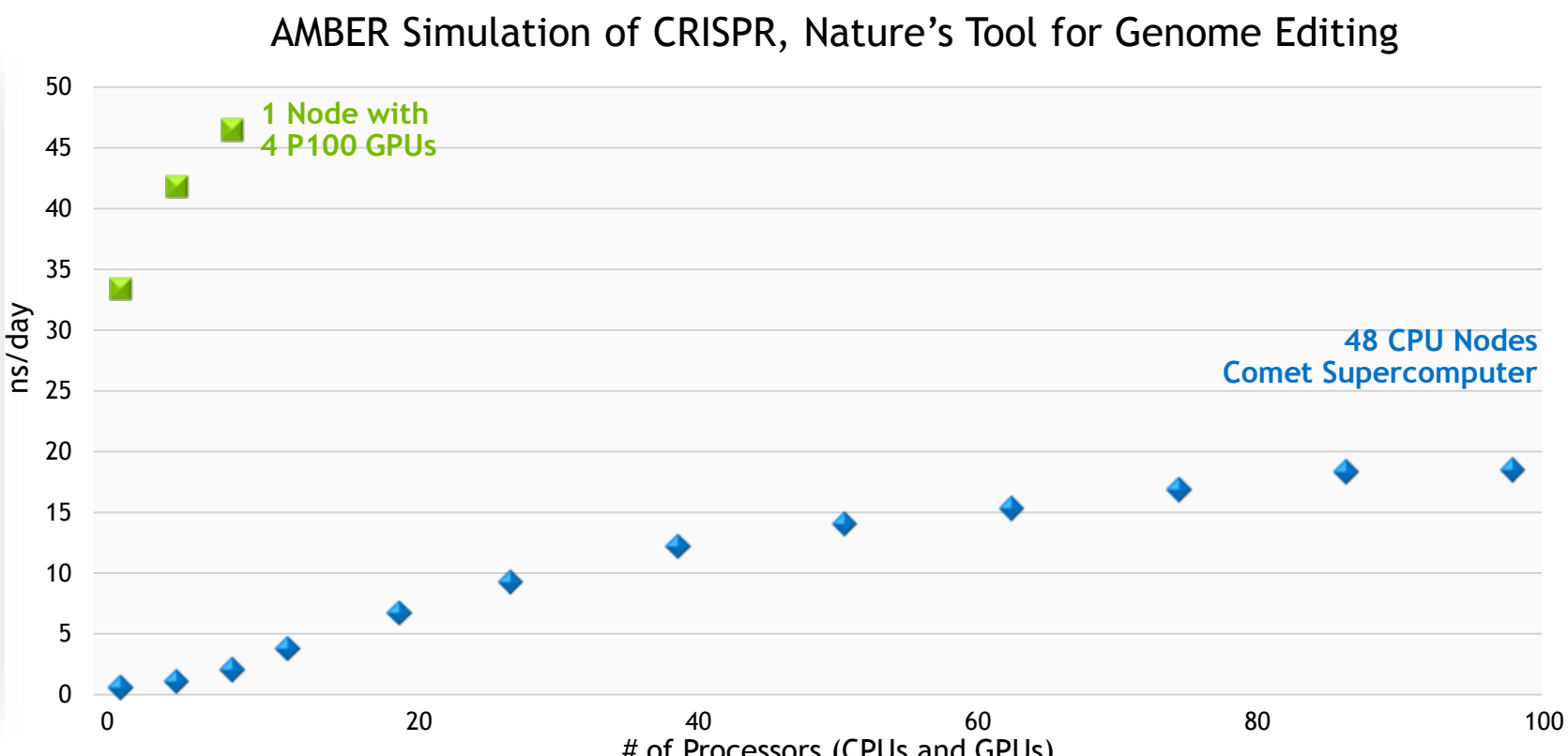


Simple Parallel Programming with Virtually Unlimited Memory Space

BIG PROBLEMS NEED FAST COMPUTERS

2.5x Faster than the Largest CPU Data Center

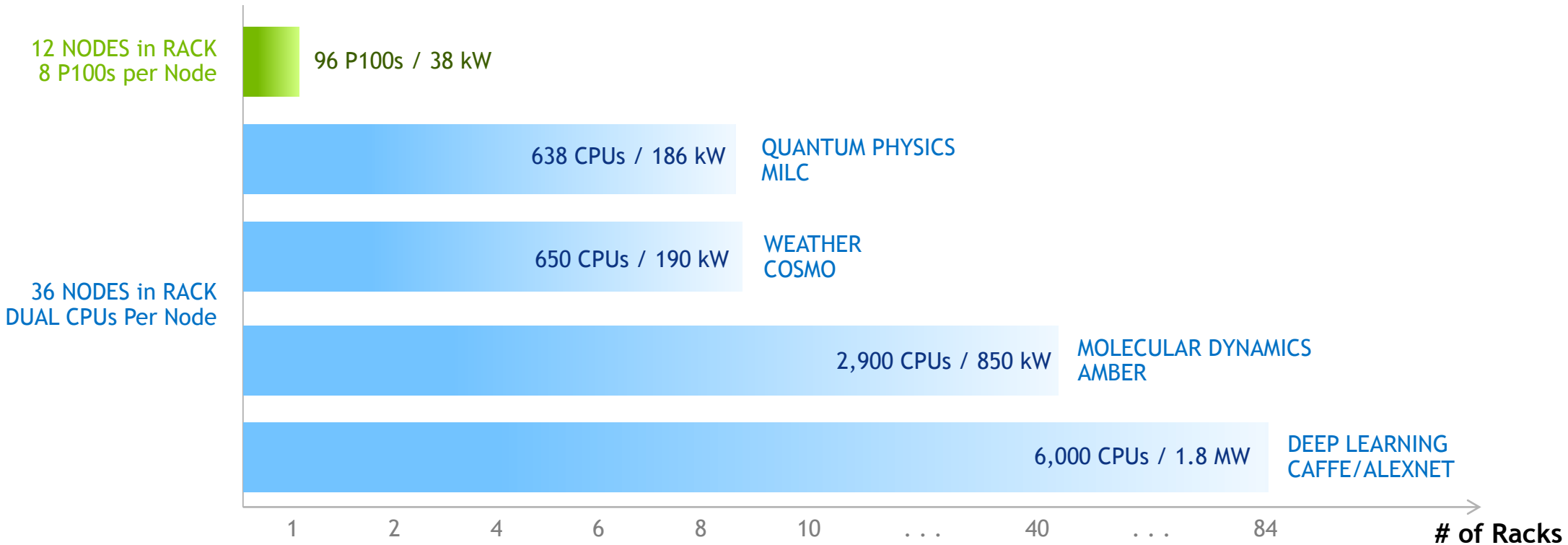
"Biotech discovery of the century"
-MIT Technology Review 12/2014



AMBER 16 Pre-release, CRSPR based on PDB ID 5f9r, 336,898 atoms
CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB

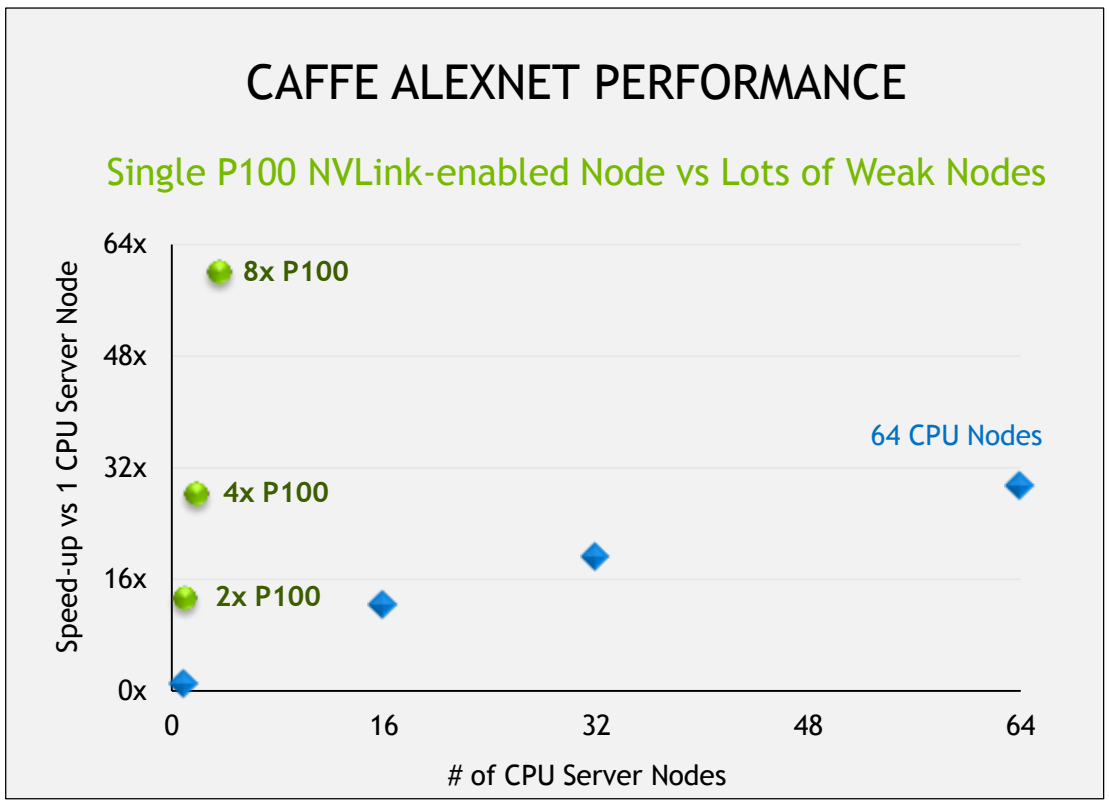
DATACENTER IN A RACK

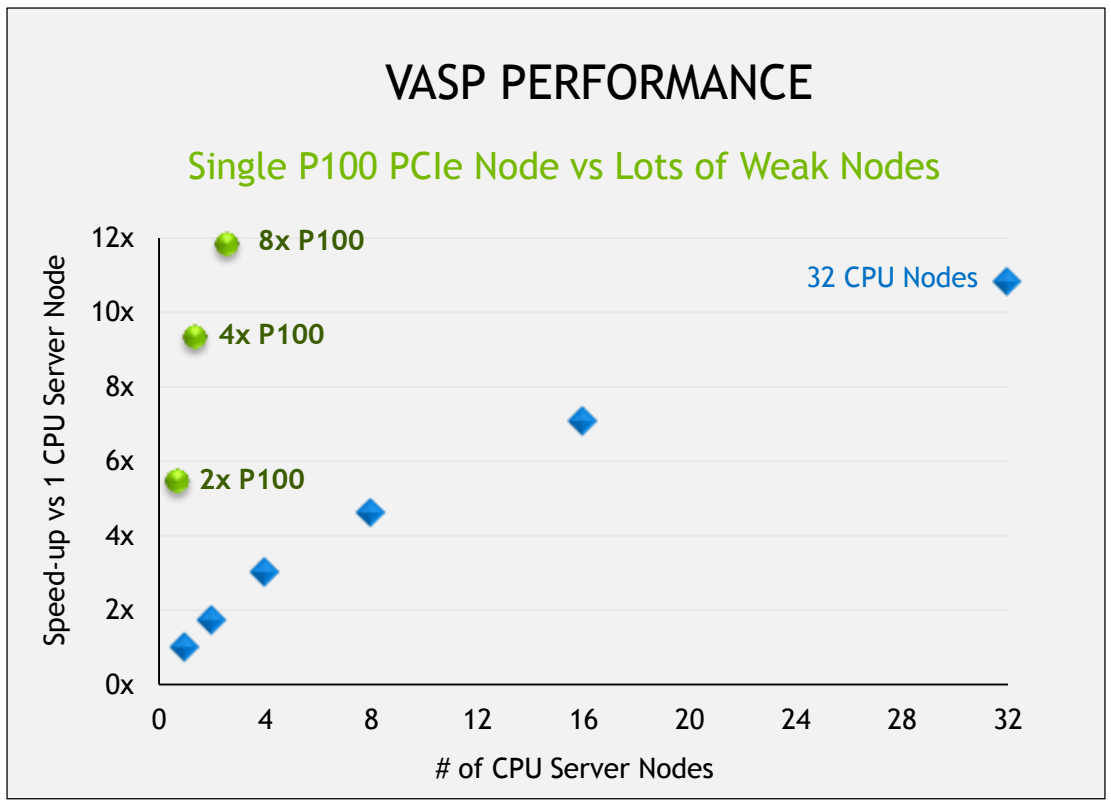
1 Rack of Tesla P100 Delivers Performance of 6,000 CPUs



EXTRAORDINARY STRONG SCALING

One Strong Node Faster Than Lots of Weak Nodes





CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB
VASP 5.4.1_05Feb16, Si-Huge Dataset. 16, 32 Nodes are estimated based on same scaling from 4 to 8 nodes
Caffe AlexNet scaling data: https://software.intel.com/en-us/articles/caffe-training-on-multi-node-distributed-memory-systems-based-on-intel-xeon-processor-e5

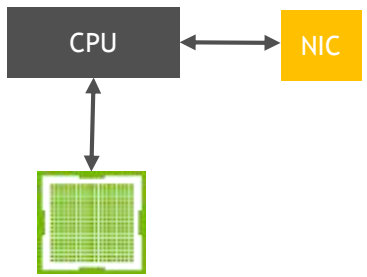
TESLA PRODUCTS DECODER

	K80	M40	M4	P100 (SXM2)	P100 (PCIE)	P40	P4
GPU	2x GK210	GM200	GM206	GP100	GP100	GP102	GP104
PEAK FP64 (TFLOPs)	2.9	NA	NA	5.3	4.7	NA	NA
PEAK FP32 (TFLOPs)	8.7	7	2.2	10.6	9.3	12	5.5
PEAK FP16 (TFLOPs)	NA	NA	NA	21.2	18.7	NA	NA
PEAK TIOPs	NA	NA	NA	NA	NA	47	22
Memory Size	2x 12GB GDDR5	24 GB GDDR5	4 GB GDDR5	16 GB HBM2	16/12 GB HBM2	24 GB GDDR5	8 GB GDDR5
Memory BW	480 GB/s	288 GB/s	80 GB/s	732 GB/s	732/549 GB/s	346 GB/s	192 GB/s
Interconnect	PCIe Gen3	PCIe Gen3	PCIe Gen3	NVLINK + PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3
ECC	Internal + GDDR5	GDDR5	GDDR5	Internal + HBM2	Internal + HBM2	GDDR5	GDDR5
Form Factor	PCIE Dual Slot	PCIE Dual Slot	PCIE LP	SXM2	PCIE Dual Slot	PCIE Dual Slot	PCIE LP
Power	300 W	250 W	50-75 W	300 W	250 W	250 W	50-75 W

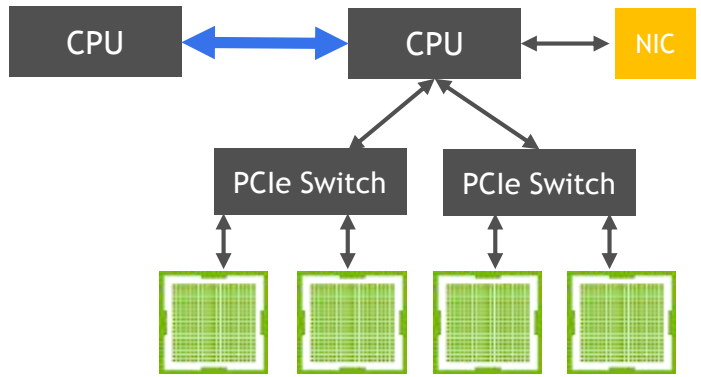
TOPOLOGY INFORMATION

SYSTEM TOPOLOGY USED FOR K80/P100 PCIE

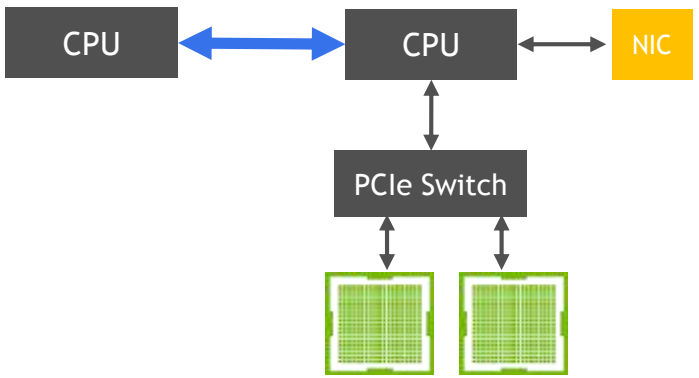
1 CPU – 1x (P100, K80)



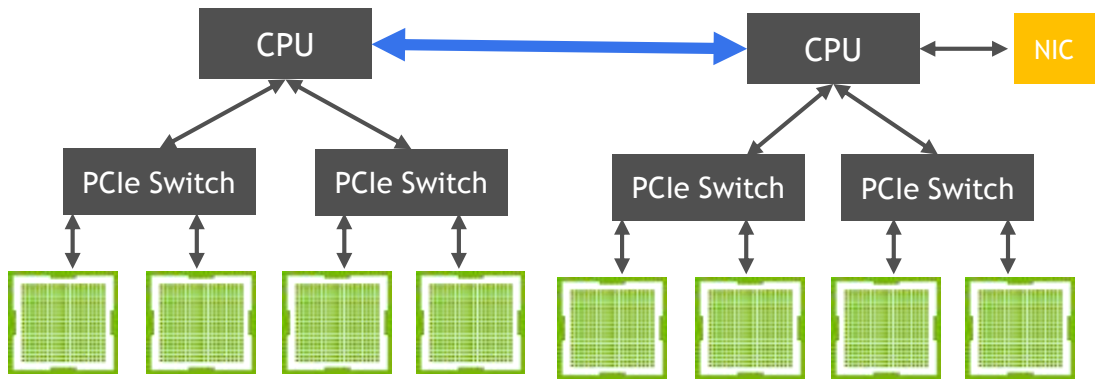
2 CPU – 4x (P100, K80)



2 CPU – 2x (P100, K80)

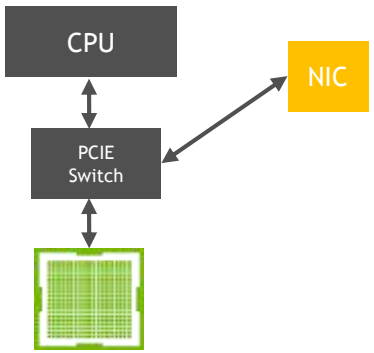


2 CPU – 8x (P100, K80)

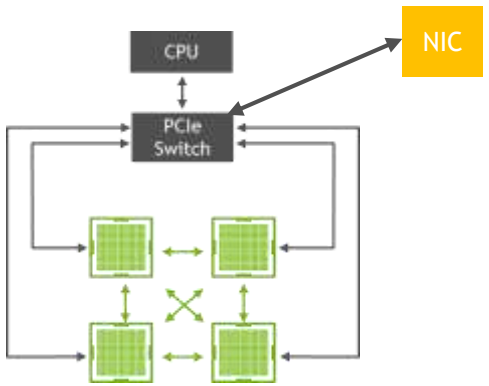


SYSTEM TOPOLOGY USED FOR P100 SXM2

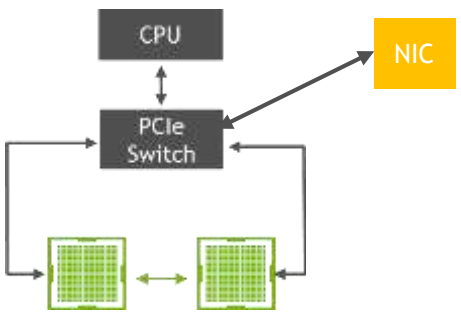
1 CPU – 1x P100 SXM2



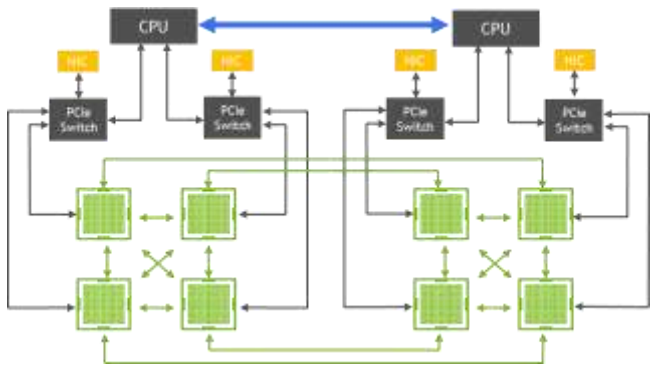
2 CPU – 4x P100 SXM2



2 CPU – 2x P100 SXM2



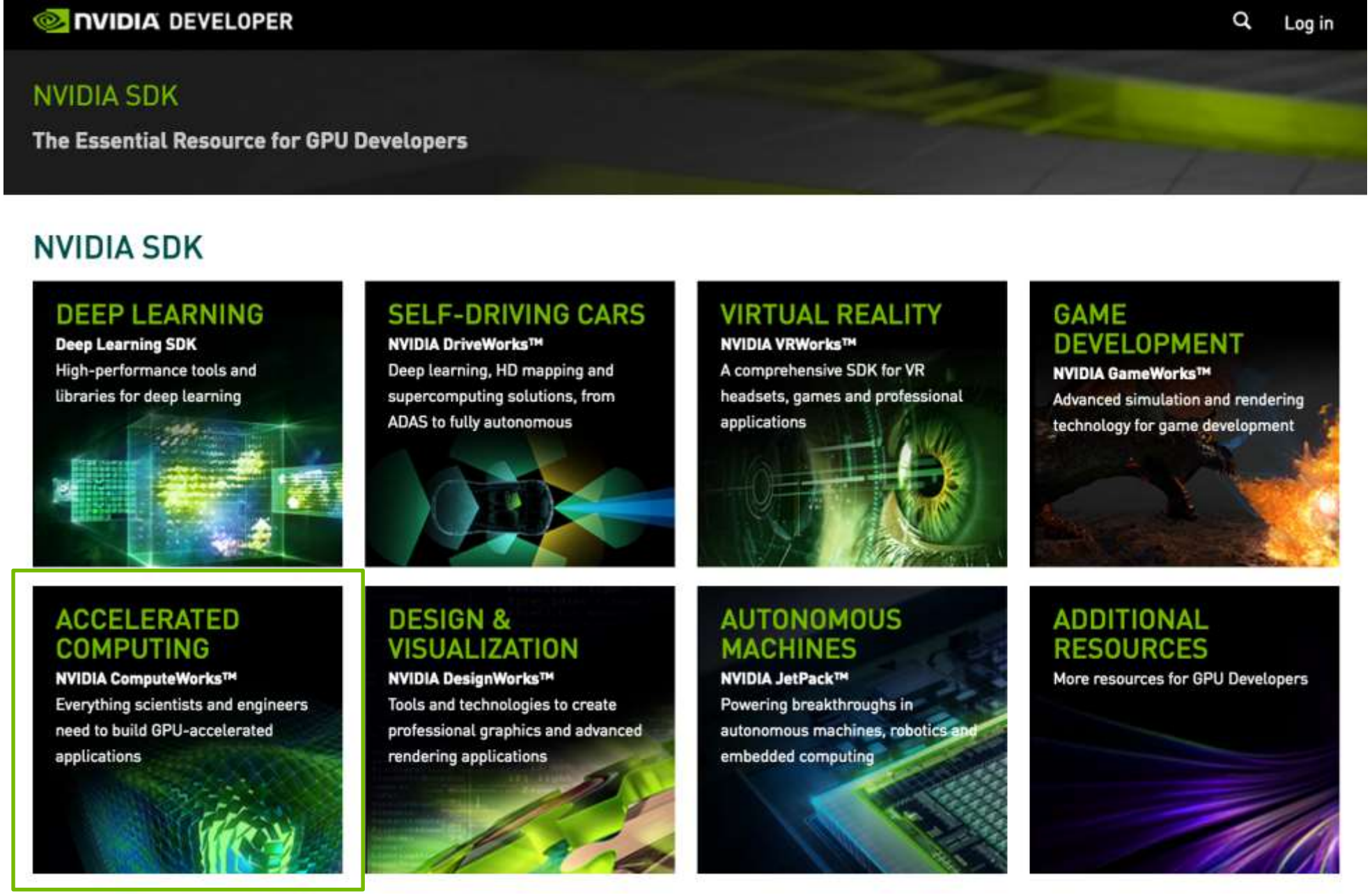
2 CPU – 8x P100 SXM2



NVLink

PCIE Gen 3

TESLA PLATFORM FOR DEVELOPERS

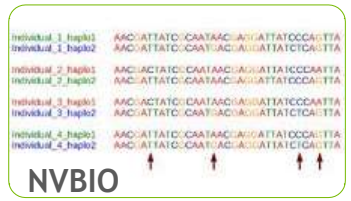

NVIDIA DEVELOPER
Log in
NVIDIA SDK
The Essential Resource for GPU Developers
NVIDIA SDK
DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning
SELF-DRIVING CARS
NVIDIA DriveWorks™
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous
VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications
GAME DEVELOPMENT
NVIDIA GameWorks™
Advanced simulation and rendering technology for game development
ACCELERATED COMPUTING
NVIDIA ComputeWorks™
Everything scientists and engineers need to build GPU-accelerated applications
DESIGN & VISUALIZATION
NVIDIA DesignWorks™
Tools and technologies to create professional graphics and advanced rendering applications
AUTONOMOUS MACHINES
NVIDIA JetPack™
Powering breakthroughs in autonomous machines, robotics and embedded computing
ADDITIONAL RESOURCES
More resources for GPU Developers

GPU ACCELERATED LIBRARIES

"Drop-in" Acceleration for Your Applications

Domain-specific
Deep Learning, GIS, EDA, Bioinformatics, Fluids

cuDNN



gpp
Geometry Performance Primitives

Sundog Software
Triton Ocean SDK

Visual Processing
Image & Video

FFmpeg

OpenCV

Linear Algebra
Dense, Sparse, Matrix

SuiteSparse
CHOLMOD

MAGMA
ICL UT

[HiPLAR]

Math Algorithms
AMG, Templates, Solvers

IMSL
RogueWave

Thrust
&
ArrayFire

developer.nvidia.com/gpu-accelerated-libraries

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

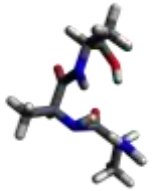
7-8x Speed-Up
5% of Code Modified

8000+

Developers

using OpenACC

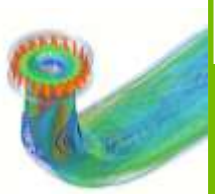
http://www.cray.com/sites/default/files/resources/OpenACC_213462.12_OpenACC_Cosmo_CS_FNL.pdf
http://www.hpcwire.com/off-the-wire/first-round-of-2015-hackathons-gets-underway
http://on-demand.gputechconf.com/gtc/2015/presentation/S5297-Hisashi-Yashiro.pdf
http://www.openacc.org/content/experiences-porting-molecular-dynamics-code-gpus-cray-xk7



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

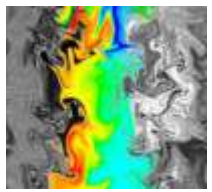
Computational Electromagnetics

2.5X speedup
60% less energy

COSMO

Climate Weather

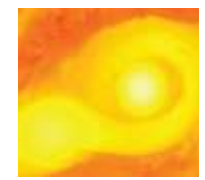
40X speedup
3X energy efficiency



CloverLeaf

CFD

4X speedup
Single CPU/GPU code



MAESTRO CASTRO

Astrophysics

4.4X speedup
4 weeks effort

PGI Community Edition

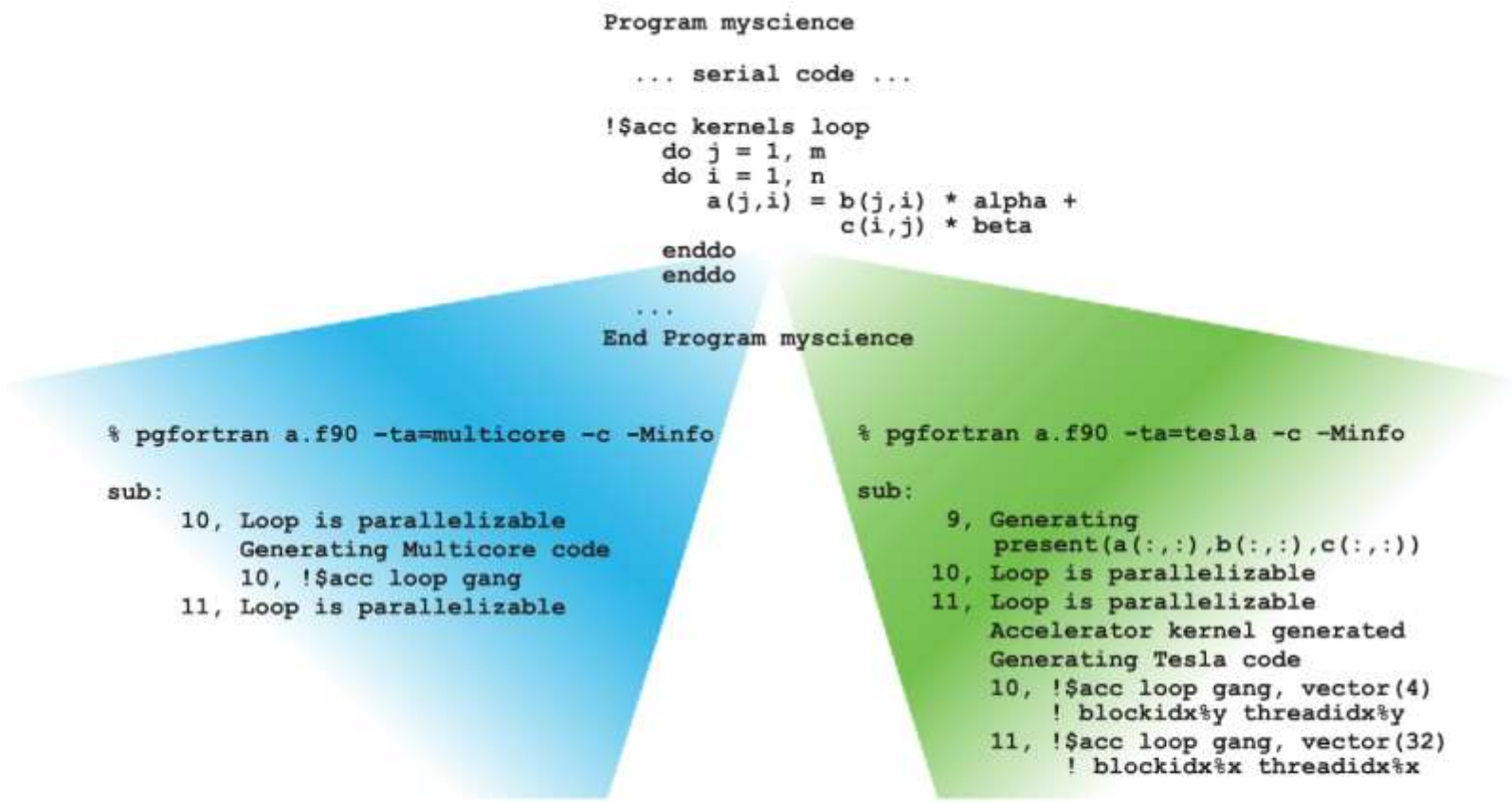
A no-cost license to a recent release of the PGI Fortran, C and C++ compilers and tools for multicore CPUs and NVIDIA Tesla GPUs, including all OpenACC, OpenMP and CUDA Fortran features. The PGI Community Edition enables development of performance-portable HPC applications with uniform source code across the most widely used parallel processors and systems.

The PGI Product Feature Comparison lists differences between the PGI Community Edition and the for-fee PGI Professional Edition. Please see the PGI Community Edition FAQ for additional information.

Download PGI Community Edition Version 16.10 (released November 14, 2016)

Platform	MD5 Checksum
Linux x86	9bb6bfb7b1052f9e6a45829ba7a24e47
Linux OpenPOWER	778aaaf5f9ac53cb857364ab94cf4bc4
macOS	be090ffc79e034845405cd59f7ce6b71

Accelerate Your Code with OpenACC



CUDA

Super Simplified Memory Management Code

CPU Code

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort(data, N, 1, compare);


  use_data(data);

  free(data);
}
```

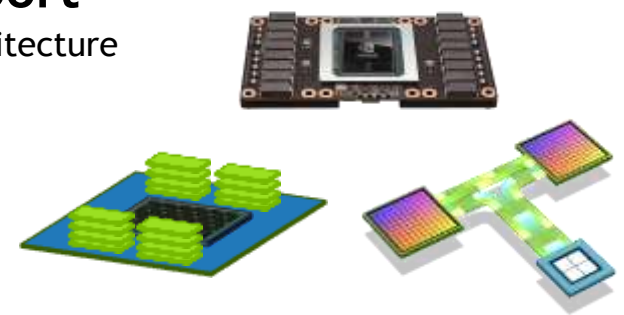
CUDA 6 Code with Unified Memory

```
void sortfile(FILE *fp, int N) {
  char *data;
  cudaMallocManaged(&data, N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  cudaFree(data);
}
```

CUDA 8 – WHAT'S NEW

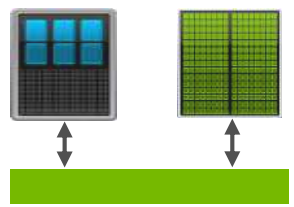
P100 Support

New Pascal Architecture
Stacked Memory
NVLINK
FP16 math



Unified Memory

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

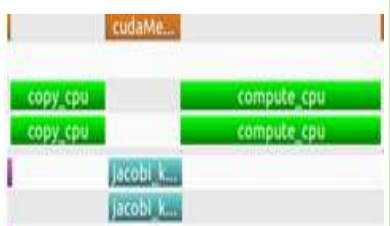


Libraries

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



NVIDIA'S GPU EDUCATORS PROGRAM

Advancing STEM Education with Accelerated Computing

The Flagship Offering: **GPU Teaching Kits** - Breaking the barriers of GPU education in academia:

- Lecture slides
- Lecture videos
- Hands-on labs/solutions
- Larger coding projects/solutions
- Quiz/exam questions/solutions
- Text and e-books

Different kits for different courses

- **Accelerated/Parallel Computing** *(available now!)*
- **Robotics** *(available now!)*
- **Machine/Deep Learning** *(available now!)*
- Computer Vision, Computer Architecture, Computational Domain Sciences, Mathematics, etc. *(future)*

Get started today!

developer.nvidia.com/educators

BENCHMARKS

LINPACK

Benchmark
Measures floating point computing power

Accelerated Features	Metric	Scalability
All	GFLOPS	Multi-GPU, Multi Node

https://www.top500.org/project/**linpack**/

Linpack 2.1

Speedup Vs Dual-Socket CPU Server

CPU Server: Dual Xeon E5-2699 v4@2.2GHz (44-cores)
GPU Servers: Dual Xeon E5-2699 v4@2.2GHz (44-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.44
Dataset: HPL.dat

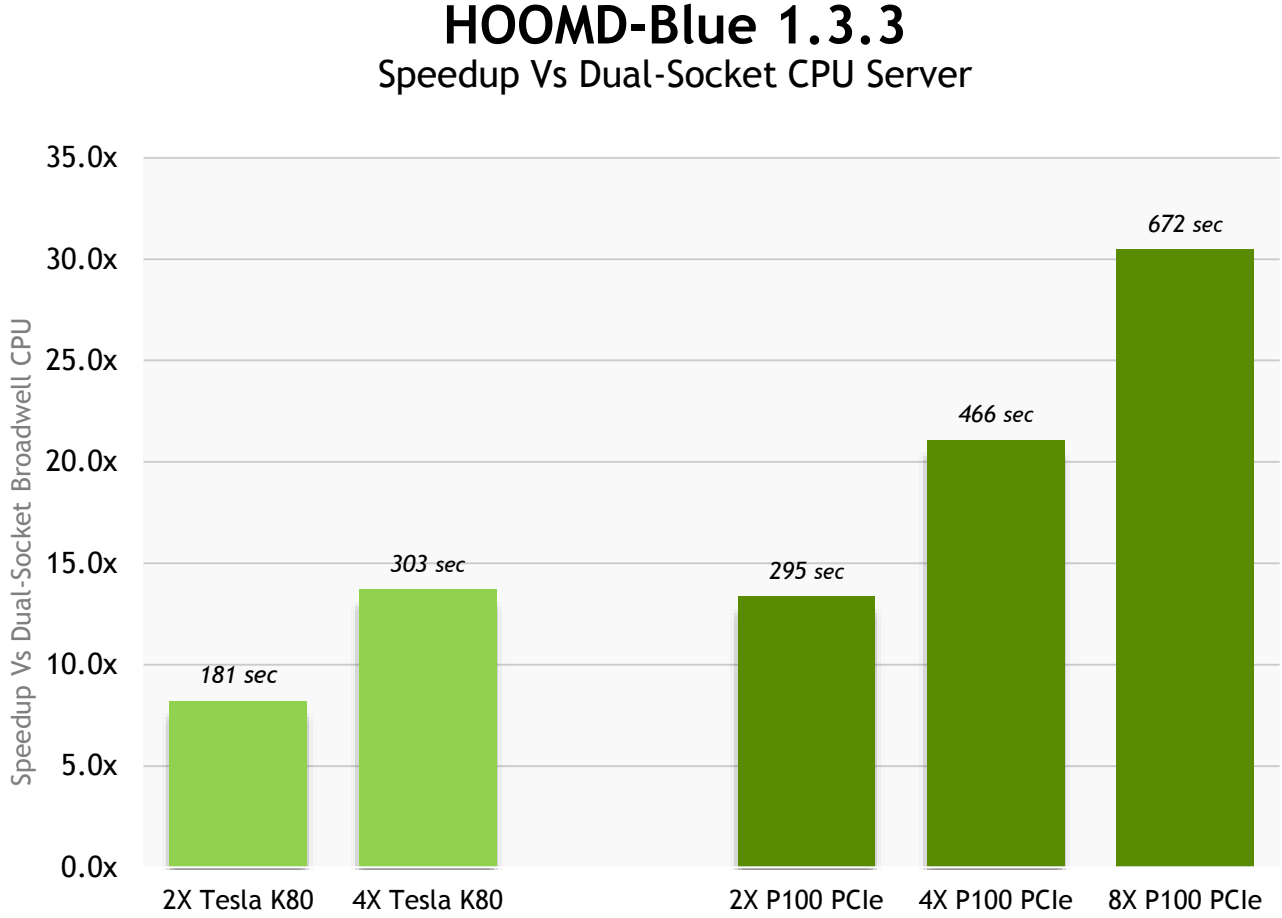
HOOMD-BLUE

Molecular Dynamics

Particle dynamics package written grounds up for GPUs

Accelerated Features	Metric	Scalability
CPU & GPU versions available	Avg. Timesteps (secs)	Multi-GPU, multi-node

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



CPU Server: Dual Xeon E5-2699 v4@2.2GHz (44-cores)
GPU Servers: Dual Xeon E5-2699 v4@2.2GHz (44-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.44
Dataset: lj_liquid_1m

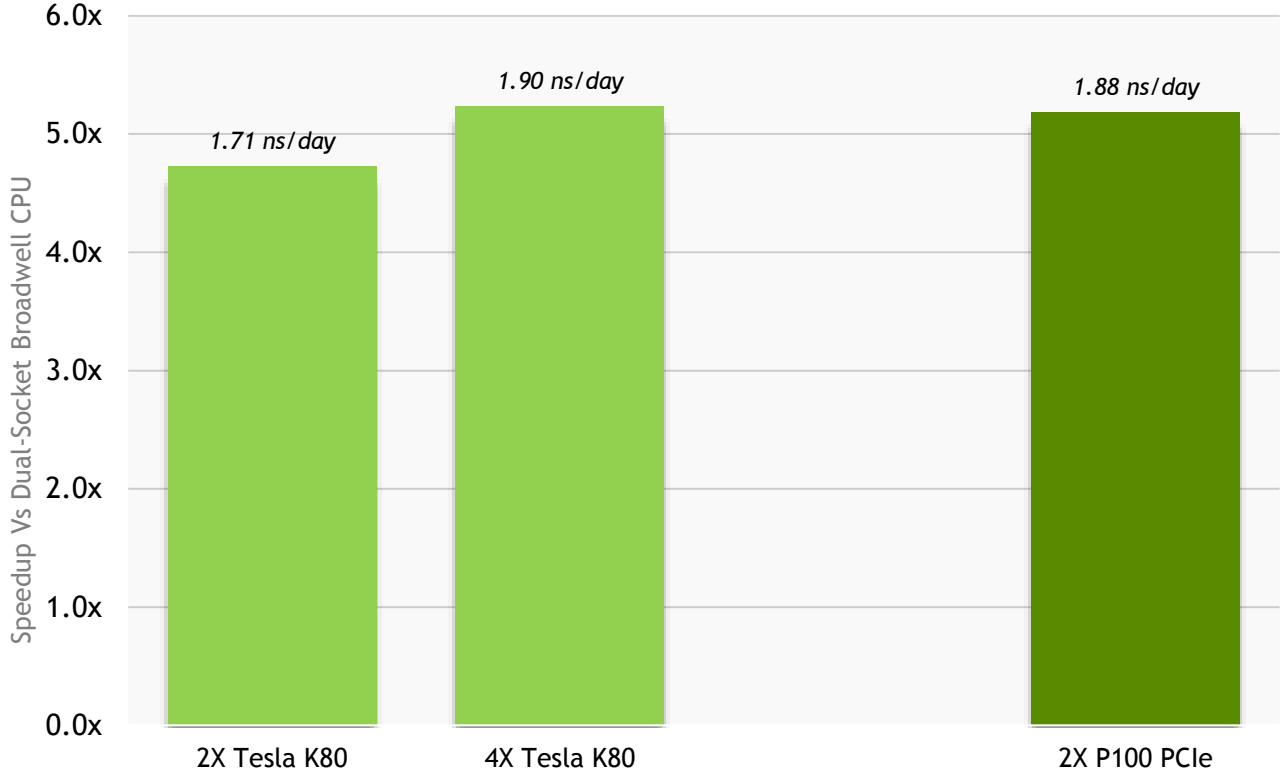
NAMD

Molecular Dynamics
Designed for high-performance simulation of large molecular systems

Accelerated Features	Metric	Scalability
Full electrostatics with PME and most simulation features	Nanoseconds per day	Up to 100M atom capable, multi-GPU, multi-node

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

NAMD 2.11
Speedup Vs Dual-Socket CPU Server

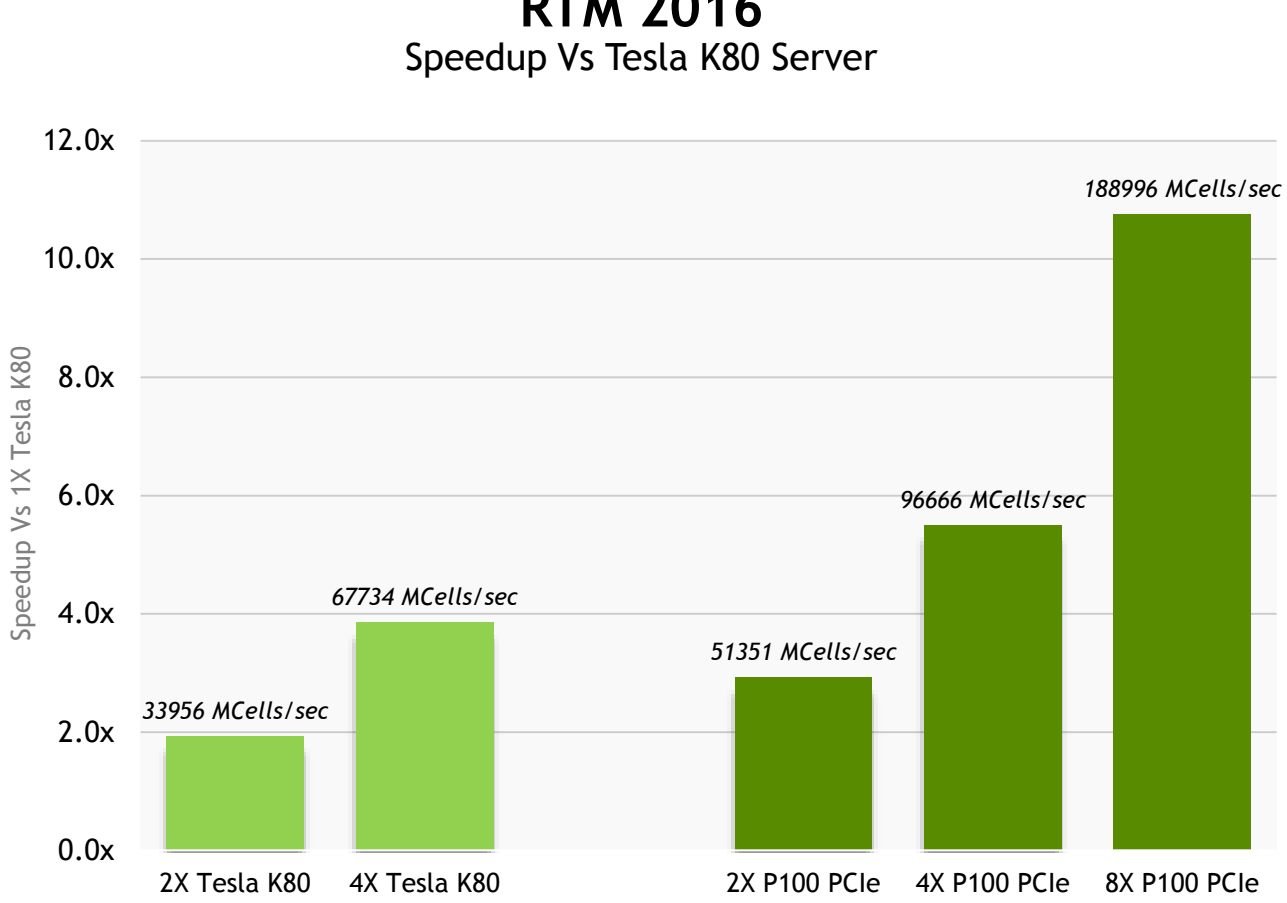


CPU Server: Dual Xeon E5-2699 v4@2.2GHz (44-cores)
GPU Servers: Dual Xeon E5-2699 v4@2.2GHz (44-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.44
Dataset: STMV

RTM

Reverse time migration (RTM) modeling is a critical component in the seismic processing workflow of oil and gas exploration

Accelerated Features	Metric	Scalability
Batch algorithm that executes in a console that has no UI, HPC application suite integration	MCells/second	Multiple GPUs, multi-nodes



CPU Server: Single Xeon E5-2699 v4@2.2GHz (22-cores)
GPU Servers: Single Xeon E5-2699 v4@2.2GHz (22-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.42
Dataset: ISO rX 2X mgpu

QUDA

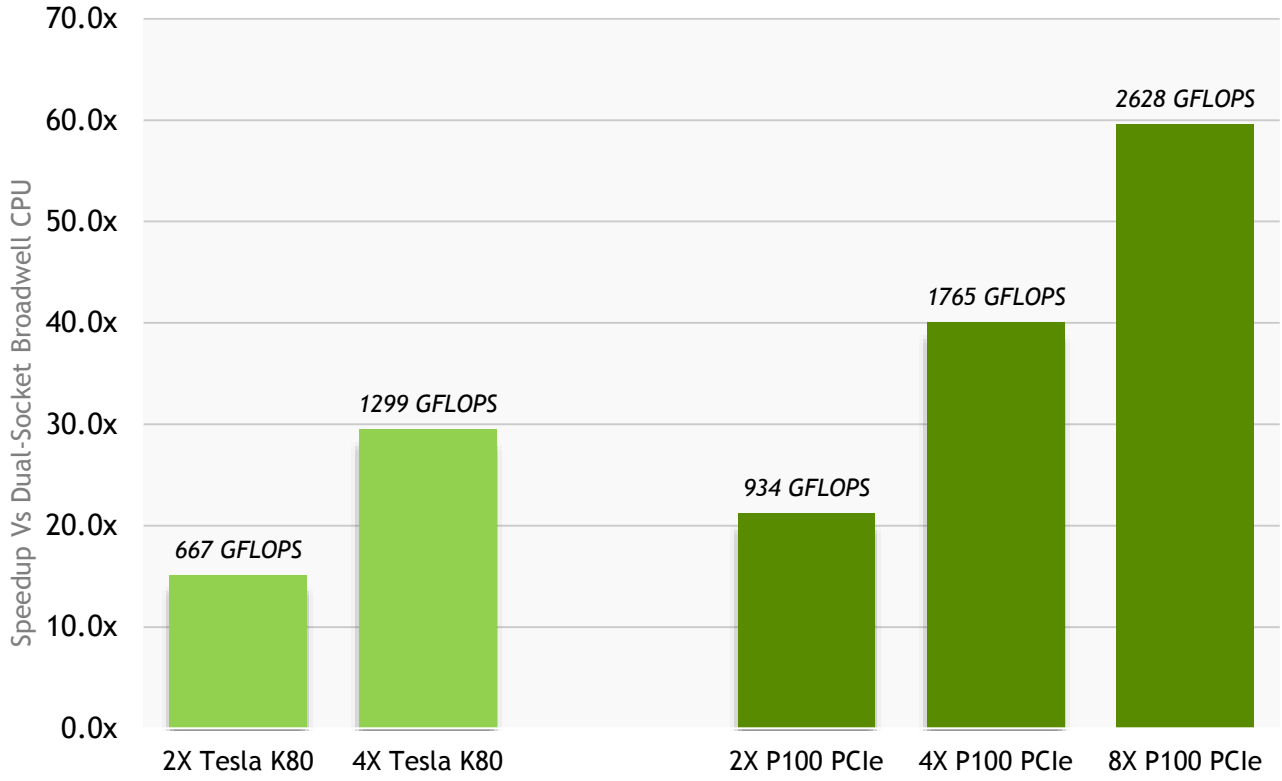
LQCD
A library for Lattice Quantum Chromo Dynamics on GPUs

Accelerated Features	Metric	Scalability
All	GFLOPS	Multi-GPU

https://lattice.github.io/quda/

QUDA 0.9

Speedup Vs Dual-Socket CPU Server



CPU Server: Dual Xeon E5-2699 v4@2.2GHz (44-cores)
GPU Servers: Dual Xeon E5-2699 v4@2.2GHz (44-cores) with Tesla K80s or P100s PCIe
CUDA Version: CUDA 8.0.44
Dataset: QUDA (GPU) - Dslash Wilson-Clover,Precision: Double; Problem Size 32x32x32x64
QPhiX (CPU) - Dslash Wilson-Clover 32x32x32x64; Precision: Double

THANK YOU

Piero Altoe: paltoe@nvidia.com