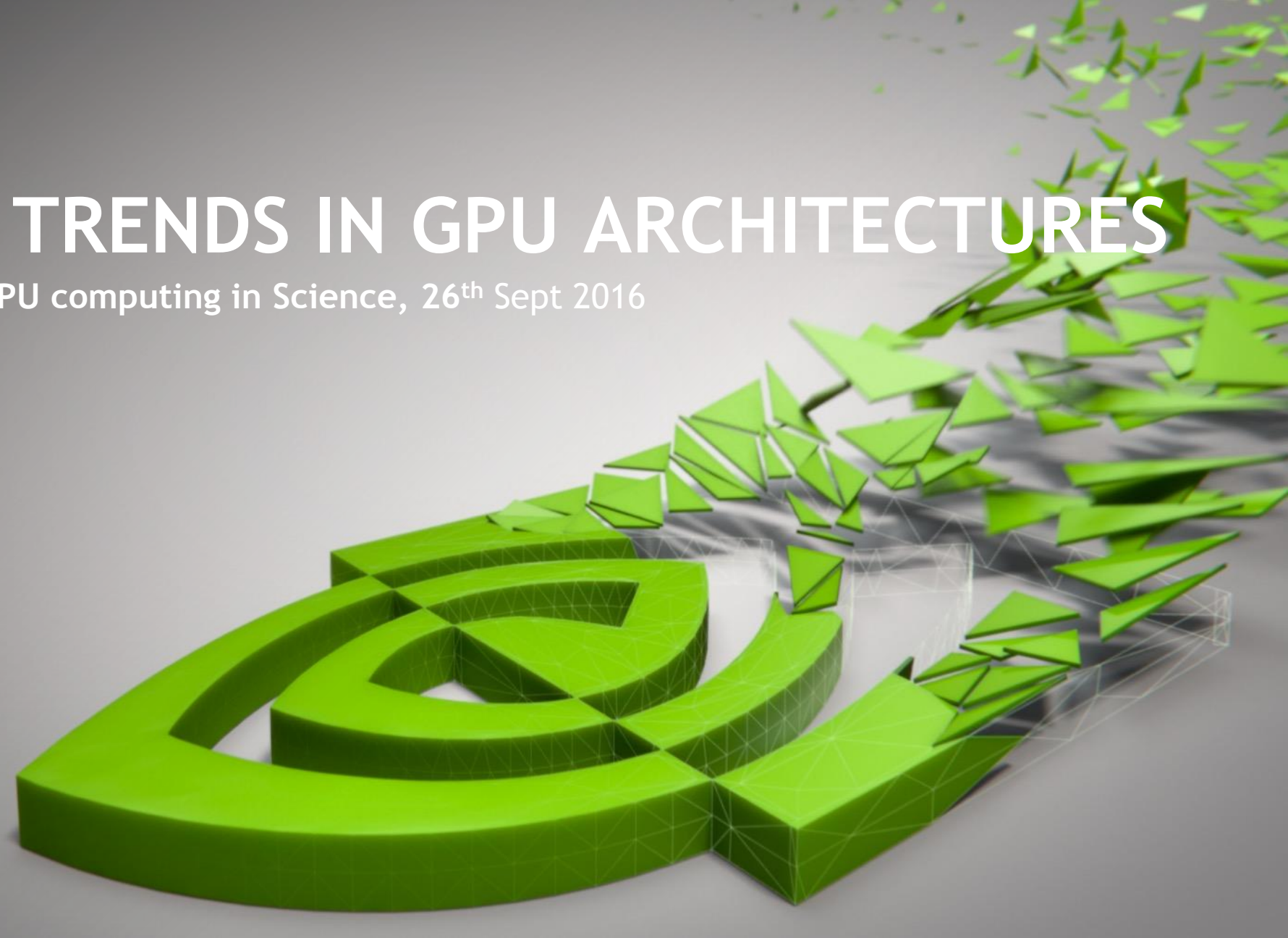
# RECENT TRENDS IN GPU ARCHITECTURES

Perspectives of GPU computing in Science, 26th Sept 2016

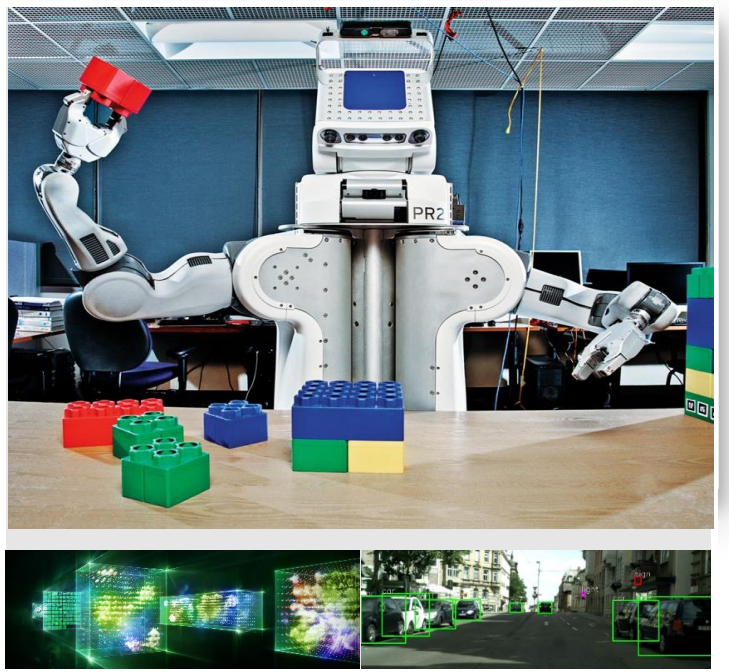
NVIDIA



# NVIDIA
# “THE AI COMPUTING COMPANY”


GPU Computing

Computer Graphics



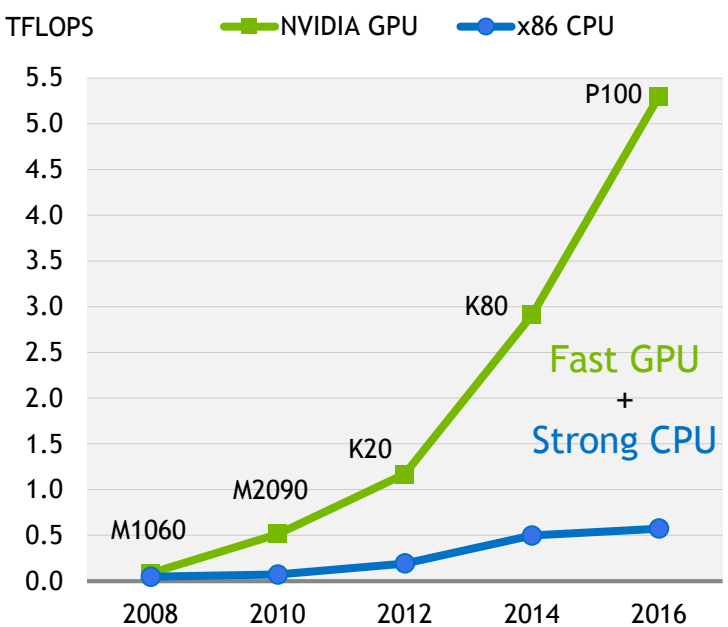

Artificial Intelligence

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

# 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

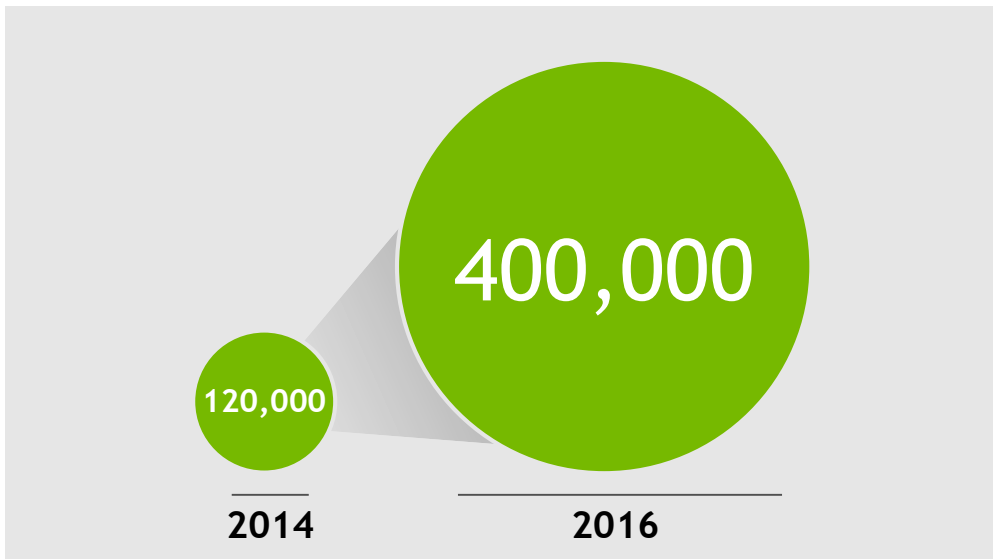
### Accessibility

阿里云计算 Alibaba Cloud Computing

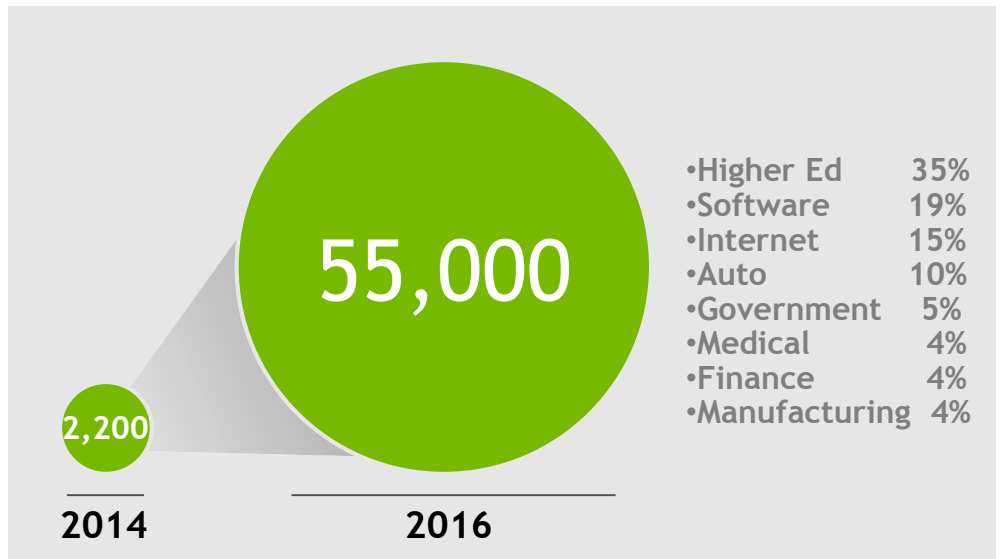
amazon web services™ IBM SOFTLAYER

Microsoft Azure


# DEDICATED TO ADVANCEMENT OF HPC AND AI



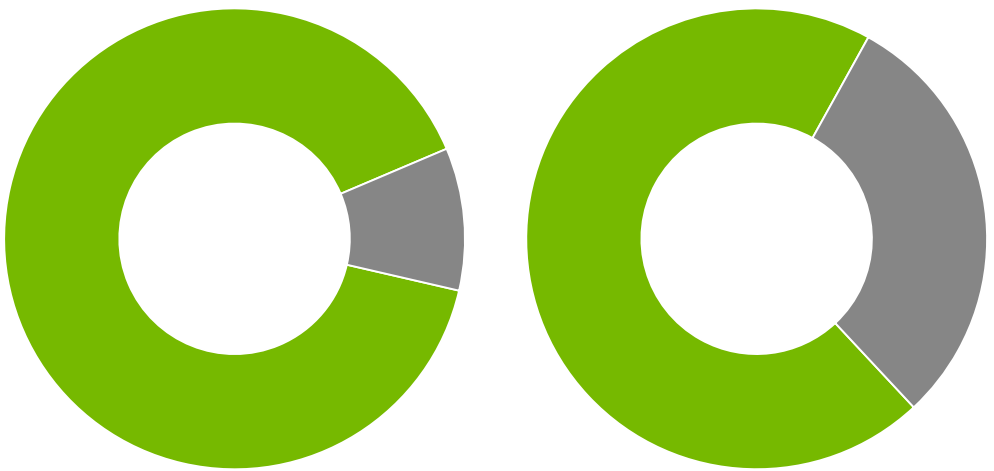

3X GPU Developers




25x Deep Learning Developers

# 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

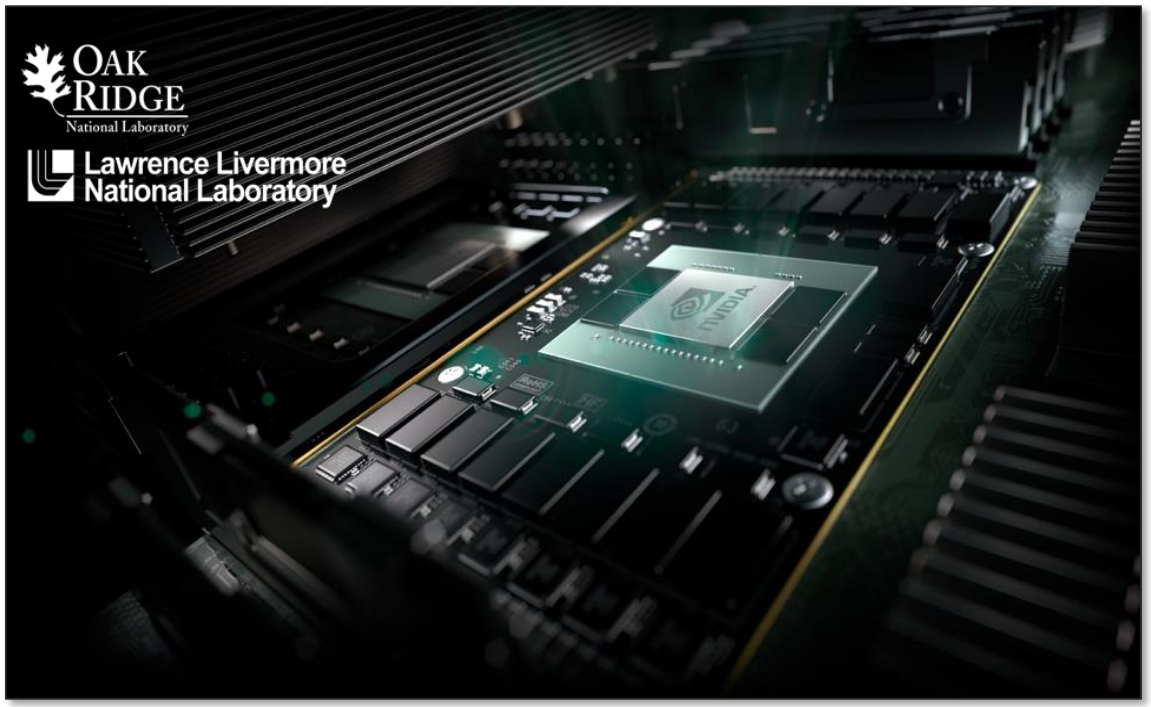
- ANSYS Fluent
- WRF
- VASP
- OpenFOAM
- CHARMM
- Quantum Espresso

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

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

# 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

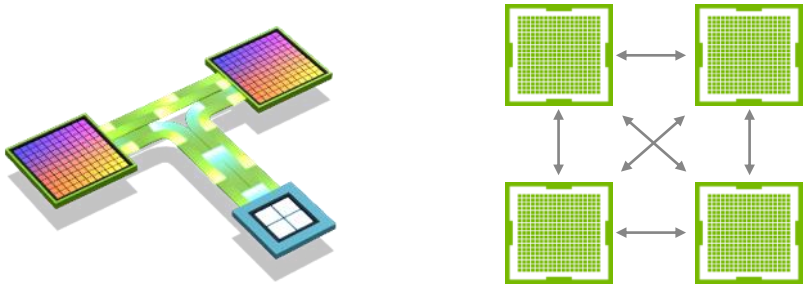
40 TFLOPS per Node, >3,400 Nodes

2017

# TESLA PLATFORM LEADS IN EVERY WAY

## PROCESSOR

## INTERCONNECT



## SOFTWARE

OpenACC
Directives for Accelerators

NVIDIA CUDA

## ECOSYSTEM

ParaView

torch
Facebook AI Research

IBM

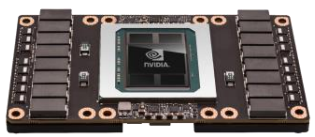
NAMD
Scalable Molecular Dynamics

DMTK

TensorFlow

# END-TO-END PRODUCT FAMILY

## HYPERSCALE HPC



Training - Tesla P100

Inference - Tesla P40 & P4

Hyperscale deployment for deep learning training & inference

## STRONG-SCALE HPC

Tesla P100 with NVLink

Data centers running HPC and DL apps scaling to multiple GPUs

## MIXED-APPS HPC

Tesla P100 with PCI-E

HPC data centers running mix of CPU and GPU workloads

## FULLY INTEGRATED DL SUPERCOMPUTER

**DGX-1**

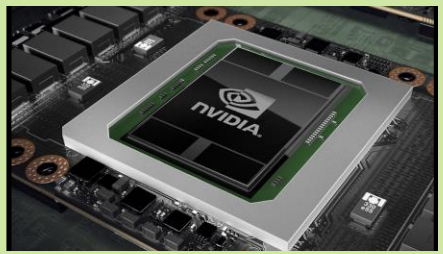
For customers who need to get going now with fully integrated solution

# TESLA PLATFORM FOR STRONG SCALING HPC

# INTRODUCING TESLA P100

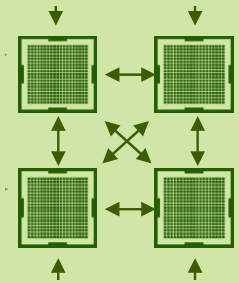
## New GPU Architecture to Enable the World's Fastest Compute Node

### Pascal Architecture



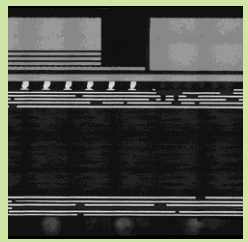

Highest Compute Performance

### NVLink



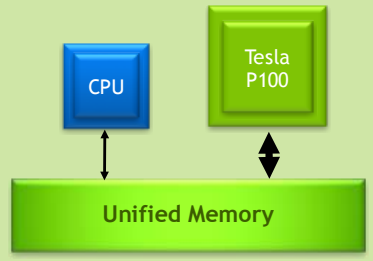
GPU Interconnect for Maximum Scalability

### CoWoS HBM2

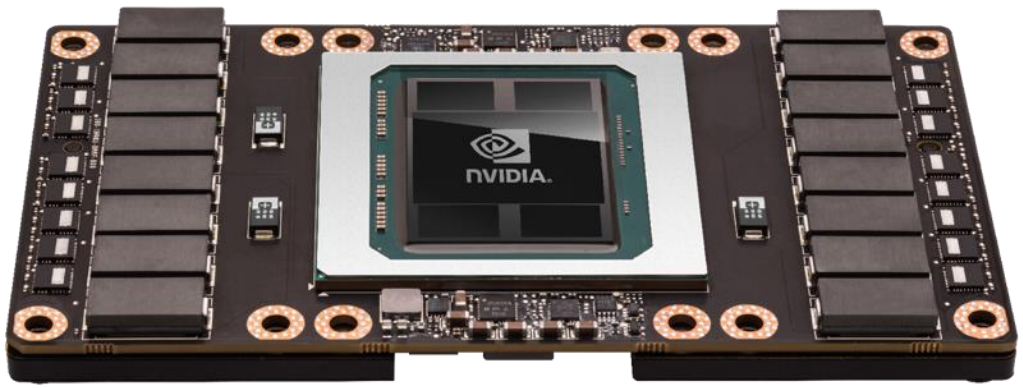


Unifying Compute & Memory in Single Package

### Page Migration Engine



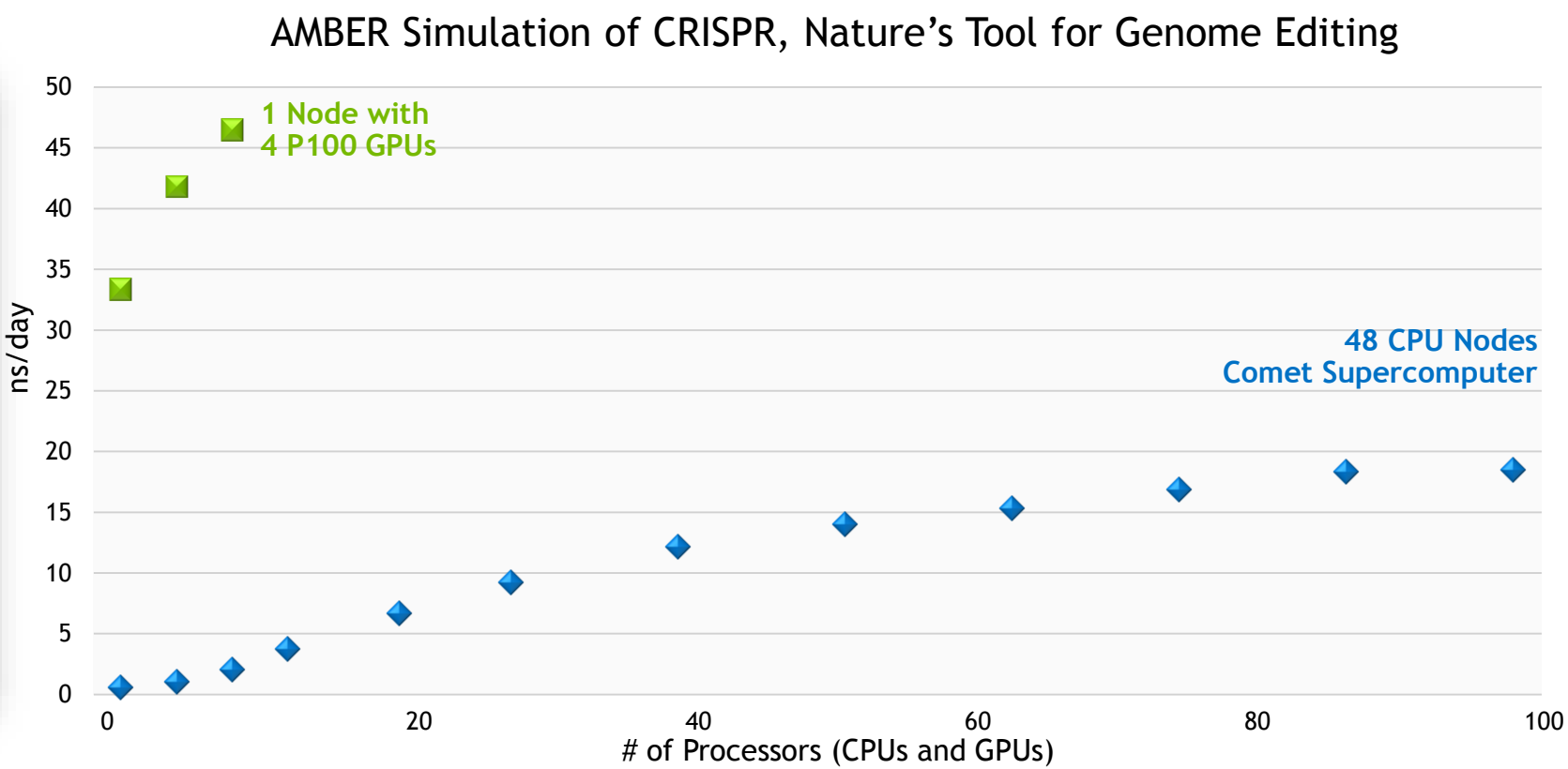

Simple Parallel Programming with Virtually Unlimited Memory



# BIG PROBLEMS NEED FAST COMPUTERS

## 2.5x Faster than the Largest CPU Data Center


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

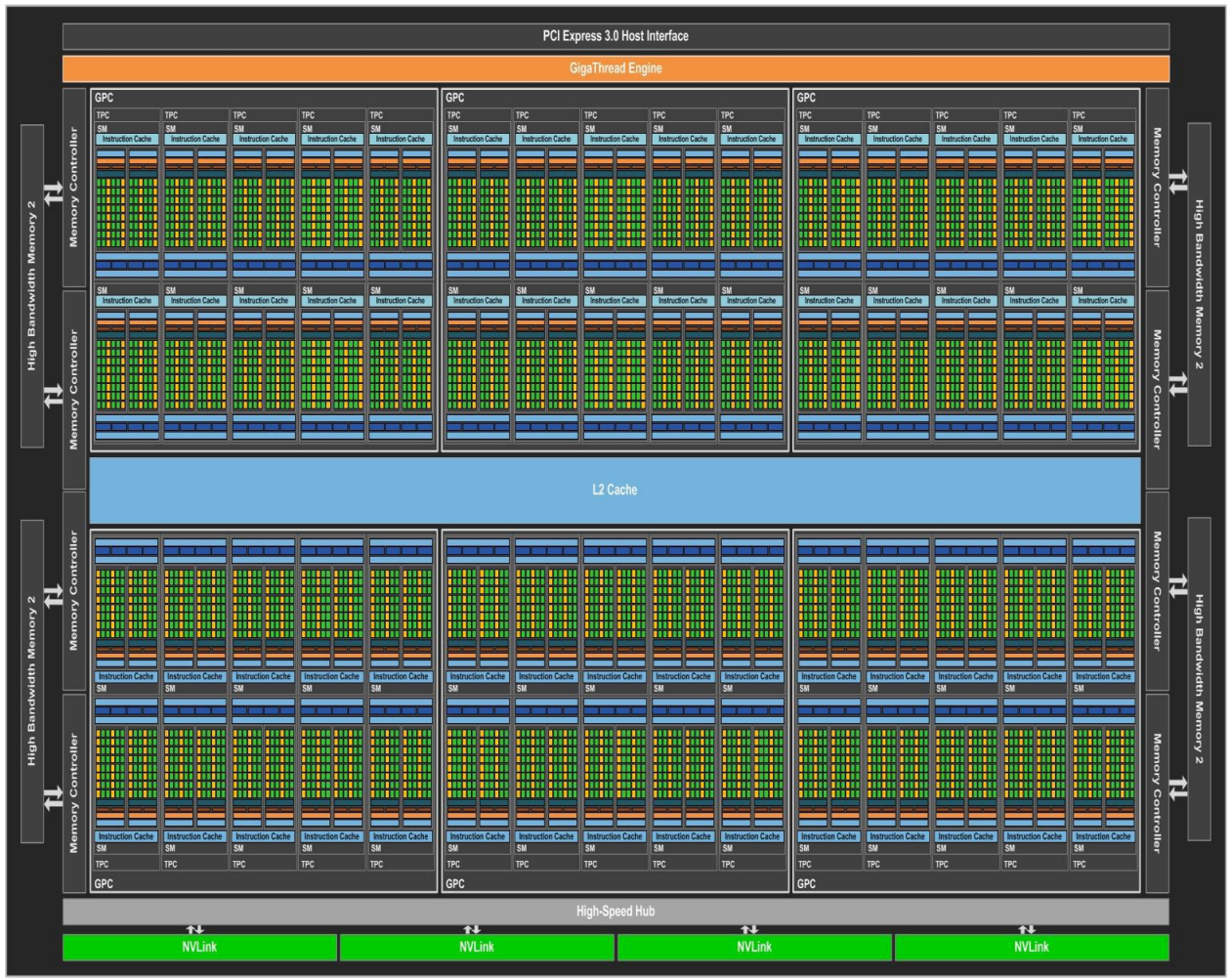



AMBER 16 Pre-release, CRSPR based on PDB ID 5f9r, 336,898 atoms
CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB

# 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<br>Unified Memory |
| **Availability** | DGX-1: Order Now<br>Cray, Dell, HP, IBM: Q1 2017 |

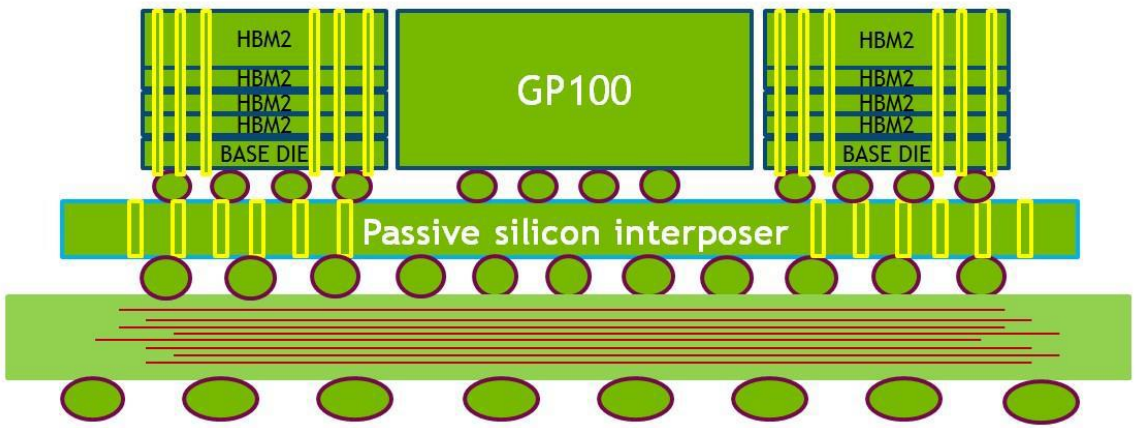
# TESLA P100 ARCHITECTURE(1)



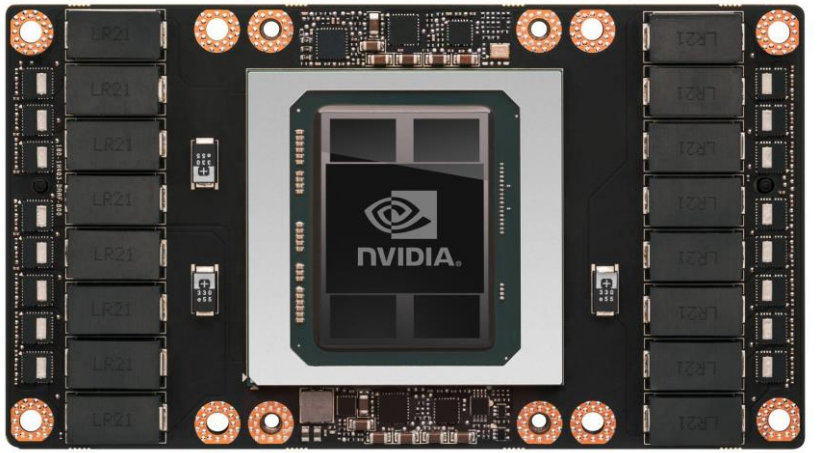

| Tesla Products | Tesla K40 | Tesla M40 | Tesla P100 |
|---|---|---|---|
| GPU | GK110 (Kepler) | GM200 (Maxwell) | GP100 (Pascal) |
| SMs | 15 | 24 | 56 |
| TPCs | 15 | 24 | 28 |
| FP32 CUDA Cores / SM | 192 | 128 | 64 |
| FP32 CUDA Cores / GPU | 2880 | 3072 | 3584 |
| FP64 CUDA Cores / SM | 64 | 4 | 32 |
| FP64 CUDA Cores / GPU | 960 | 96 | 1792 |
| Base Clock | 745 MHz | 948 MHz | 1328 MHz |
| GPU Boost Clock | 810/875 MHz | 1114 MHz | 1480 MHz |
| Peak FP32 GFLOPs[1] | 5040 | 6840 | 10600 |
| Peak FP64 GFLOPs[1] | 1680 | 210 | 5300 |
| Texture Units | 240 | 192 | 224 |
| Memory Interface | 384-bit GDDR5 | 384-bit GDDR5 | 4096-bit HBM2 |
| Memory Size | Up to 12 GB | Up to 24 GB | 16 GB |
| L2 Cache Size | 1536 KB | 3072 KB | 4096 KB |
| Register File Size / SM | 256 KB | 256 KB | 256 KB |
| Register File Size / GPU | 3840 KB | 6144 KB | 14336 KB |

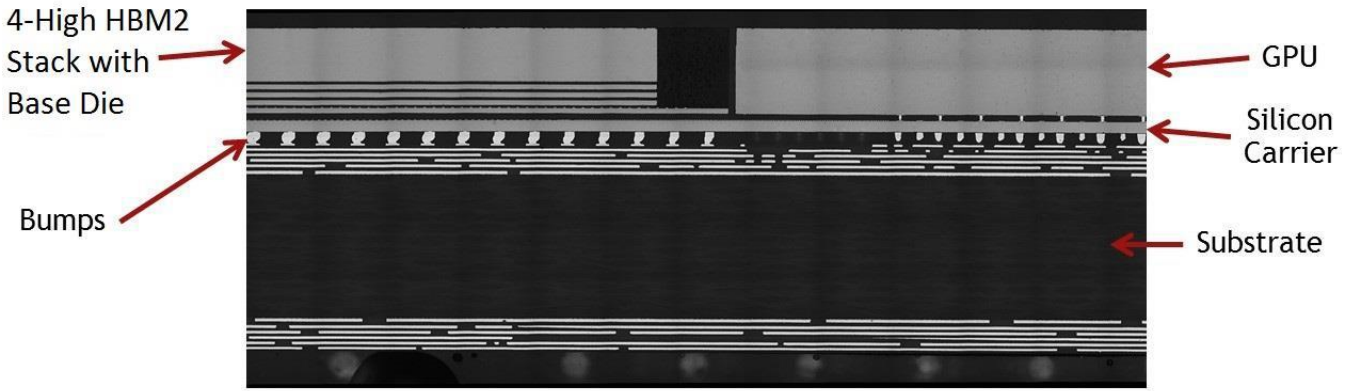
# TESLA P100 ARCHITECTURE (2)

# TESLA P100 PHYSICAL VIEW

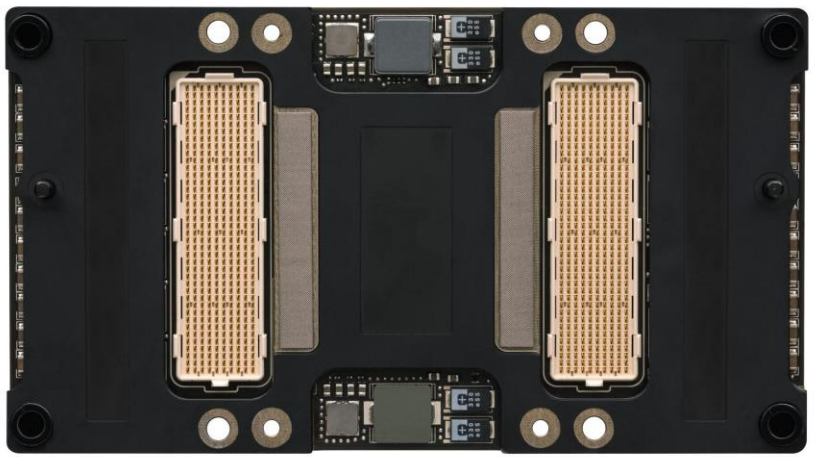



Front



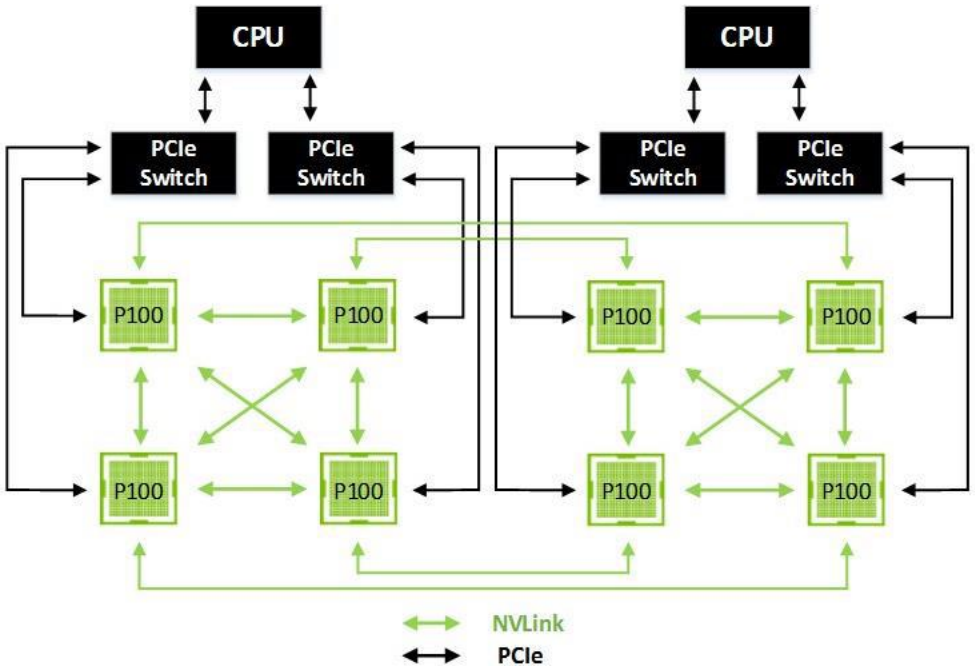



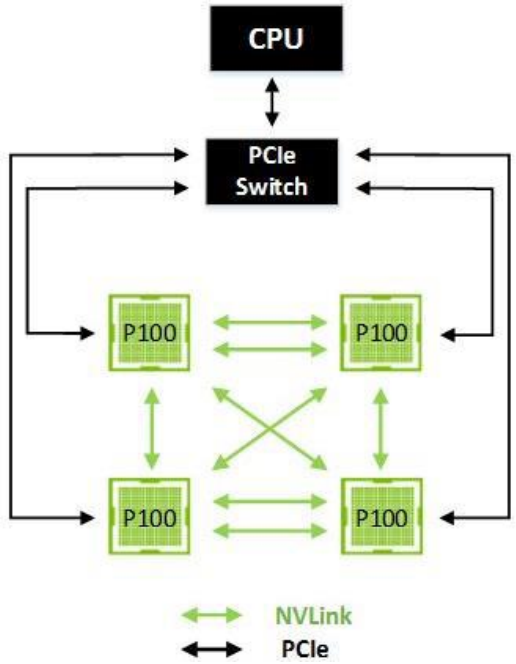

Back

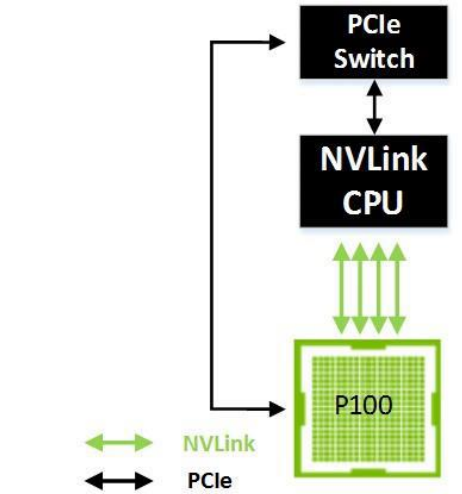


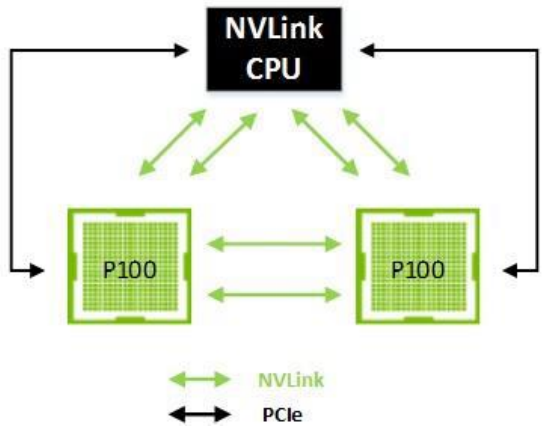
# HIGH SPEED COMMUNICATION BUS NVLINK

# TESLA P100 NVLINK CONFIGURATIONS



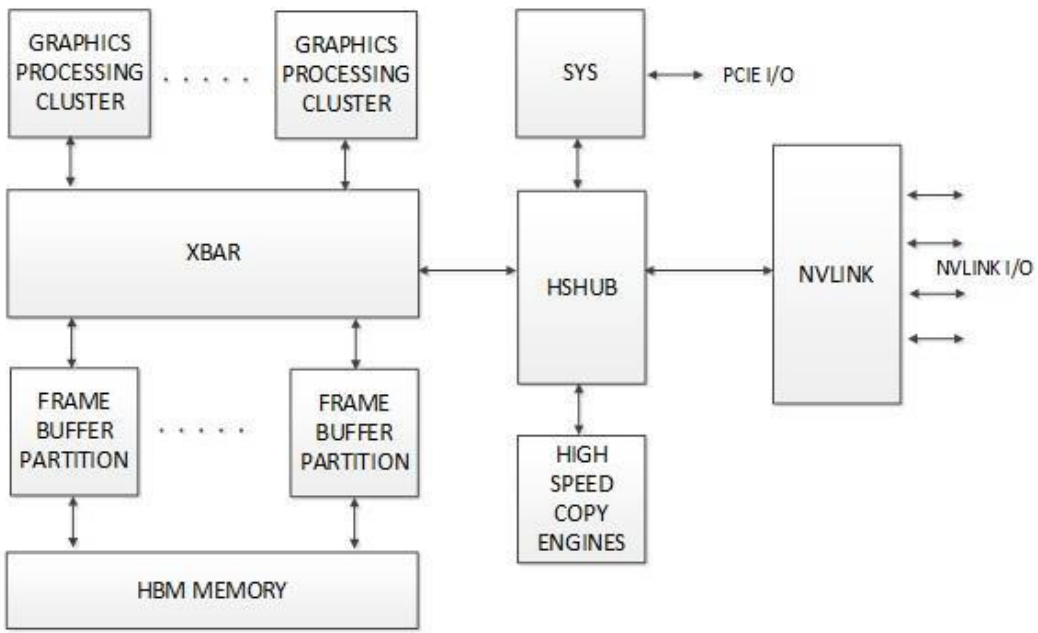



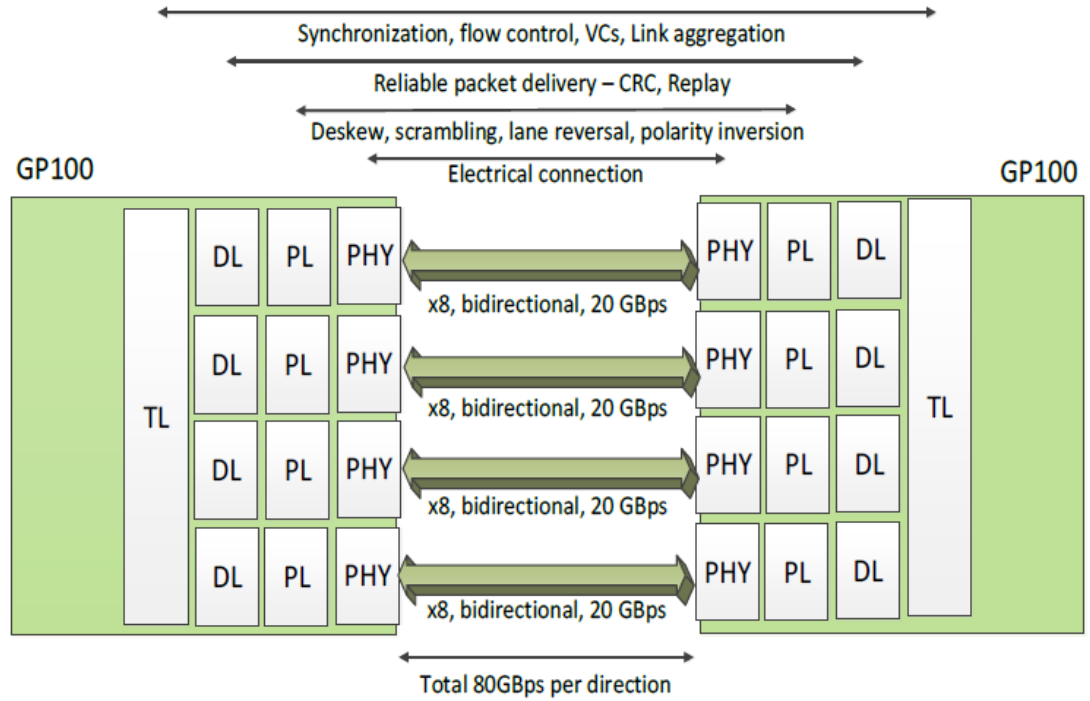




Numerous topologies are possible, and different configurations can be optimized for different applications. In this section, we discuss the following NVLink configurations:

- GPU-to-GPU NVLink Connectivity
- CPU-to-GPU NVLink Connectivity

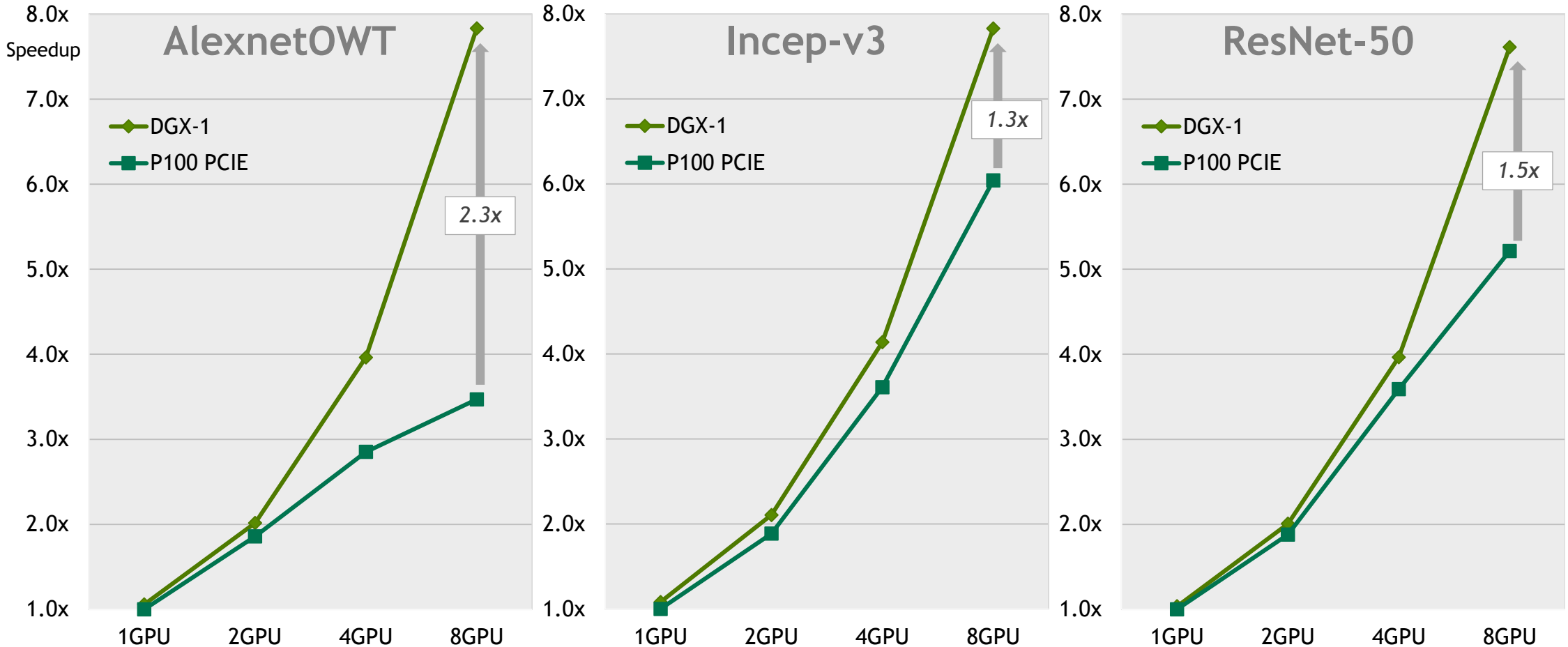
# TESLA P100 NVHS




NVLink Layers and Links; Physical Layer (PHY), Data Link Layer (DL),
Transaction Layer (TL)

# NVLINK ENABLES LINEAR MULTI-GPU SCALING




Deepmark test with NVCaffe. AlexnetOWT use batch 128, Incep-v3/ResNet-50 use batch 32, weak scaling,
P100 and DGX-1 are measured, FP32 training, software optimization in progress, CUDA8/cuDNN5.1, Ubuntu 14.04

# TESLA P100 TECH INFO

P100 WHITE PAPER:
http://www.nvidia.com/object/pascal-architecture-whitepaper.html

# NVIDIA DGX-1

# NVIDIA DGX-1

## WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER


170 TFLOPS FP16

8x Tesla P100 16GB

NVLink Hybrid Cube Mesh

Accelerates Major AI Frameworks

Dual Xeon

7 TB SSD Deep Learning Cache

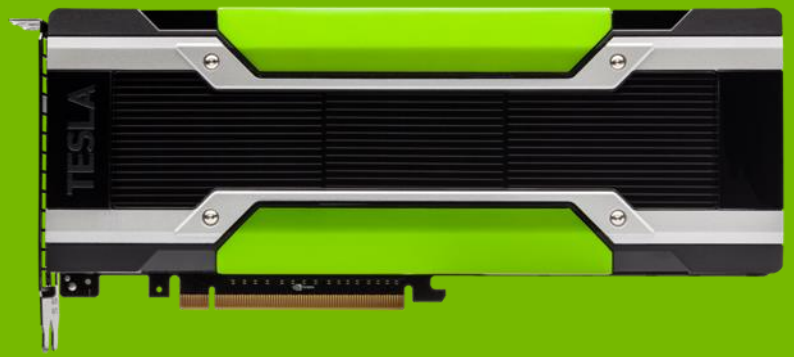
Dual 10GbE, Quad IB 100Gb

3RU – 3200W

# TESLA PLATFORM FOR MIXED APPS HPC

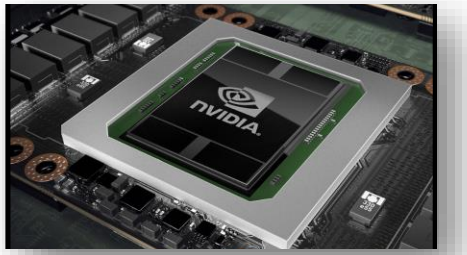
# TESLA P100

## MOST ADVANCED DATA CENTER GPU FOR MIXED-APP HPC



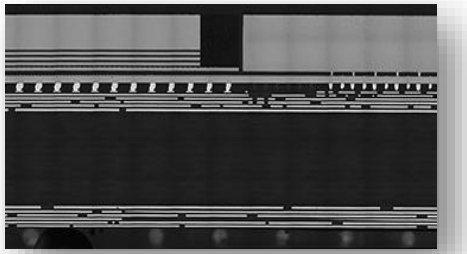

Tesla P100 for PCIe-based Servers

### PASCAL



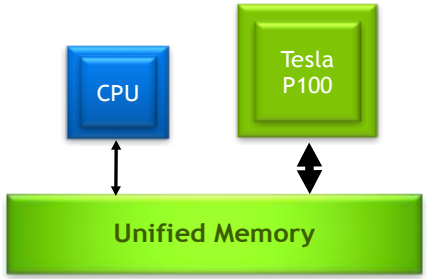

18.7 TF HP · 9.3 TF SP · 4.7 TF DP
New Deep Learning Instructions
More Registers & Cache per SM

### CoWoS with HBM2



Up to 720 GB/Sec Bandwidth
Up to 16 GB Memory Capacity
ECC with Full Performance & Capacity

### PAGE MIGRATION ENGINE



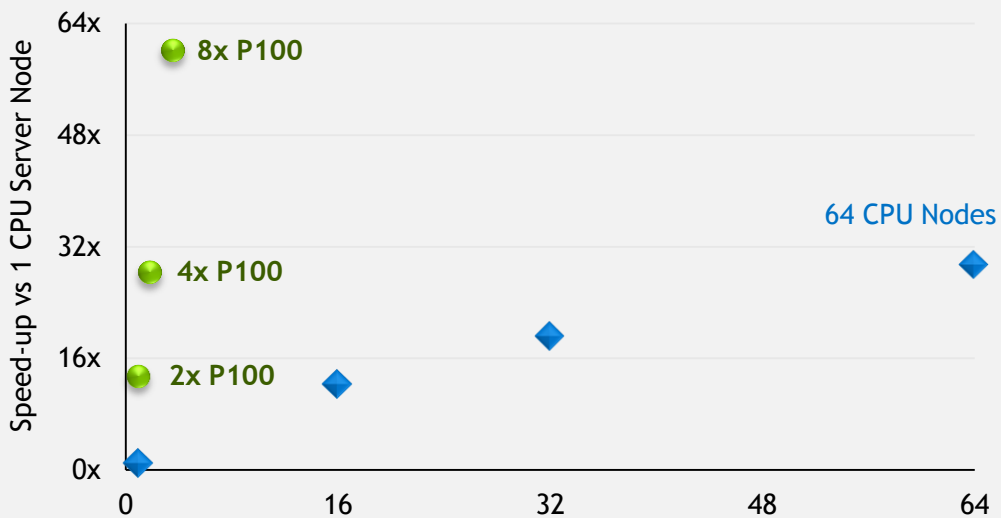

Simpler Parallel Programming
Virtually Unlimited Data Size
Performance w/ data locality

# EXTRAORDINARY STRONG SCALING

## One Strong Node Faster Than Lots of Weak Nodes

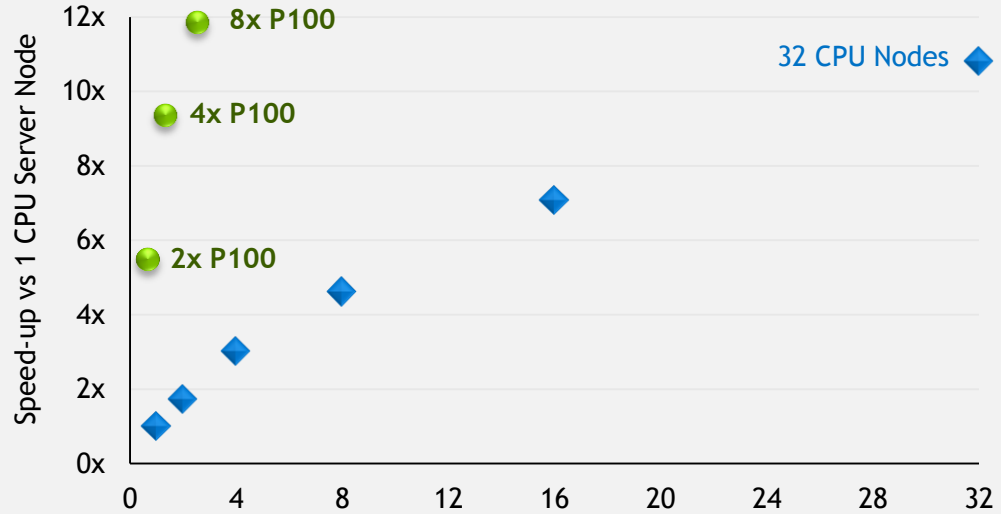
### CAFFE ALEXNET PERFORMANCE

Single P100 NVLink-enabled Node vs Lots of Weak Nodes




### VASP PERFORMANCE

Single P100 PCIe Node vs Lots of Weak Nodes

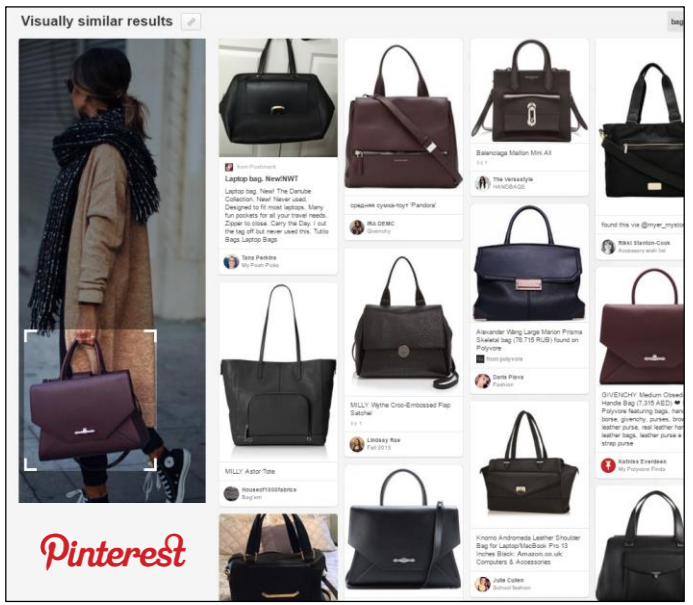



CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB
VASP 5.4.1_05Feb16, Si-Huge Dataset. 16, 32 Nodes are estimated based on same scaling from 4 to 8 nodes
Caffe AlexNet scaling data: https://software.intel.com/en-us/articles/caffe-training-on-multi-node-distributed-memory-systems-based-on-intel-xeon-processor-e5

# TESLA PLATFORM FOR HYPERSCALE HPC

# AI IS EVERYWHERE

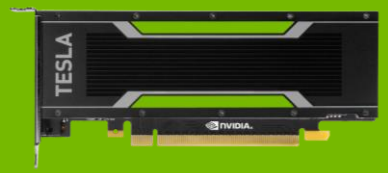

“Find where I parked my car”



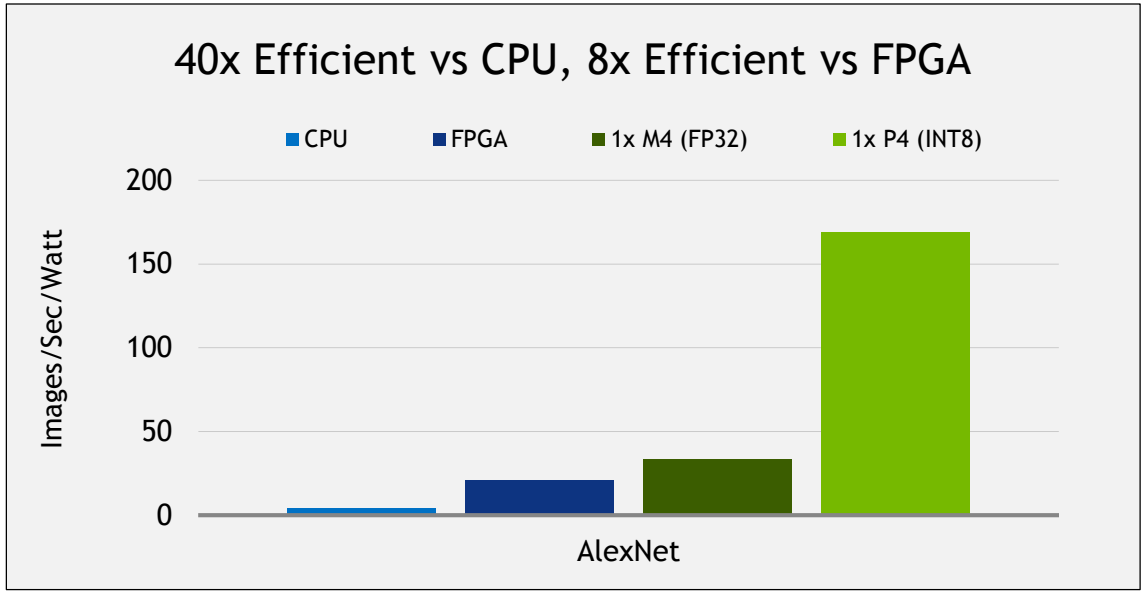

“Find the bag I just saw
in this magazine”


“What movie should
I watch next?”

# TESLA P4



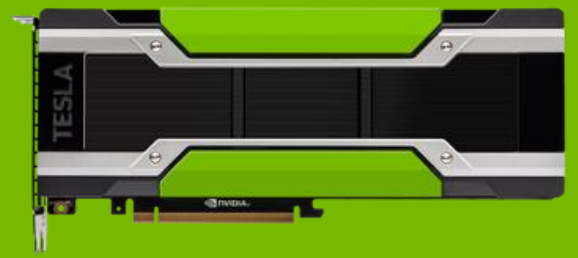

Maximum Efficiency for Scale-out Servers

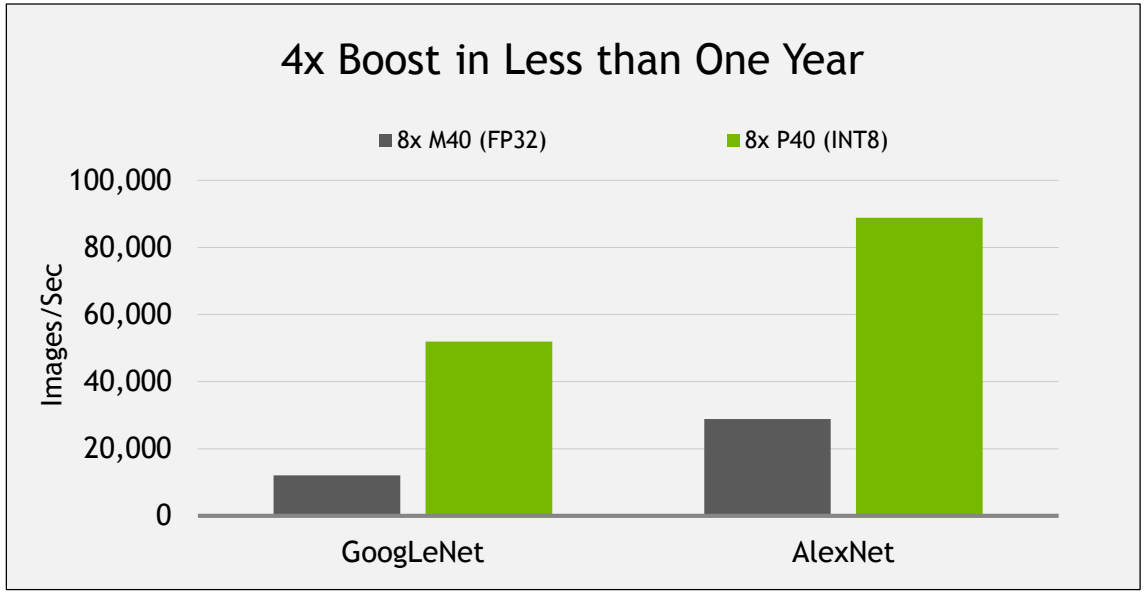

| P4 | |
|---|---|
| **# of CUDA Cores** | 2560 |
| **Peak Single Precision** | 5.5 TeraFLOPS |
| **Peak INT8** | 22 TOPS |
| **Low Precision** | 4x 8-bit vector dot product with 32-bit accumulate |
| **Video Engines** | 1x decode engine, 2x encode engine |
| **GDDR5 Memory** | 8 GB @ 192 GB/s |
| **Power** | 50W & 75 W |

*AlexNet, batch size = 128, CPU: Intel E5-2690v4 using Intel MKL 2017, FPGA is Arria10-115*
*1x M4/P4 in node, P4 board power at 56W, P4 GPU power at 36W, M4 board power at 57W, M4 GPU power at 39W, Perf/W chart using GPU power*

# TESLA P40




Highest Throughput for Scale-up Servers


| P40 | |
|---|---|
| **# of CUDA Cores** | 3840 |
| **Peak Single Precision** | 12 TeraFLOPS |
| **Peak INT8** | 47 TOPS |
| **Low Precision** | 4x 8-bit vector dot product with 32-bit accumulate |
| **Video Engines** | 1x decode engine, 2x encode engines |
| **GDDR5 Memory** | 24 GB @ 346 GB/s |
| **Power** | 250W |

*GoogLeNet, AlexNet, batch size = 128, CPU: Dual Socket Intel E5-2697v4*

# TESLA PASCAL FAMILY

# 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 |

# TESLA PLATFORM FOR DEVELOPERS

# NVIDIA SDK

**The Essential Resource for GPU Developers**

## NVIDIA SDK

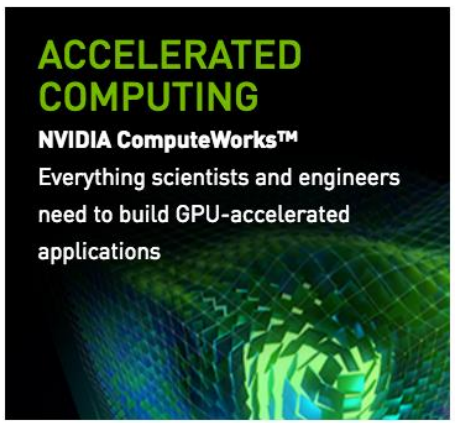
### DEEP LEARNING
**Deep Learning SDK**
High-performance tools and libraries for deep learning

### 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

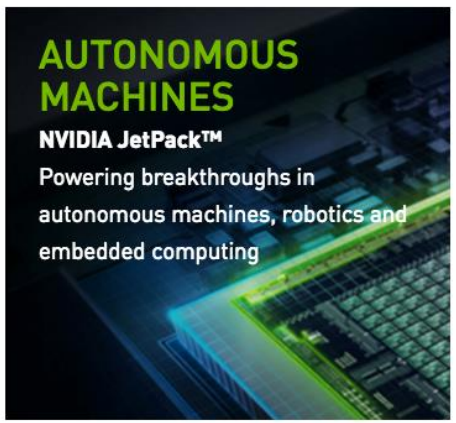
### GAME DEVELOPMENT
**NVIDIA GameWorks™**
Advanced simulation and rendering technology for game development

### 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



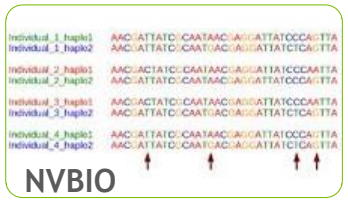
### ADDITIONAL RESOURCES
More resources for GPU Developers

# 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 Software

Thrust
&
ArrayFire


developer.nvidia.com/gpu-accelerated-libraries

# NVIDIA GPU EDUCATORS PROGRAM

### Equipping Educators with Teaching Materials and GPU Computing Tools

The NVIDIA GPU Educators Program provides teaching materials and real GPU resources for use in university classrooms and labs to empower today's students with the accelerated computing skills they'll need tomorrow.

Join now


### GPU Teaching Kits

Access academic teaching resources, collaborative opportunities, and expert support.

Learn more

### Existing Course Material

Learn more about real university classes, labs, and MOOCs currently using GPUs.

Learn more

### Academic Textbooks

Suggested academic and reference textbooks for university courses written by industry experts.

Learn more

### GPU Access and Development Tools

Recommended development tools, CUDA downloads, and other resources.

Learn more

### Training Material and Code Samples

Tutorials, seminars, training slides, and code samples that help teach an array of parallel programming concepts.

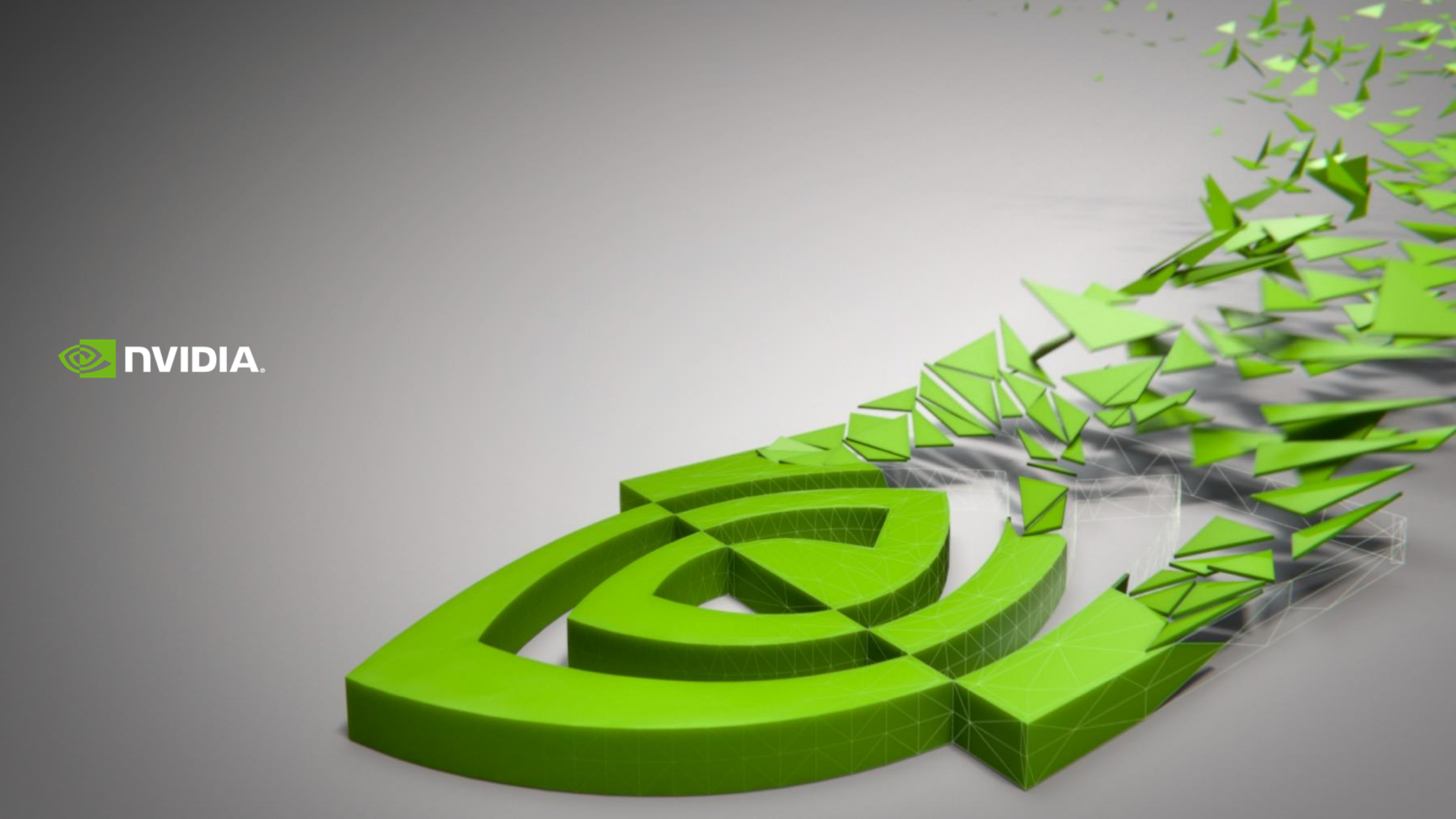
Learn more

### Community Forum and Support

Please visit our forum and email us with any questions, feedback, or suggestions.

Learn more

https://developer.nvidia.com/educators


NVIDIA