Low-Latency Accelerated Computing on GPUs

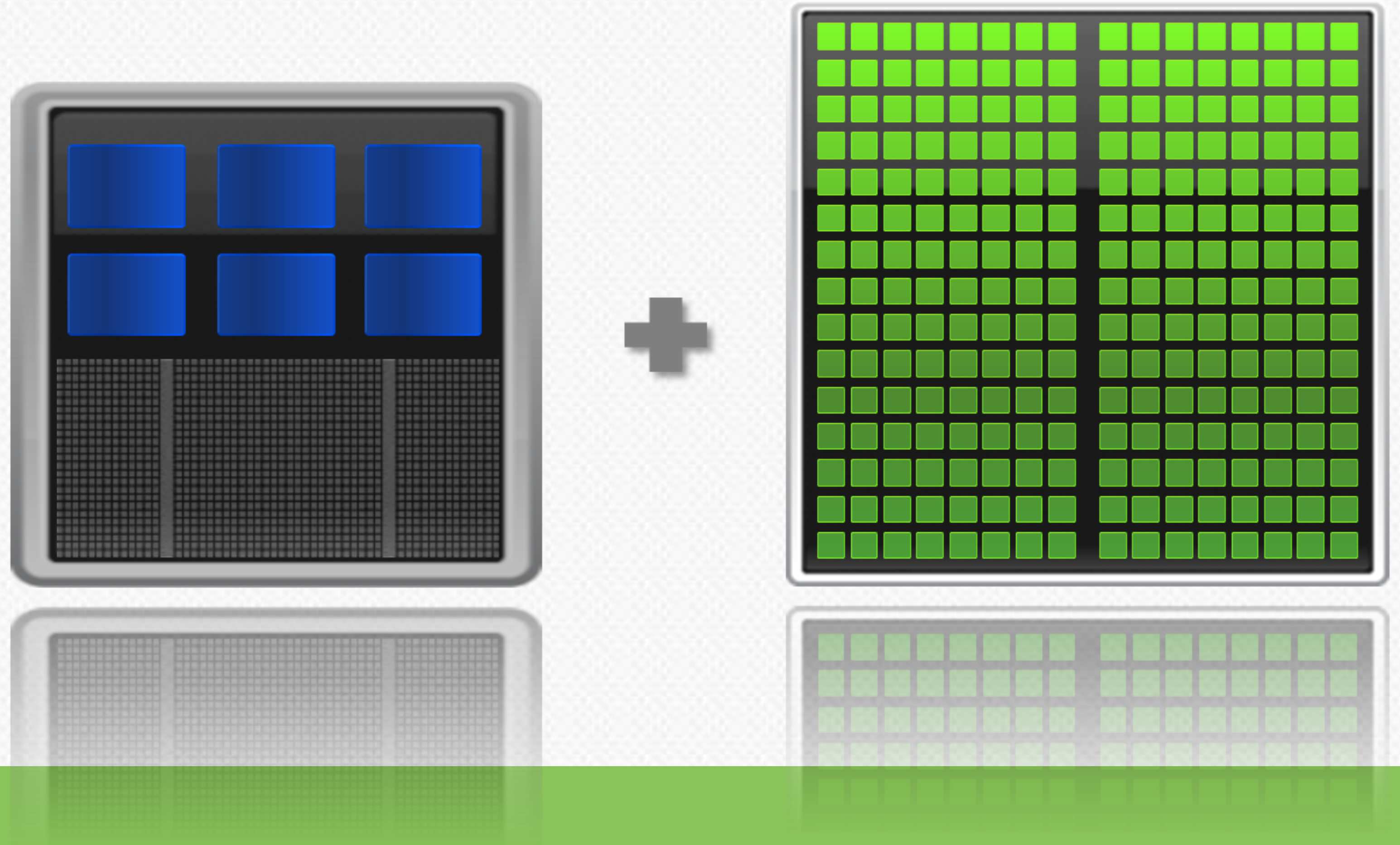
Dr. Christoph Angerer
DevTech, NVIDIA

Accelerated Computing

High Performance & High Energy Efficiency for Throughput Tasks

CPU
Serial Tasks

GPU Accelerator
Parallel Tasks



Accelerating Insights

“ 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

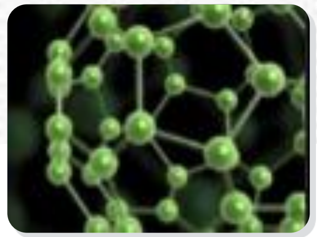
4 kWatts
$33,000

Deep learning with COTS HPC systems, A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro ICML 2013

From HPC to Enterprise Data Center

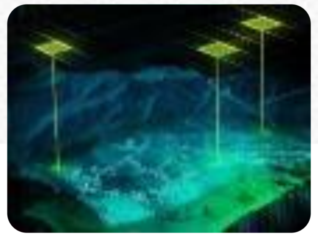
Oil & Gas

Schlumberger
BR PETROBRAS
Eni
Chevron
Statoil



Higher Ed

HARVARD School of Engineering and Applied Sciences
STANFORD UNIVERSITY
Georgia Tech
ETH
UNIVERSITY OF CAMBRIDGE



Government

Air Force Research Laboratory
Raytheon
NASA
Naval Research Laboratory

Supercomputing

CSCS
NCSA
Tokyo Institute of Technology
OAK RIDGE National Laboratory
Lawrence Livermore National Laboratory

Finance

J.P.Morgan
BARCLAYS
STANDARD LIFE
BNP PARIBAS
MUREX

Consumer Web

Baidu 百度
salesforce
SHAZAM
amazon.com
Yandex

Tesla: Platform for Accelerated Datacenters

Partner Ecosystem

Data Center Infrastructure

- Optimized Systems
- Communication Solutions
- Infrastructure Management

Development

- Programming Languages
- Development Tools
- Software Applications

NVIDIA

GPU Accelerators | **Interconnect** | **System Management** | **Compiler Solutions** | **Profile and Debug** | **Libraries**

Enterprise Services

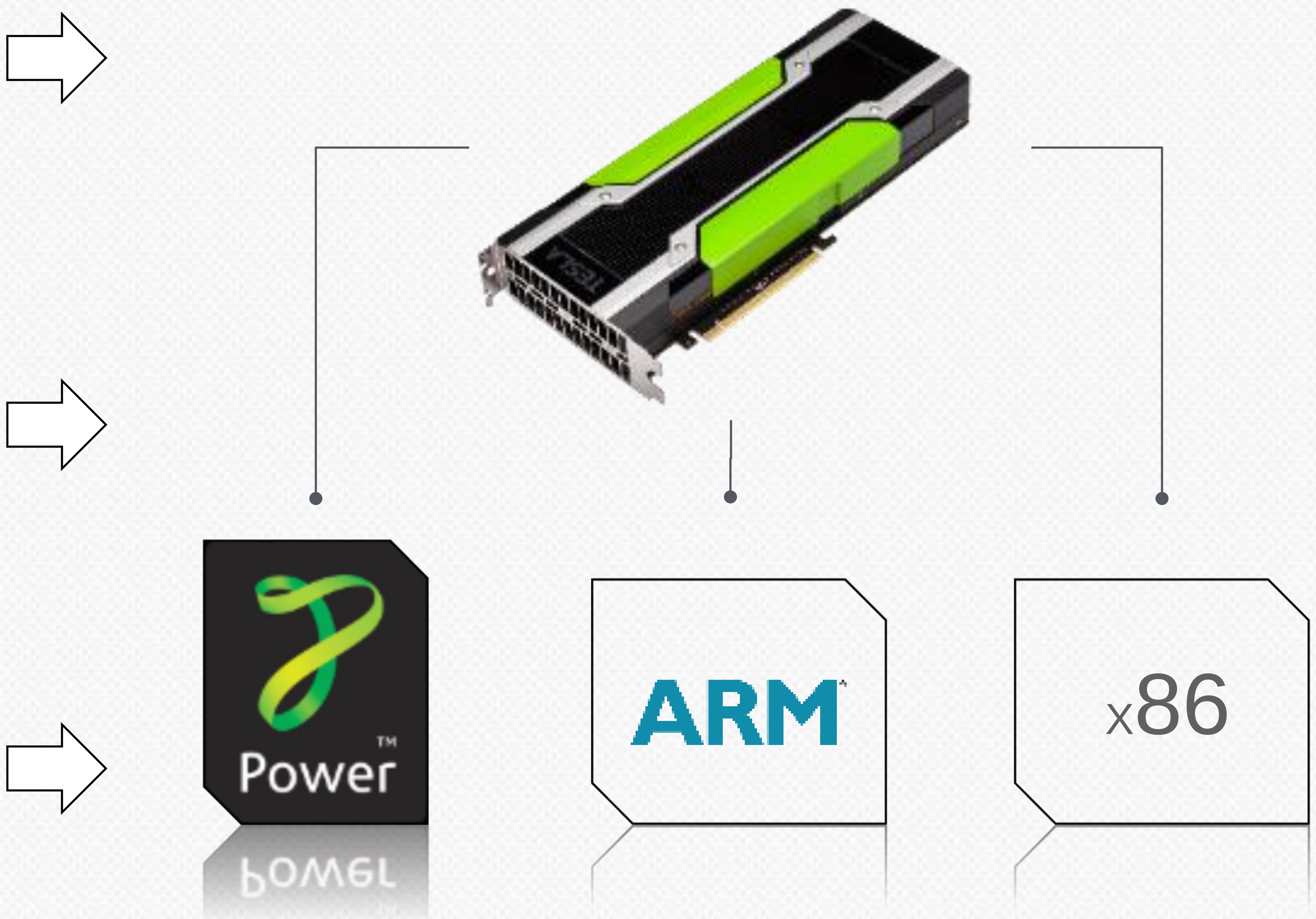
Tesla Accelerated Computing Platform

Common Programming Models Across Multiple CPUs

Libraries
AmgX, cuDNN, cuBLAS, OpenCV, Thrust

Compiler Directives
OpenACC

Programming Languages
C/C++, Fortran, python, Java



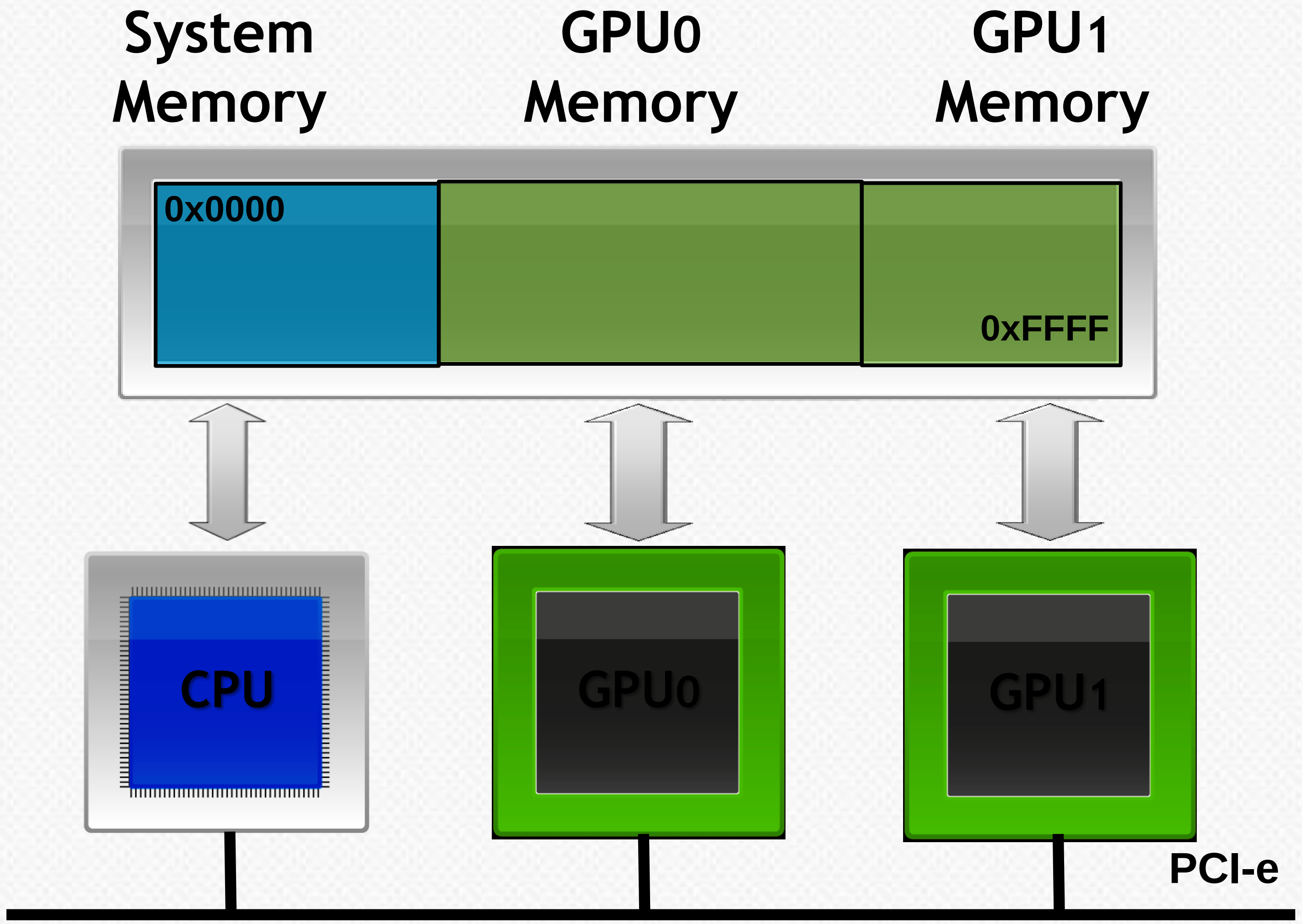
GPU Roadmap


Normalized Performance
20
16
12
8
4
0
Pascal
Unified Memory
3D Memory
NVLink
Maxwell
DX12
Kepler
Dynamic
Parallelism
Fermi
FP64
Tesla
CUDA
2008
2010
2012
2014
2016

GPUDirect

Multi-GPU: Unified Virtual Addressing

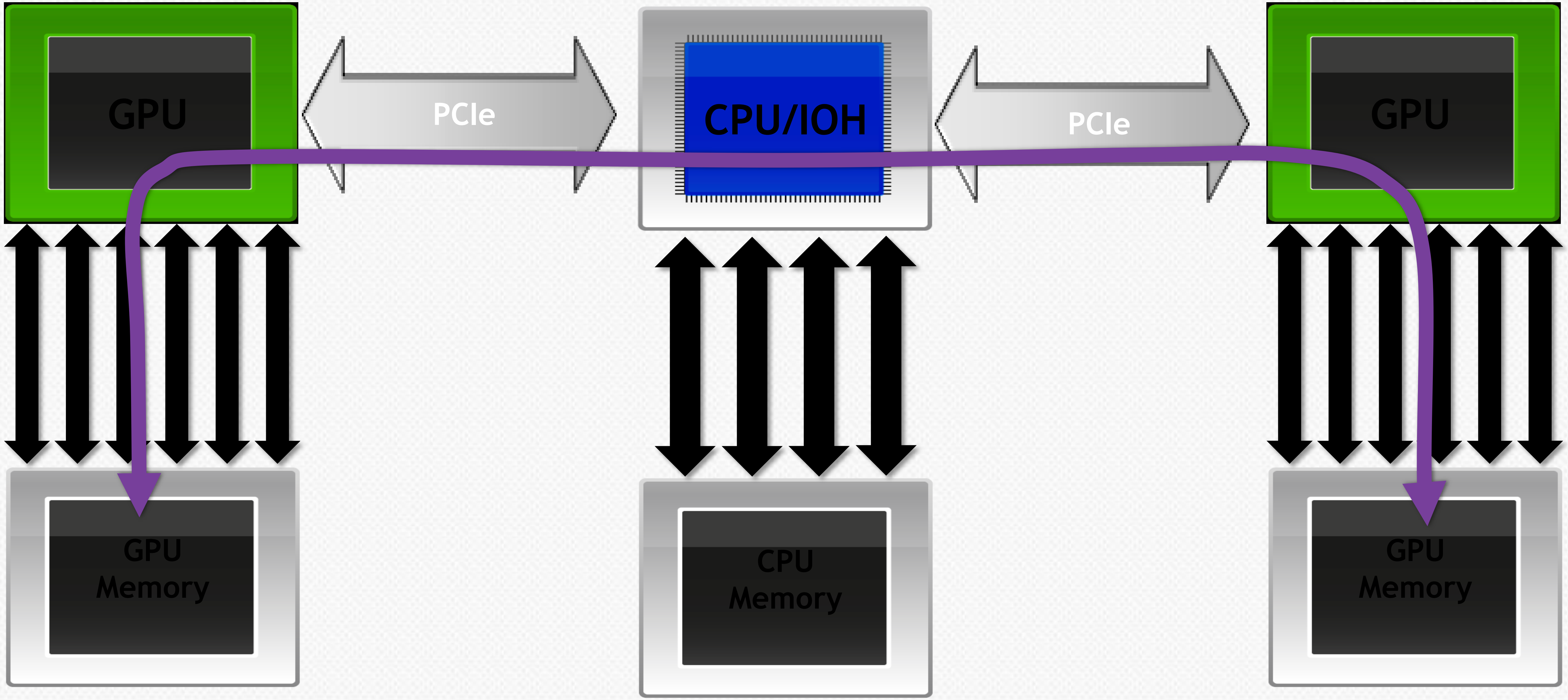
Single Partitioned Address Space



GPUDirect Technologies

- GPUDirect Shared GPU-Sysmem for Inter-node copy optimization
  - How: Use GPUDirect-aware 3rd party network drivers
- GPUDirect P2P Transfers for on-node GPU-GPU memcpy
  - How: Use CUDA APIs directly in application
  - How: use P2P-aware MPI implementation
- GPUDirect P2P Access for on-node inter-GPU LD/ST access
  - How: Access remote data by address directly in GPU device code
- GPUDirect RDMA for Inter-node copy optimization
  - What: 3rd party PCIe devices can read and write GPU memory
  - How: Use GPUDirect RDMA-aware 3rd party network drivers and MPI implementations or custom device drivers for other hardware

GPUDirect P2P: GPU-GPU Direct Access



P2P Goal: Improve *intra-node* programming model

- **Improve CUDA programming model**

- **How?**
- **Transfer data between two GPUs quickly/easily**

int main() {
double *cpuTmp, *gpu0Data, gpu1Data;

setup (gpu0Data, gpu1Data);

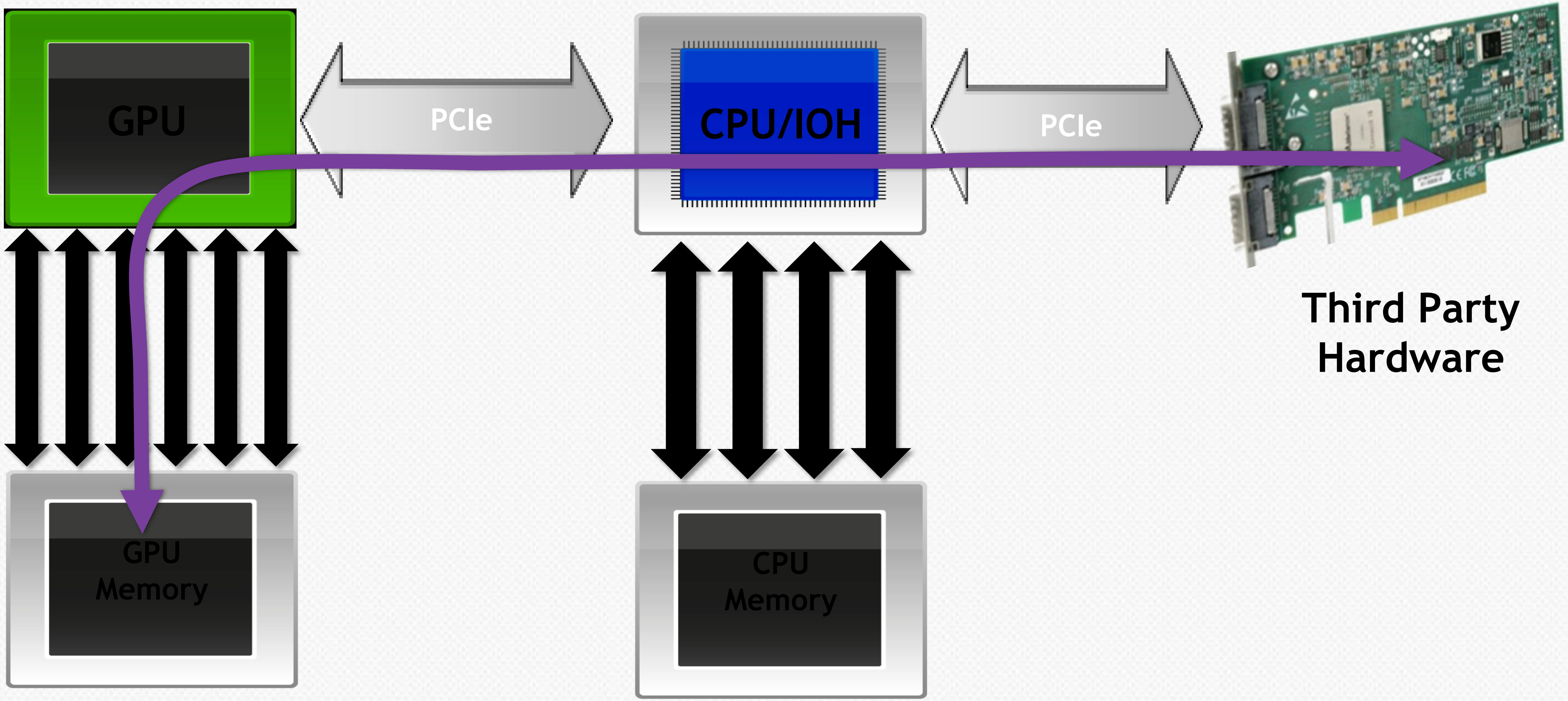
cudaSetDevice (0);
kernel <<< … >>> (gpu0Data);
~~**cudaMemcpy (cpuTmp, gpu0Data);**~~
~~**cudaMemcpy (gpu1Data, cpuTmp);**~~
cudaSetDevice (1);
kernel <<< … >>> (gpu0Data);
}

GPUDirect P2P: Common use cases

- MPI implementation optimization for Intra-Node communication

- HPC applications that fit on a single node
- You get much better efficiency with GPUDirect P2P (compared to MPI)
- Can use between ranks with cudaIpc() APIs

GPUDirect RDMA



GPUDirect RDMA Goal

- **Inter-node Latency and Bandwidth**

- **How?**
- **Transfer data between GPU and third party device (e.g. NIC) with possibly zero host-side copies**

```
int main() {
    double *cpuData, *cpuTmp,
*gpuData;

    setup (gpuData);

    kernel <<< … >>> (gpuData);
    cudaDeviceSynchronize ();
    cudaMemcpy (cpuTmp, gpuData);
    memcpy (cpuData, cpuTmp);
    MPI_Send (gpuData)
}
```

GPUDirect RDMA: What does it get you?

- **Latency Reduction**

- MPI_Send latency of 25µs with Shared GPU-Sysmem*
- No overlap possible
- Bidirectional transfer is difficult

- MPI_Send latency of 5µs with RDMA
- Does not affect running kernels
- Unlimited concurrency
- RDMA possible!

- MPI-3 One sided of 3µs

GPUDirect RDMA: Common use cases

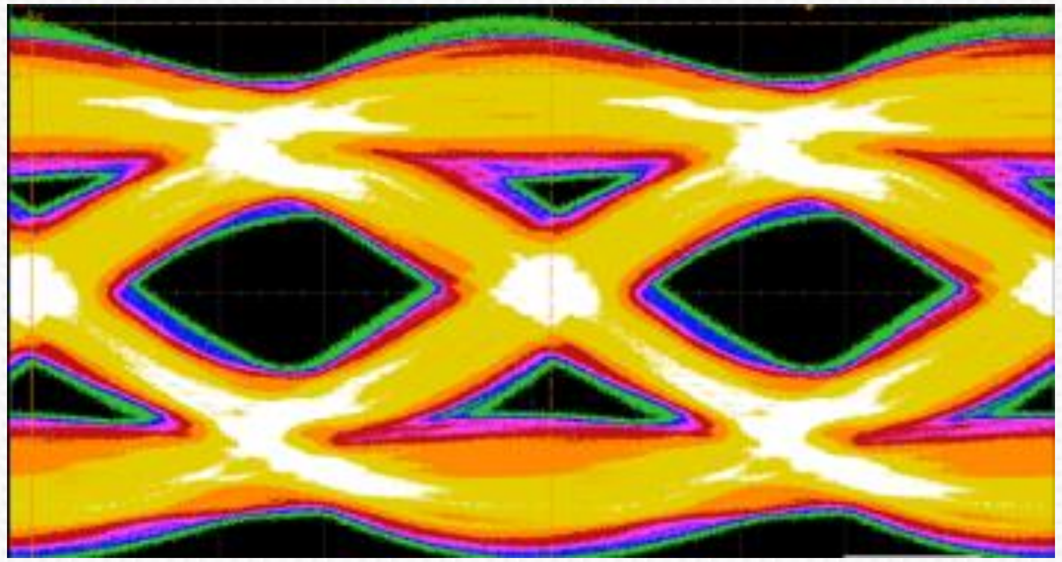
- Inter-Node MPI communication
- Transfer data between GPU and a remote Node
- Use CUDA-aware MPI

- Interface with third party hardware
- Requires adopting GPUDirect-Interop API in vendor driver stack

- Limitation
- GPUDirect RDMA does not work with CUDA Unified Memory today

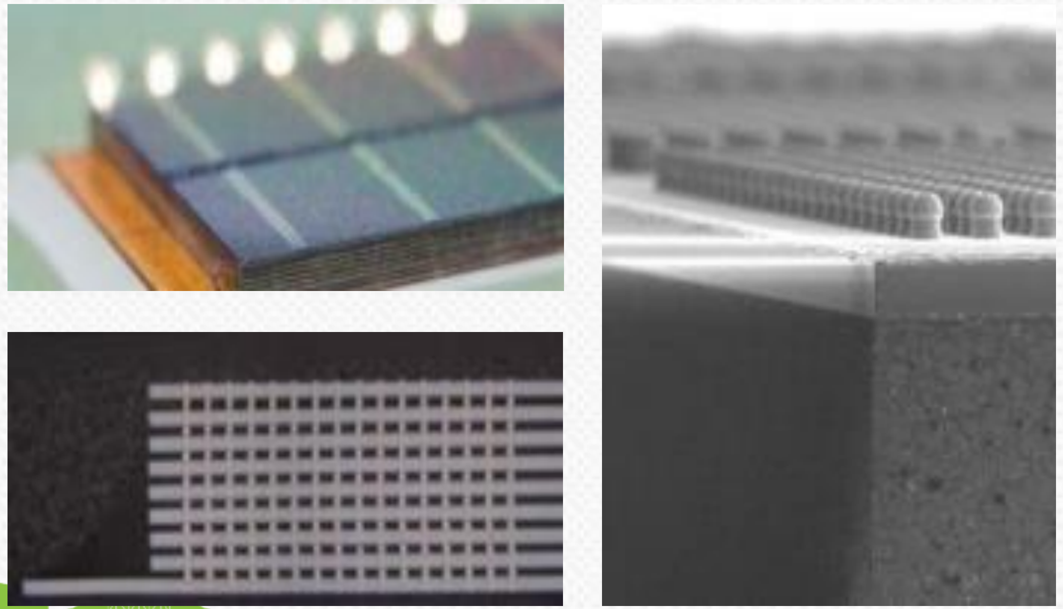
NVLINK

Pascal GPU Features
NVLINK and Stacked Memory



NVLINK

- GPU high speed interconnect
- 80-200 GB/s

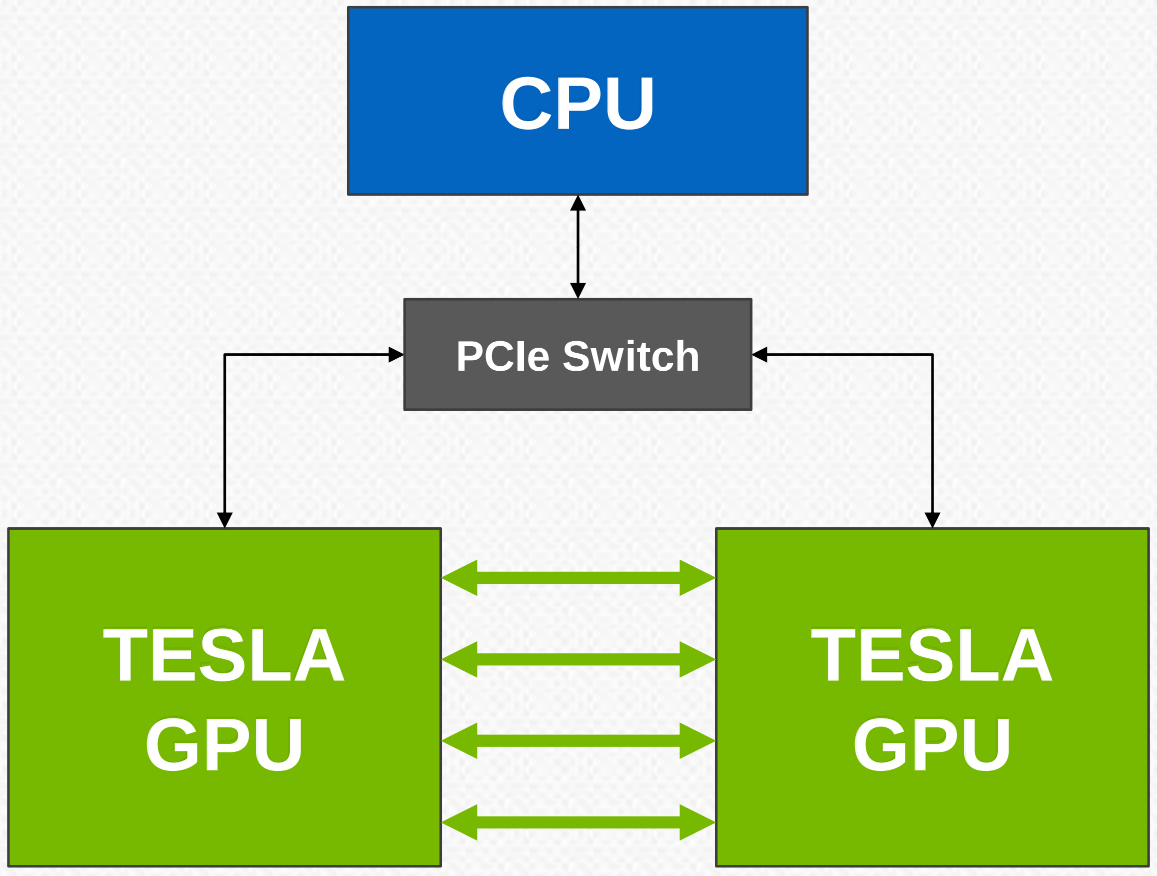


3D Stacked Memory

- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

NVLink Unleashes Multi-GPU Performance

GPUs Interconnected with NVLink



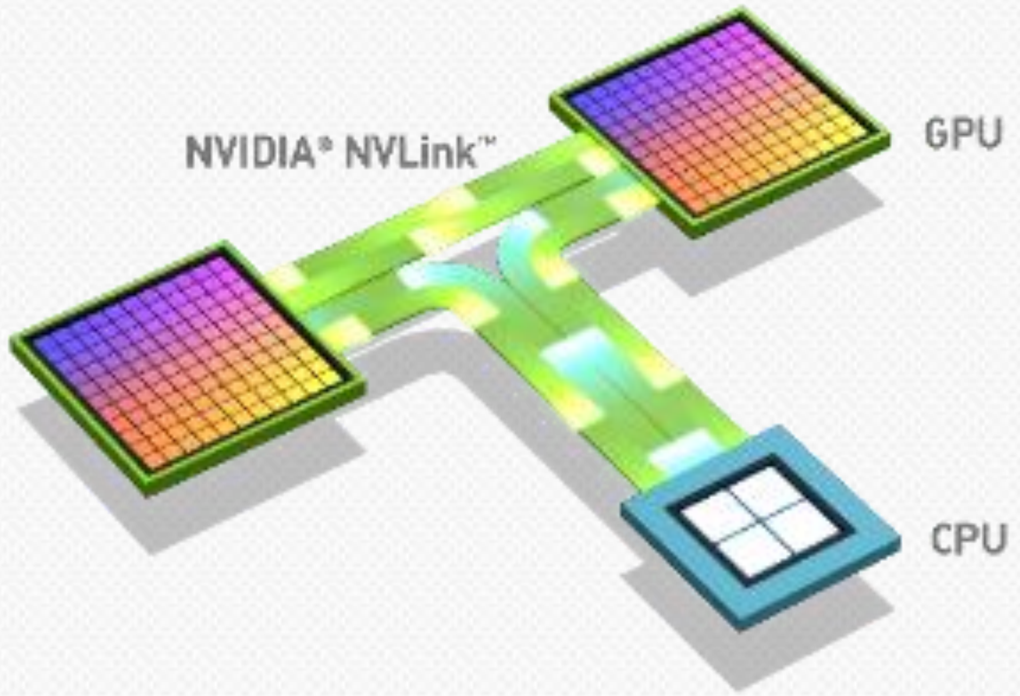
5x Faster than PCIe Gen3 x16

Over 2x Application Performance Speedup

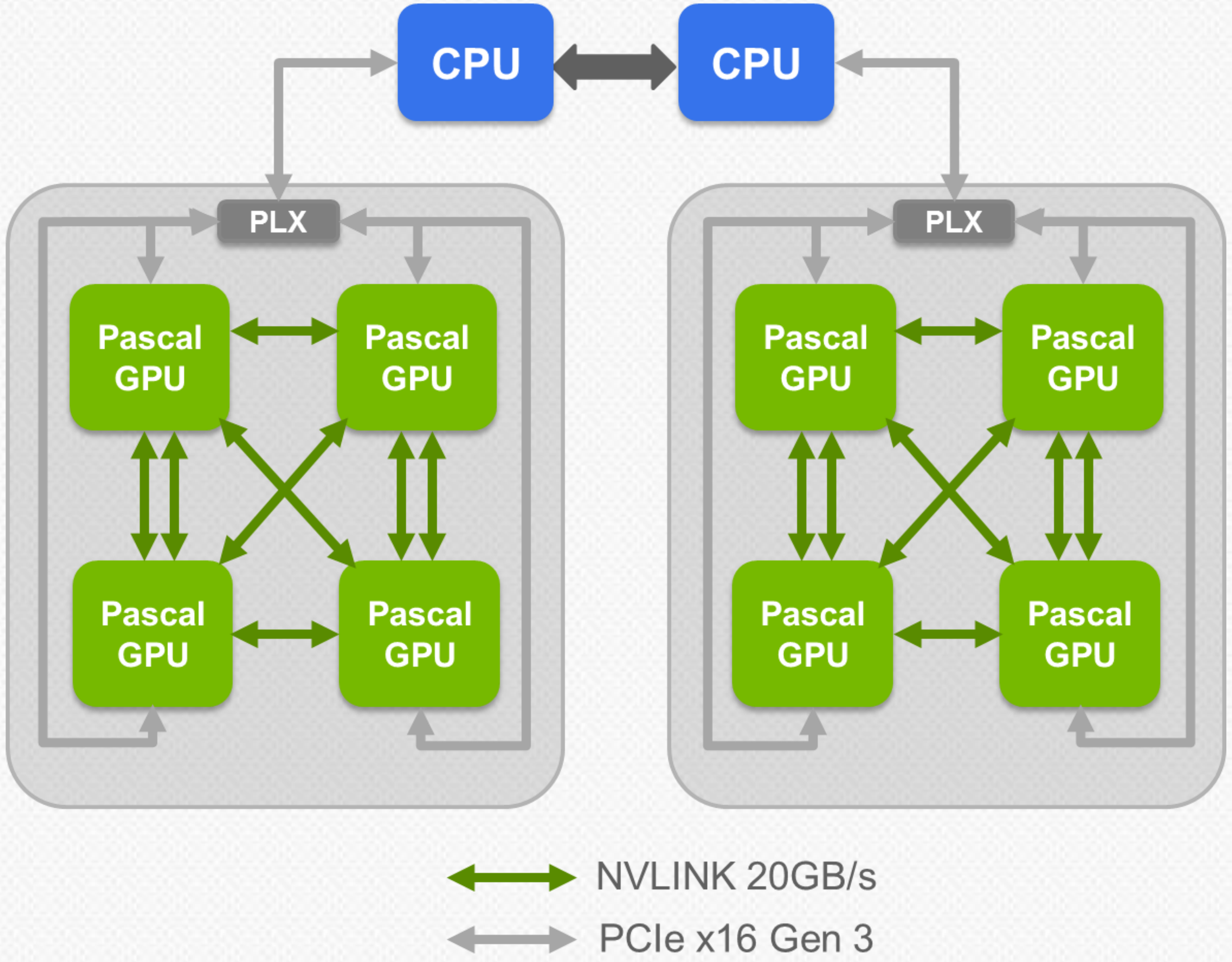
When Next-Gen GPUs Connect via NVLink Versus PCIe

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

NVLink High-Speed GPU Interconnect

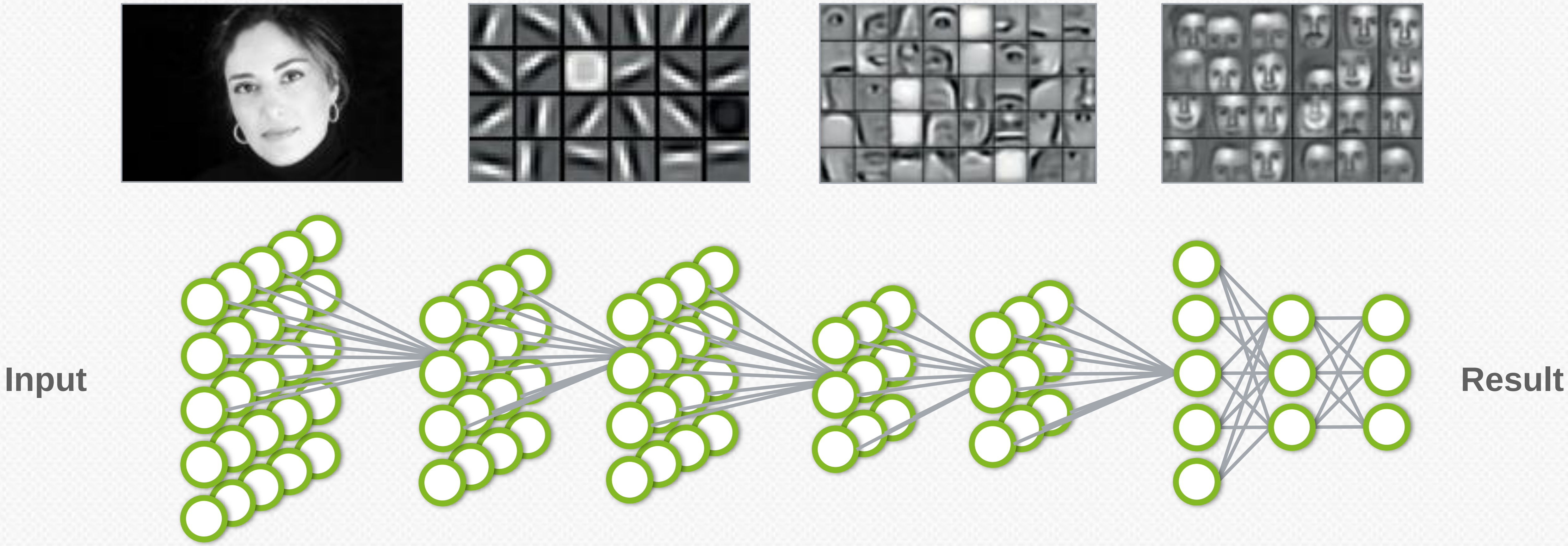


Example: 8-GPU Server with NVLink



Machine Learning

Machine Learning using Deep Neural Networks



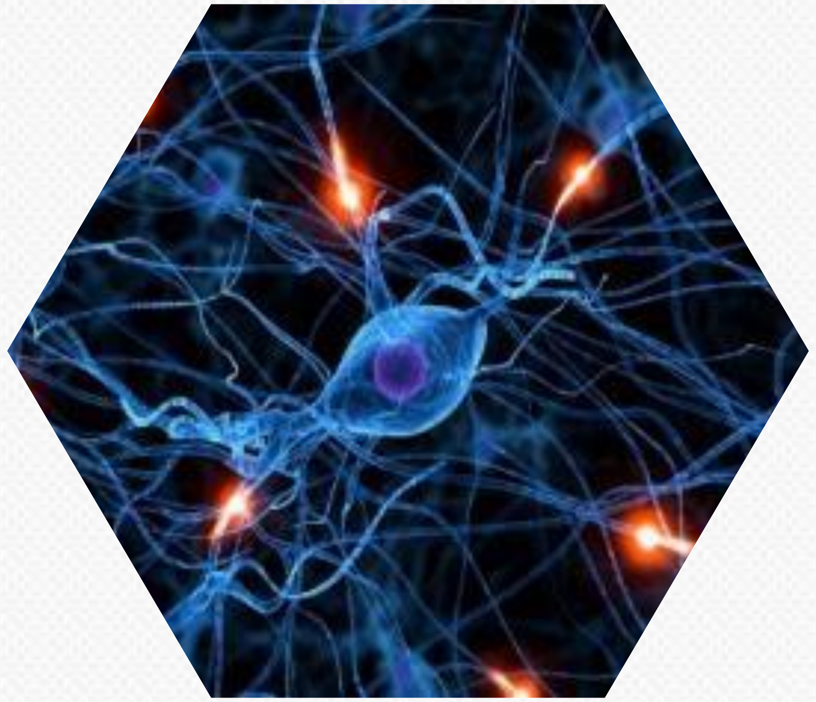
Hinton et al., 2006; Bengio et al., 2007; Bengio & LeCun, 2007; Lee et al., 2008; 2009

Visual Object Recognition Using Deep Convolutional Neural Networks
Rob Fergus (New York University / Facebook) http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2985

3 Drivers for Deep Learning

More Data

Better Models



Powerful GPU Accelerators

Broad use of GPUs in Deep learning

Early Adopters

Adobe

Image Analytics for Creative Cloud

Baidu

Speech/Image Recognition

flickr

Image Classification

IBM

Hadoop

NETFLIX

Recommendation

Yandex

Search Rankings

Use Cases

Image Detection

Face Recognition

Gesture Recognition

Video Search & Analytics

Speech Recognition & Translation

Recommendation Engines

Indexing & Search

Talks @ GTC

facebook

NYU

STANFORD UNIVERSITY

DARPA

DENSO

Carnegie Mellon University

MIT Massachusetts Institute of Technology

Berkeley University of California

What is Next?

Analyzing Unstructured Data

Anomaly Detection

Behavior Prediction

Diagnostic Support

Language Analysis

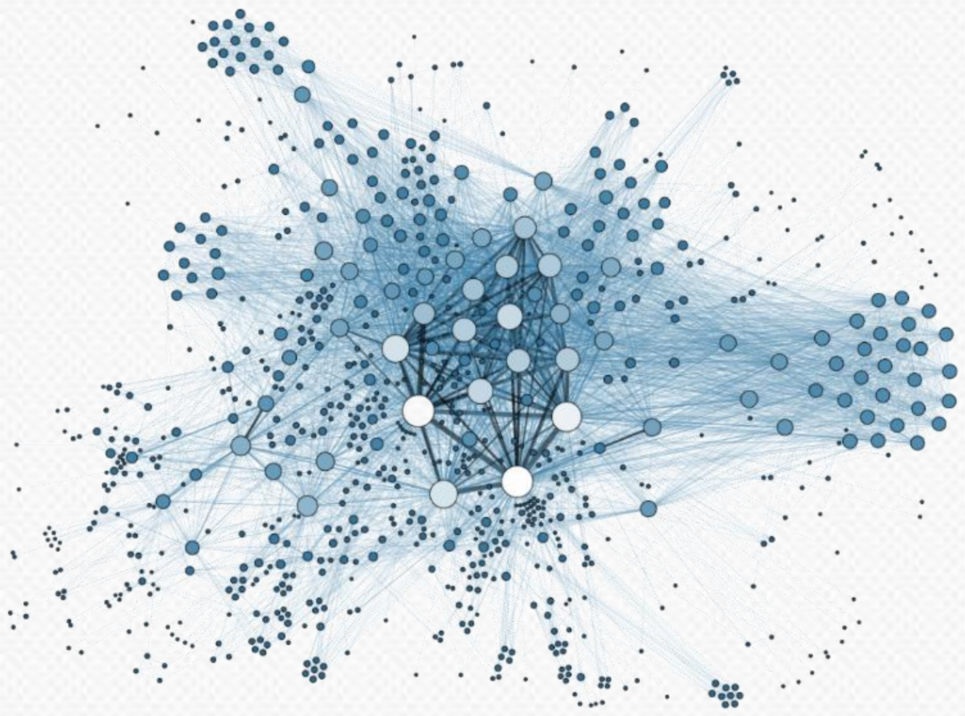
….

“Any product that excites you over the next five years and makes you think: ‘That is magical, how did they do that?’, is probably based on this [deep learning].”

Steve Jurvetson, Partner DFJ Venture

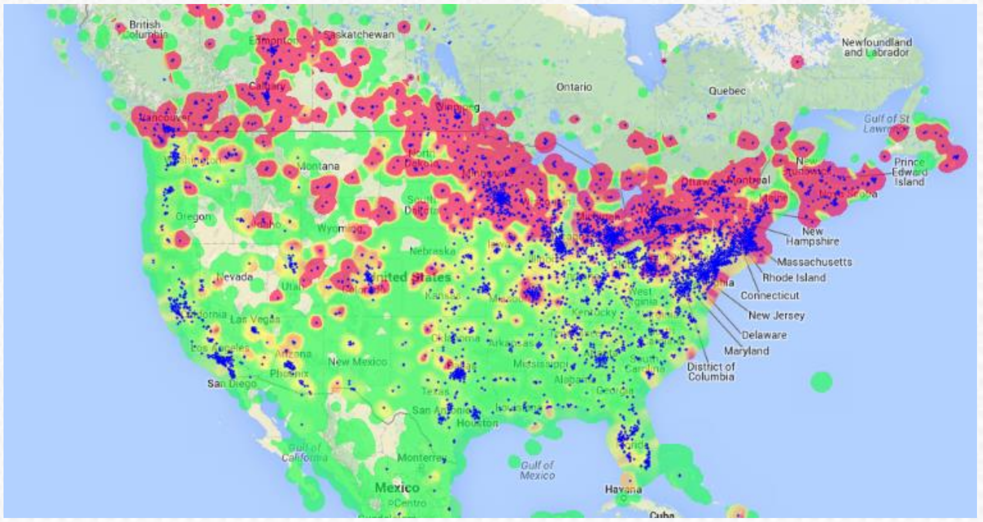
Beyond Deep Learning

Graph Analytics



Database Acceleration

Real Time Analytics



GTC 2015 had many
Deep Learning Sessions

Check GTC on-demand
http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php

Adobe	Google
Alibaba	iFlytek, Ltd
Baidu	NUANCE
Carnegie Mellon	Stanford Univ
Facebook	UC Berkeley
Flickr / Yahoo	Univ of Toronto

NVIDIA in OCP

Unlocking Access to the Tesla Platform

Engaging with OEMs, end customers and technology partners to include NVIDIA Accelerators in the OCP Platform

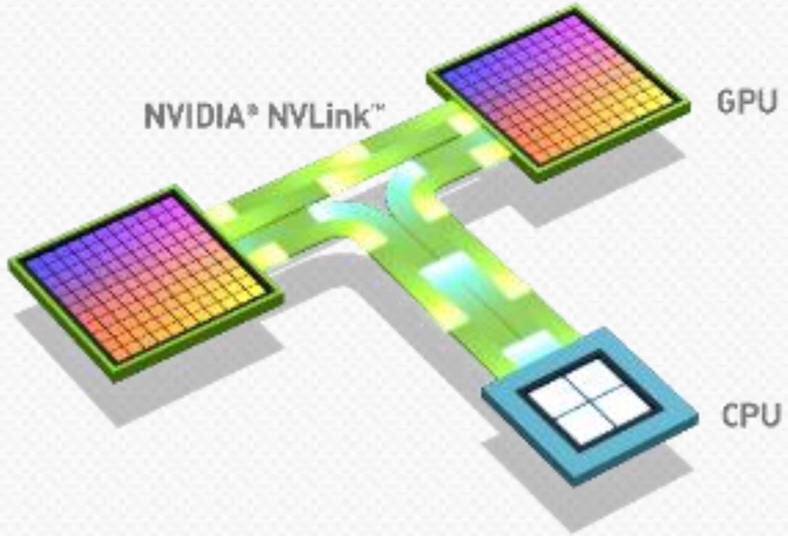
NVLINK based designs for maximum performance

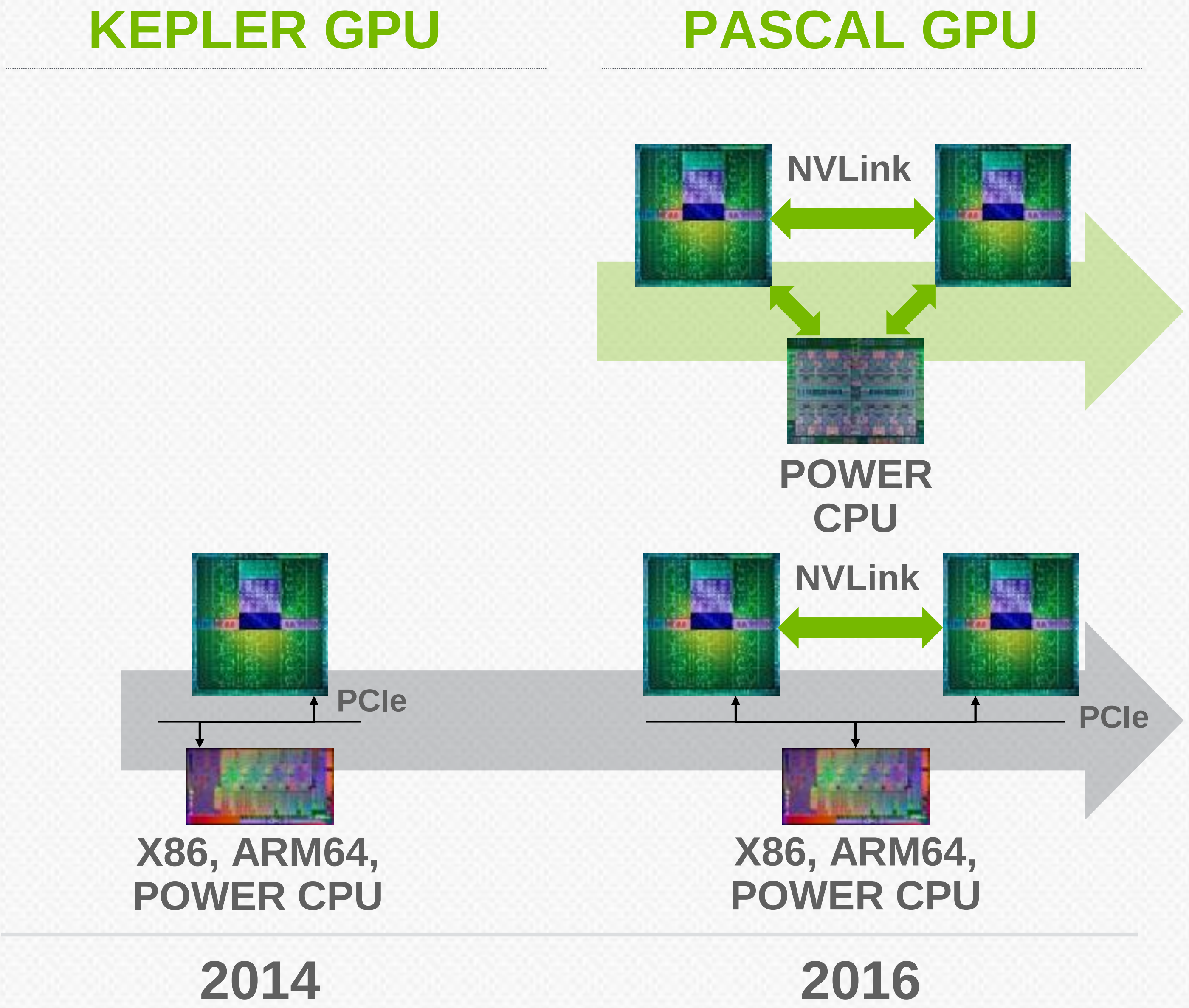
Standard PCIe designs for scale out

Enabling NVLink GPU-CPU connections

NVLink

High-Speed GPU Interconnect





Thanks

cangerer@nvidia.com