CSC 447: Parallel Programming for Multi-Core and Cluster Systems

Introduction to GPU and CUDA

Instructor: Haidar M. Harmanani

Spring 2018

GPU Introduction

- GPUs are massively parallel architectures that have become commodities
  - Highly parallel (100s of processor cores)
  - Very fast (>900 GFLOPS of peak performance)
  - Operates in a SIMD manner
    - A key restriction
  - Multiple processors operate in lock-step (same instruction) but on different data

GPU Introduction

- A GPU is similar to a symmetrical multiprocessor system on a single processor
  - A number of SMs exist on a single GPU and share a common global memory space.
  - SM is a processor in its own right, capable of running up multiple blocks of threads, typically 256, 512, 0r 1024 threads per block.

Acceleration

- Used to accelerate image/stream processing, data compression, numerical algorithms, and CAD algorithms.
- Inexpensive, off-the-shelf cards like the NVIDIA Quadro FX / 280 GTX GPU achieve impressive performance
  - 933 GFLOPs peak performance
  - 240 SIMD cores partitioned into 30 Multiprocessors (MPs)
  - 4GB (Quadro) and 1GB (GTX 280) device memory with 142 GB/s bandwidth
  - 1.4 GHz GPU operating frequency

GPU Introduction

- GPU hardware consists of a number of key blocks:
  - Memory (global, constant, shared)
  - Streaming multiprocessors (SMs)
  - Streaming processors (SPs)
- There are multiple SPs in each SM
  - Number depends on the GPU generation
- Global memory is provided through GDDR on the graphics card

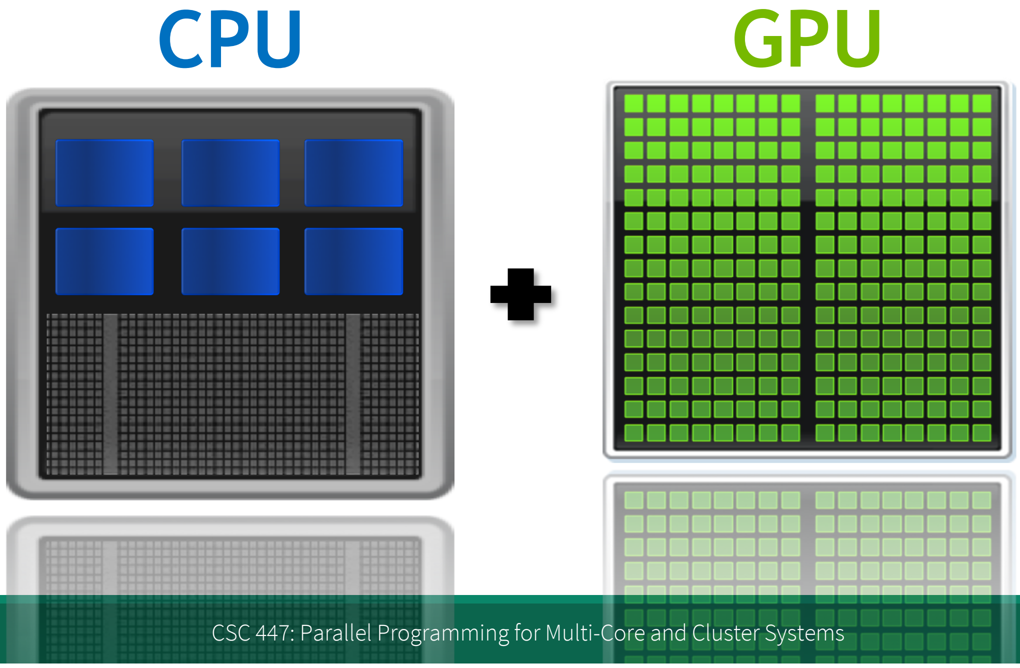
GPU Generations

- Tesla (Compute Capability 1)
- Fermi (Compute Capability 2)
- Kepler (Compute Capability 3)
- Maxwell (Compute Capability 5)
- Pascal (Compute Capability 6)
  - TITAN Xp, Titan X, GeForce GTX 1080 Ti, GTX 1080, GTX 1070 Ti, GTX 1070, GTX 1060, GTX 1050 Ti, GTX 1050, GT 1030, MX150
- Volta (Latest Generation, Compute Capability 7)
  - TITAN V

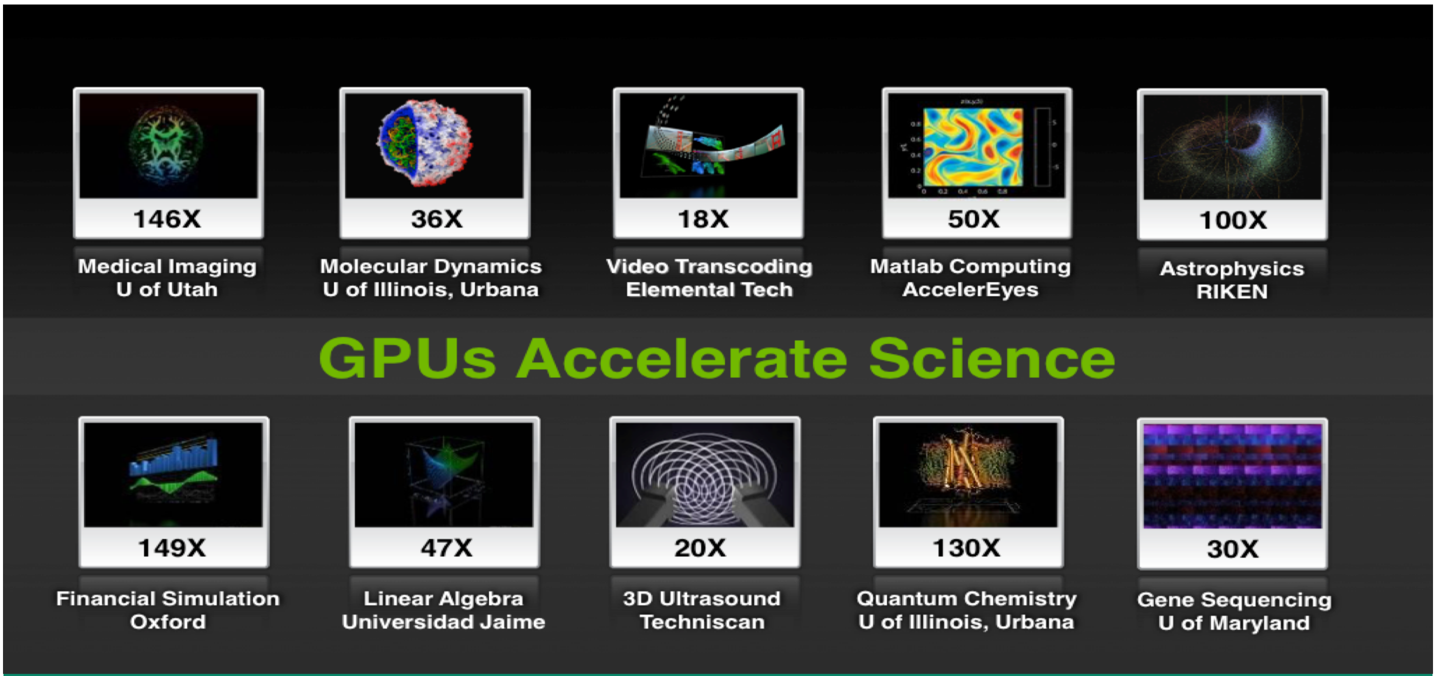
GPU Generations

- Tesla is the brand name for NVidia's GPGPU line of cards as well as the name for the 1st generation microarchitecture

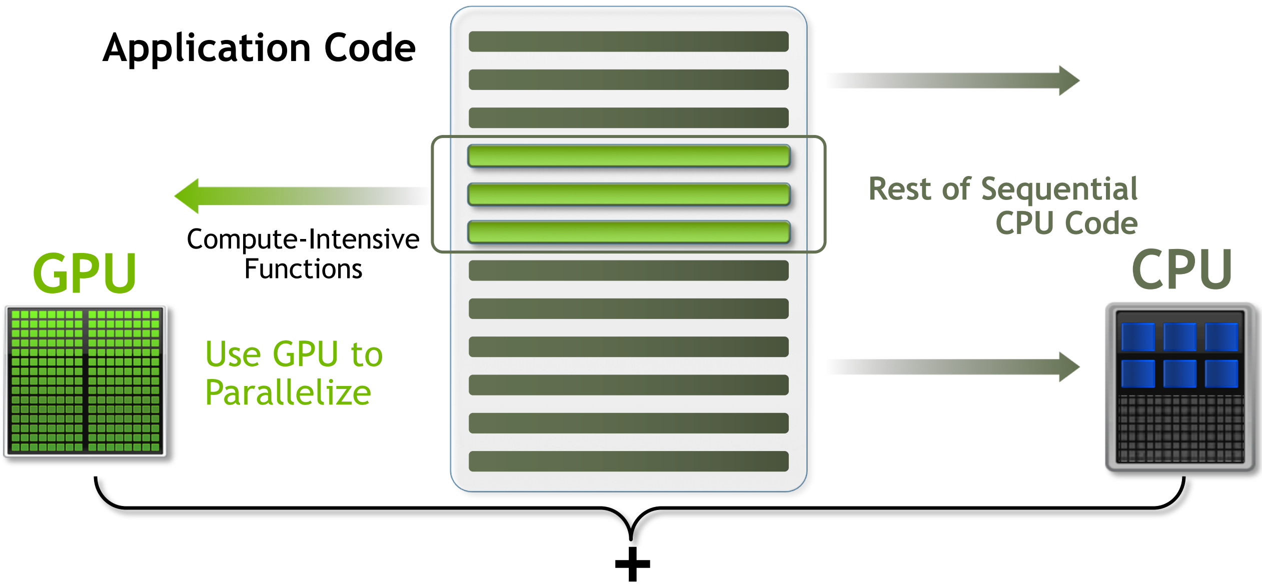
GPUs: Accelerate Science Applications



GPUs: Accelerate Science Applications

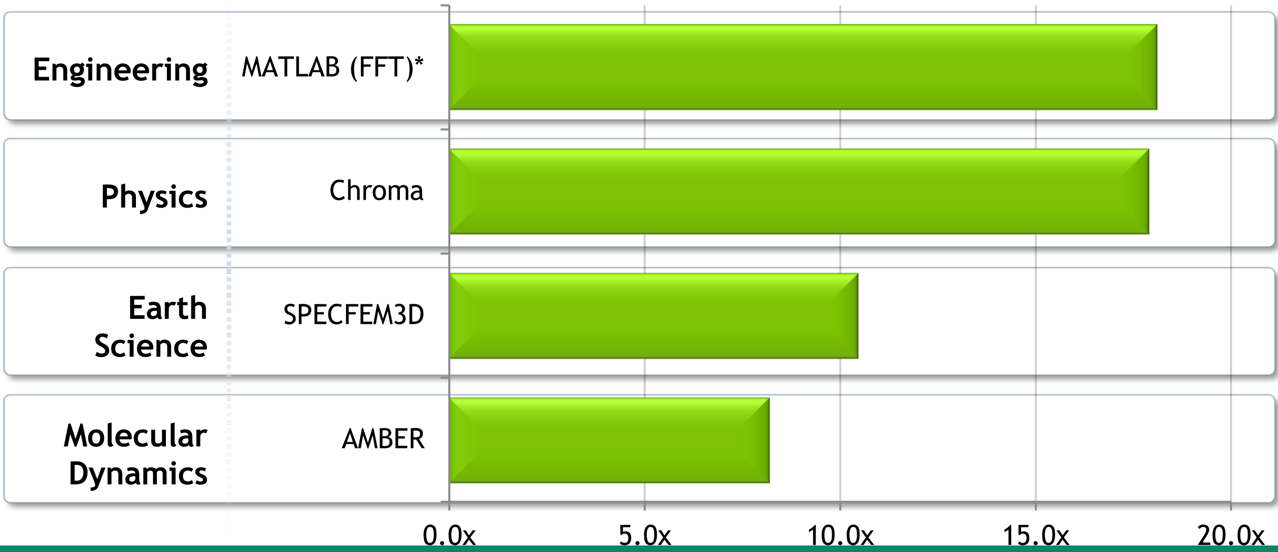


Small Changes, Big Speed-up



Fast Performance on Scientific Applications

Tesla K20X Speed-Up over Sandy Bridge CPUs

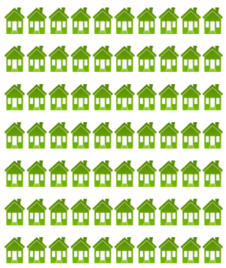


Why Computing Perf/Watt Matters?

2.3 PFlops

7.0 Megawatts

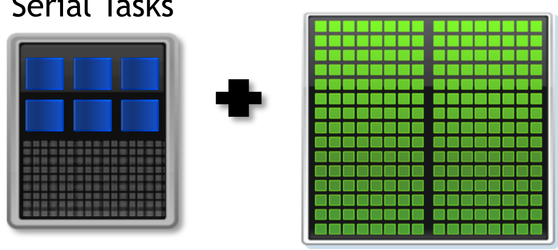
7000 homes



7.0 Megawatts

CPU
Optimized for Serial Tasks

GPU Accelerator
Optimized for Many Parallel Tasks



10x performance/socket
> 5x energy efficiency

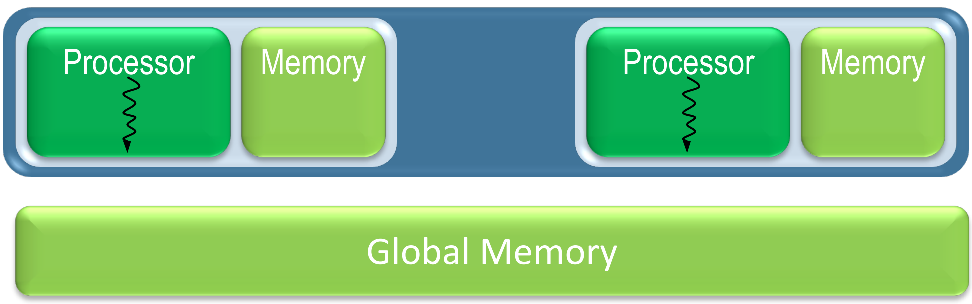
Traditional CPUs are not economically feasible

Era of GPU-accelerated computing is here

GPU Hardware

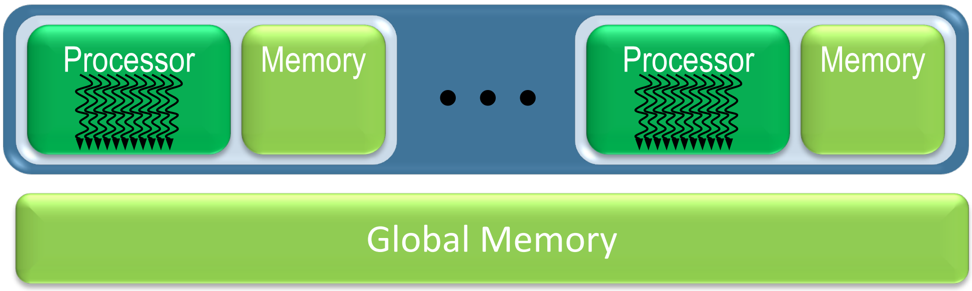
- A GPU has multiple streaming multiprocessors (SM) that contain
  - memory registers for threads to use
  - several memory caches
    - shared memory
    - constant cache
    - texture memory
    - L1 cache
  - Thread schedulers
  - Several CUDA cores
    - A core consists of an Arithmetic logic unit (ALU) that handles integer and single precision calculations and a Floating point unit (FPU) that handles double precision calculations
  - Special function units (SFU) for transcendental functions (e.g. log, exp, sin, cos, sqrt)

Generic Multicore Chip



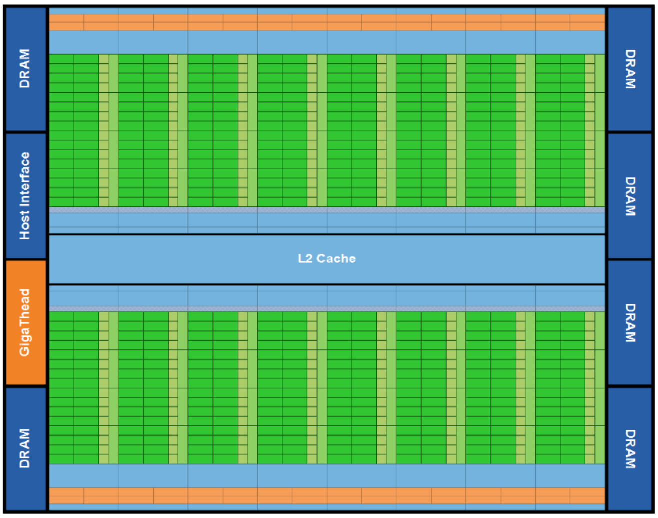
- Handful of processors each supporting ~1 hardware thread
- On-chip memory near processors (cache, RAM, or both)
- Shared global memory space (external DRAM)

Generic Manycore Chip



- Many processors each supporting many hardware threads
- On-chip memory near processors (cache, RAM, or both)
- Shared global memory space (external DRAM)

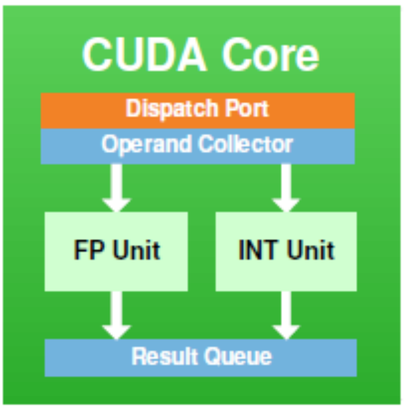
Inside a GPU



Streaming Multiprocessor Architecture

- Multiple CUDA cores per SM
  - 32-128 cores!
- 2:1 ratio SP:DP floating-point performance
- Dual Thread Scheduler
- 64 KB of RAM for shared memory and L1 cache (configurable)

CUDA Core



Tesla C-Series Workstation GPUs



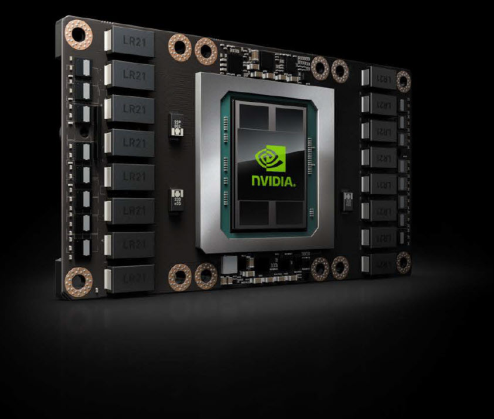
	Tesla C1060	Tesla C2050	Tesla C2070
Architecture	Tesla 10-series GPU	Tesla 20-series GPU	
Number of Cores	240	448	
Caches	16 KB Shared Memory / 8 cores	64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache	
Floating Point Peak Performance	933 Gigaflops (single) 78 Gigaflops (double)	1030 Gigaflops (single) 515 Gigaflops (double)	
GPU Memory	4 GB	3 GB 2.625 GB with ECC on	6 GB 5.25 GB with ECC on
Memory Bandwidth	102 GB/s (GDDR3)	144 GB/s (GDDR5)	
System I/O	PCIe x16 Gen2	PCIe x16 Gen2	
Power	188 W (max)	237 W (max)	225 W (max)
Available	Available now	Available now	Available now

Lab Hardware


- nVidia GeForce GTX 460
  - 1.95 billion transistors
  - 336 CUDA Cores at 1350 MHz
  - 7 Streaming Multiprocessors
  - 1024 MB Memory with 115.2 GB/sec bandwidth
- nVidia Tesla C2070
  - 3 billion transistors
  - 448 CUDA Cores at 1.15 GHz
  - 6GB dedicated Memory with 144 GB/sec bandwidth
  - Double Precision floating point performance (peak): 515 GFLOPs
  - Single Precision floating point performance (peak): 1.03 TFLOPs

Lab Hardware

- nVidia GeForce GTX 980
  - Maxwell
  - 5.2 billion transistors
  - 16 Streaming Multiprocessors
    - 128 cores each
  - 2048 CUDA Cores at 1126 MHz
  - 4GB Memory with 224.5 GB/sec bandwidth



TESLA P100

THE MOST ADVANCED
HYPERSCALE DATACENTER GPU EVER BUILT

150B XTORS | 5.3TF FP64 | 10.6TF FP32 | 21.2TF FP16 | 14MB SM RF | 4MB L2 Cache

GTX Titan: For High Performance Gaming Enthusiasts

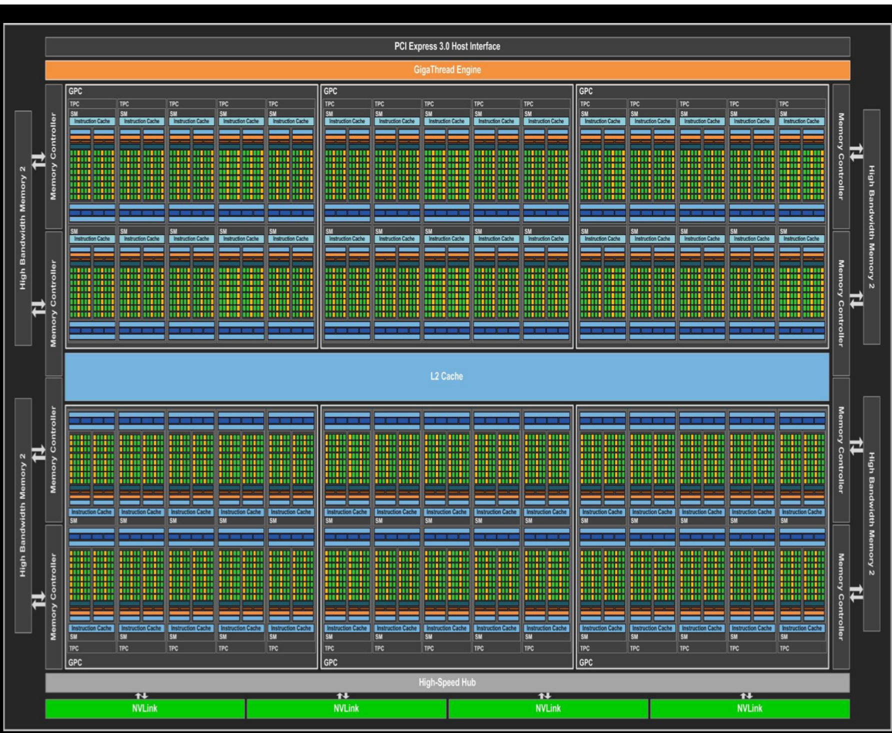
CUDA Cores	2688
Single Precision	~4.5 Tflops
Double Precision	~1.27 Tflops
Memory Size	6GB
Memory B/W	288GB/s

Pascal

- Titan Z
  - 5,760 cores, 12GB (6GB x2) of 7Gb/s GDDR5 memory
  - 8 TFLOPS
  - $3000
- Titan X
  - 8 billion transistors
  - 3072 CUDA cores
  - 7 TFLOPS SP/0.2 TFLOPS DP
  - 12 GB Memory
  - Geared towards to game development at 4K resolution
  - $1000
- Titan Xp
  - 30 SM, 3840 cores
  - 12 billion transistors
  - 12 GB GDDR5X
  - 12 TFLOPS

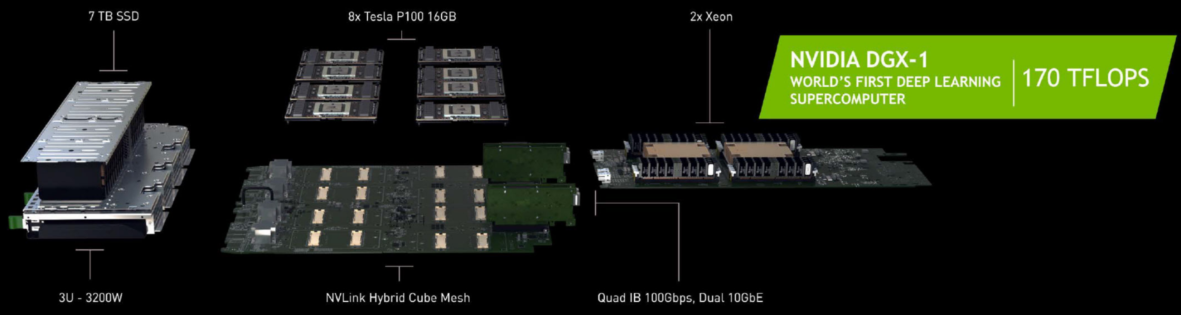
NVIDIA Tesla Graphics Card	Tesla K40 (PCI-Express)	Tesla M40 (PCI-Express)	Tesla P100 (PCI-Express)	Tesla P100 (PCI-Express)	Tesla P100 (Mezzanine)
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GP100 (Pascal)	GP100 (Pascal)
Process Node	28nm	28nm	16nm	16nm	16nm
Transistors	7.1 Billion	8 Billion	15.3 Billion	15.3 Billion	15.3 Billion
GPU Die Size	551 mm2	601 mm2	610 mm2	610 mm2	610 mm2
SMs	15	24	56	56	56
CUDA Cores Per SM	192	128	64	64	64
CUDA Cores (Total)	2880	3072	3584	3584	3584
FP64 CUDA Cores / SM	64	4	32	32	32
FP64 CUDA Cores / GPU	960	96	1792	1792	1792
Base Clock	745 MHz	948 MHz	TBD	TBD	1328 MHz
Boost Clock	875 MHz	1114 MHz	1300MHz	1300MHz	1480 MHz
FP64 Compute	1.68 TFLOPs	0.2 TFLOPs	4.7 TFLOPs	4.7 TFLOPs	5.30 TFLOPs
Texture Units	240	192	224	224	224
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2	4096-bit HBM2
Memory Size	12 GB GDDR5	24 GB GDDR5	12 GB HBM2	16 GB HBM2	16 GB HBM2
L2 Cache Size	1536 KB	3072 KB	4096 KB	4096 KB	4096 KB
TDP	235W	250W	250W	250W	300W

Pascal GP100 Block Diagram

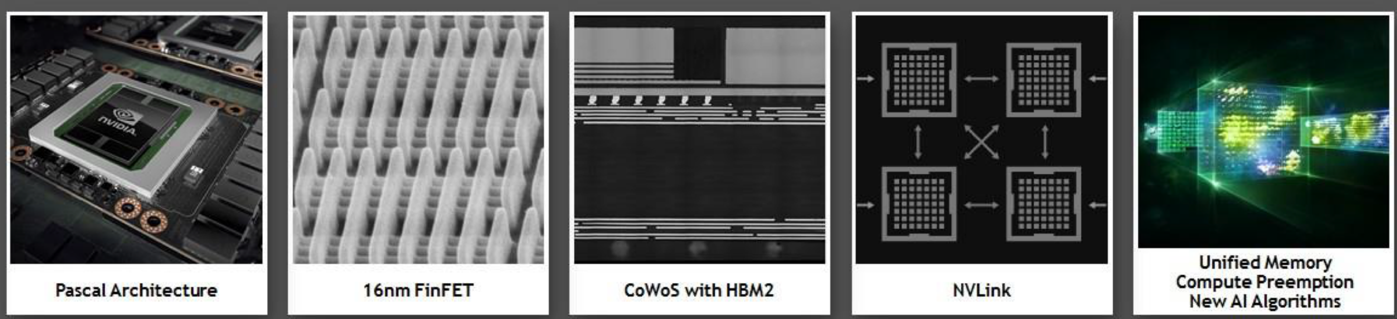


Pascal GP100 Block Diagram

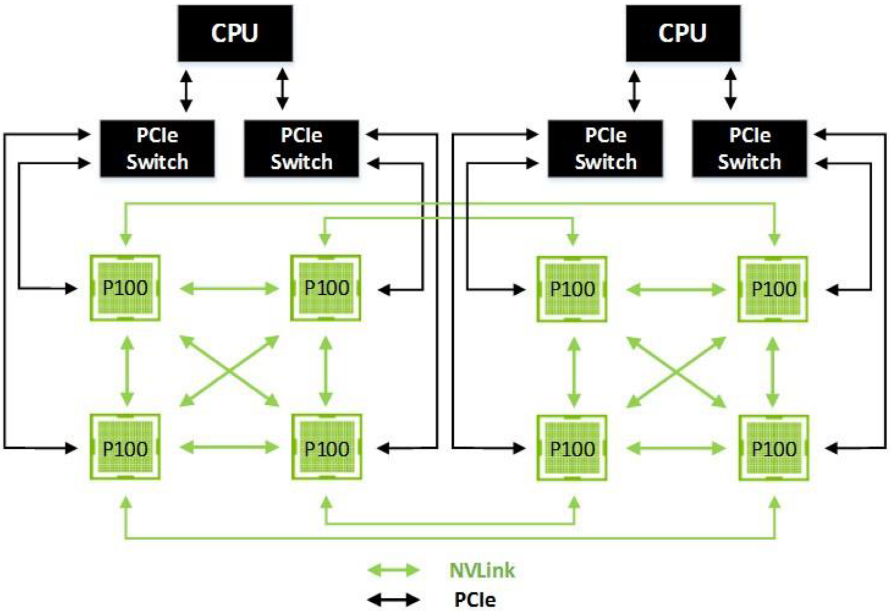
The Pascal GP100 Streaming Multiprocessor



- Up to 8 Tesla P100 boards and costs $129,000 US
  - Up to 170 teraflops of half-precision (FP16) peak performance
  - Eight Tesla P100 GPU accelerators, 16GB memory per GPU
  - NVLink Hybrid Cube Mesh
  - 7TB SSD DL Cache
  - Dual 10GbE, Quad InfiniBand 100Gb networking
  - 3U – 3200W

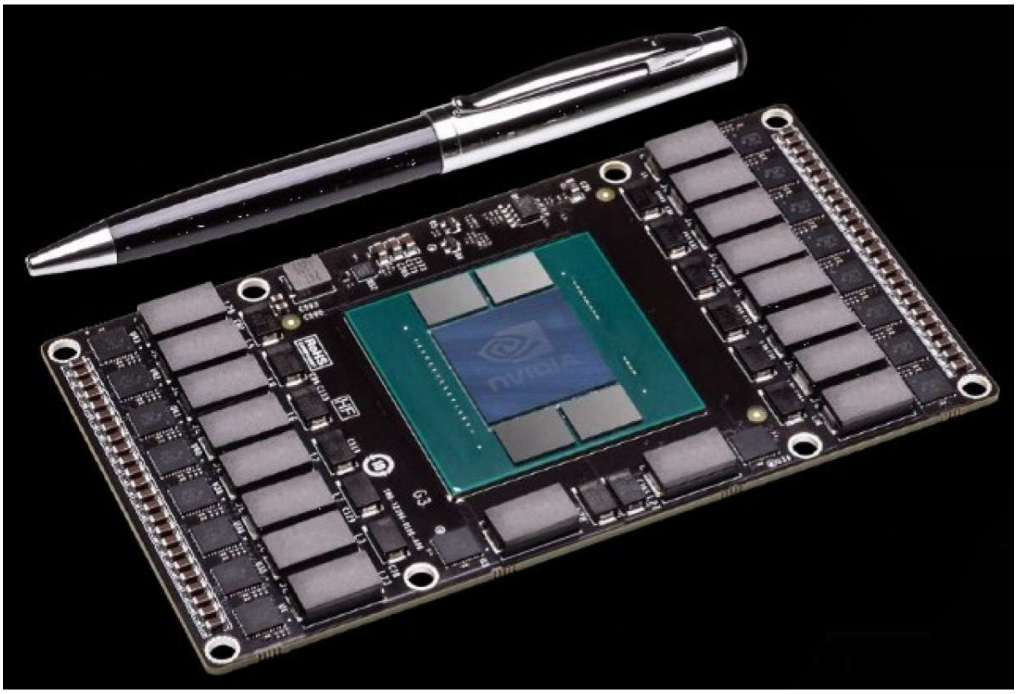


Tesla P100 New Technologies

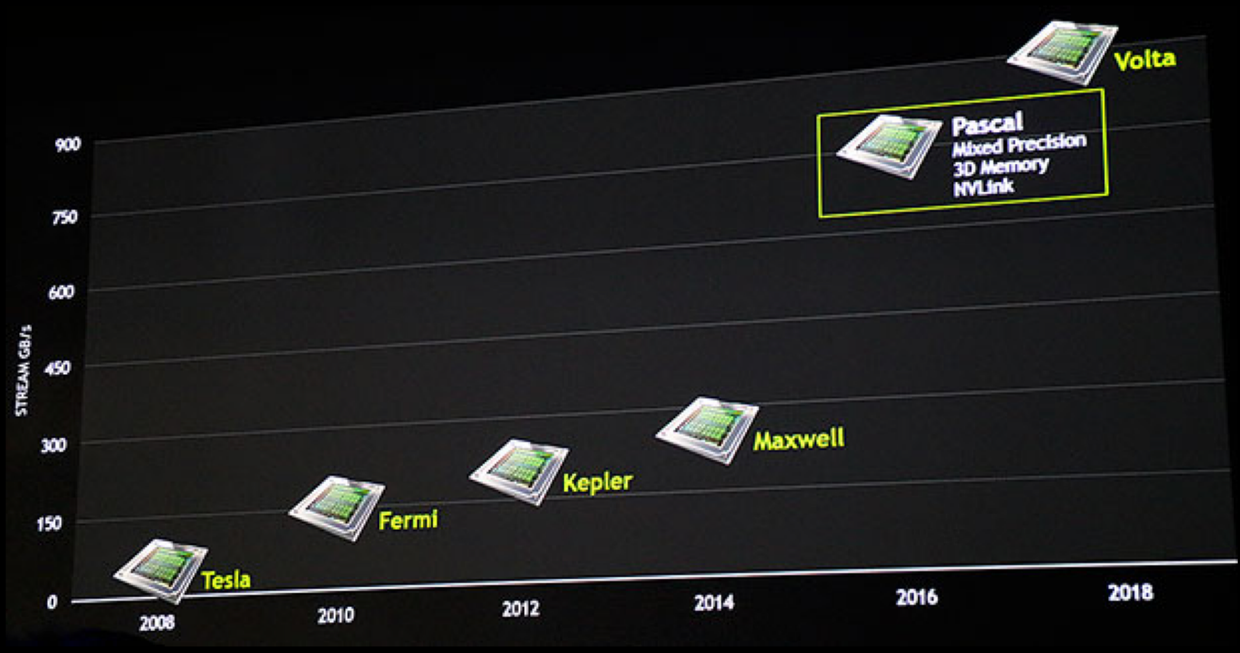


GPU-to-GPU data transfers at up to 160 Gigabytes/second of bidirectional bandwidth—5x the bandwidth of PCIe Gen 3 x16

NVLink Connecting Eight Tesla P100 Accelerators in a Hybrid Cube Mesh Topology



Pascal GP104 Block Diagram



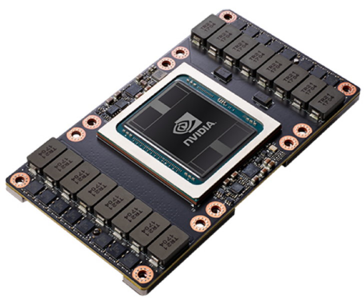
Volta
Pascal
Mixed Precision
3D Memory
NVLink
Maxwell
Kepler
Fermi
Tesla
STREAM GB/s
900
750
600
450
300
150
0
2008
2010
2012
2014
2016
2018

Volta's Titan V

- Volta architecture
- 80 SM
- 5120 single precision cores
- 110 TFLOPS!
- Defacto GPU for deep learning
- $3000 GPU

NVIDIA's Most Powerful GPU

- Tesla V100
  - 640 cores
  - Multiple V100 GPUs connected using NVLink
  - 112 teraflops
  - $10,000



Other Simpler GPU Compute Capabilities

Product	Compute Capability
Tesla C2050/C2070	2.0
Tesla K80	3.7
Tesla K10	3.0
Quadro K2200	5.0
Tegra K1	3.2
Jetson TK1	3.2

GPU	Kepler GK110	Maxwell GM200	Pascal GP100
Compute Capability	3.5	5.2	6.0
Threads / Warp	32	32	32
Max Warps / Multiprocessor	64	64	64
Max Threads / Multiprocessor	2048	2048	2048
Max Thread Blocks / Multiprocessor	16	32	32
Max 32-bit Registers / SM	65536	65536	65536
Max Registers / Block	65536	32768	65536
Max Registers / Thread	255	255	255
Max Thread Block Size	1024	1024	1024
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB

Closing Remarks: Machine Learning

- A lot of data and a lot of burning questions
- Need techniques that minimize software engineering effort
  - simple algorithms, teach computer how to learn from data
  - don't spend time hand-engineering algorithms or high-level features from the raw data

Closing Remarks: Deep Learning

- The modern reincarnation of Artificial Neural Networks from 1980's and1990's
- A collection of trainable mathematical units which collaborate to compute a complicated function
- Compatible with supervised, unsupervised, and reinforcement learning

- Widely used at Google and Microsoft for image processing
- Harnesses the power of multiprocessors and GPUs