Parallel Computing with GPUs

GPU Architectures
Part 3 – GPU Hardware

The University Of Sheffield.

Dr Paul Richmond

http://paulrichmond.shef.ac.uk/teaching/COM4521/

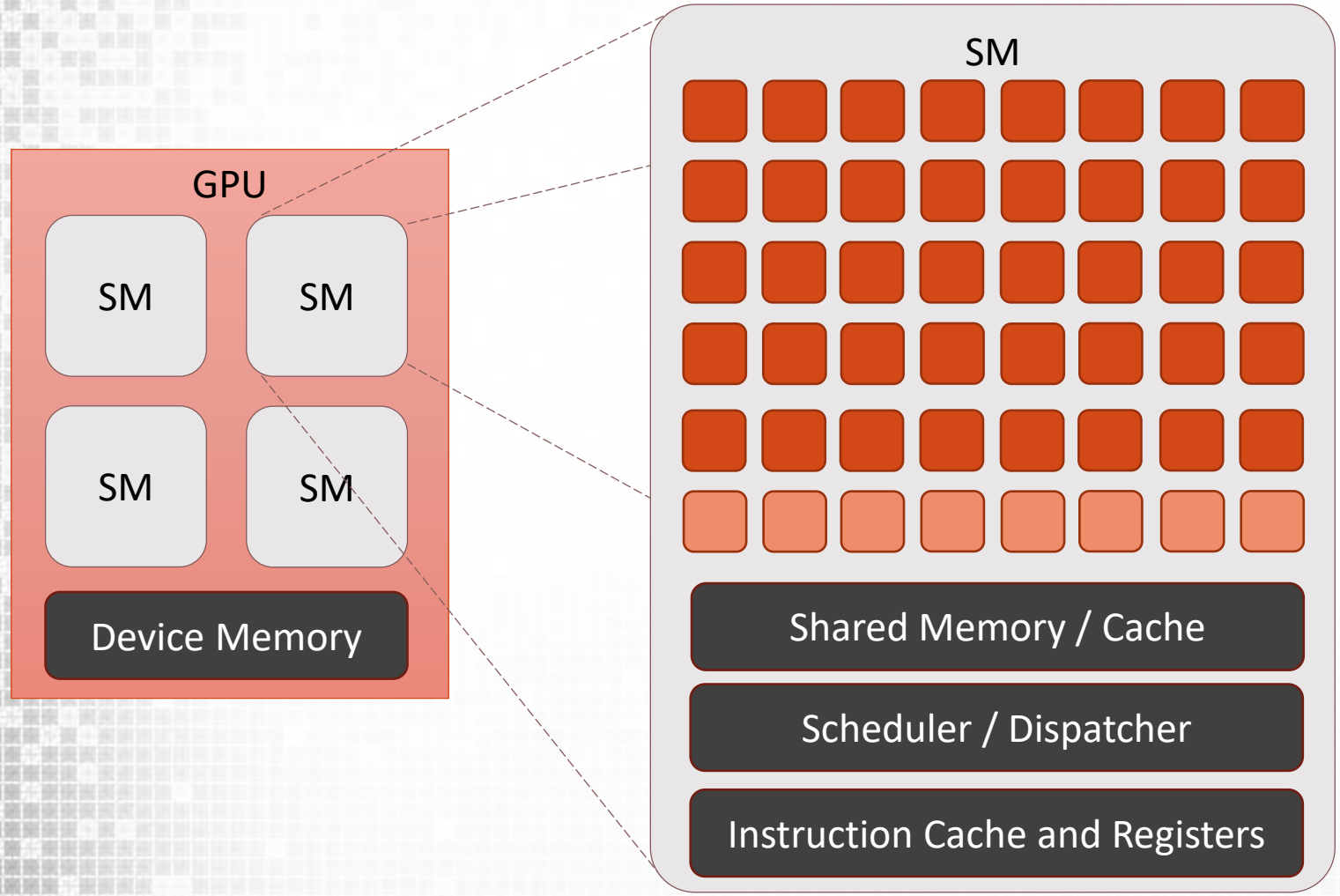
This Lecture (learning objectives)

- NVIDIA GPU Hardware
  - Explain the NVIDIA hardware model and key terminology
  - Compare the hardware variants and identify changing architectural characteristics
  - Give examples of GPU usage at different system scales

NVIDIA GPU Range

- GeForce
  - Consumer range
  - Gaming oriented for mass market
- Quadro Range
  - Workstation and professional graphics
- Tesla
  - Number crunching boxes
  - Much better support for double precision
  - Faster memory bandwidth
  - Better Interconects
- Mobile …

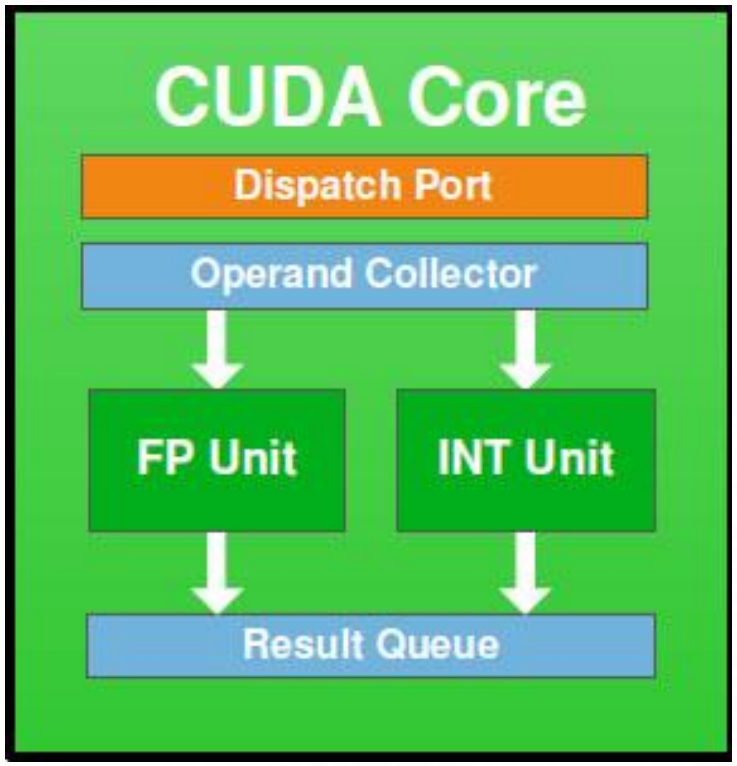
Hardware Model



- NVIDIA GPUs have a 2-level hierarchy
  - Each Streaming Multiprocessor (SMP) has multiple vector “CUDA” cores
  - The number of SMs varies across different hardware implementations
  - The design of SMPs varies between GPU families
  - The number of cores per SMP varies between GPU families

NVIDIA CUDA Core

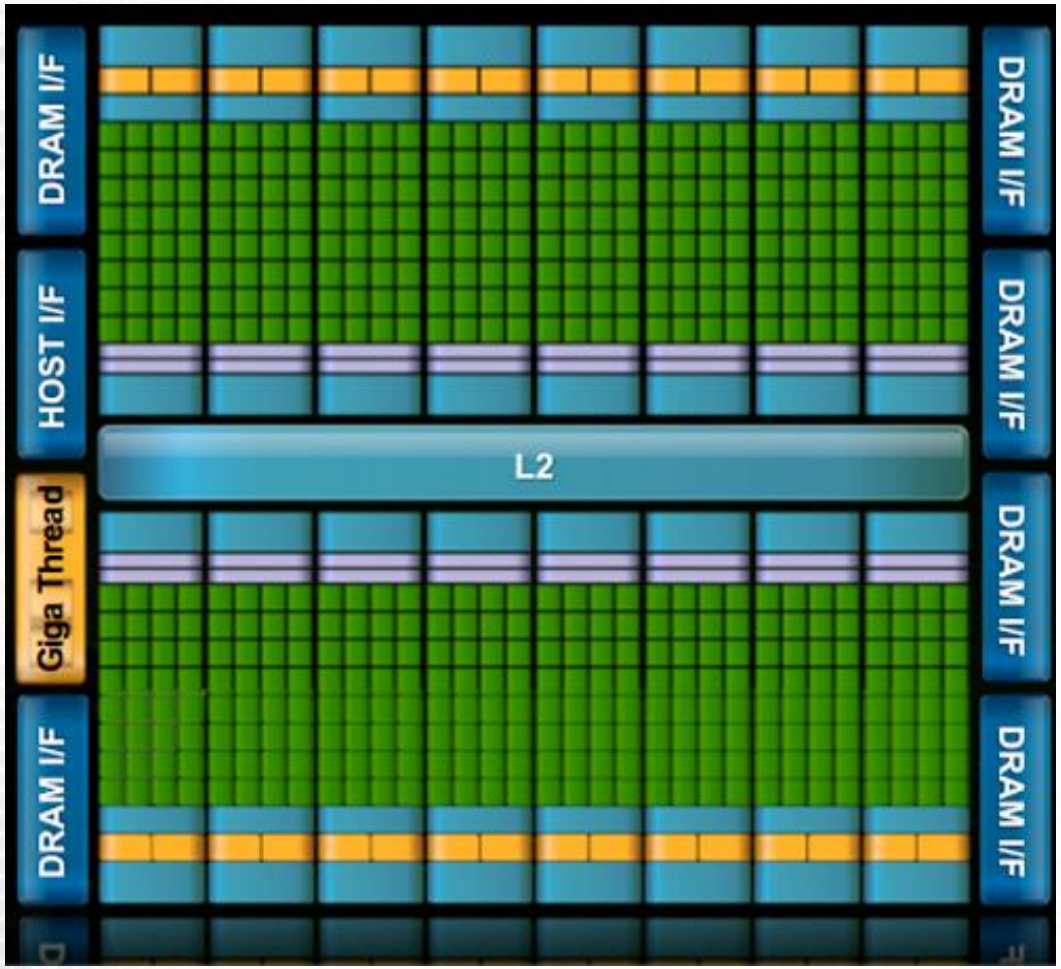
- CUDA Core
  - Vector processing unit
  - Either FP32, FP64
    - Volta onwards: INT32 and Tensor Cores
  - Works on a single operation by initiating instructions on clock cycles



Tesla Range Specifications

	"Kepler" K20	"Kepler" K40	"Maxwell" M40	Pascal P100	Volta V100	Ampere A100
32bit CUDA cores	2496	2880	3072	3584	5120	6912
Chip Variant	GK110	GK110B	GM200	GP100	GV100	GA100
Cores per SM	192	192	128	64	64	64
Single Precision Performance	3.52 Tflops	4.29 Tflops	7.0 Tflops	9.5 Tflops	15.4 Tflops	**19.5 Tflops**
Double Precision Performance	1.17 TFlops	1.43 Tflops	*0.21 Tflops*	4.7 Tflops	7.8Tflops	**9.7 Tflops**
Memory Bandwidth	208 GB/s	288 GB/s	288GB/s	720GB/s	900GB/s	**1555 GB/s**
Memory	5 GB	12 GB	12GB	12/16GB	16/32GB	40 GB

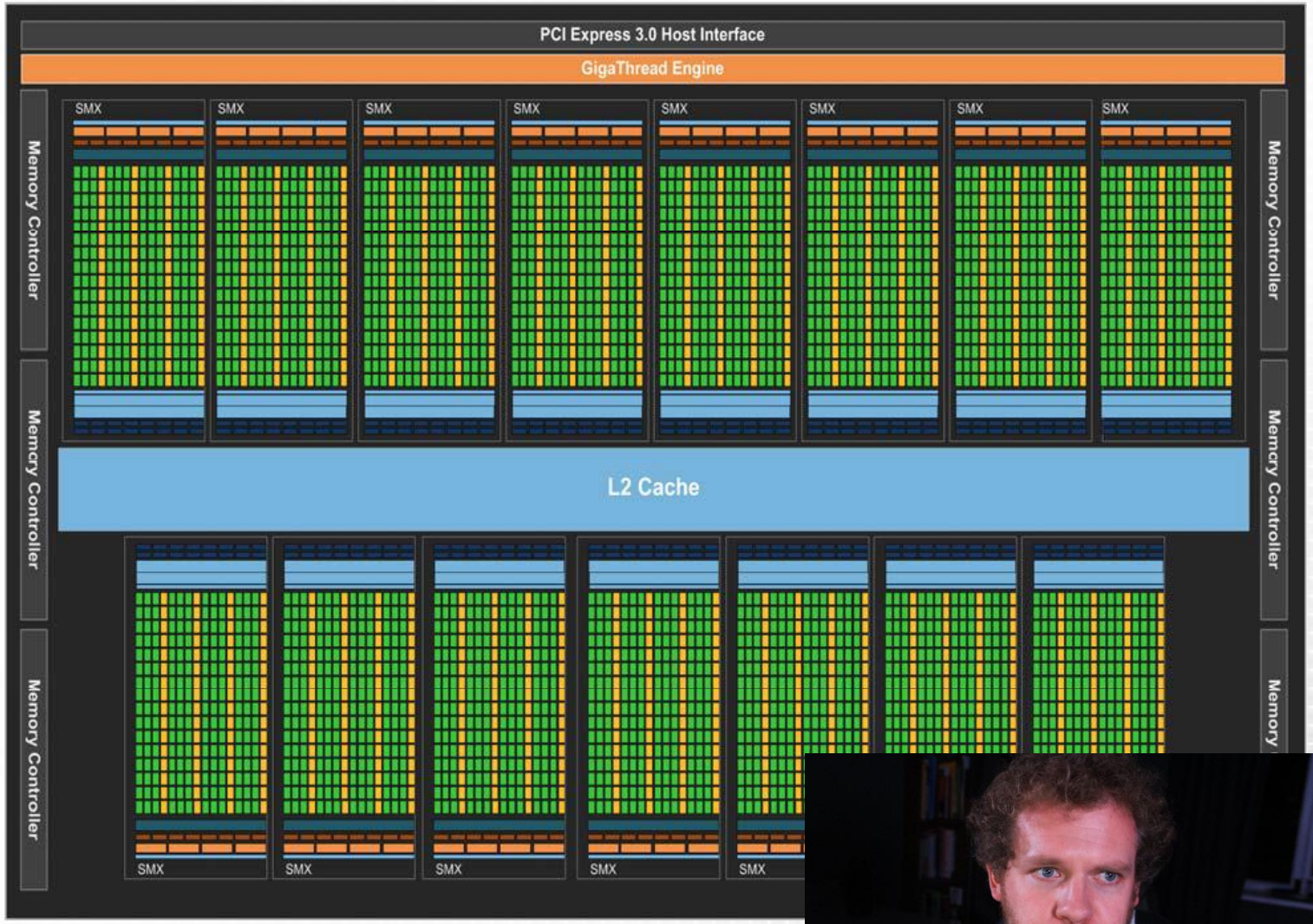
Fermi Family of Tesla GPUs



- Chip partitioned into Streaming Multiprocessors (SMPs)
- 32 vector cores per SMP
- Not cache coherent. No communication possible across SMPs.

Kepler Family of Tesla GPUs

- Streaming Multiprocessor Extreme (SMX)
- Huge increase in the number of cores per SMX
  - Smaller 28nm processes
- Increased L2 Cache
- Cache coherency at L2 not at L1



Maxwell Family Tesla GPUs

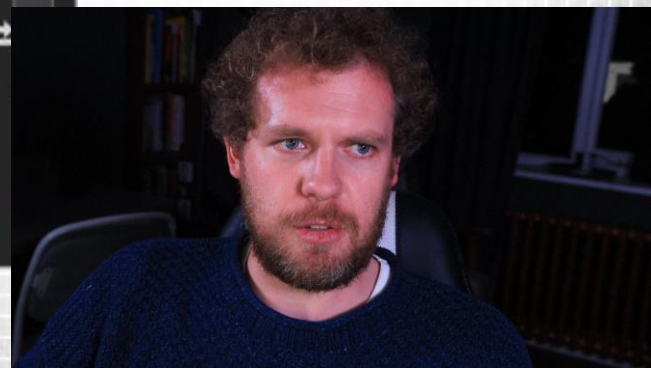
- Streaming Multiprocessor Module (SMM)
- SMM Divided into 4 quadrants (GPC)
  - Each has own instruction buffer, registers and scheduler for each of the 32 vector cores
- SMM has 90% performance of SMX at 2x energy efficiency
  - 128 cores vs. 192 in Kepler
  - BUT small die space = more SMMs
- 8x the L2 cache of Kepler (2MB)
- 20nm transistor size

Pascal P100 GPU

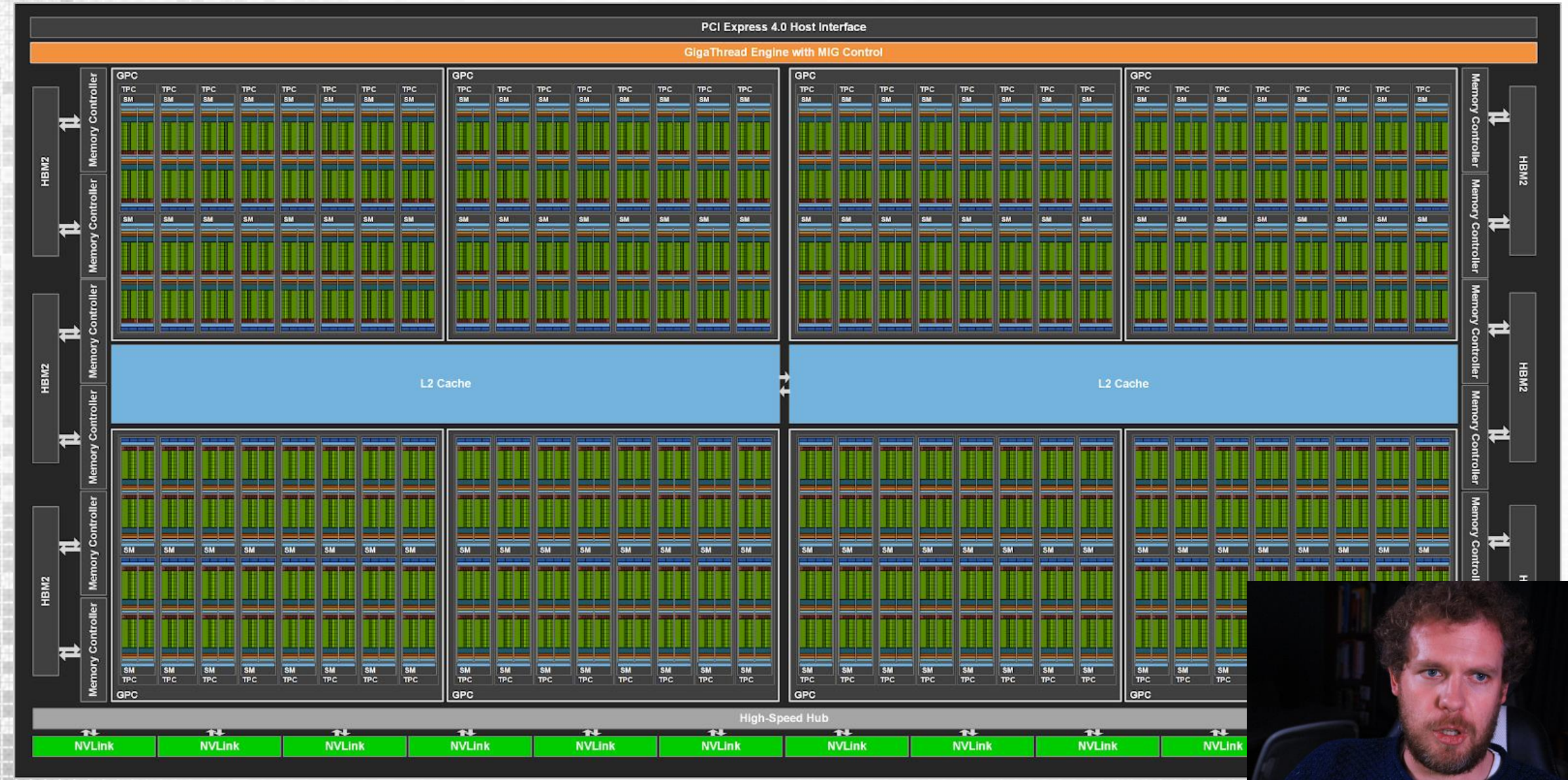
- Many more SMPs
- More GPCs
- Each CUDA core is more efficient
  - More registers available
- Same die size as Maxwell
  - Transistor shrink to 16nm
- Memory bandwidth improved drastically
  - NVLink

Volta V100 GPU

- Massive silicone size
  - Die shrink (12nm)
- CUDA cores and tensor cores



Ampere A100


PCI Express 4.0 Host Interface
GigaThread Engine with MIG Control
GPC
TPC
SM
L2 Cache
L2 Cache
HBM2
Memory Controller
High-Speed Hub
NVLink

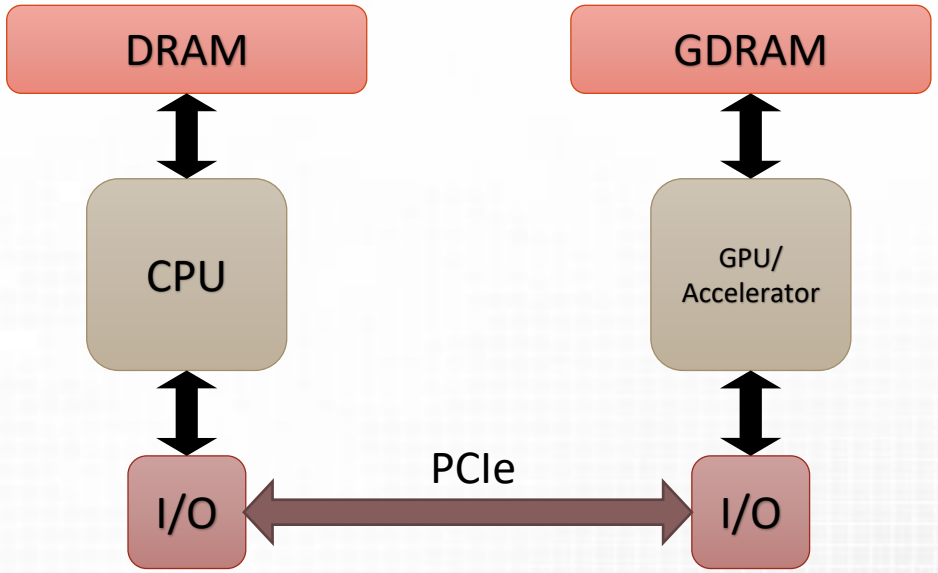
Warp Scheduling

- GPU Threads are always executed in groups called warps (32 threads)
  - **Warps are transparent to users**

- SMPs have zero overhead warp scheduling
  - Warps with instructions ready to execute are eligible for scheduling
  - Eligible warps are selected for execution on priority (context switching)
  - All threads (in a warp) execute the same instruction (SIMD) when executed on the vector processors (CUDA cores)
- The specific way in which warps are scheduled varies across families
  - Fermi, Kepler and Maxwell have different numbers of warp schedulers and dispatchers

Accelerated Systems

- CPUs and Accelerators are used together
  - GPUs cannot be used instead of CPUs
  - GPUs perform compute heavy parts
- Communication is via PCIe bus
  - PCIe 3.0: up to 8 GB per second throughput
  - NVLINK: 5-12x faster than PCIe 3.0

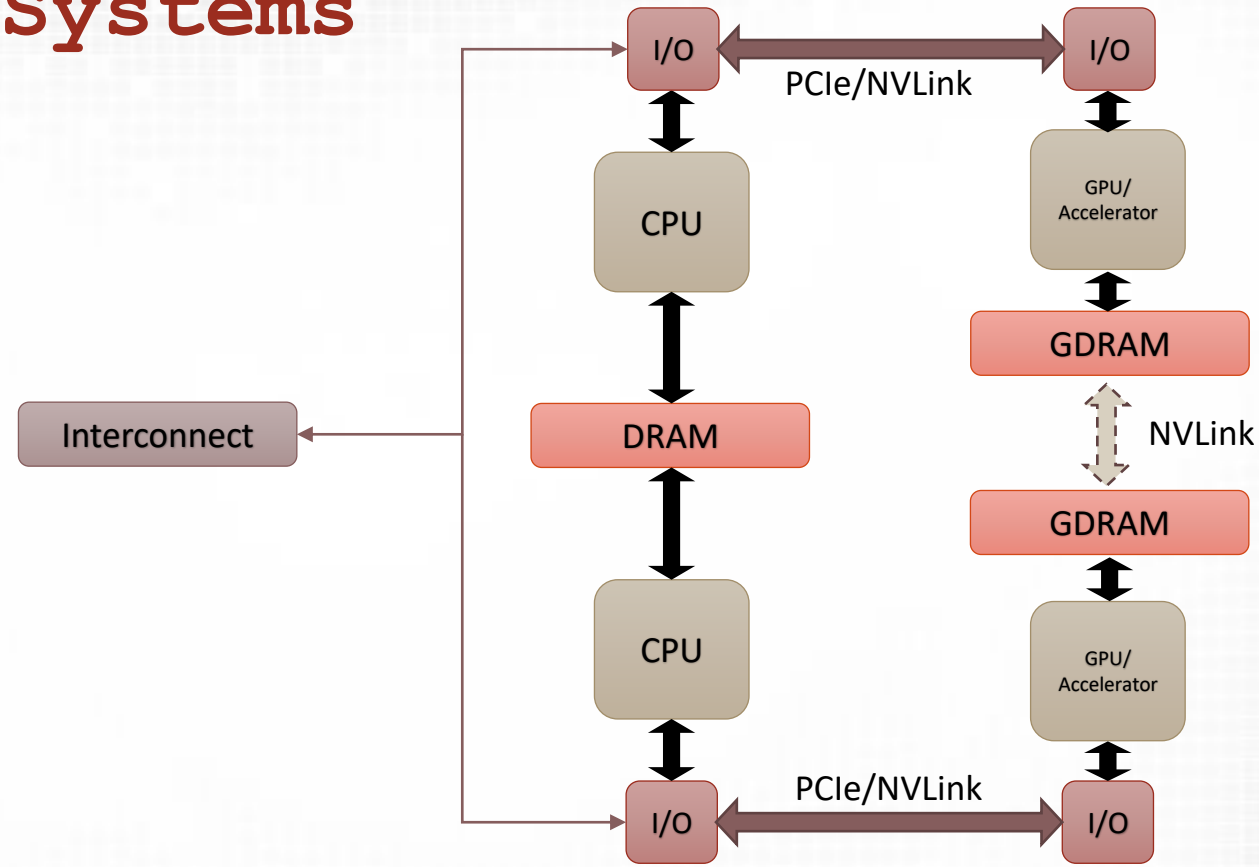


Simple Accelerated Workstation

- Insert your accelerator into PCI-e
- Make sure that
  - There is enough space
  - Your power supply unit (PSU)is up to the job
  - You install the latest GPU drivers

Larger Accelerated Systems

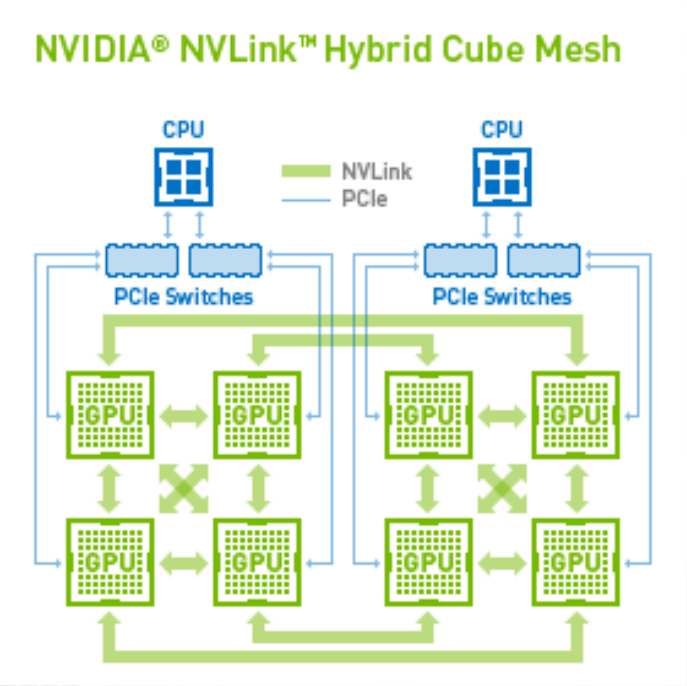
- Can have multiple CPUs and Accelerators within each “Shared Memory Node”
  - CPUs share memory but accelerators do not!

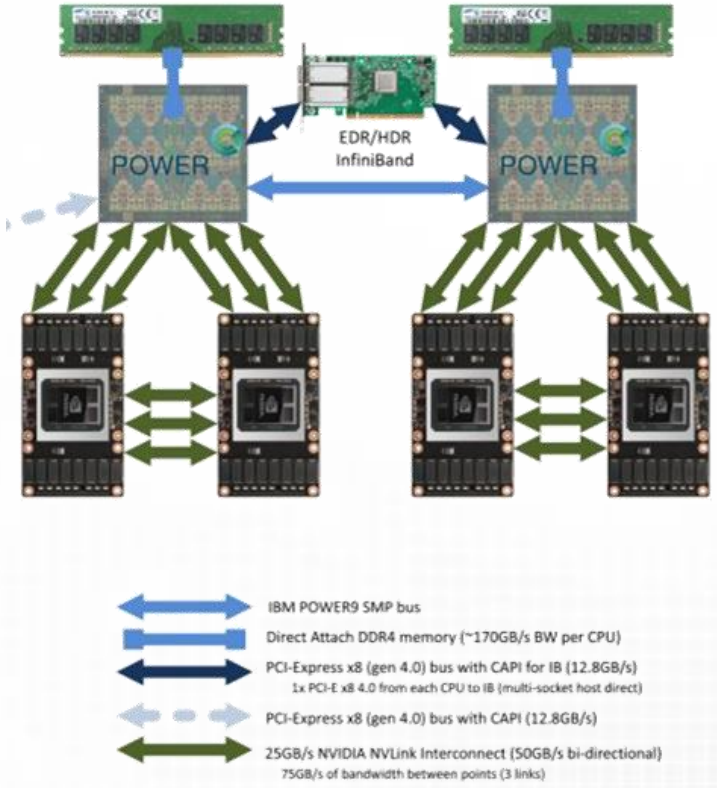


GPU Workstation Server

- Multiple Servers can be connected via interconnect
- Several vendors offer GPU servers
- For example 2 multi core CPUs + 4 GPUS
- Make sure your case and power supply are upto the job!

DGX, Power Systems and Supercomputers





Summary

- NVIDIA GPU Hardware
  - Explain the NVIDIA hardware model and key terminology
  - Compare the hardware variants and identify changing architectural characteristics
  - Give examples of GPU usage at different system scales