Heterogeneous Computing & GPU Introduction

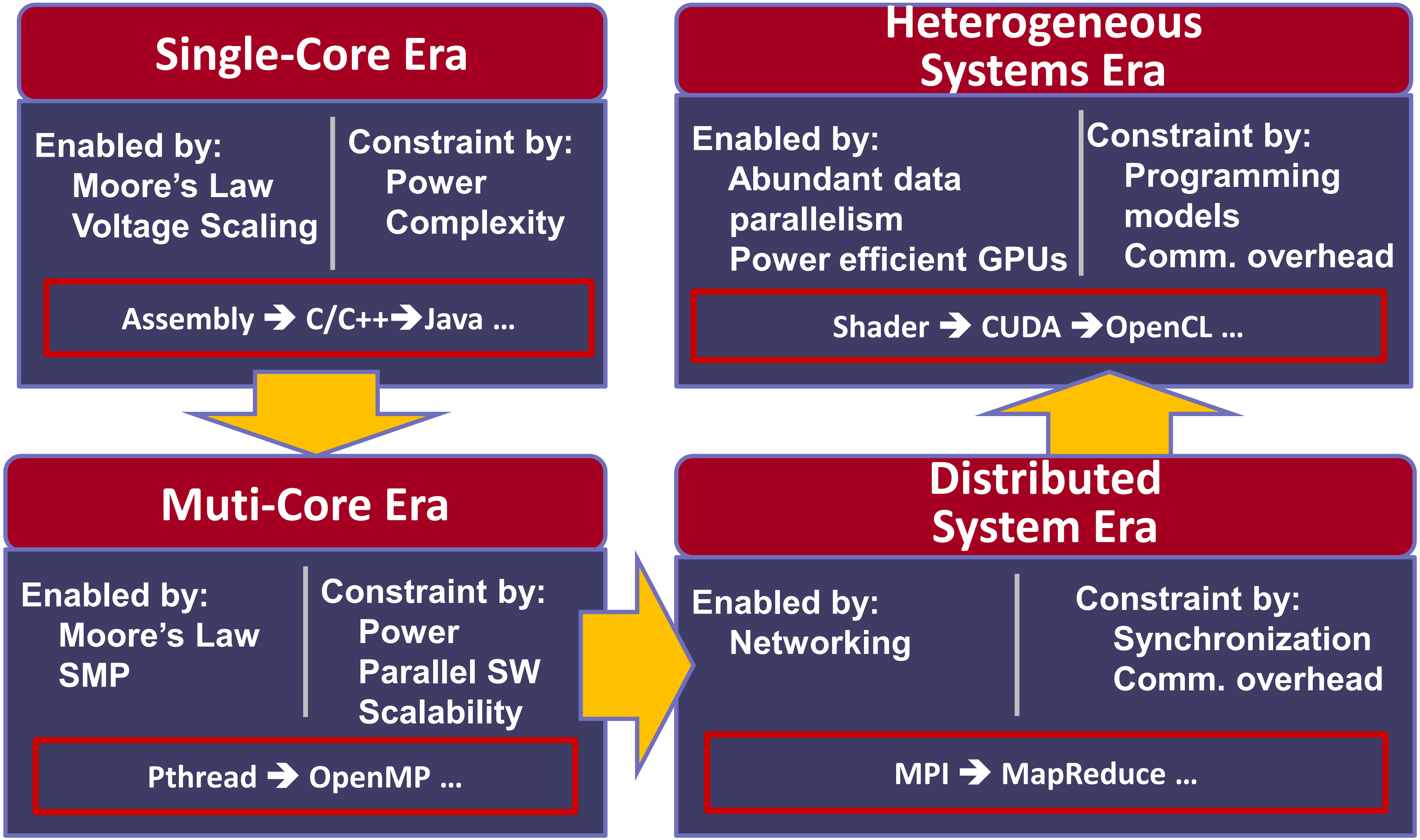
National Tsing Hua University

2025, Fall Semester

Outline

- Heterogeneous Computing
- GPU

Trend of Parallel Computers



Single-Core Era

Enabled by:
Moore's Law
Voltage Scaling

Constraint by:
Power
Complexity

Assembly → C/C++→Java …

Muti-Core Era

Enabled by:
Moore's Law
SMP

Constraint by:
Power
Parallel SW
Scalability

Pthread → OpenMP …

Distributed System Era

Enabled by:
Networking

Constraint by:
Synchronization
Comm. overhead

MPI → MapReduce …

Heterogeneous Systems Era

Enabled by:
Abundant data parallelism
Power efficient GPUs

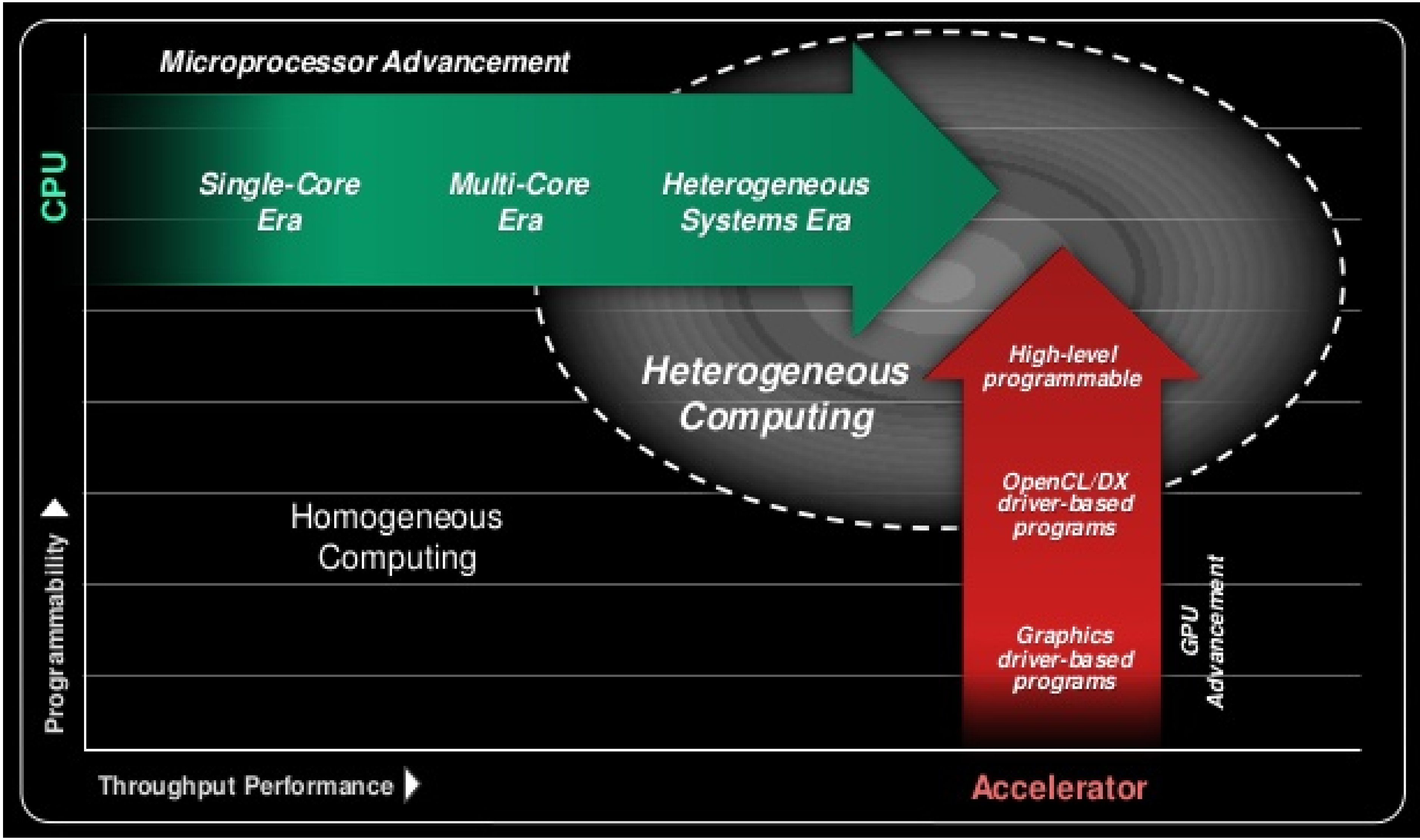
Constraint by:
Programming models
Comm. overhead

Shader → CUDA →OpenCL …

Heterogeneous Computing

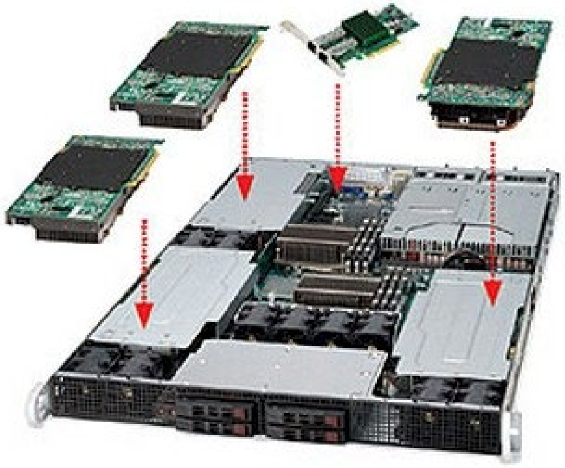
- Heterogeneous computing is an integrated system that consists of different types of (programmable) computing units.
  - DSP (digital signal processor)
  - FPGA (field-programmable gate array)
  - ASIC (application-specific integrated circuit)
  - GPU (graphics processing unit)
  - Co-processor (Intel Xeon Phi)
- A system can be a cell phone or a supercomputer

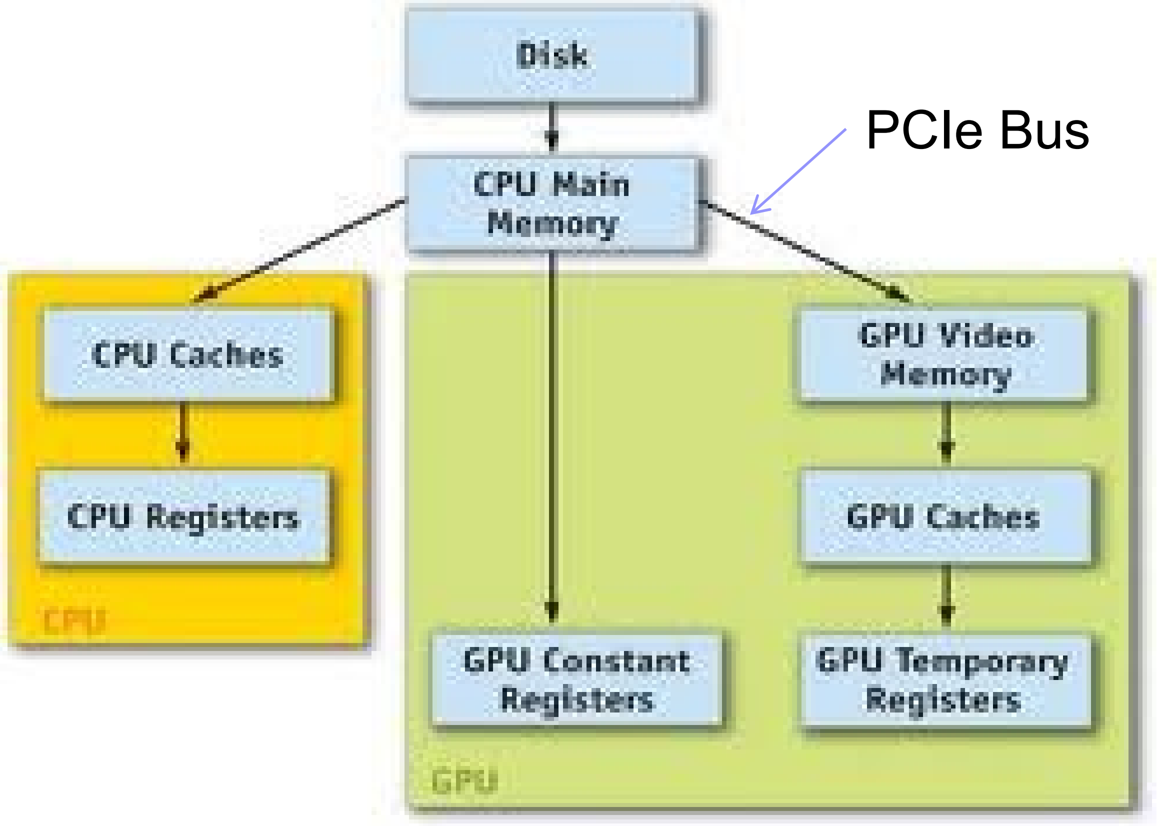
Shift of Computing Paradigm



GPU Servers

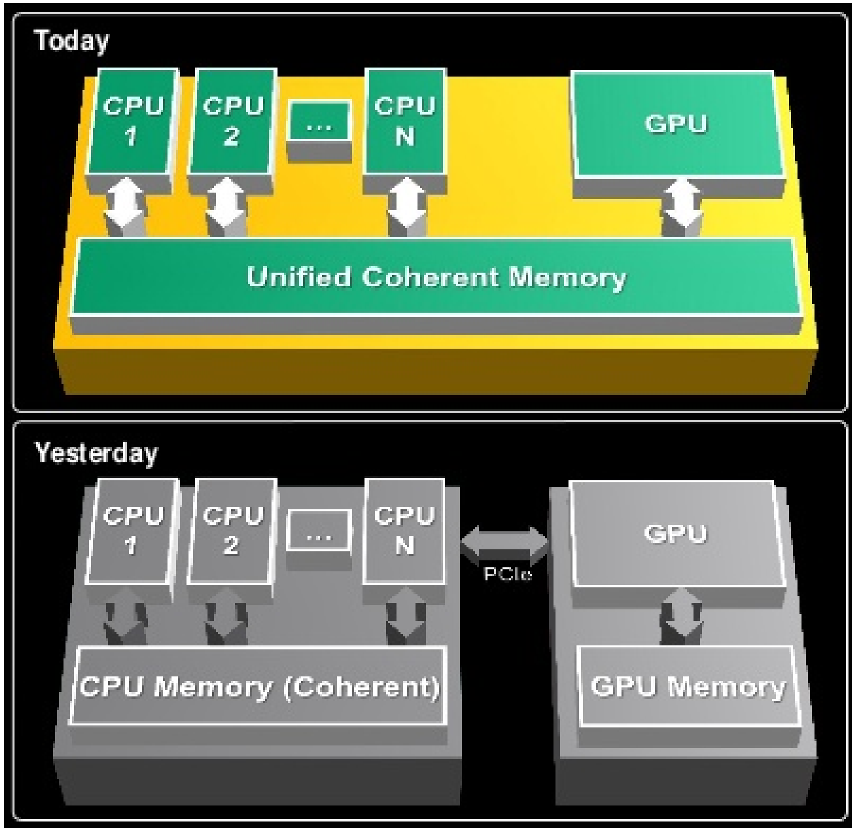
- Same HW architecture as commodity server, but memory copy between CPU and GPU becomes the main bottleneck





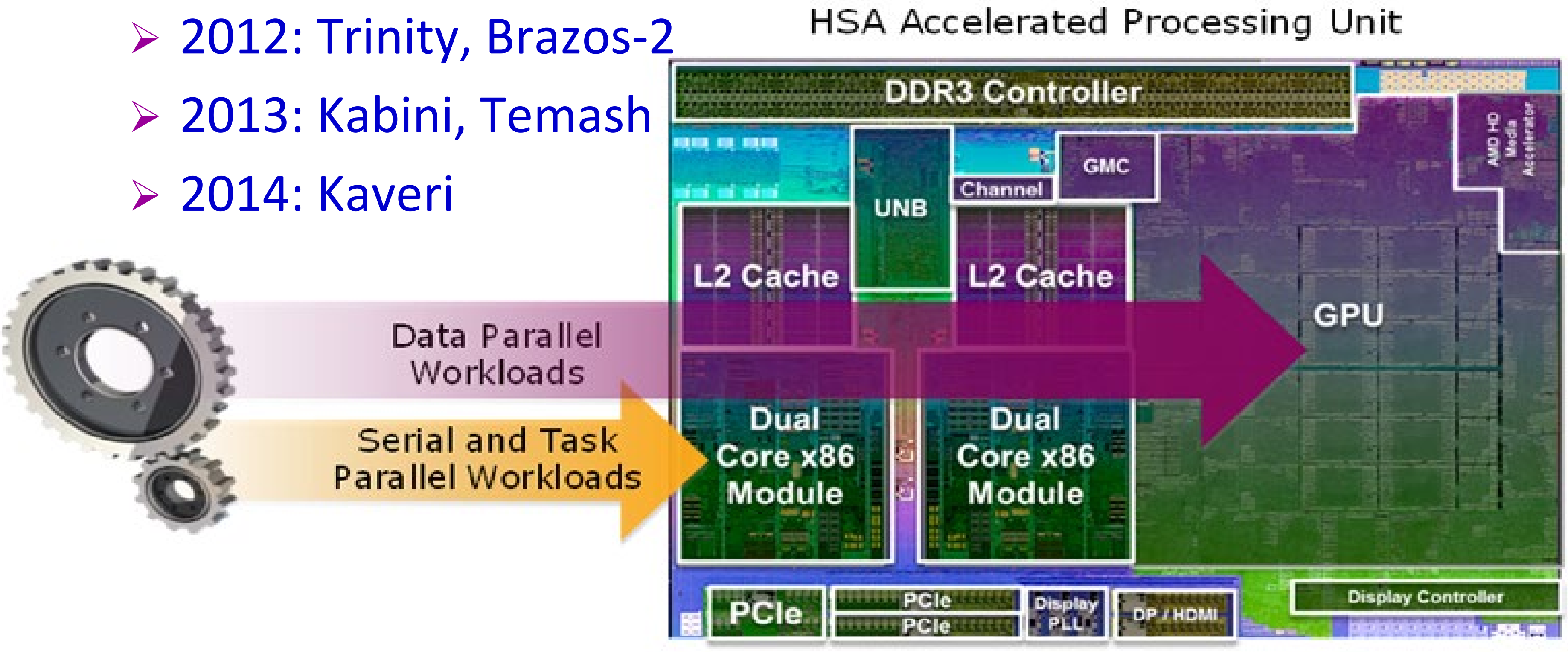
Heterogeneous System Architecture (HSA)

- Aim to provide a common system architecture for designing higher-level programming models for all devices
- Unified coherent memory
  - Single virtual memory address space
  - Prevent memory copy



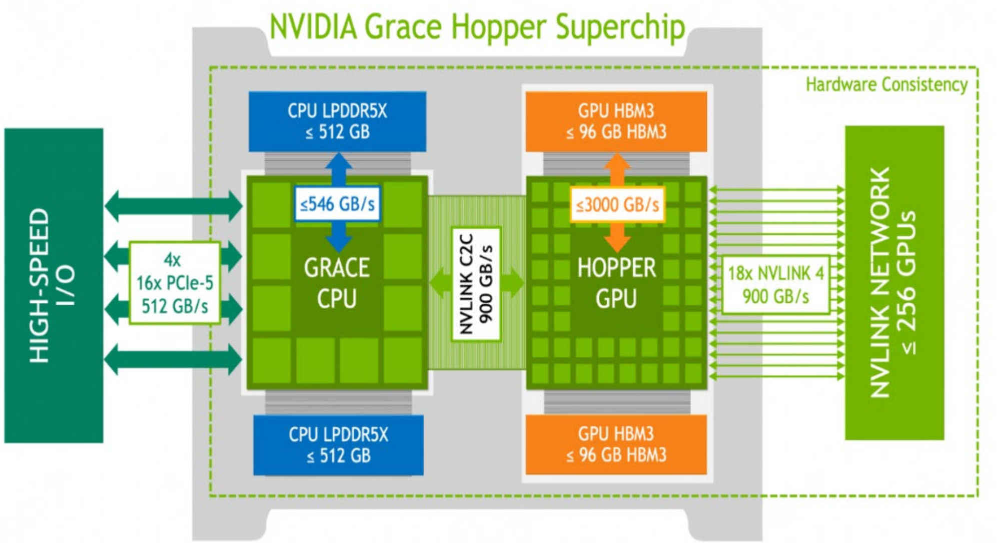
AMD Accelerated Processing Unit (APU)

- A.k.a ***Fusion:*** a series of 64-bit microprocessors from AMD designed to act as a CPU and GPU on a single chip
  - 2011: Llano, Brazos
  - 2012: Trinity, Brazos-2
  - 2013: Kabini, Temash
  - 2014: Kaveri



NVIDIA Grace Hopper Superchip

- Massive Bandwidth for Compute Efficiency
- Using NVIDIA® NVLink®-C2C to deliver a CPU+GPU coherent memory model for accelerated AI and HPC applications.
- High-speed I/O
- HBM3 memory
- On-chip fabrics
- System-on-chip (SoC)
- ARM-based processors



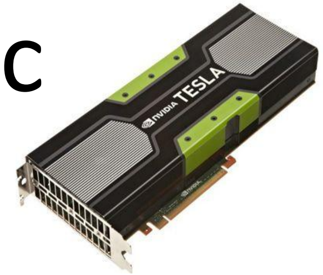
Outline

- Heterogeneous Computing
- GPU

GPU (Graphic Processing Unit)

- A specialized chip designed for rapidly display and visualization
  - SIMD architecture
- Massively multithreaded manycore chips
  - NVIDIA Tesla products have up to 5120 scalar processors
  - Over 12,000 **concurrent** threads
  - Over 470 GFOLPS sustained performance
- Two major vendors: NVIDIA and ATI (now AMD)

GPGPU (General-Purpose Graphic Processing Unit)



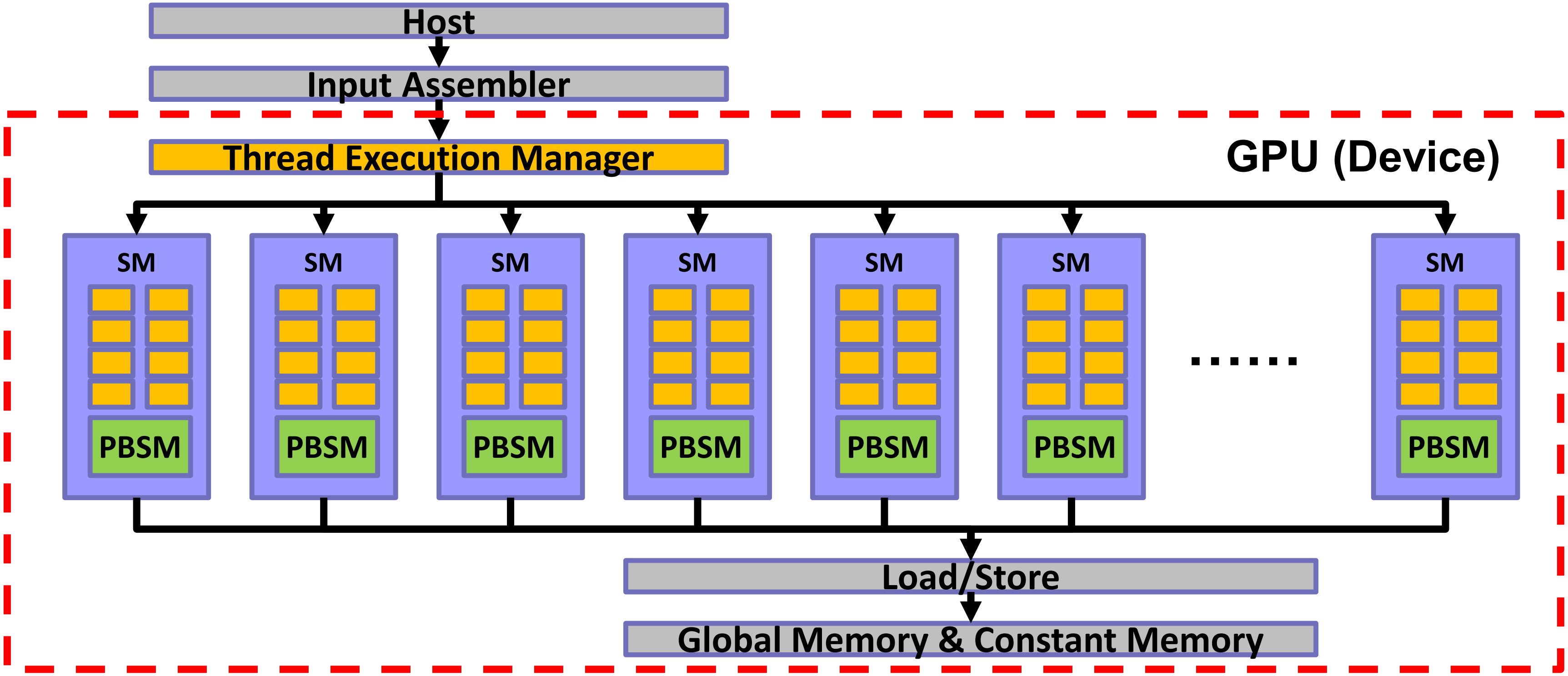
- **Expose** the horse power of GPUs for general purpose computations
  - Exploit **data parallelism** for solving embarrassingly parallel tasks and numeric computations
  - Users across science & engineering disciplines are achieving 100x or better speedups on GPUs
- Programmable
  - Early GPGPU: using the libraries in computer graphics, such as OpenGL or DirectX, to perform the tasks other than the original hardware designed for.
  - Now **CUDA** and openCL provides an extension to C and C++ that enables parallel programming on GPUs

Manycore GPU – Block Diagram

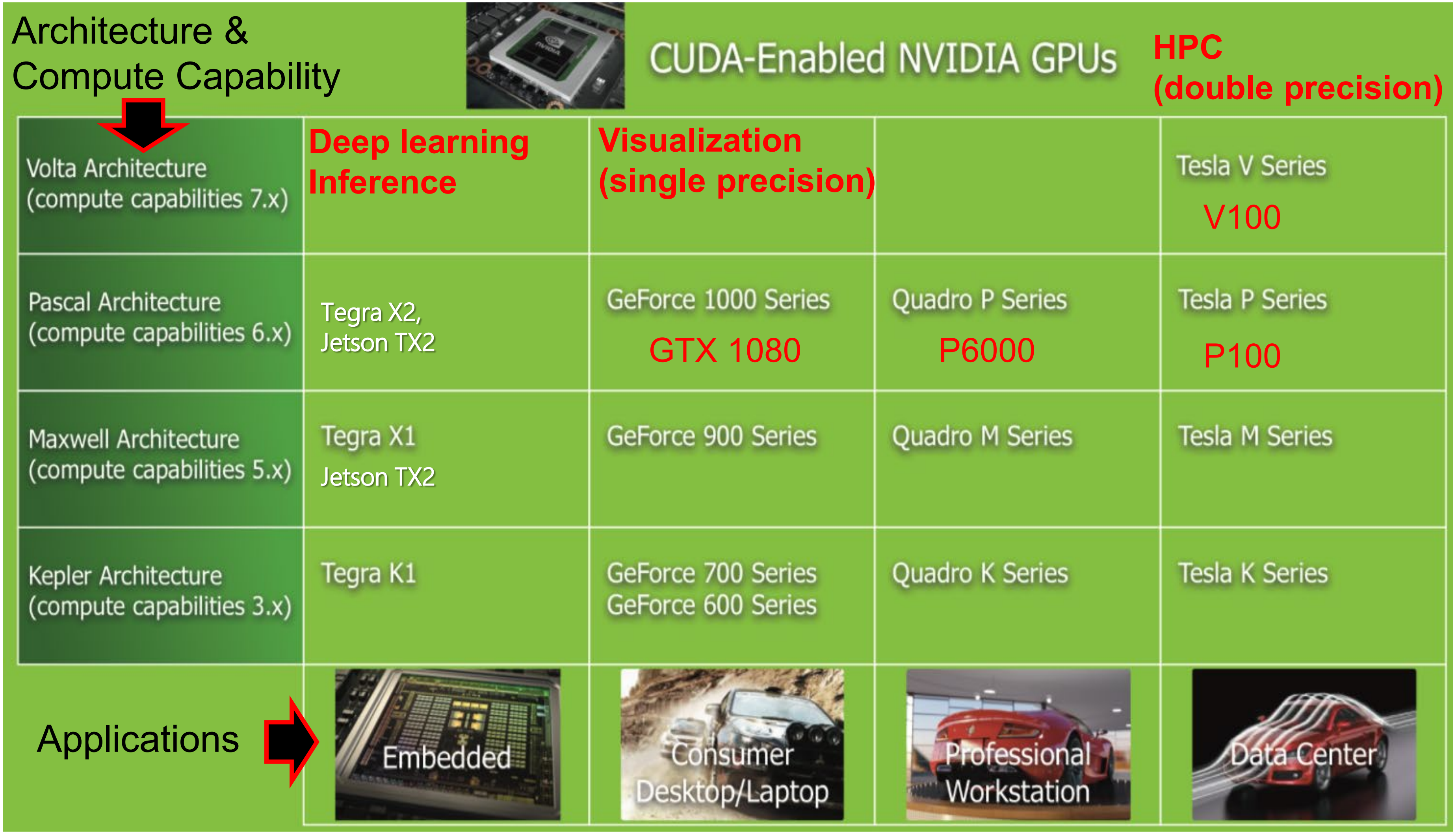
- Consist of multiple stream multi-processors (SM)
- Memory hierarchic:
  - global memory ➔ PBSM/shared memory ➔ local register

Slow, but large & shared

Fast, but small & local



NVIDIA CUDA-Enabled GPUs Products



Architecture & Compute Capability

CUDA-Enabled NVIDIA GPUs

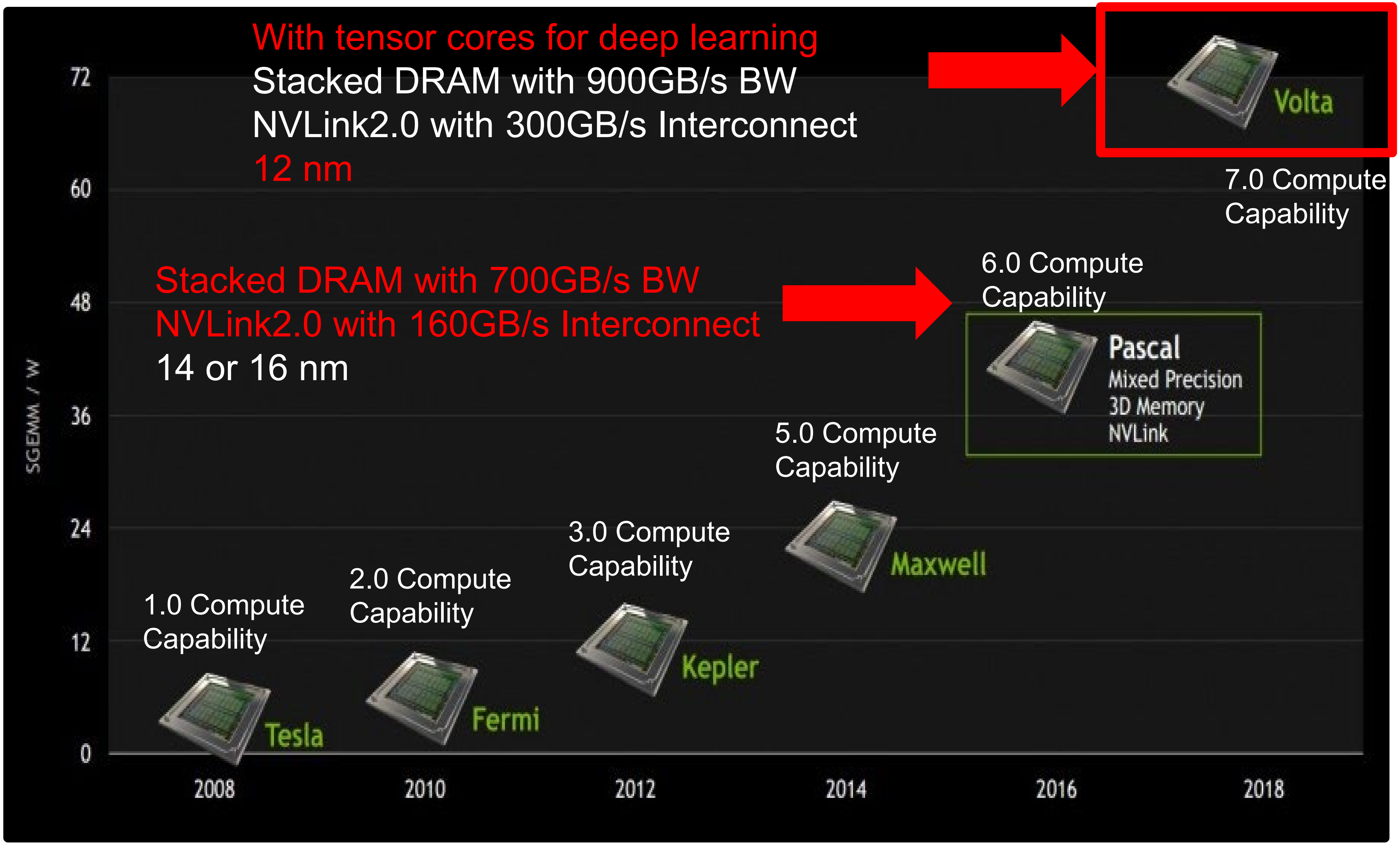
HPC (double precision)

Volta Architecture (compute capabilities 7.x)	**Deep learning Inference**	**Visualization (single precision)**		Tesla V Series V100
Pascal Architecture (compute capabilities 6.x)	Tegra X2, Jetson TX2	GeForce 1000 Series GTX 1080	Quadro P Series P6000	Tesla P Series P100
Maxwell Architecture (compute capabilities 5.x)	Tegra X1 Jetson TX2	GeForce 900 Series	Quadro M Series	Tesla M Series
Kepler Architecture (compute capabilities 3.x)	Tegra K1	GeForce 700 Series GeForce 600 Series	Quadro K Series	Tesla K Series
Applications	Embedded	Consumer Desktop/Laptop	Professional Workstation	Data Center

NVIDIA GPU HW Specification

		Tesla K40	Tesla P100	Tesla V100	GeForce GTX1080
Launch Date		2013 Oct	2016 Jun	2017 Jun	2016 May
Architecture		Kepler	Pascal	Volta	Pascal
CUDA Cores		2888	3584	5120	2560
Core Clock		745MHz	1126MHz	1370MHz	1607MHz
GPU Memory Bandwidth		288GB/s	732GB/s	900GB/s	320GB/s
GPU Memory Size		12GB	16GB	16GB	8GB
Interconnect Bandwidth	PCIe3x16	32GB/s	32GB/s	32GB/s	32GB/s
	NV Link	----	160GB/s	300GB/s	---
Single Precision		4.29 TFLOPS	9.3 TFLOPS	14 TFLOPS	8.8 TFLOPS
Double Precision		1.43 TFLOPS	4.7 TFLOPS	7.0 TFLOPS	0.2 TFLOPS
TDP		235W	250W	250W	180W
Compute Capability		3.5	6.0	7.0	6.1
Launch Price (USD)		$5499	$7374/$9428(NV)	8GPU: 150K	$550

NVIDIA GPU Architecture Roadmap



With tensor cores for deep learning
Stacked DRAM with 900GB/s BW
NVLink2.0 with 300GB/s Interconnect
12 nm

Stacked DRAM with 700GB/s BW
NVLink2.0 with 160GB/s Interconnect
14 or 16 nm

7.0 Compute Capability

6.0 Compute Capability

5.0 Compute Capability

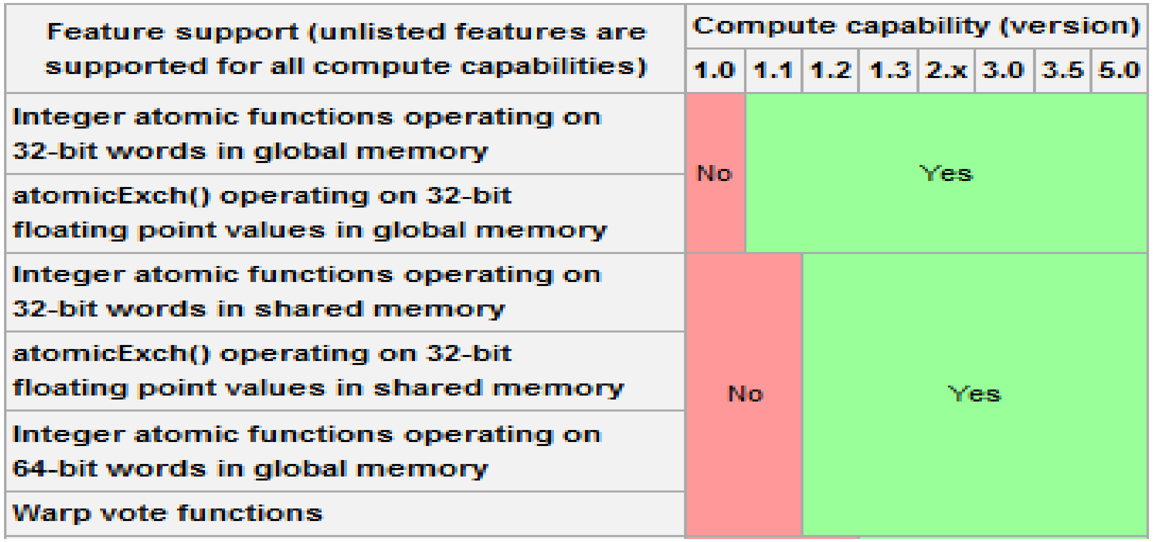
3.0 Compute Capability

2.0 Compute Capability

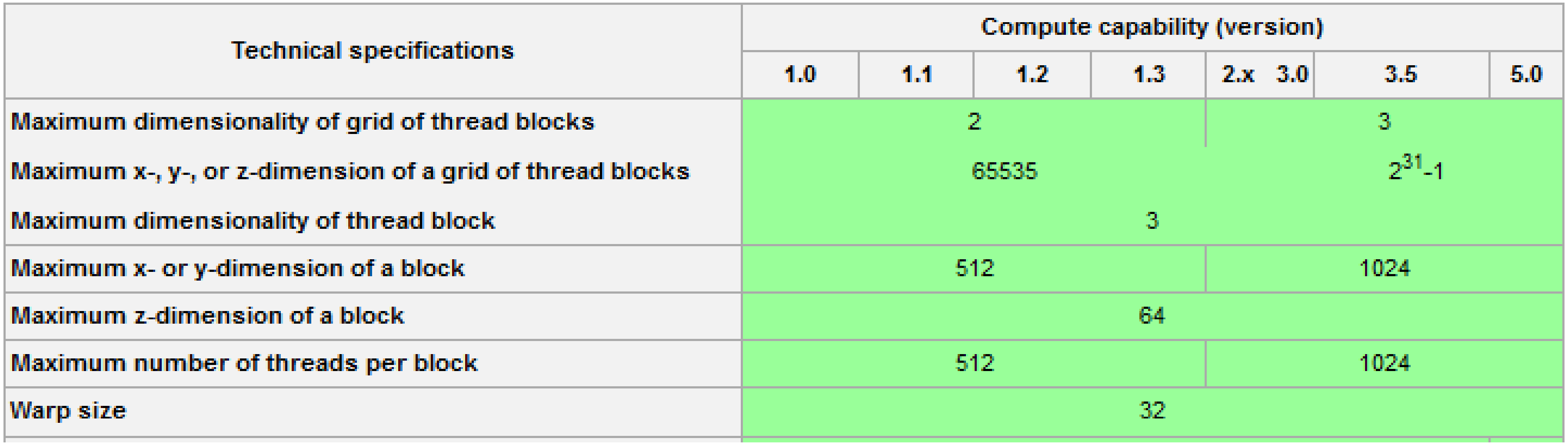
1.0 Compute Capability

GPU Compute Capability

- Programming ability of a GPU device



Feature support (unlisted features are supported for all compute capabilities)	Compute capability (version)							
	1.0	1.1	1.2	1.3	2.x	3.0	3.5	5.0
Integer atomic functions operating on 32-bit words in global memory	No	Yes						
atomicExch() operating on 32-bit floating point values in global memory								
Integer atomic functions operating on 32-bit words in shared memory	No		Yes					
atomicExch() operating on 32-bit floating point values in shared memory								
Integer atomic functions operating on 64-bit words in global memory								
Warp vote functions								

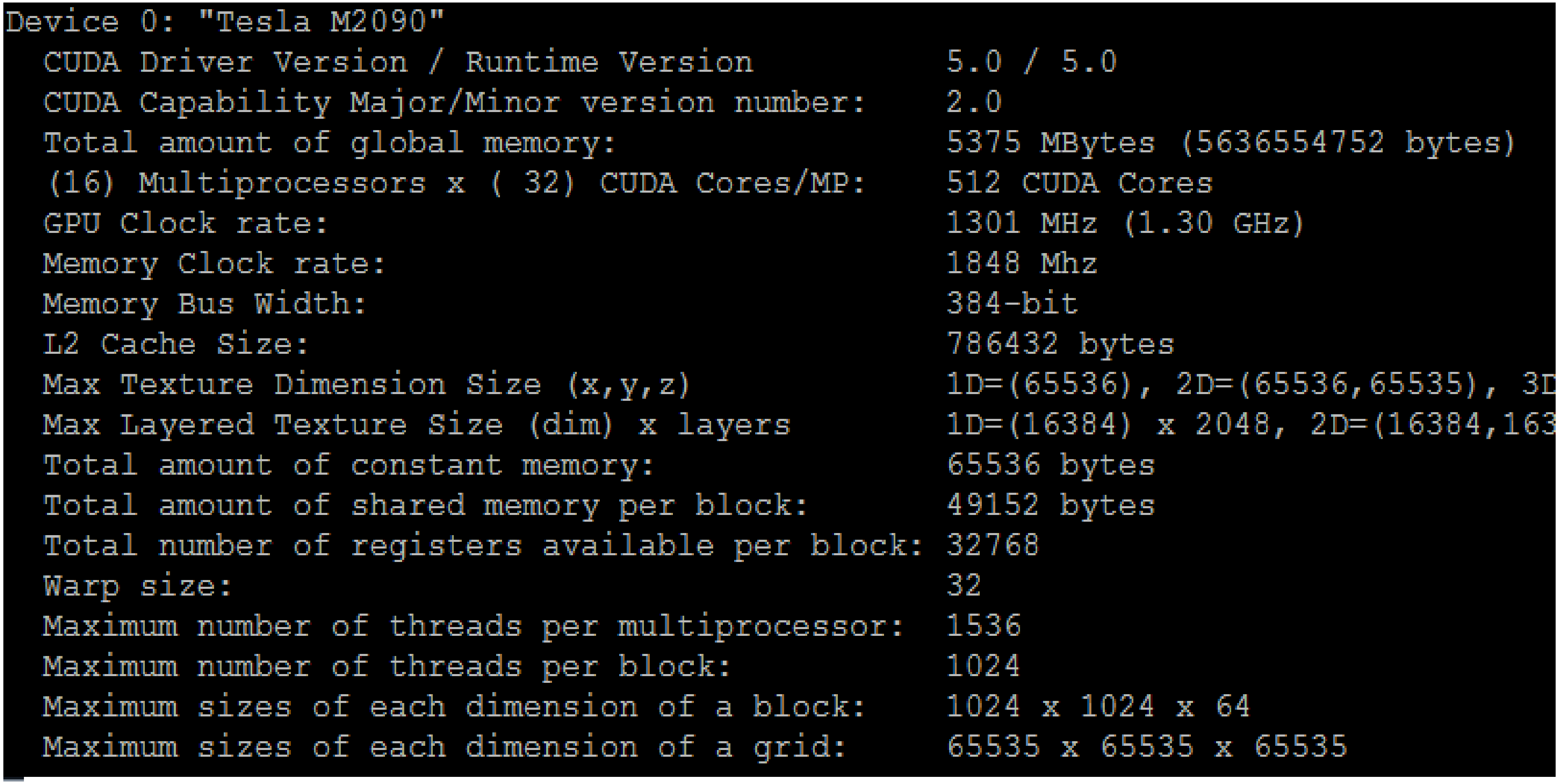


Technical specifications	Compute capability (version)							
	1.0	1.1	1.2	1.3	2.x	3.0	3.5	5.0
Maximum dimensionality of grid of thread blocks	2				3			
Maximum x-, y-, or z-dimension of a grid of thread blocks	65535					2^{31}-1		
Maximum dimensionality of thread block	3							
Maximum x- or y-dimension of a block	512				1024			
Maximum z-dimension of a block	64							
Maximum number of threads per block	512				1024			
Warp size	32							

source: http://en.wikipedia.org/wiki/CUDA

CUDA SDK Device Query

- deviceQuery.cpp



```
Device 0: "Tesla M2090"
  CUDA Driver Version / Runtime Version          5.0 / 5.0
  CUDA Capability Major/Minor version number:    2.0
  Total amount of global memory:                 5375 MBytes (5636554752 bytes)
  (16) Multiprocessors x ( 32) CUDA Cores/MP:    512 CUDA Cores
  GPU Clock rate:                                1301 MHz (1.30 GHz)
  Memory Clock rate:                             1848 Mhz
  Memory Bus Width:                              384-bit
  L2 Cache Size:                                 786432 bytes
  Max Texture Dimension Size (x,y,z)             1D=(65536), 2D=(65536,65535), 3D
  Max Layered Texture Size (dim) x layers        1D=(16384) x 2048, 2D=(16384,163
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total number of registers available per block: 32768
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  1536
  Maximum number of threads per block:           1024
  Maximum sizes of each dimension of a block:    1024 x 1024 x 64
  Maximum sizes of each dimension of a grid:     65535 x 65535 x 65535
```

CUDA Toolkits

- Software Development Kit(SDK) for CUDA Programming
  - The CUDA-C and CUDA-C++ compiler, nvcc
  - Tools: IDE, Debugger, Profilers, Utilities
  - Library: BLAS, CUDA Device Runtime, FFT, …
  - Sample Code
  - Documentation

CUDA SDK Version	Compute Capability	Architecture
6.5	1.X	Tesla
7.5	2.0-5.x	Fermi, Kepler, Maxwell
8.0	2.0-6.x	Fermi, Kepler, Maxwell, Pascal
9.0	3.0-7.x	Kepler, Maxwell, Pascal, Volta

http://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html#major-components

Reference

- Cyril Zeller, NVIDIA Developer Technology slides