

IN5050 – GPU & CUDA

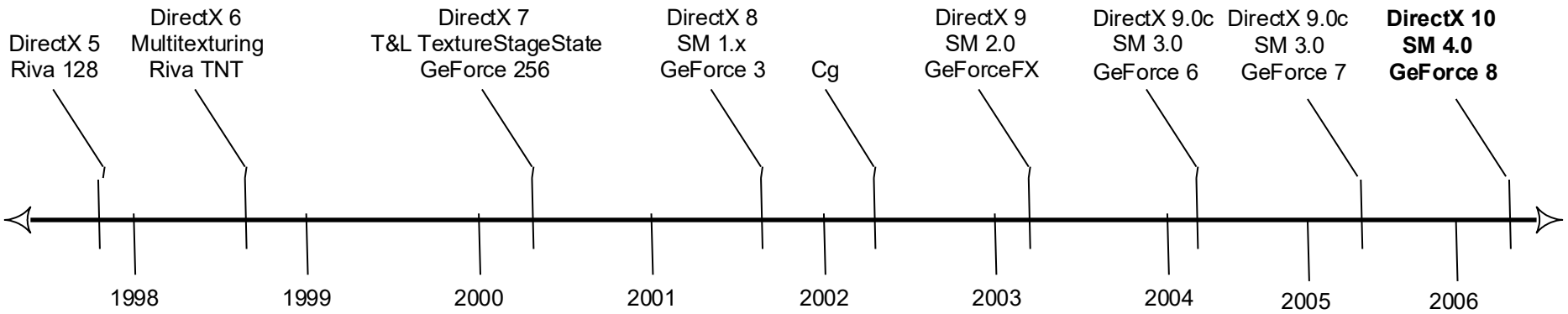
Håkon Kvale Stensland

Simula Research Laboratory / Department for Informatics

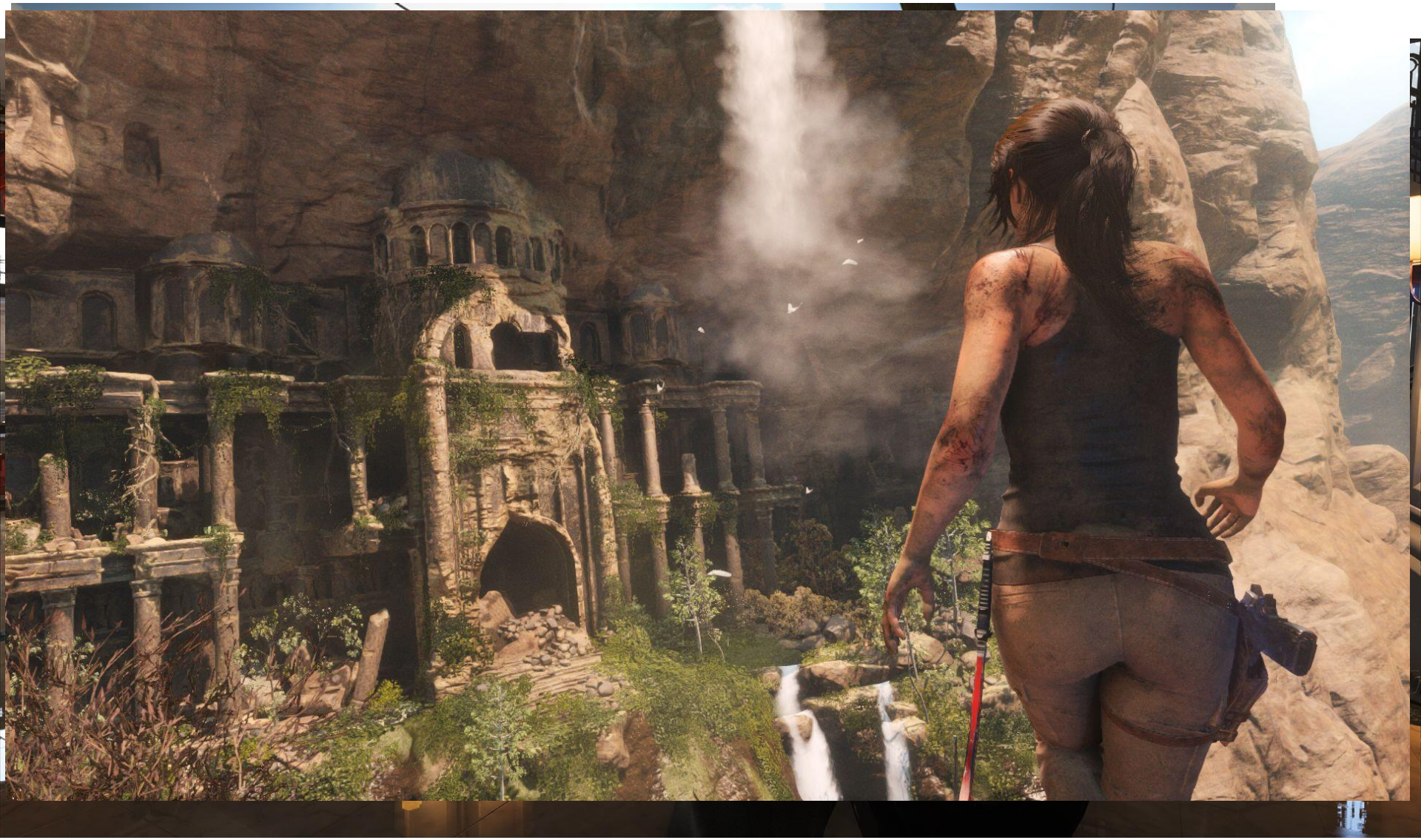
PC Graphics Timeline

- Challenges:
  - Render infinitely complex scenes
  - And extremely high resolution
  - In 1/60th of one second (60 frames per second)

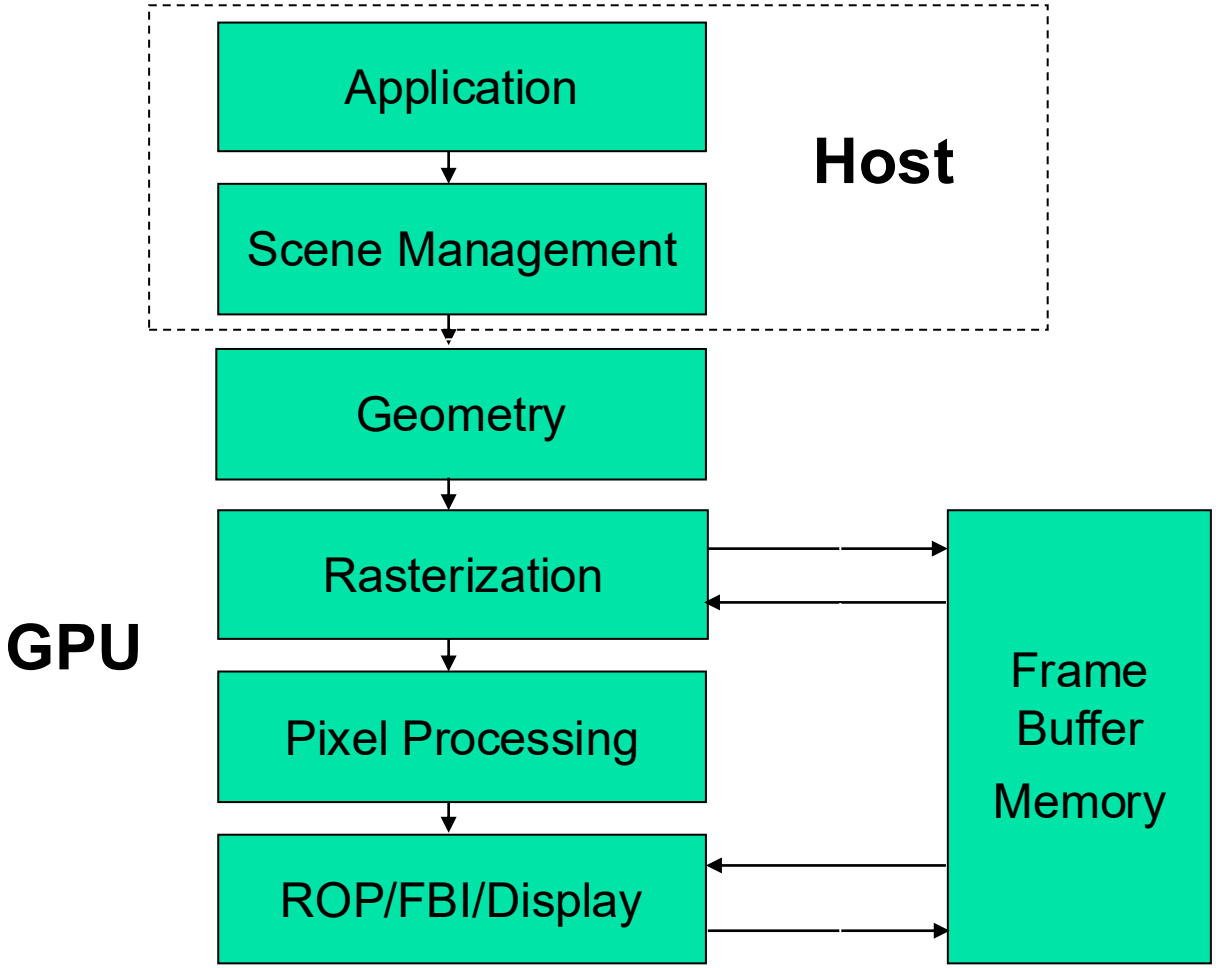
- Graphics hardware has evolved from a simple hardwired pipeline to a highly programmable multiword processor



GPU – Graphics Processing Units

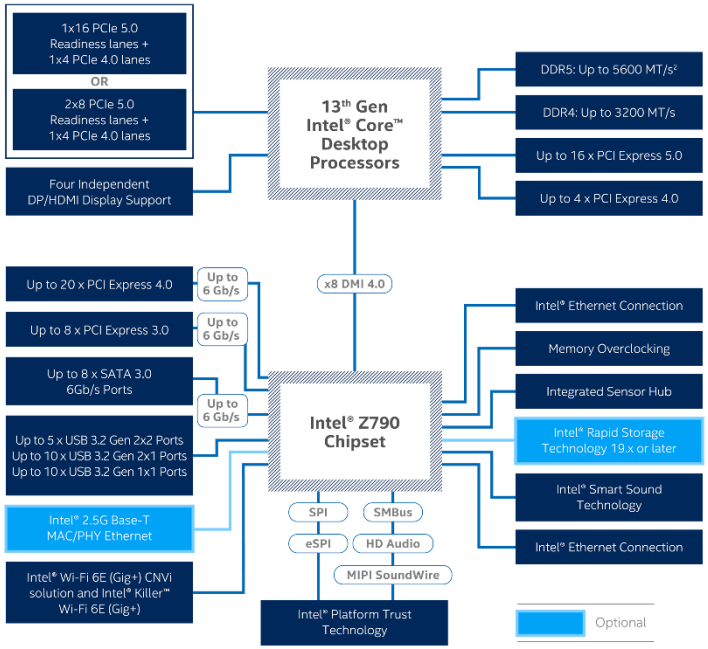


Basic 3D Graphics Pipeline

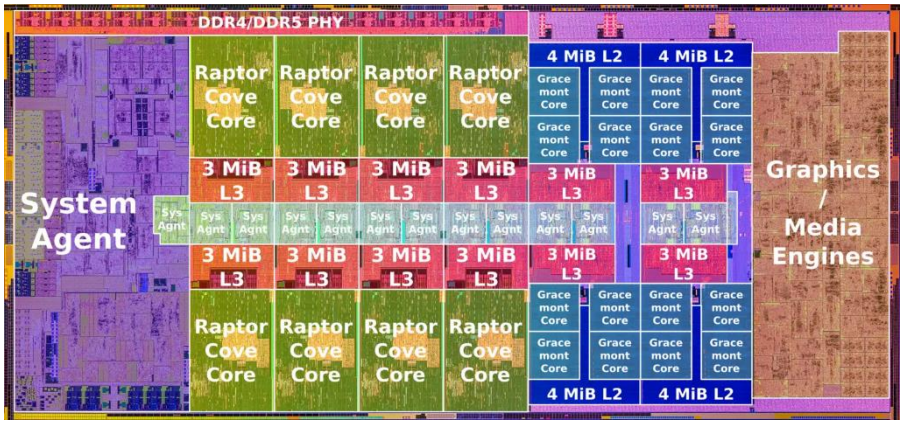


Graphics in the PC Architecture

- PCIe (PCI Express) between processor and chipset
  - Memory Control now integrated in CPU
- The old "Northbridge" integrated onto CPU
  - **PCI Express 5.0 x16 bandwidth at 128 GB/s (64 GB in each direction)**
- "Southbridge" (Z790) handles all other peripherals

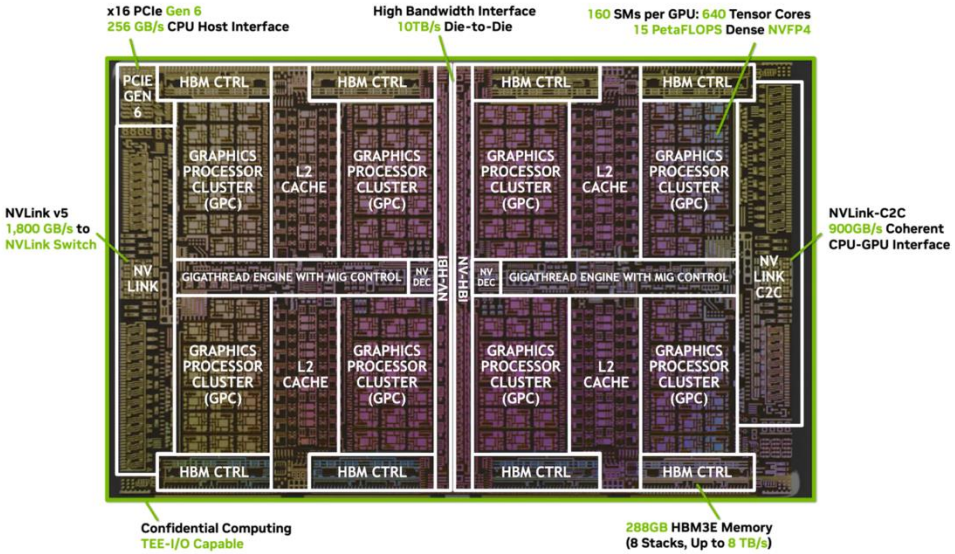


- **Most mainstream CPUs now come with integrated GPU**
  - Same capabilities as discrete GPU's
  - Less performance (limited by die space and power)



Intel «Raptor Lake» (13th Gen) CPU

High-end «Graphics» Hardware



NVIDIA Blackwell GB300

- NVIDIA Blackwell Architecture
- The latest generation GPU, codenamed GB300

- 104 (x2) **billion** transistors
- 20480 Processing cores (SP)
  - 160 SMs per full GPU
  - Mixed precision
  - 5th Gen Tensor cores
  - 5th Gen NVLink interconnect
  - PCI Express 6.0
  - 65 MB L2 Cache
  - 288 GB HBM3e memory
  - Supports SR-IOV
  - Compute version 10.0

GPUs not always for Graphics

RTX 5090 (GB102)

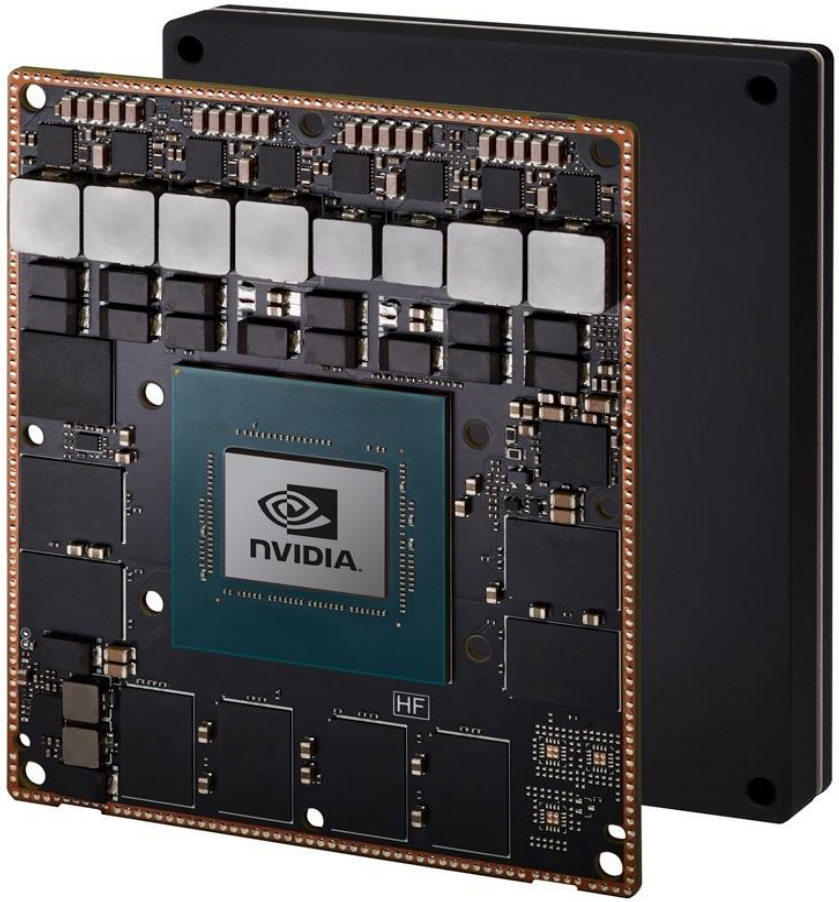


B200 Accelerator



- GPUs are now common in HPC
- Largest supercomputer in November 2025 is the **El Capitan** at Lawrence Livermore National Laboratory
  - 11136 nodes
    - Four AMD MI300A APU's
  - Theoretical: ***1,809 exaflops***
- *Before:* Dedicated compute card released after graphics model
- *Now:* Nvidia's Blackwell architecture (B100/B200) released in March 2024. Graphics architecture (GB102, GB103, GB107) released January 2025.

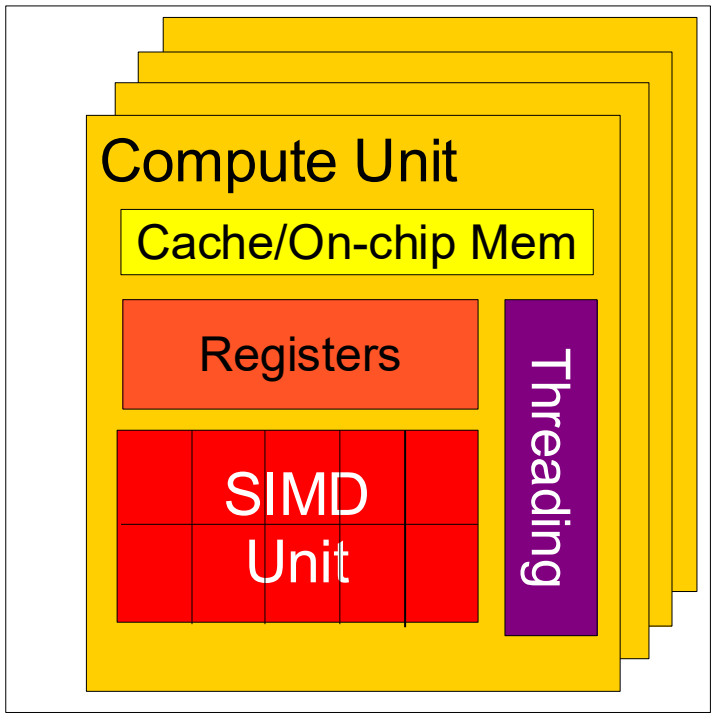
Lab Hardware (Home Exam 1 & 2)



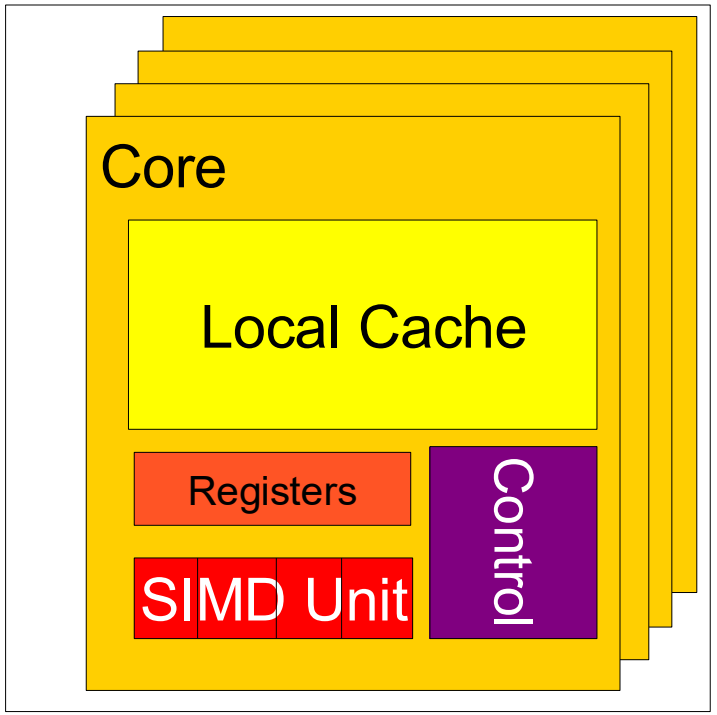
- **nVIDIA Jetson AGX Xavier**
  - **Volta GPU Architecture**
  - **Codename of GPU is GV10B**
- *No desktop or mobile counterpart, similarities with a shrunken TU117*
  - 512 Processing cores (8 Volta SM)
  - 64 Tensor Cores
  - 16/32 GB Memory with 137 GB/sec bandwidth (LPDDR4X)
  - 512 kB Level 2 cache
  - 1,4 TFLOPS theoretical FP32 performance.
  - 2,8 TFLOPS theoretical FP16 performance.
  - Compute version 7.2

CPU and GPU Design Philosophy

GPU
Throughput Oriented Cores

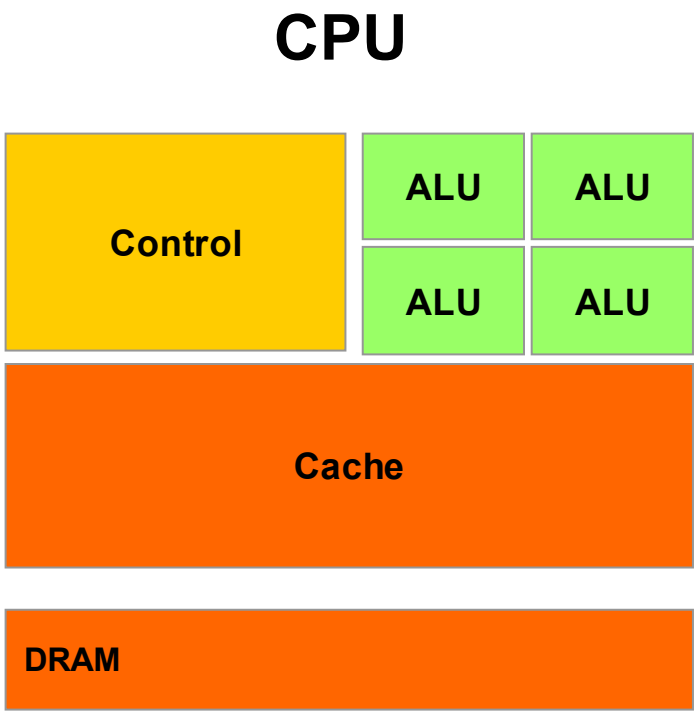


CPU
Latency Oriented Cores



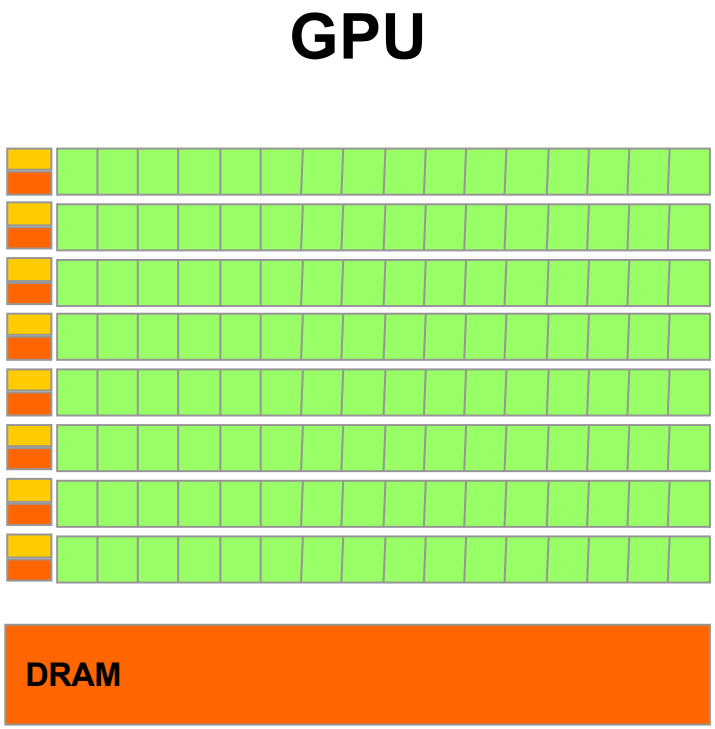
CPUs: Latency Oriented Design

- Large caches
  - Convert long latency memory accesses to short latency cache accesses
- Sophisticated control
  - Branch prediction for reduced branch latency
  - Data forwarding for reduced data latency
- Powerful ALU
  - Reduced operation latency



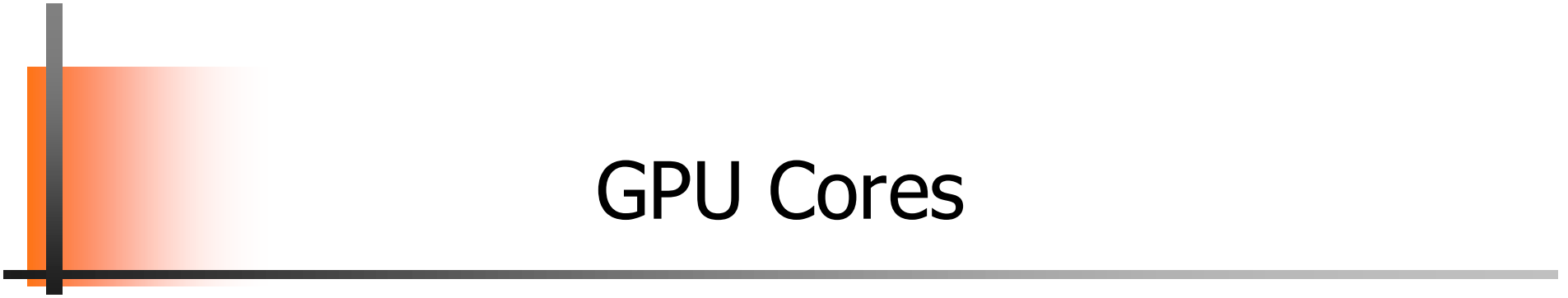
GPUs: Throughput Oriented Design

- Small caches
  - To boost memory throughput
- Simple control
  - No branch prediction
  - No data forwarding
- Energy efficient ALUs
  - Many, long latency but heavily pipelined for high throughput
- Require massive number of threads to tolerate latencies



Think both about CPU and GPU…

- CPUs for sequential parts where latency matters
  - CPUs can be 10+X faster than GPUs for sequential code
- GPUs for parallel parts where throughput wins
  - GPUs can be 10+X faster than CPUs for parallel code

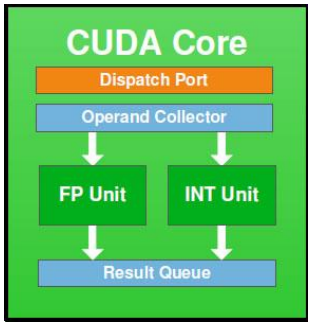


GPU Cores

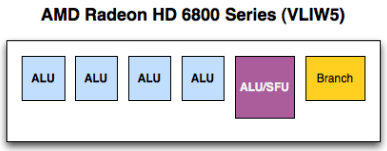
The PR departments favorite metric!

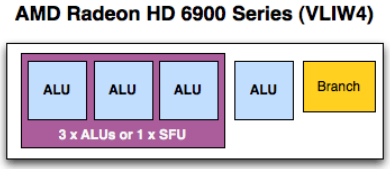
The Core: The basic processing block

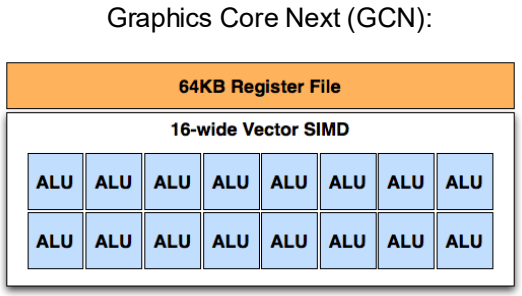
- *The nVIDIA Approach:*
  - Called Stream Processor and later CUDA core. Works on a single operation.



- *The AMD Approach:*
  - *VLIW5:* The GPU work on up to five operations
  - VLIW4: The GPU work on up to four operations
  - GCN: 16-wide SIMD vector unit
  - RDNA & CDNA: Same as GCN

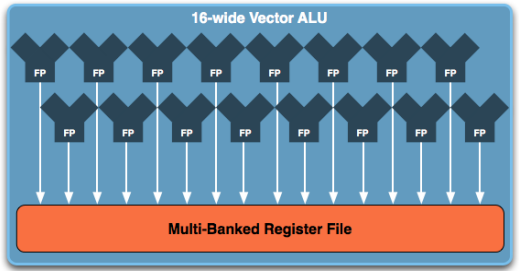


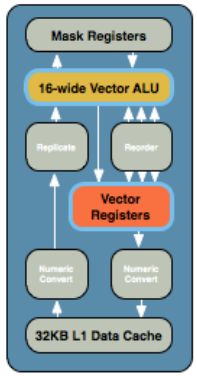




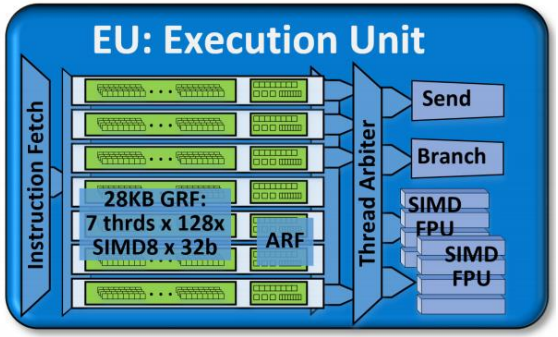
The Core: The basic processing block

- The (failed) Intel Approach:
  - 512-bit SIMD units in x86 cores
  - Failed because of complex x86 cores and software ROP pipeline
  - Used in Xeon Phi, and basis for AVX-512





- The (new) Intel Approach:
  - Used in Intel Graphics Technology (GT)
  - Intel Xe architecture expands on the GT architecture.
  - 128 SIMD-8 32-bit registers



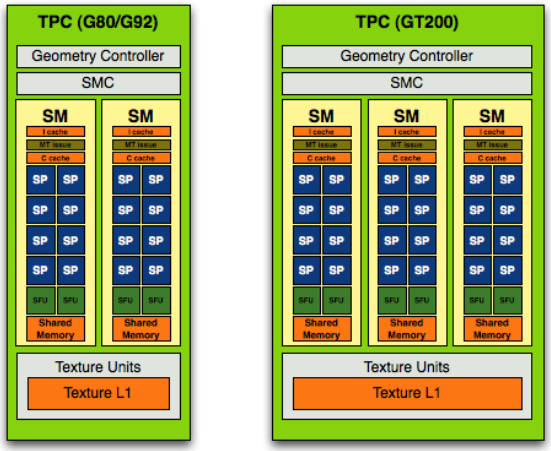


Streaming Multiprocessor (SM)

The real processing element in an NVIDIA GPU

The nVIDIA GPU Architecture Evolving

- **Streaming Multiprocessor (SM) 1.x on the Tesla Architecture**
  - 8 CUDA Cores (Core)
  - 2 Super Function Units (SFU)
- Dual schedulers and dispatch units
  - 1 to 512 or 768 threads active
- Local register (32k)
- 16 KB shared memory
- 2 operations per cycle



- **Streaming Multiprocessor (SM) 2.0 on the Fermi Architecture (GF1xx)**
  - 32 CUDA Cores (Core)
  - 4 Super Function Units (SFU)
- Dual schedulers and dispatch units
  - 1 to 1536 threads active
- Local register (32k)
- 64 KB shared memory / Level 1 cache
- 2 operations per cycle



The nVIDIA GPU Architecture Evolving

- **Streaming Multiprocessor (SMX) 3.x on the Kepler Architecture (Graphics)**
  - 192 CUDA Cores (CC)
  - 8 DP CUDA Cores (DP Core)
  - 32 Super Function Units (SFU)
- Four (simple) schedulers and eight dispatch units
  - 1 to 2048 threads active
- Local register (32k)
- 64 KB shared memory / Level 1 cahce
- 1 operation per cycle



- **Streaming Multiprocessor (SMM) on the Maxwell & Pascal Architecture**
  - 128 CUDA Cores (Core)
  - 4 DP CUDA Cores (DP Core)
  - 32 Super Function Units (SFU)
- Four schedulers and eight dispatch units
  - 1 to 2048 threads active
- Local register (64k)
- 64 KB shared memory
- 24 KB Level 1 / Texture Cache
- 1 operation per cycle



Volta Streaming Multiprocessor (Volta SM)

- **Streaming Multiprocessor (Volta SM) on Volta**
  - 64 CUDA Cores (Core)
  - 32 DP CUDA Cores (DP Core)*
  - 16 Super Function Units (SFU)
  - 8 Tensor Cores (GEMM)
- Four schedulers and eight dispatch units
  - 1 to 2048 active threads
  - Software controlled scheduling
- Local register (64k)
- 128 KB Level 1 / Shared Memory
  - Unified Data Cache
- 1 operation per cycle
- **GV100 / GV10B**





GPGPU

Foils adapted from nVIDIA

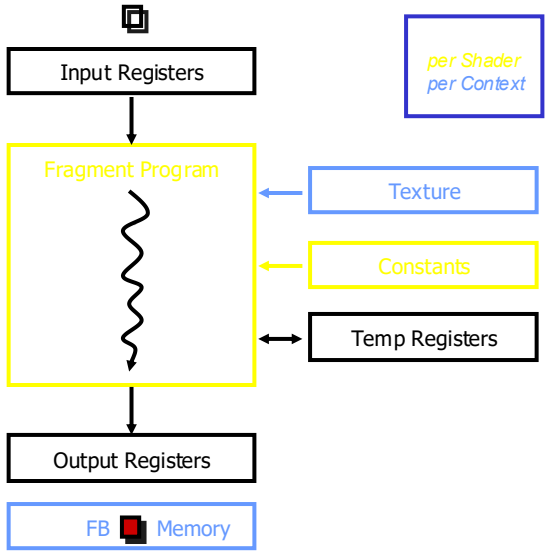
What is really GPGPU?

- Idea:
  - Potential for very high performance at low cost
  - Architecture well suited for certain kinds of parallel applications ***(data parallel)***
  - Demonstrations of ***30-100X*** speedup over CPU

- Early challenges:
  - Architectures very customized to graphics problems (e.g., vertex and fragment processors)
  - Programmed using graphics-specific programming models or libraries

Previous GPGPU use, and limitations

- Working with a Graphics API
  - Special cases with an API like Microsoft Direct3D or OpenGL
- Addressing modes
  - Limited by texture size
- Shader capabilities
  - Limited outputs of the available shader programs
- Instruction sets
  - No integer or bit operations
- Communication is limited
  - Between pixels





NVIDIA CUDA

Foils adapted from nVIDIA

nVIDIA CUDA



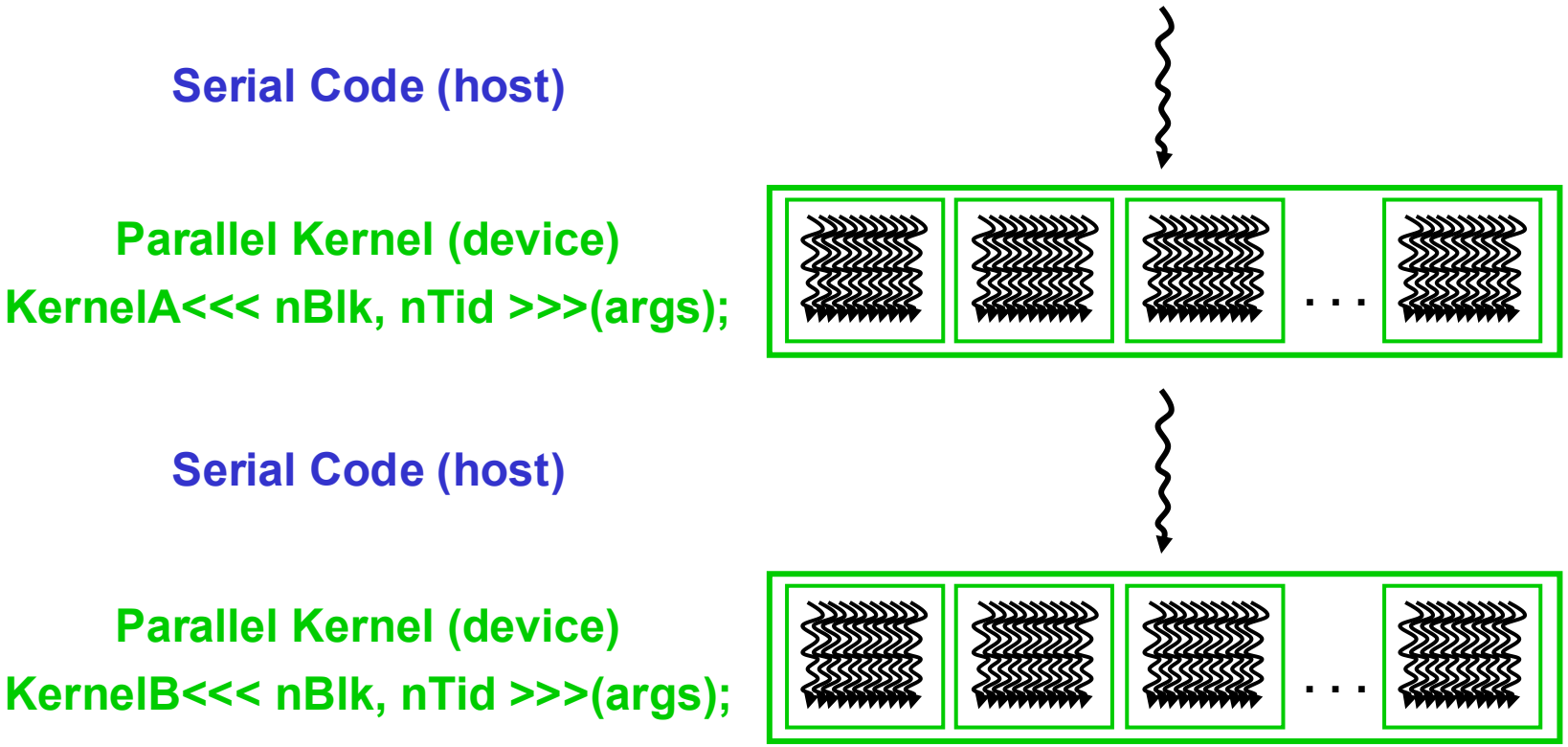
- "Compute Unified Device Architecture"
- General purpose programming model
  - User starts several batches of threads on a GPU
  - GPU is in this case a dedicated super-threaded, massively data parallel co-processor
- Software Stack
  - Graphics driver, language compilers (Toolkit), and tools (SDK)
- Graphics driver loads programs into GPU
  - All drivers from nVIDIA now support CUDA
  - Interface is designed for computing (no graphics ☺)
  - "Guaranteed" maximum download & readback speeds
  - Explicit GPU memory management

The CUDA Programming Model

- The GPU is viewed as a compute device that:
  - Is a coprocessor to the CPU, referred to as the host
  - Has its own DRAM called device memory
  - Runs **many** threads in parallel
- Data-parallel parts of an application are executed on the device as kernels, which run in parallel on many threads
- Differences between GPU and CPU threads
  - GPU threads are extremely lightweight
    - Very little creation overhead
  - GPU needs 1000s of threads for full efficiency
    - Multi-core CPU needs only a few

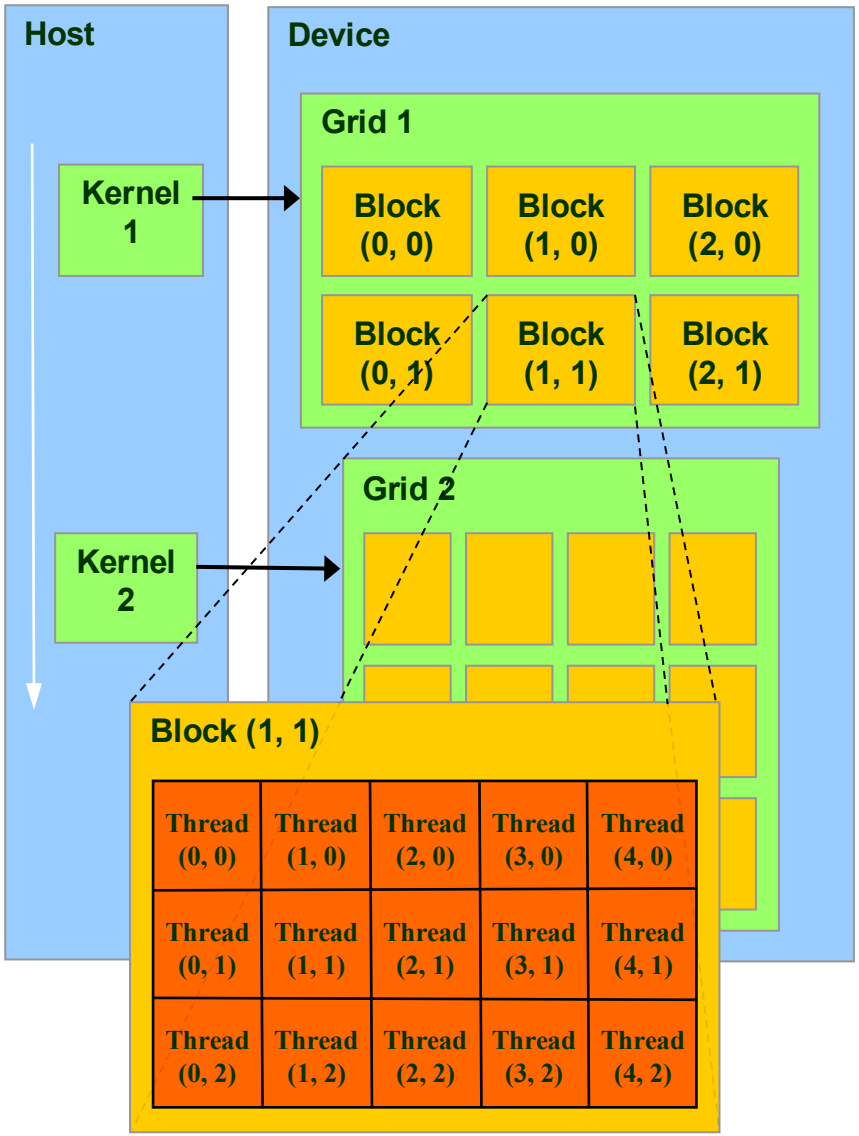
CUDA C – Execution Model

- Integrated host + device C program
  - Serial or modestly parallel parts in **host** C code
  - Highly parallel parts in **device** SPMD kernel C code



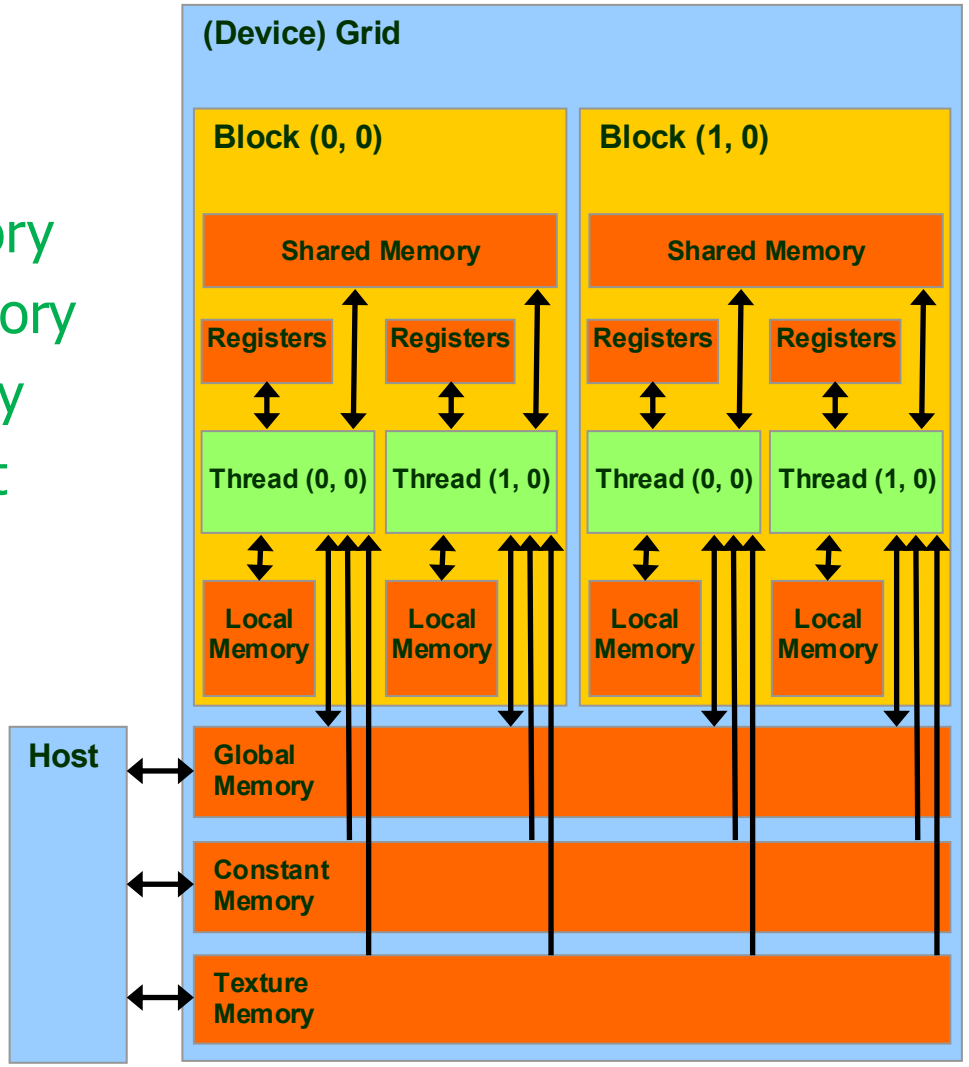
Thread Batching: Grids and Blocks

- A kernel is executed as a grid of thread blocks
  - All threads share data memory space
- A thread block is a batch of threads that can cooperate with each other by:
  - Synchronizing their execution
    - Non synchronous execution is very bad for performance!
  - Efficiently sharing data through a low latency shared memory
- Two threads from two different blocks cannot cooperate



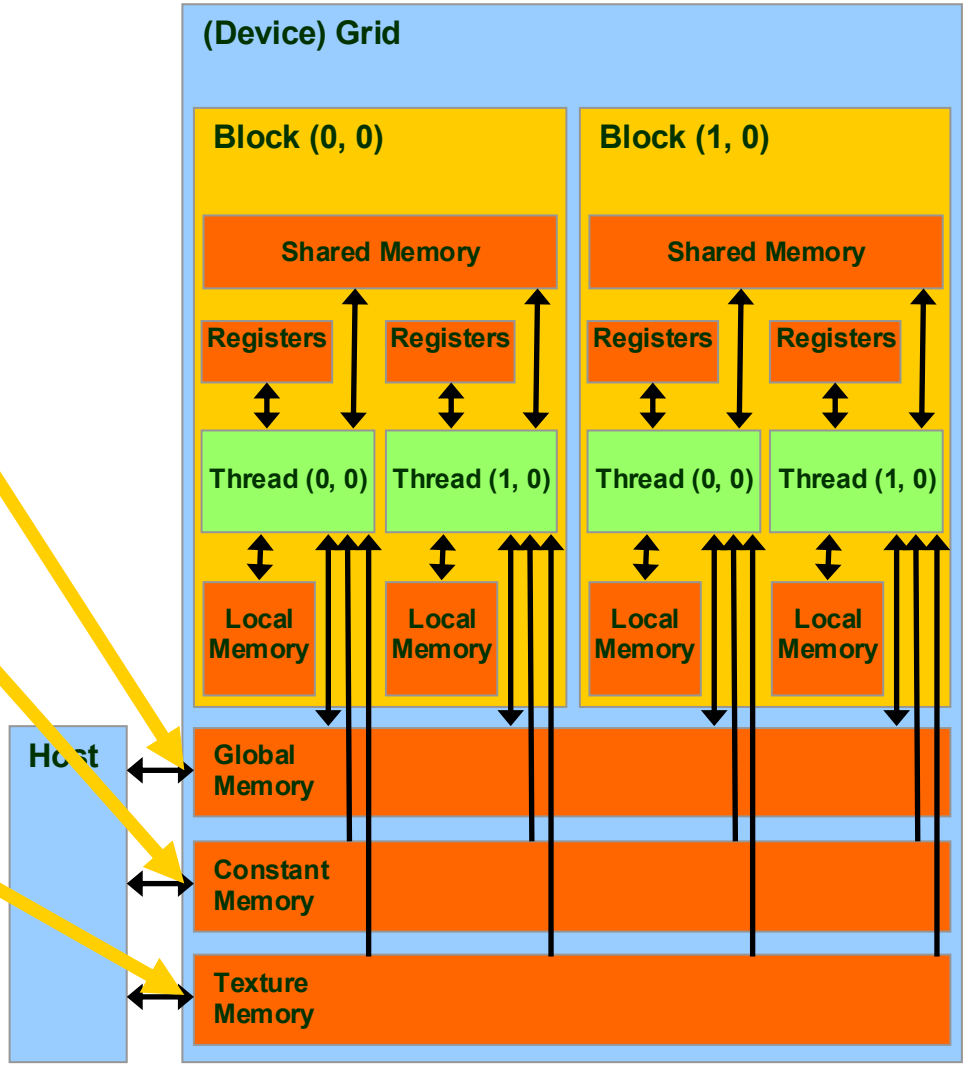
CUDA Device Memory Space Overview

- Each thread can:
  - R/W per-thread registers
  - R/W per-thread local memory
  - R/W per-block shared memory
  - R/W per-grid global memory
  - Read only per-grid constant memory
  - Read only per-grid texture memory
- The host can R/W global, constant, and texture memories



Global, Constant, and Texture Memories

- Global memory:
  - Main means of communicating R/W Data between host and device
  - Contents visible to all threads
- Texture and Constant Memories:
  - Constants initialized by host
  - Contents visible to all threads



Access Times

- Register – Dedicated HW – Single cycle
- Shared Memory – Dedicated HW – Single cycle
- Local Memory – DRAM, no cache* – "Slow"
- Global Memory – DRAM, no cache* – "Slow"
- Constant Memory – DRAM, cached, 1…10s…100s of cycles, depending on cache locality
- Texture Memory – DRAM, cached, 1…10s…100s of cycles, depending on cache locality

* Can be cached in L2 or L1 cache on GPU, however this can not be controlled by the programmer.

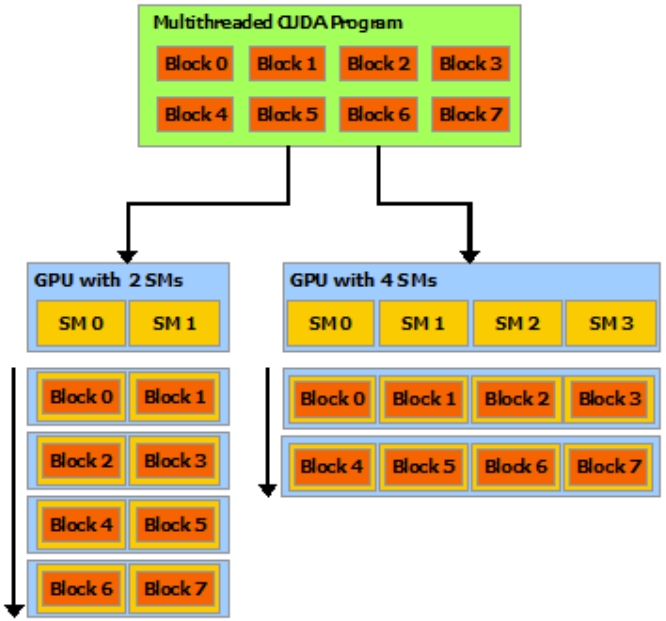
Terminology Recap

- device = GPU = Set of multiprocessors
- Multiprocessor = Set of processors & shared memory
- Kernel = Program running on the GPU
- Grid = Array of thread blocks that execute a kernel
- Thread block = Group of SIMD threads that execute a kernel and can communicate via shared memory

Memory	Location	Cached	Access	Who
Local	Off-chip	No (Has L2 cache)	Read/write	One thread
Shared	On-chip	N/A - resident	Read/write	All threads in a block
Global	Off-chip	No (Has L2 cache)	Read/write	All threads + host
Constant	Off-chip	Yes	Read	All threads + host
Texture	Off-chip	Yes	Read/write	All threads + host

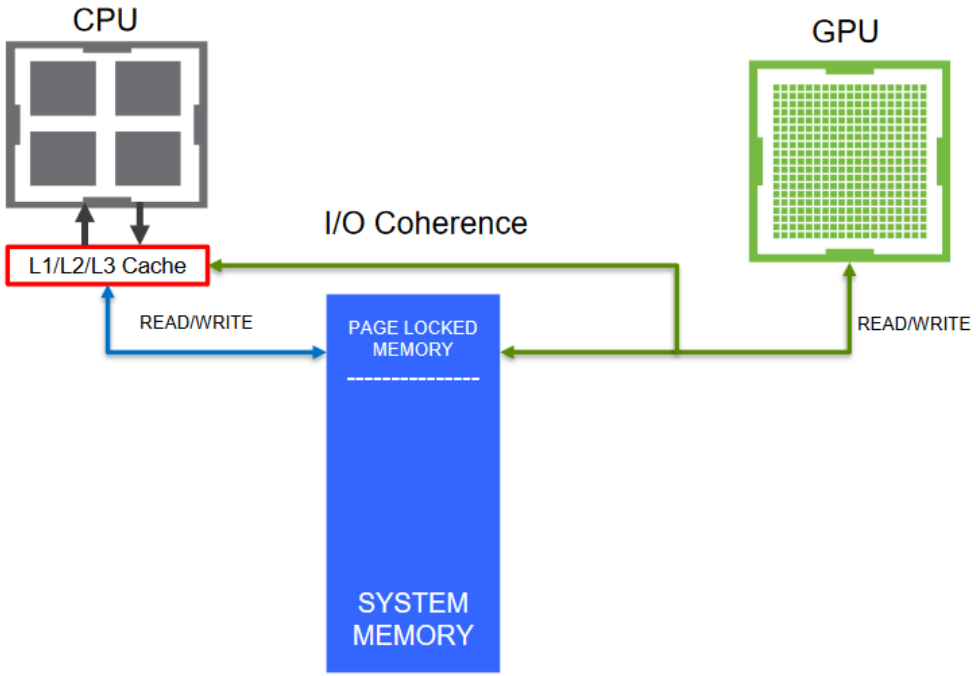
Scalability

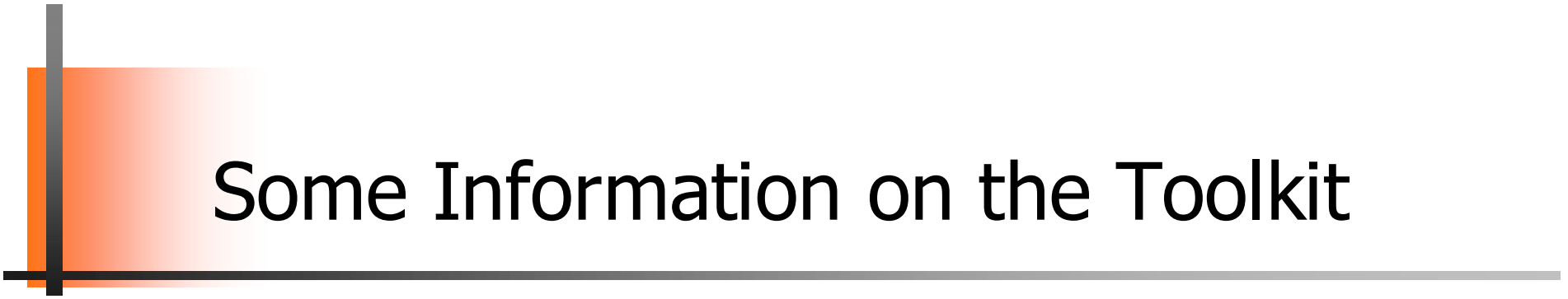
- GPU is built around an array of Streaming Multiprocessors (SMs)
- CUDA has three key abstractions:
  - Hierarchy of thread groups
  - Shared memories
  - Barrier synchronization
- The CUDA Runtime will scale the program to the available resources
  - JIT Compile



Data movement between CPU and GPU

- For Home Exam 1 & 2, you are developing for a system were CPU and GPU share memory!
- Xavier supports I/O Coherency:
  - Hardware cache coherency between CPU and GPU
  - GPU can snoop CPU cache hierarchy
- Enables memory models such as Managed Memory and Pinned Host Memory





Some Information on the Toolkit

Compilation

- Any source file containing CUDA language extensions must be compiled with nvcc
- nvcc is a compiler driver
  - Works by invoking all the necessary tools and compilers like cudacc, g++, etc.
- nvcc can output:
  - Either C code
    - That must then be compiled with the rest of the application using another tool
  - Or object code directly

Linking & Profiling

- Any executable with CUDA code requires two dynamic libraries:
  - The CUDA runtime library (`cudart`)
  - The CUDA core library (`cuda`)

- Several tools are available to optimize your application
  - NVIDIA Nsight Systems
  - NVIDIA Nsight Compute
  - CUDA Occupancy Calculator [1]

[1] https://docs.nvidia.com/cuda/archive/11.4.4/cuda-occupancy-calculator

Before you start...

- Four lines must be added to your group users *.bashrc* file

```
PATH=$PATH:/usr/local/cuda-11.4/bin
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/cuda-11.4/lib64:/lib

export PATH
export LD_LIBRARY_PATH
```

- Two CUDA versions are available on the Tegra machines:
  - CUDA 11.4
  - CUDA 12.2
- Code samples is installed with CUDA

  /usr/local/cuda/samples
- Copy and build in your user's home directory

Some usefull resources

NVIDIA CUDA Programming Guide 11.4

https://docs.nvidia.com/cuda/archive/11.4.4/cuda-c-programming-guide/

NVIDIA CUDA C Best Practices Guide 11.4

https://docs.nvidia.com/cuda/archive/11.4.4/cuda-c-best-practices-guide/

Tuning CUDA Applications for Volta

https://docs.nvidia.com/cuda/archive/11.4.4/volta-tuning-guide/

CUDA for Tegra – Application Note

https://docs.nvidia.com/cuda/archive/11.4.4/cuda-for-tegra-appnote/

Parallel Thread Execution ISA – Version 7.4

https://docs.nvidia.com/cuda/archive/11.4.4/parallel-thread-execution/