GPU COMPUTING

@ HPC & BigData, UvA, January 2019

Ana Lucia Varbanescu (UvA)
a.l.varbanescu@uva.nl

HPC = high performance computing

- Big Science = Big Data, Big Simulation, Complex models…
- Challenges
  - Compute and storage
  - Efficiency: Performance vs. Energy

- Traditional HPC: complex machines, on-demand
- Modern HPC: more and more based on existing machines, put together in dense clusters/datacenters

- **HPC is expanding to new application domains**
  - **More machines**
  - **More tools & programming models**

Parallelism ⇔ HPC

- Parallelism is **mandatory** for **high performance**
  - Yesterday: **clusters (and grids)**
  - Today: **multi-/many-core processors**
  - Tomorrow: **massive multi-scale heterogeneous parallelism = clusters using different types of multi-/many-cores**

>95% of computing systems today are parallel!

Main challenges: learn to program parallel machines and learn to use them efficiently!

BASICS OF PARALLEL PROGRAMMING

Parallel processing & programming

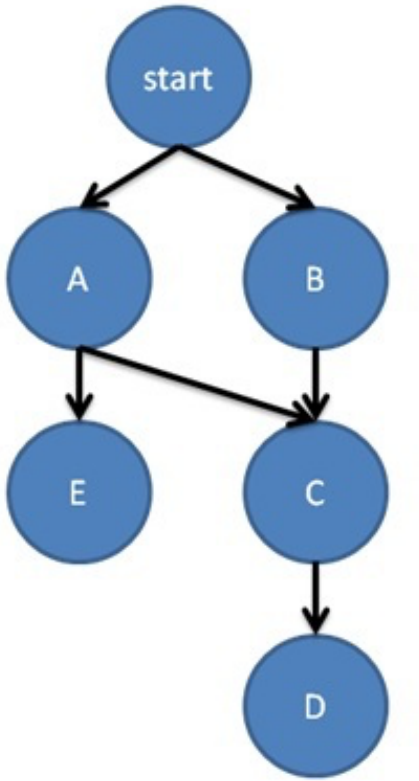
- Application = collection of tasks
- Parallel processing = enable (some of) the tasks to execute in parallel
  - The more the better?
  - Which tasks?
- Parallel machines
  - Enable parallel execution of tasks on hardware resources
- Parallel programming = *marry* parallel processing with parallel machines :D
  - Two problems:
    - “Detect” parallelism => parallel models of computation, algorithms, …
    - Implement parallelism => programming models & tools

Some terminology

- Task
  - A logically discrete section of computational work.
  - A parallel program consists of multiple tasks running in parallel

- Application = DAG of tasks/operations
  - Operations = nodes
  - Dependencies = edges
    - Dependency = child-after-parent

- Two metrics:
  - The **amount of work** to be executed
  - The **span** of the application
    - Also called critical-path length or computational depth



More terminology

- Computation
  - The actual **work** to be executed by the tasks
- Communication
  - Tasks must **exchange** information
  - Different mechanisms depend on machine model
    - Shared variables
    - Message passing
- Synchronization
  - A task execution **depends** on another task's progress
  - Different mechanisms
    - Barriers
    - Locks, mutexes, semaphores
    - Message passing

More terminology

- Example:

Assume a processor P with 10 cores and an application A with:

  - 10 tasks: how long does A take?
  - 20 tasks: how long does A take?
  - 33 tasks: how long does A take?

- Task **scheduling** = the order in which the threads are executed on the machine
  - User-based: programmer decides
  - OS-based: OS decides (e.g., Linux, Windows)
  - Hardware-based: hardware decides (e.g., GPUs)

From abstract tasks to … practice!

- Processes => multi-processing
  - Thread of control with own resources
  - No shared memory
  - Communication through messages or files
  - Heavy-weight (expensive to set up and destroy)

- Threads => multi-threading
  - Thread of control with limited local memory and shared address space
  - Communication through shared memory
  - Light-weight (much cheaper to set up and destroy)

Multithreading

- An execution model
  - Focuses on concurrency
  - Allows multiple threads to exist within the context of one process.
  - Threads share the process's resources, but are able to execute independently.
- Programming models/abstractions
  - Allow programs to express multiple threads and manage them
- Concurrency and parallelism
  - All threads are designed concurrent
  - Only some of them may run in parallel – why ?

Multithreading challenges

- Parallel computation (= compute things in the same time)
  - Split the work over multiple tasks
  - Challenging: granularity and dependencies
- Communication (= as little as possible)
  - Uses shared memory
  - Challenging: shared data and race conditions
- Synchronization (= as little as possible)
  - Uses locks, barriers, …
  - Challenging: locking resources and deadlocks/livelocks

Race Condition / Data Race

Definition:

- A race condition / data race exists if the behavior (the meaning) of a program depends on the execution order of program parts (threads) whose temporal behavior is beyond control.

- In practice, it often happens when multiple threads use the same (shared) data.

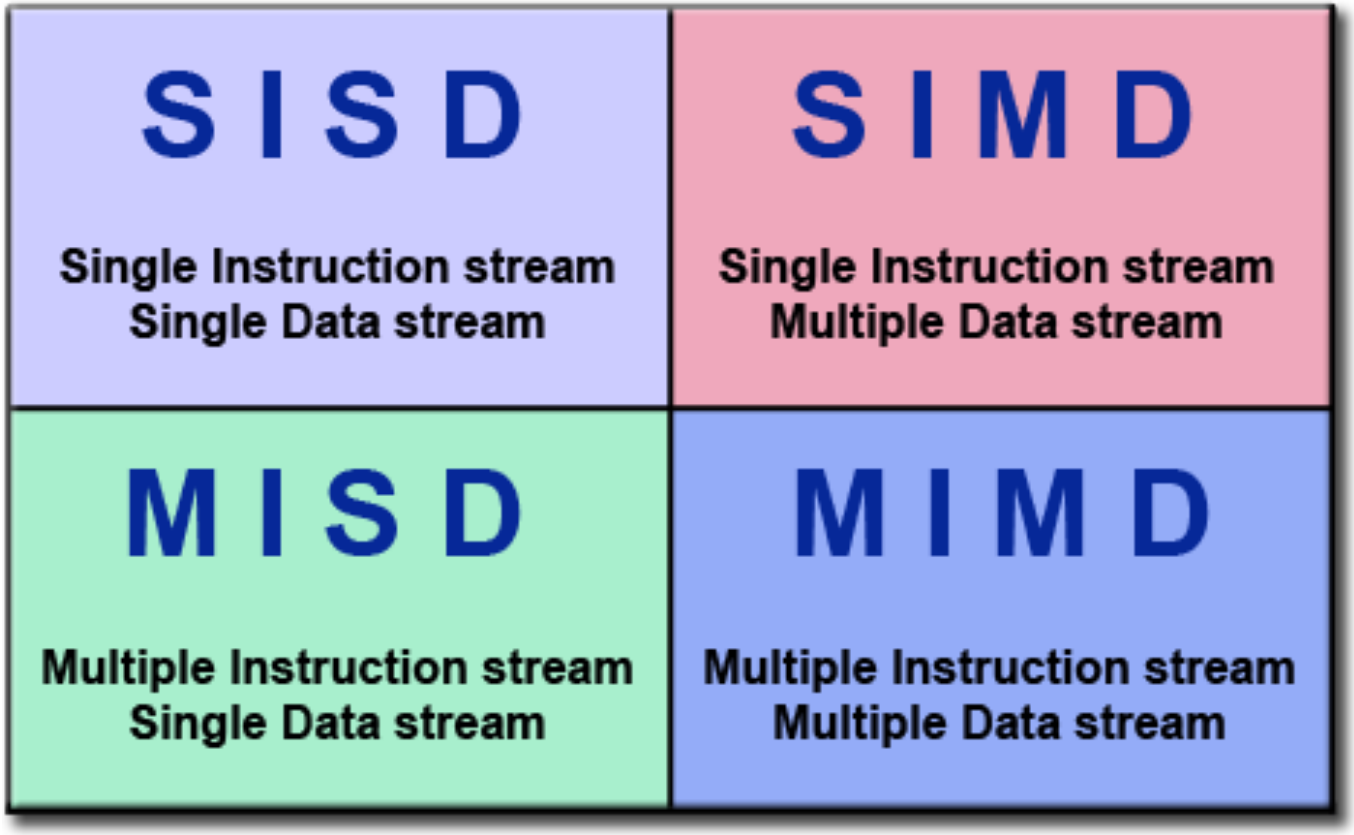
- Most solutions are based on
  - Locking
  - Replication

Locks & Deadlocks

- Locking = a thread locks a resource for single use.
- Deadlock = threads are blocked waiting for each other and no progress is being made.
- Livelock = threads “oscillate” between states with regard to one another and no progress is being made.

BASIC MODELS OF PARALLEL SYSTEMS

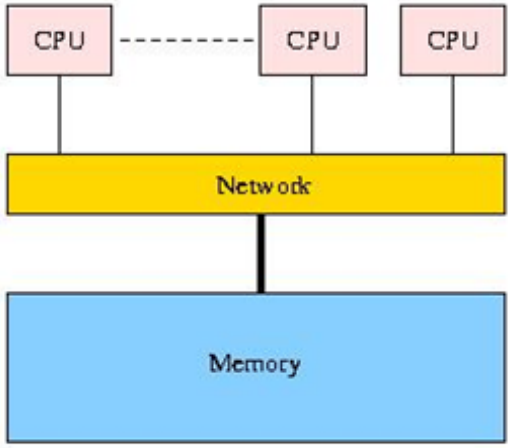
Flynn's taxonomy (1966)

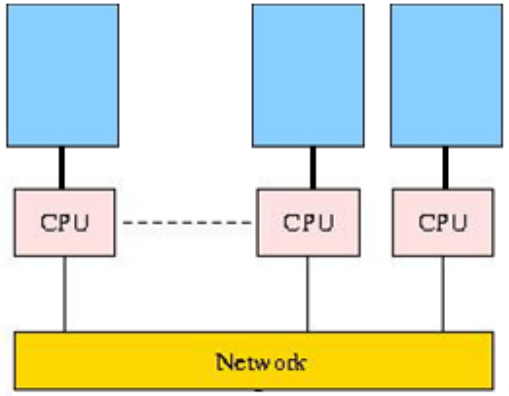


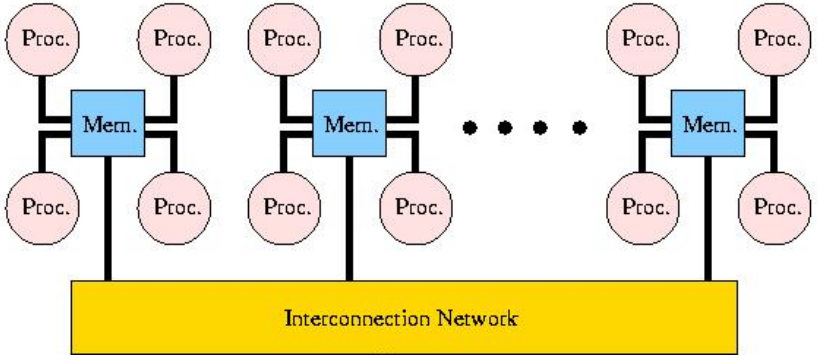
More classifications:
https://computing.llnl.gov/tutorials/parallel_comp/parallelClassifications.pdf

Parallel Machine Models

- Shared Memory
  - Multiple compute nodes
  - One single shared address space
  - Typical example: multi-cores
- Distributed Memory
  - Multiple compute nodes
  - Multiple, local (disjoint) address spaces
  - **Virtual shared memory:** software/hardware layer "emulates" shared memory
  - Typical example: clusters
- Hybrids
  - Multiple compute nodes
    - Typically heterogeneous
  - Mixed address space(s)
    - Soma shared, some global memory



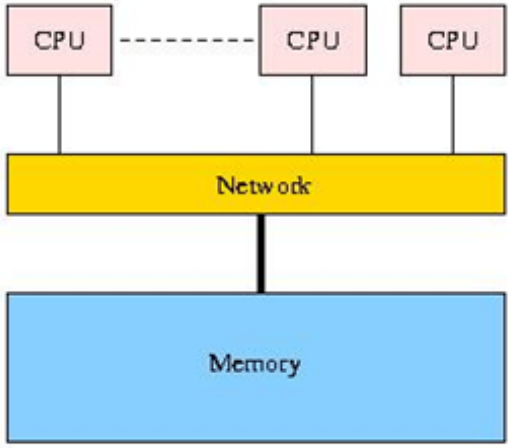




Parallel Machine Models

- Shared Memory

Programming: multi-threading
Programming models: OpenMP, pthreads, TBB, …

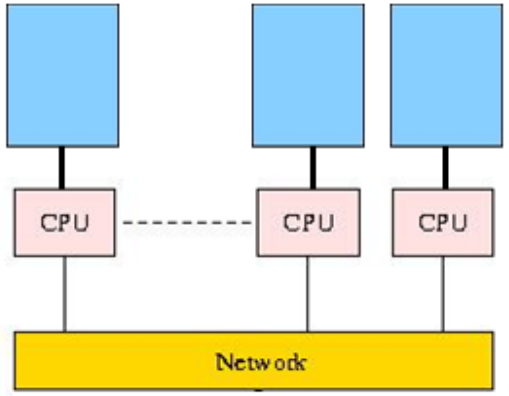


- Distributed Memory

Programming: message passing
Programming models: MPI, Big-data models, …

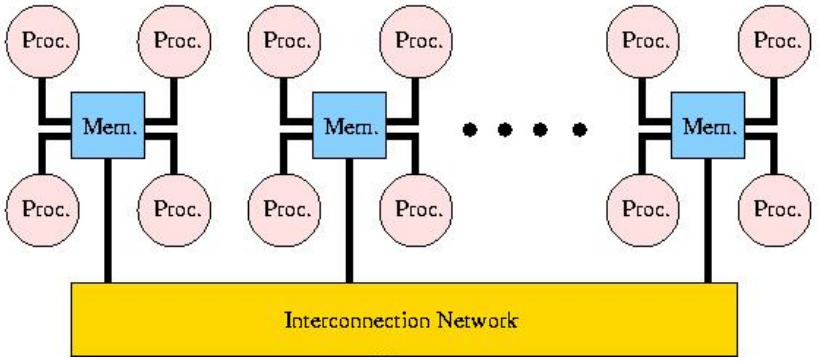
emulates "shared memory"

- Typical example: clusters



- Hybrids
  - Multiple compute nodes

Programming: very diverse, depending on the hardware configuration



Examples

- Multi-core CPUs ?
  - Shared memory with respect to system memory
  - Hybrid when taking caches into account
- Clusters ?
  - Distributed memory
  - Could be shared if middleware for virtual shared space is provided
- Supercomputers ?
  - Usually hybrid
- GPUs ?
- Architectures with GPUs?
  - Distributed for traditional, off-chip GPUs
  - Shared for new APUs

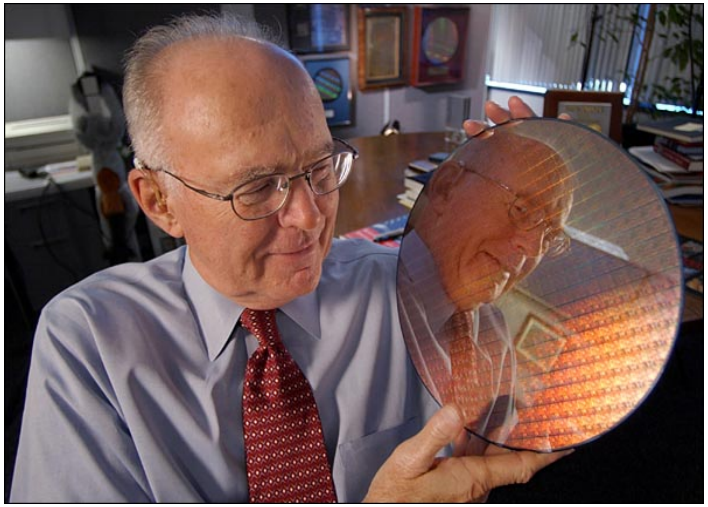
Major issues

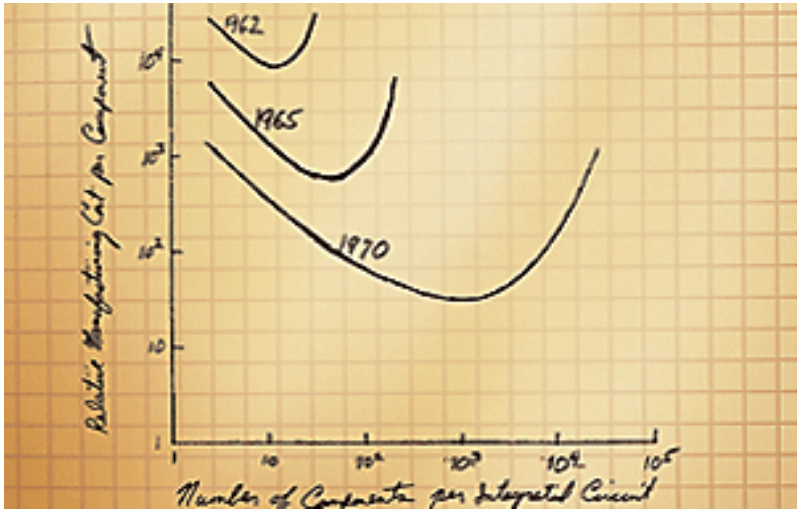
- Shared Memory model
  - Scalability problems (interconnect)
  - Programming challenge: RD/WR Conflicts
- Distributed Memory model
  - Data distribution is mandatory
  - Programming challenge: remote accesses, consistency
- Virtual Shared Memory model
  - Significant virtualization overhead
  - Easier programming
- Hybrid models
  - Local/remote data more difficult to trace

MULTI-/MANY-CORES

Moore's Law

- Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.





"The complexity for minimum component costs has increased at a rate of roughly a factor of two per year ... Certainly over the short term this rate can be expected to continue, if not to increase...." Electronics Magazine 1965

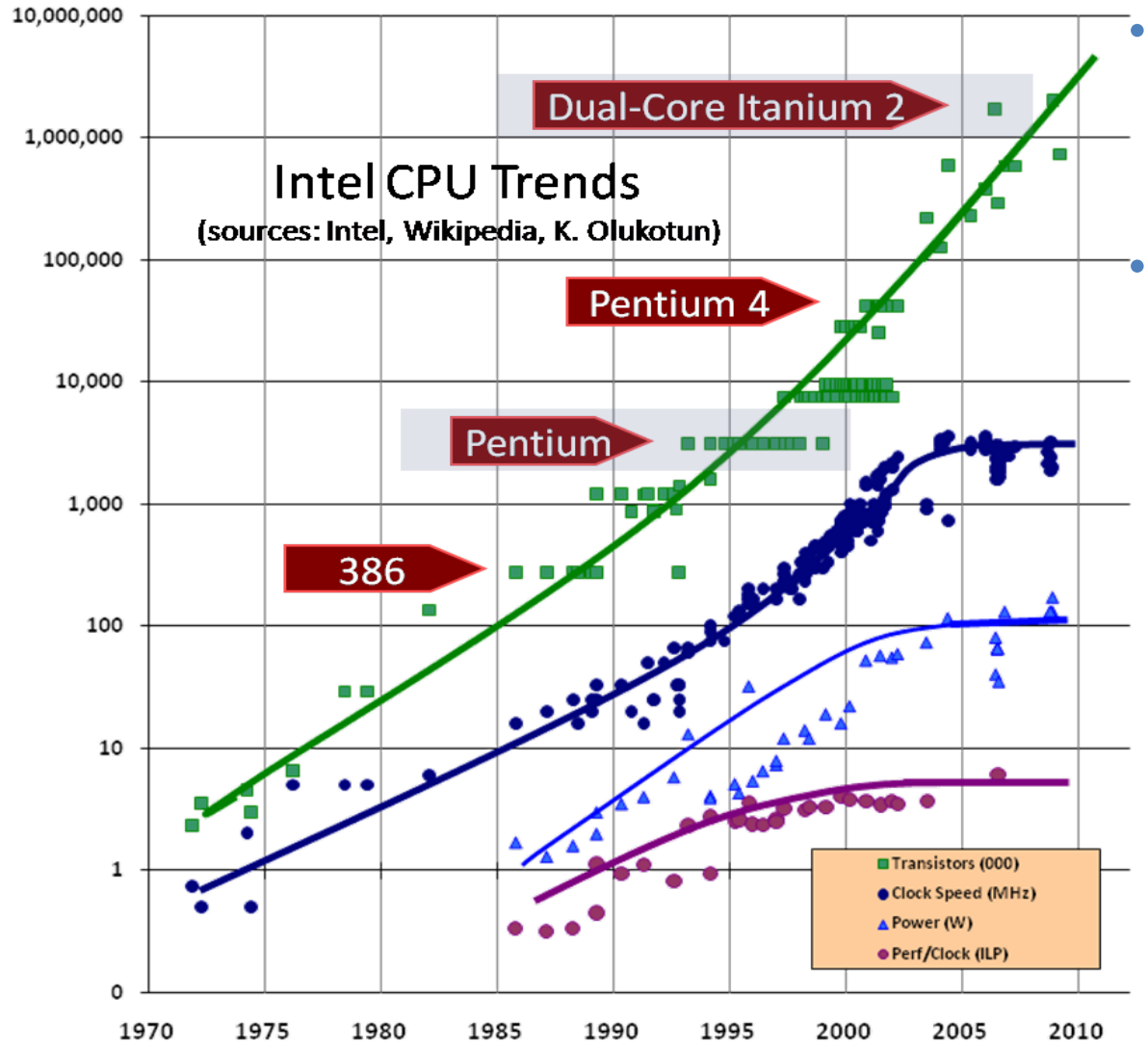
Traditionally …

- More transistors = more functionality
- Improved technology = faster clocks = more speed

- Thus, every 18 months, we obtained better and faster processors.

- They were all sequential: they execute one operation per clock cycle.

Revolution in Processors

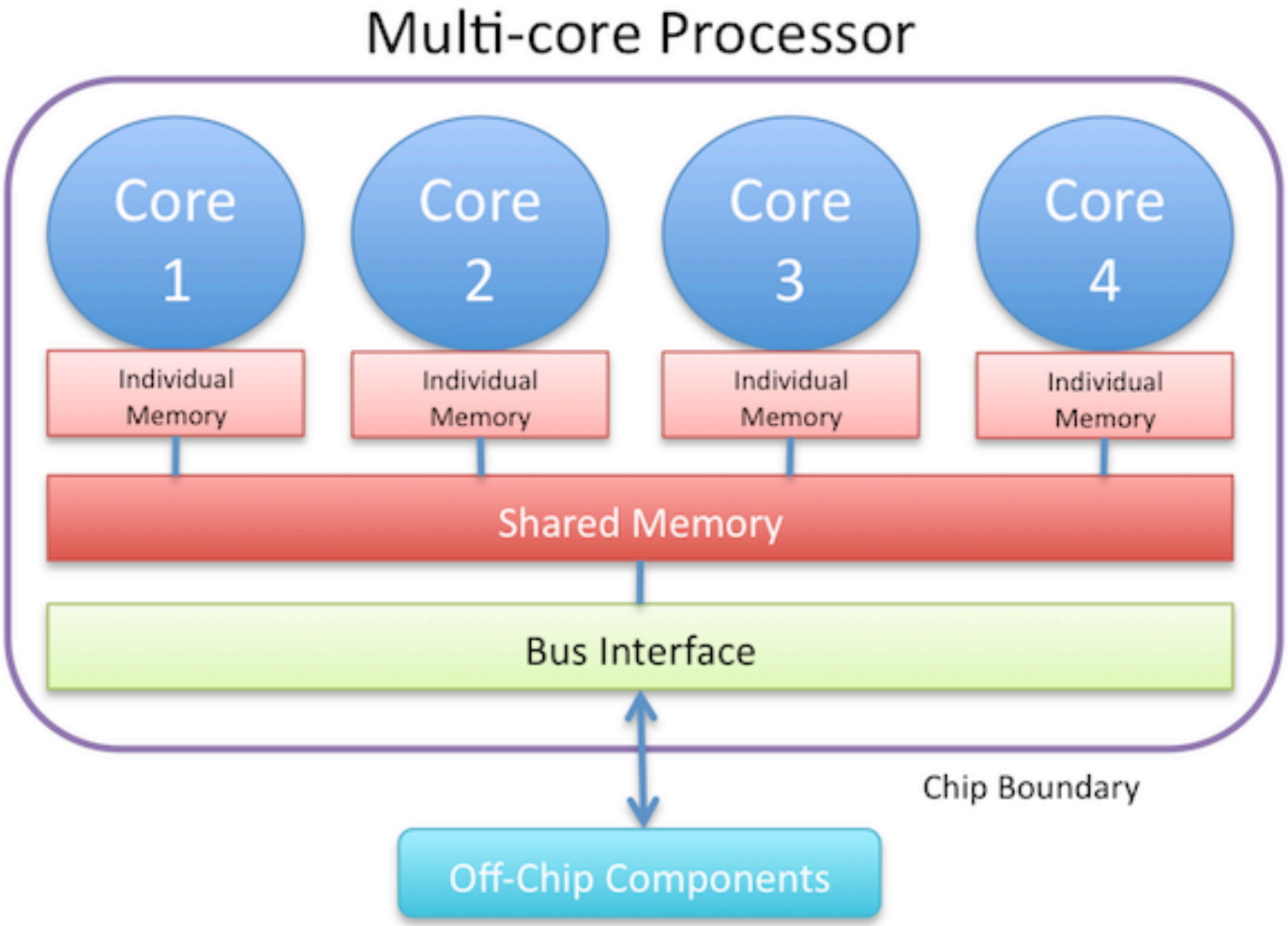


- Chip density is continuing to increase about 2x every 2 years
- BUT
  - Clock speed is not
  - Performance per cycle is not
  - Power is not

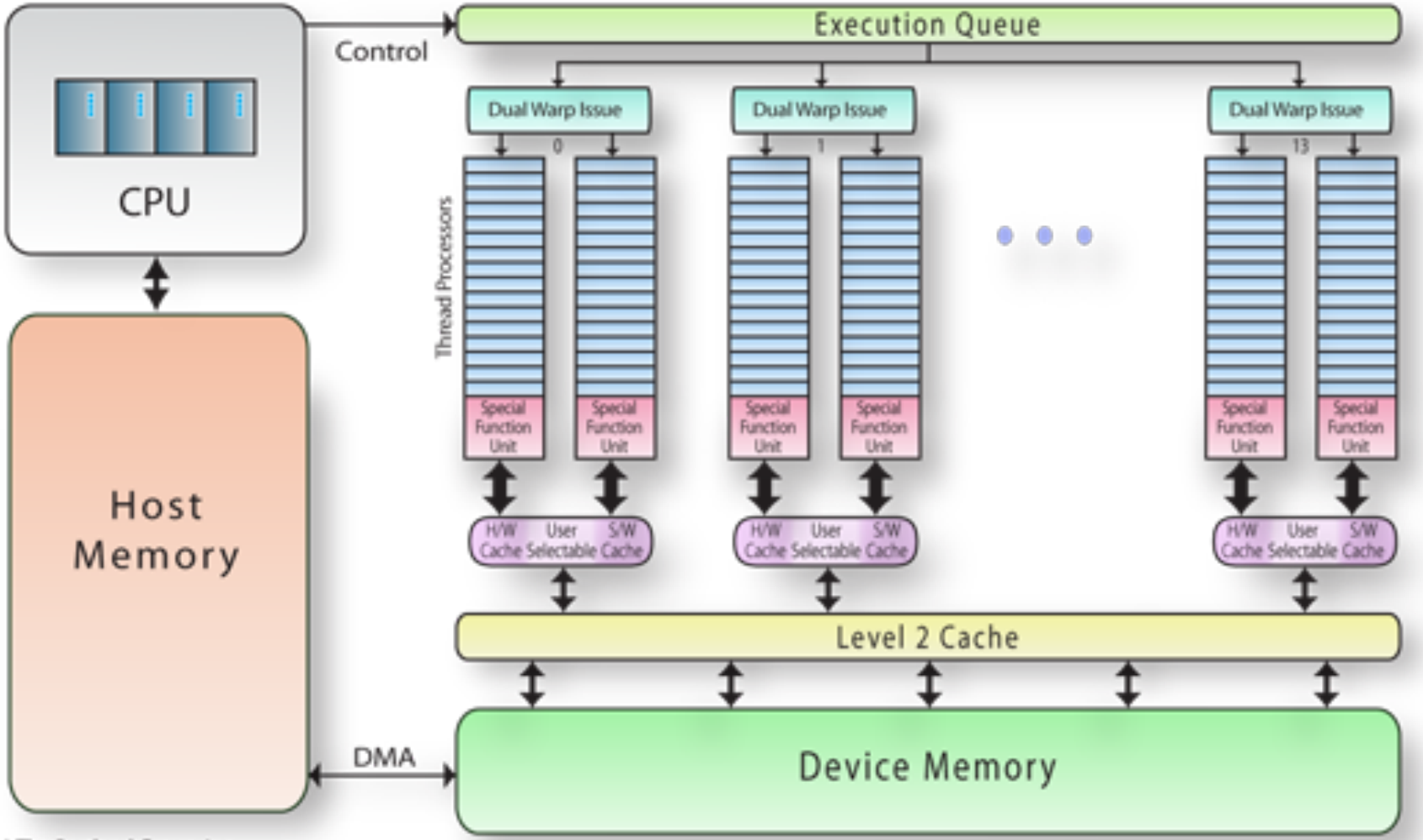
New ways to use transistors

Improve PERFORMANCE by using parallelism on-chip: multi-core (CPUs) and many-core processors (GPUs).

Generic multi-core CPU



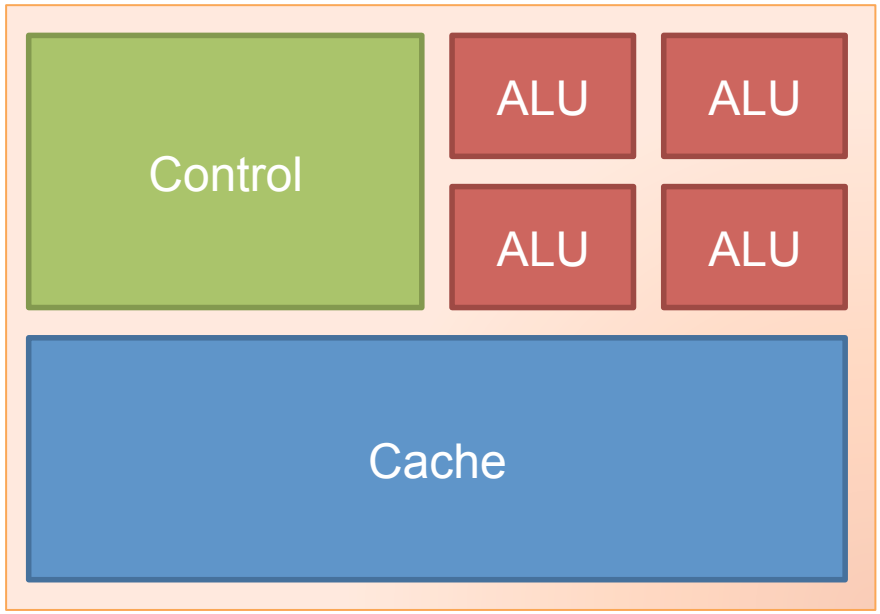
Generic GPU



CPU vs. GPU

- Which one is better?

CPU vs. GPU

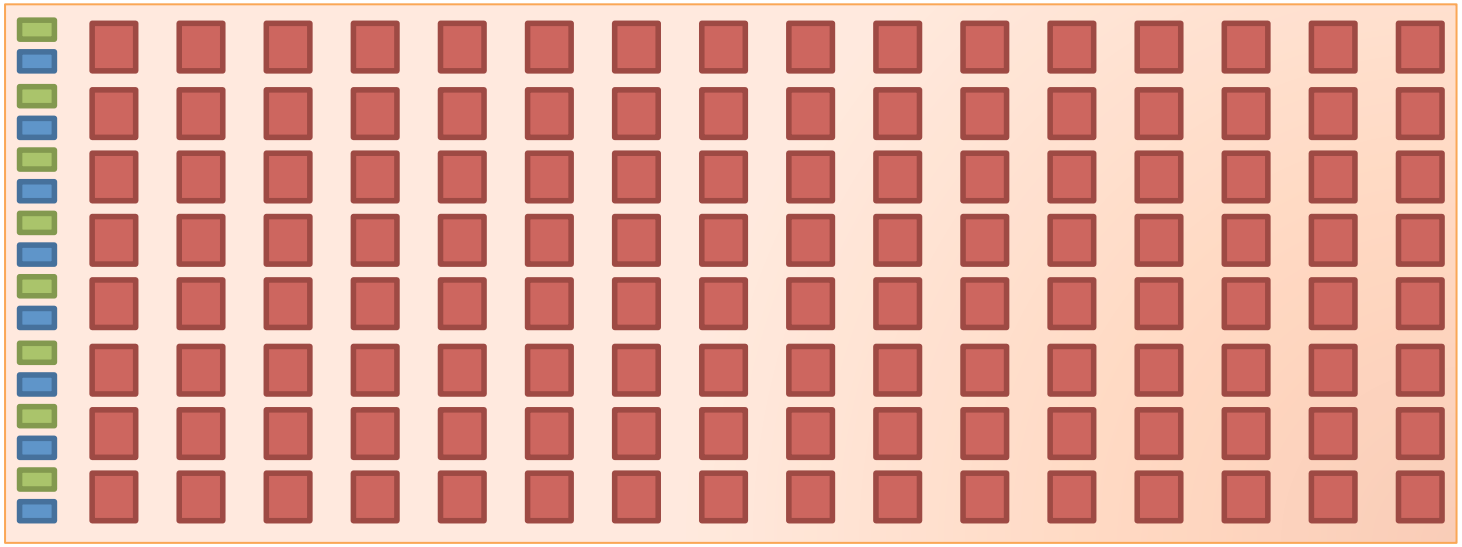


CPU

Few complex cores
Lots of on-chip memory
Lots of control logic

GPU

many
simple cores,
little memory,
little control



Why so different?

- Different goals produce different designs!
  - CPU must be good at everything
  - GPUs focus on massive parallelism
    - Less flexible, more specialized
- CPU: minimize latency experienced by 1 thread
  - big on-chip caches
  - sophisticated control logic
- GPU: maximize throughput of all threads
  - # threads in flight limited by resources => lots of resources (registers, etc.)
  - multithreading can hide latency => no big caches
  - share control logic across many threads

CPU vs. GPU

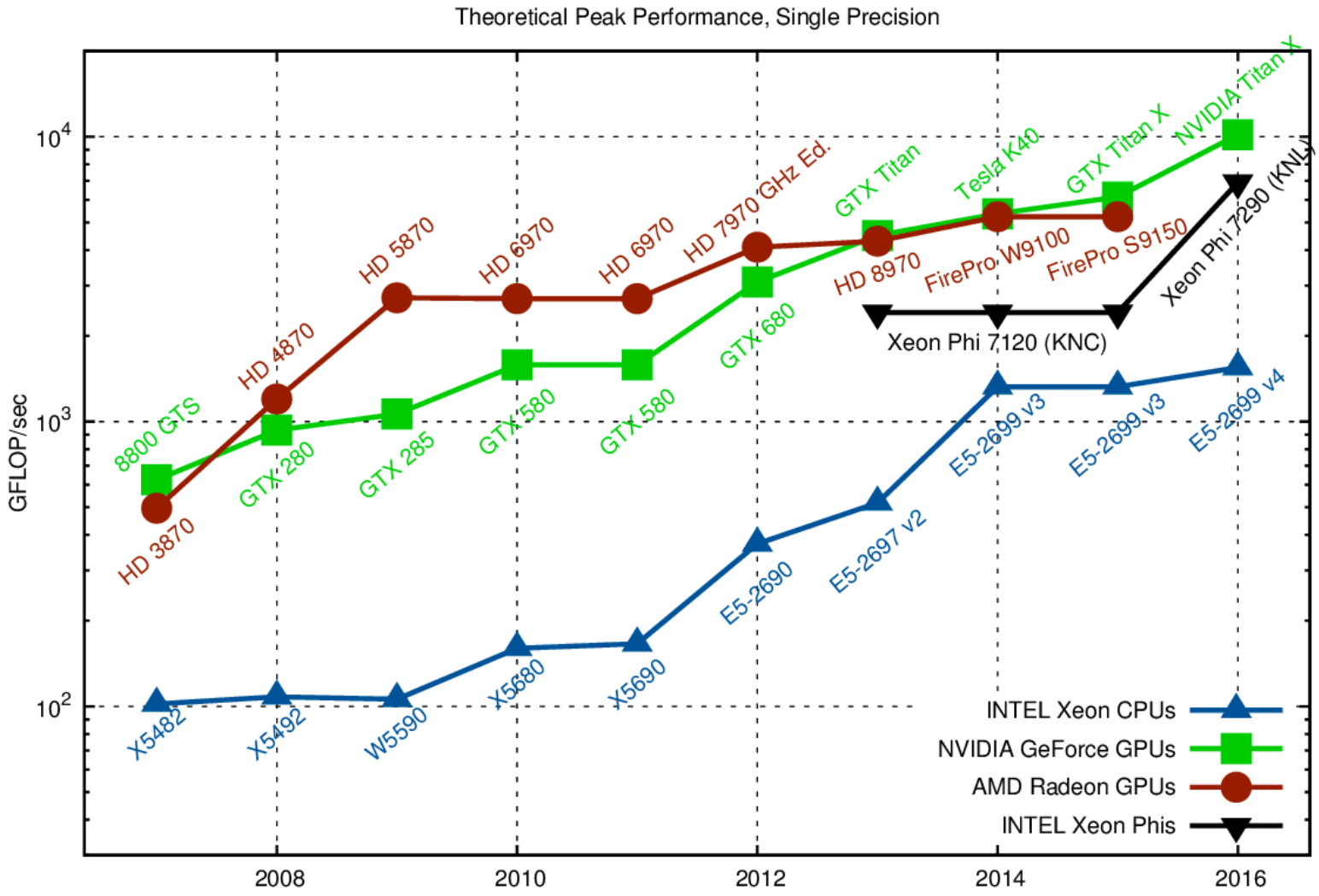
- Movie
- The Mythbusters
  - Jamie Hyneman & Adam Savage
  - Discovery Channel
- Appearance at NVIDIA's NVISION 2008

Hardware Performance metrics

- Clock frequency [GHz] = absolute hardware speed
  - Memories, CPUs, interconnects
- **Operational speed [GFLOPs]**
  - Operations per second
  - **single** AND **double** precision
- **Memory bandwidth [GB/s]**
  - Memory operations per second
    - Can differ for read and write operations !
  - Differs a lot between different memories on chip
- Power [Watt]
  - The rate of consumption of energy
- Derived metrics
  - FLOP/Byte, FLOP/Watt

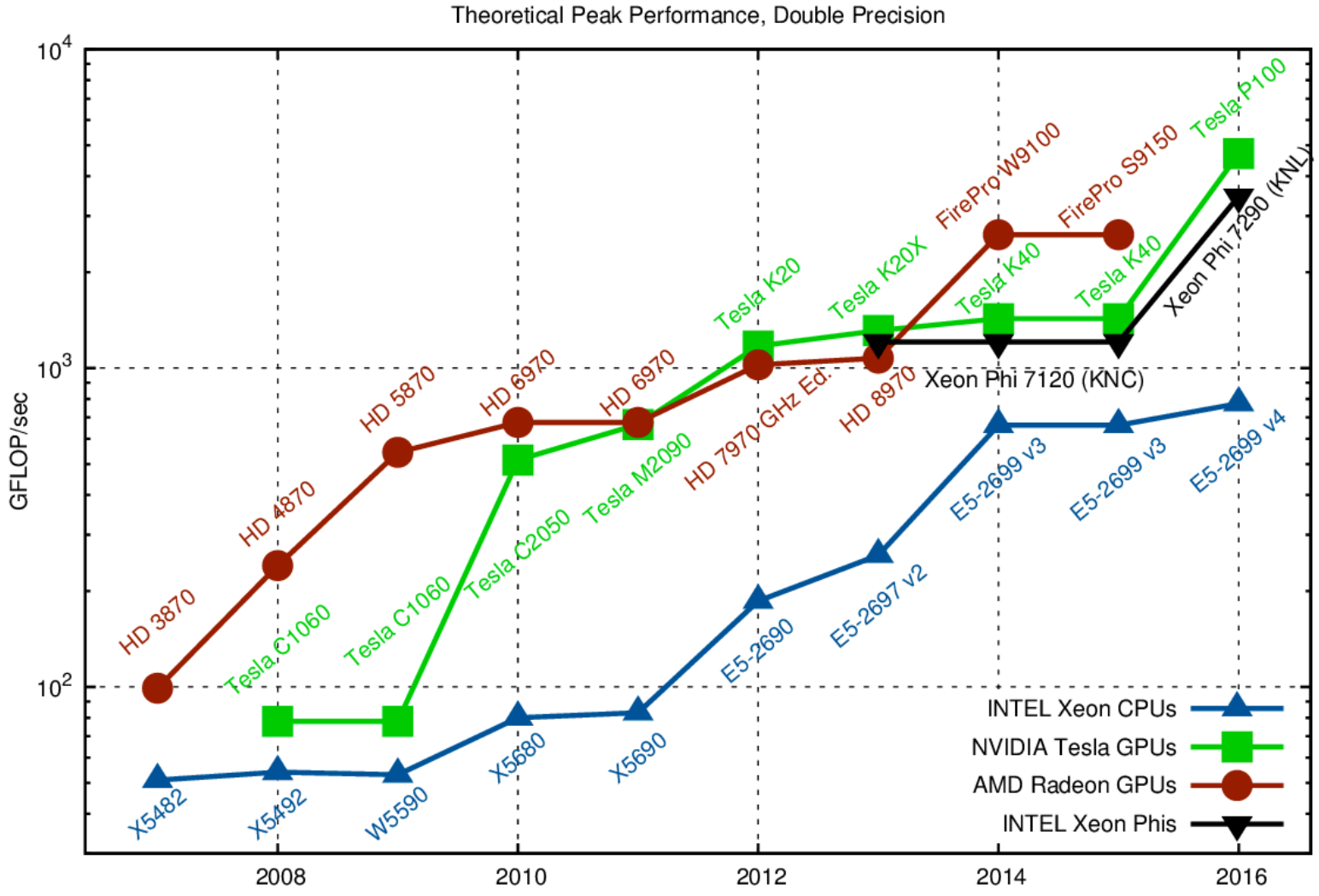
Name	FLOPS
yottaFLOPS	10^{24}
zettaFLOPS	10^{21}
exaFLOPS	10^{18}
petaFLOPS	10^{15}
teraFLOPS	10^{12}
gigaFLOPS	10^{9}
megaFLOPS	10^{6}
kiloFLOPS	10^{3}

multi vs *many* cores (SP-FLOPs)



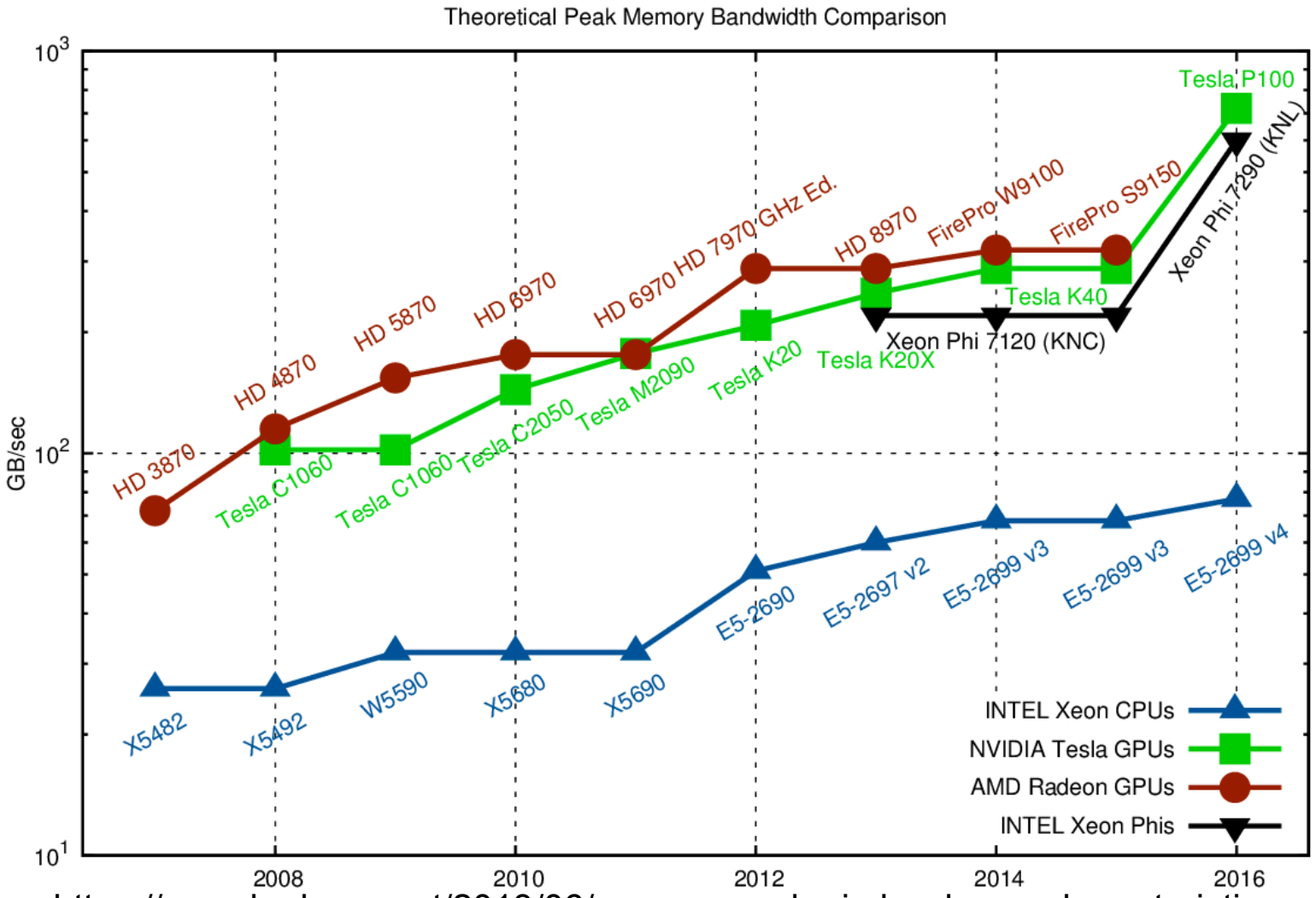
https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

multi vs *many* cores (DP-FLOPs)



https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

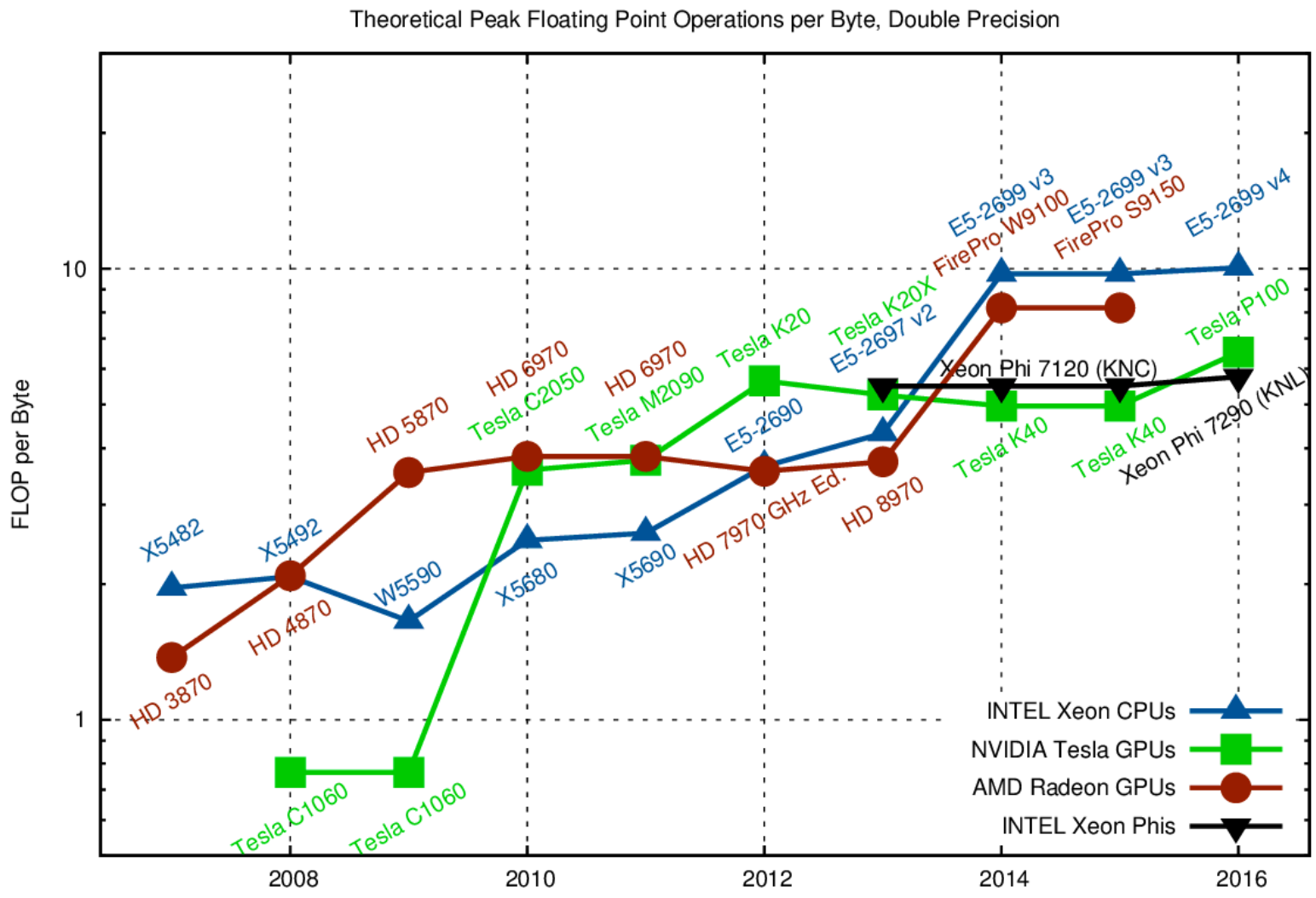
multi vs *many* cores (GB/s)



https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

Balance ?

FLOPs/Byte !

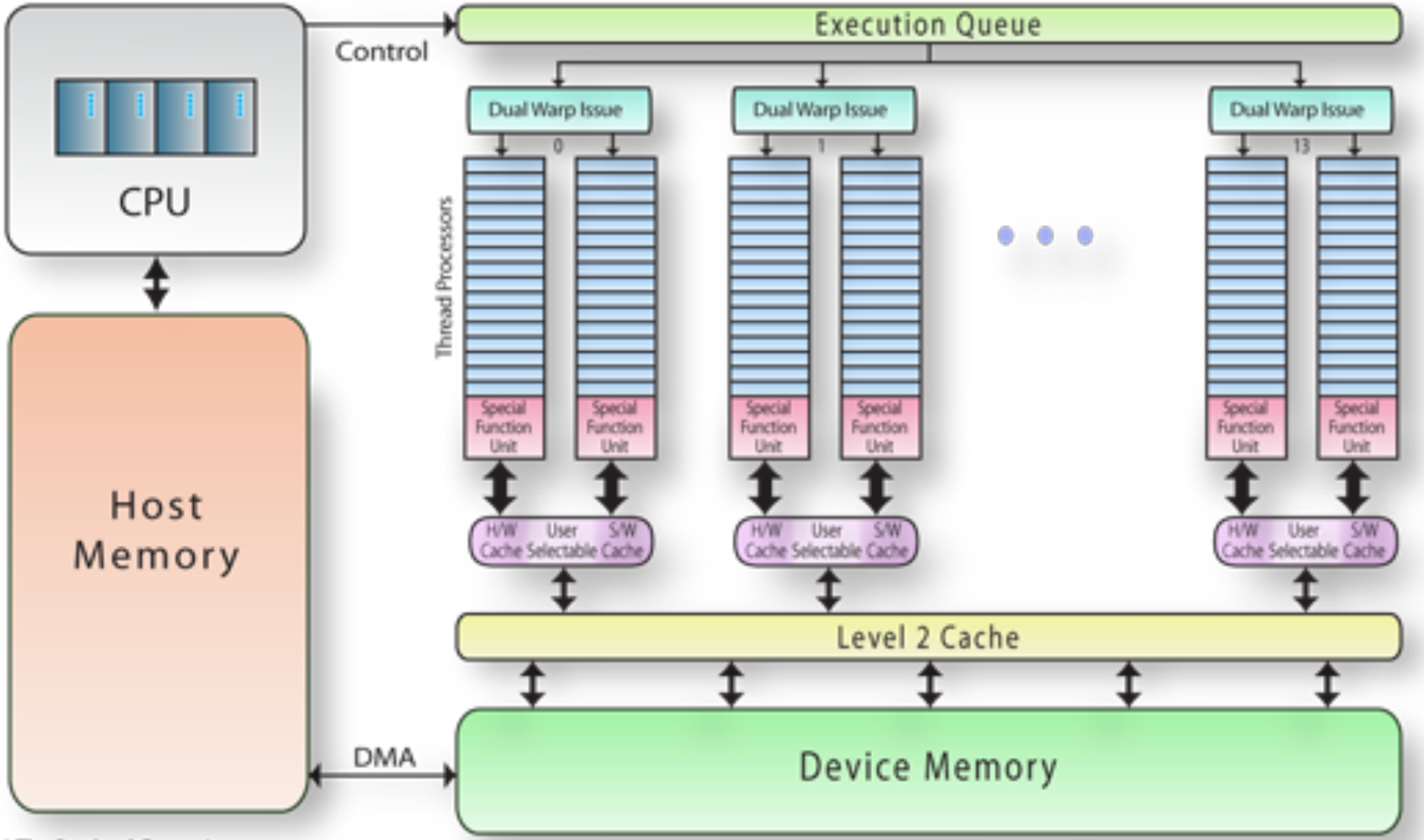


https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/

GPGPU ?!

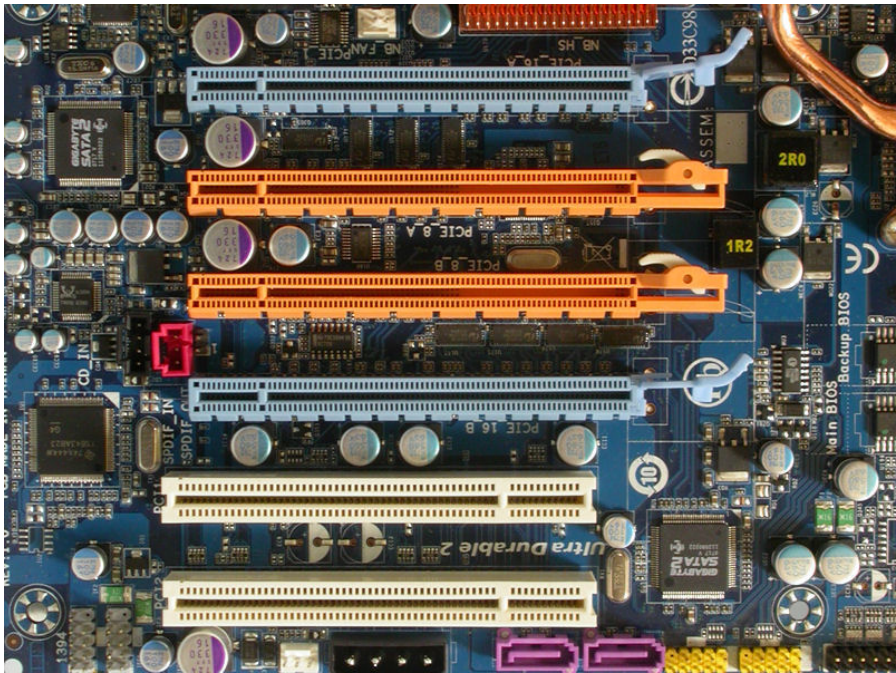
Massive parallelism => massive performance

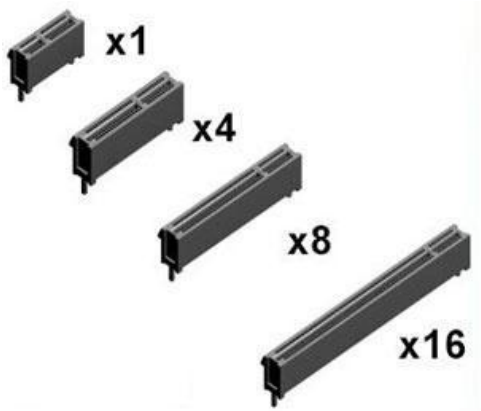
A GPU Architecture



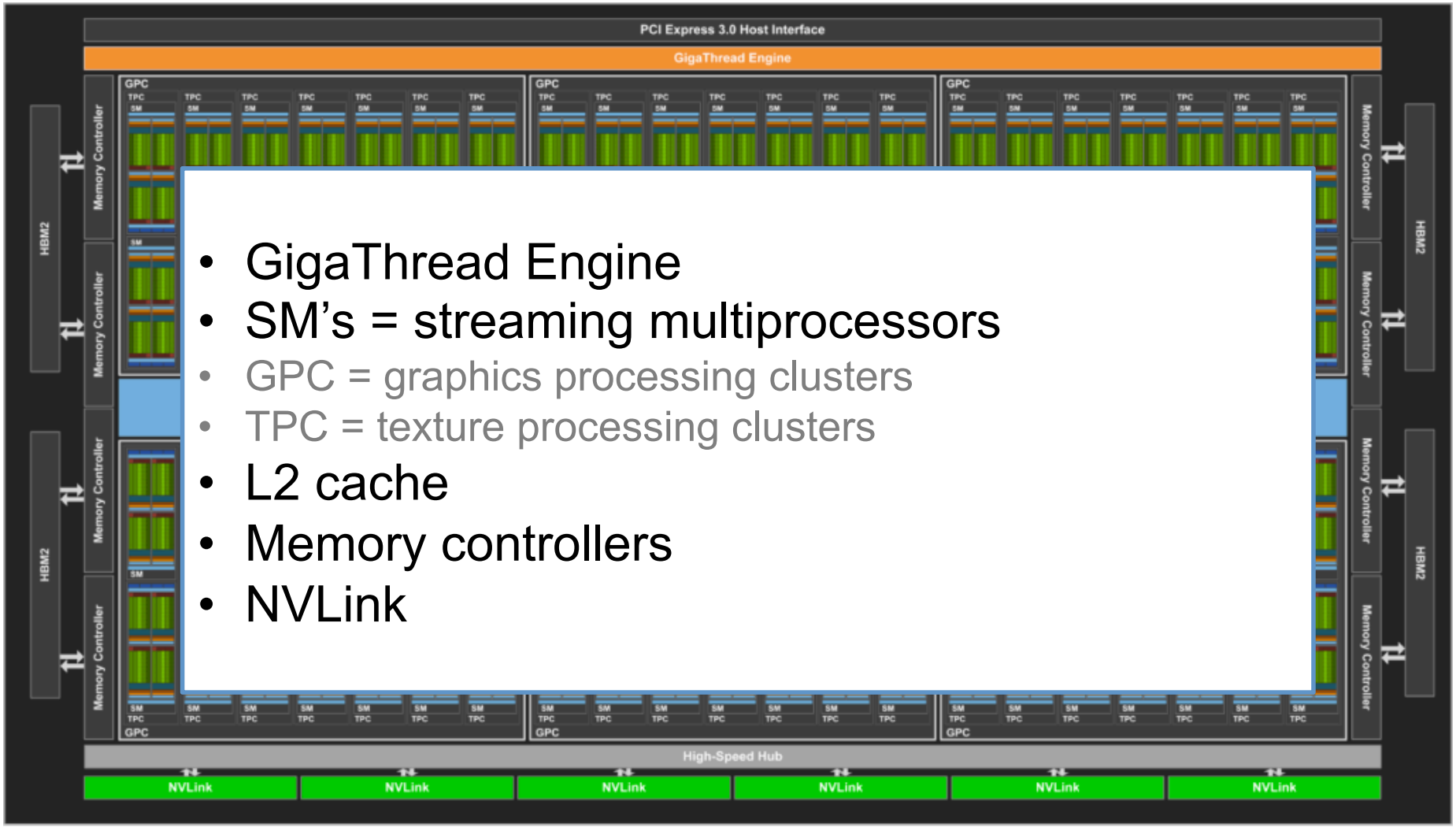
GPU Integration into the host system

- Typically PCI Express 3.0
- Theoretical speed 8 GB/s
  - Effective ≤ 6 GB/s
  - In reality: 4 – 6 GB/s
- V3.0 recently available
  - Double bandwidth
  - Less protocol overhead





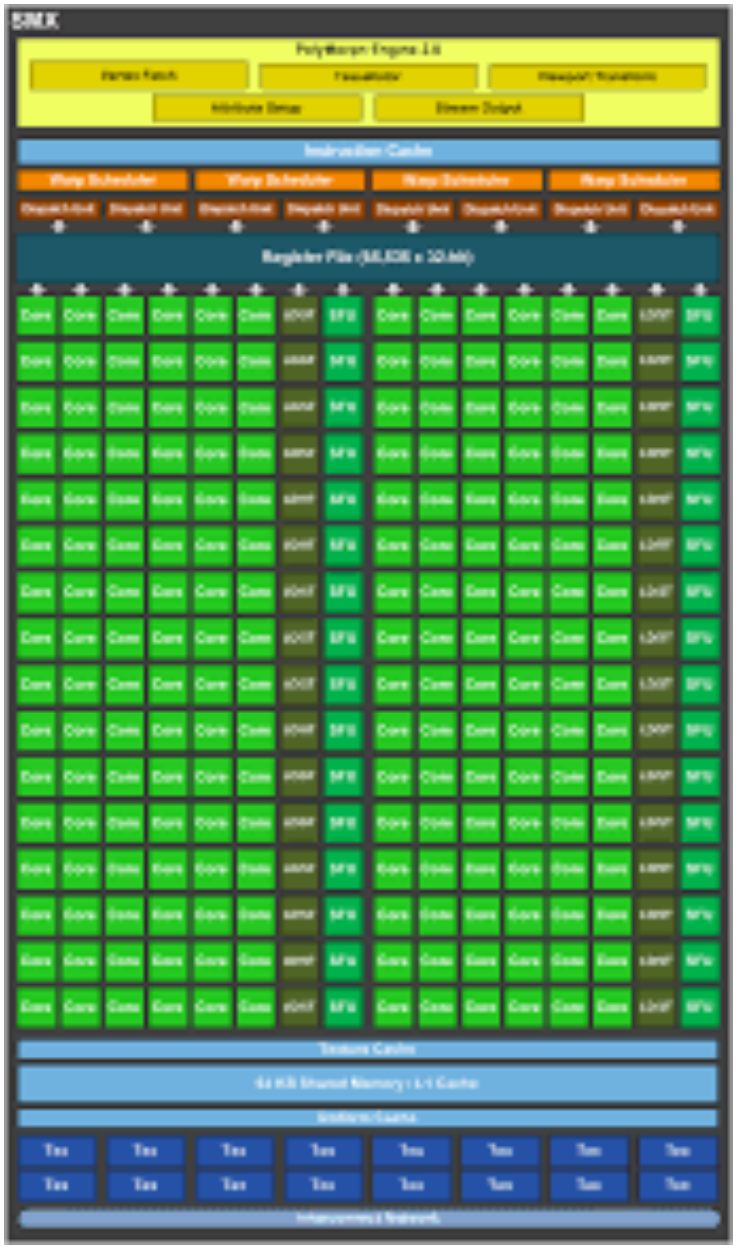
Inside an NVIDIA GPU architecture



- GigaThread Engine
- SM's = streaming multiprocessors
- GPC = graphics processing clusters
- TPC = texture processing clusters
- L2 cache
- Memory controllers
- NVLink

Inside an SM

- Different types of cores
  - CUDA Cores (INT/FP32)
  - LD/ST
  - Special function units
  - DP Units (Pascal)
  - Tensor units (Volta)
- Register file
- Warp scheduler
- Data caches
- Instruction buffers/caches
- Texture units



NVIDIA GPUs (8+ years)

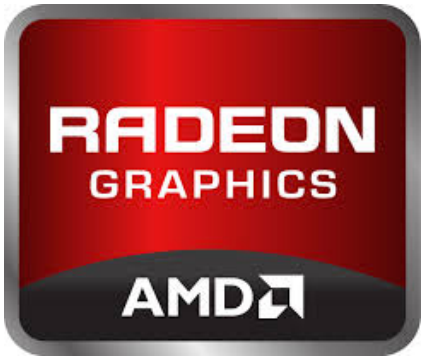
	Fermi	Kepler	Maxwell	Pascal	Volta
GPU	GTX480	GK180	GM200	GP100	GV100
Compute capability (CC)	2.x	3.5	5.2	6.0	7.0
SMs					80
TPC					40
FP3					64
FP6					32
Cloc					30
Peak FP32 [TFLOPs]					15.7
Peak FP64 [TFLOPs]	0.168	1.68	.21	5.3	7.8

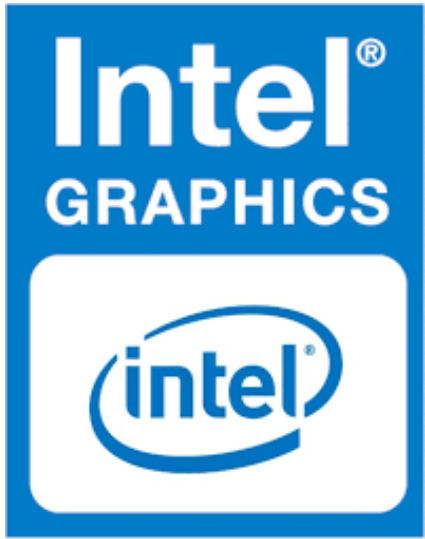
Tesla K40m uses GK110B, which has:

- 2880 cores (15 SMX x 192 cores/SMX)
- 700-900 MHz
- 12 GB RAM
- PCIe 3

Other players on the market

- **AMD** (former ATI)
  - Much better performance
  - Programmed using OpenCL (standard!)
  - Poorer software drivers and infrastructure (so far)
  - A lot less libraries and tools
  - Much smaller community effort
- **arm** (formerly ARM ☺)
  - Low-power devices (mobile platforms mostly)
  - Programmed using OpenCL
  - Lower performance than ATI and Intel, by choice
- **Intel**
  - To support own CPUs with integrated graphics
  - Programmed using OpenCL





All GPUs …

- Have a similar architecture
  - Massively parallel
  - Simple cores
  - Complex memory system
- Are programmed in a similar way
  - Fine-grain (SIMD/SIMT) parallelism
- Programming models ?
  - OpenCL is the de-facto standard for GPU programming
  - Lots of efforts for C++
  - Many other libraries and models on top of CUDA / OpenCL

Evolution in numbers

GPU / Form Factor	Kepler	Maxwell	Pascal	Pascal
	GK110 / PCIe	GM200 / PCIe	GP100 / SXM2	GP100 / PCIe
SMs	15	24	56	56
FP32 CUDA Cores / SM	192	128	64	64
FP32 CUDA Cores / GPU	2880	3072	3584	3584
FP64 CUDA Cores / SM	64	4	32	32
FP64 CUDA Cores / GPU	960	96	1792	1792
Base Clock	745 MHz	948 MHz	1328 MHz	1126 MHz
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1303 MHz
Single precision GFLOPS	5040	6844	10608	9340
Double precision GFLOPS	1680	213	5304	4670

Evolution in numbers

GPU / Form Factor	Kepler GK110 / PCIe	Maxwell GM200 / PCIe	Pascal GP100 / SXM2	Pascal GP100 / PCIe
Texture Units	240	192	224	224
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	3072-bit HBM2 (12GB) 4096-bit HBM2 (16GB)
Memory Bandwidth	288 GB/s	288 GB/s	732 GB/s	549 GB/s (12GB) 732 GB/s (16GB)
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	12 GB or 16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	4096 KB
Register File Size / SM	256 KB	256 KB	256 KB	256 KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	14336 KB
TDP	235 Watts	250 Watts	300 Watts	250 Watts
Transistors	7.1 billion	8 billion	15.3 billion	15.3 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2	610 mm^2
Manufacturing Process	28-nm	28-nm	16-nm	16-nm

PROGRAMMING MANY-CORES

Programming many-cores

= parallel programming:

- Choose/design algorithm
- Parallelize algorithm
  - Expose enough layers of parallelism
  - Minimize communication, synchronization, dependencies
  - Overlap computation and communication
- Implement parallel algorithm
  - Choose parallel programming model
  - (?) Choose many-core platform
- Tune/optimize application
  - Understand performance bottlenecks & expectations
  - Apply platform specific optimizations
  - (?) Apply application & data specific optimizations

Programming GPUs

- Low-level models
  - CUDA
  - OpenCL
  - and their variations:
    - pyCL, pyCUDA, jCUDA, …
- …. and increasing level of abstraction
  - SyCL (based on C++)
  - OpenACC (pragma-based, portable)
  - OpenMP (pragma-based, portable)
- … and going domain-specific
  - TensorFlow – machine learning/deep learning
  - Forma – DSL for image processing and stencils
  - … many more …

PROGRAMMING GPUS IN CUDA

Kernel = the parallel program

Device code = manage the parallel program

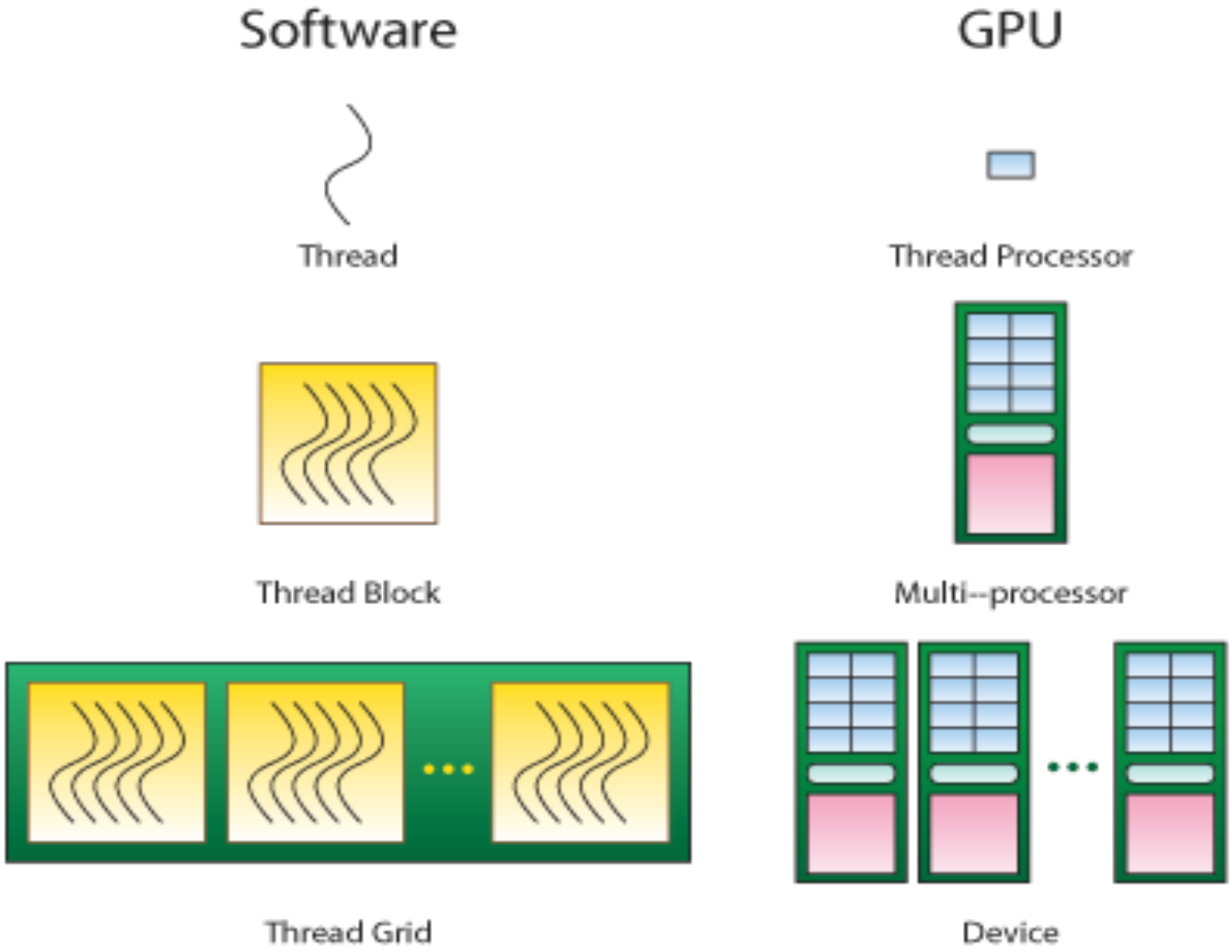
CUDA

- CUDA: Scalable parallel programming
  - C/C++ extensions
    - Other wrappers exist
- Straightforward mapping onto hardware
  - Hierarchy of threads (to map to cores)
    - Configurable at logical level
  - Various memory spaces (to map to physical spaces)
    - Usable via variable scopes
- Scale to 1000s of cores & 100,000s of threads
  - GPU threads are lightweight
  - GPUs need 1000s of threads for full utilization

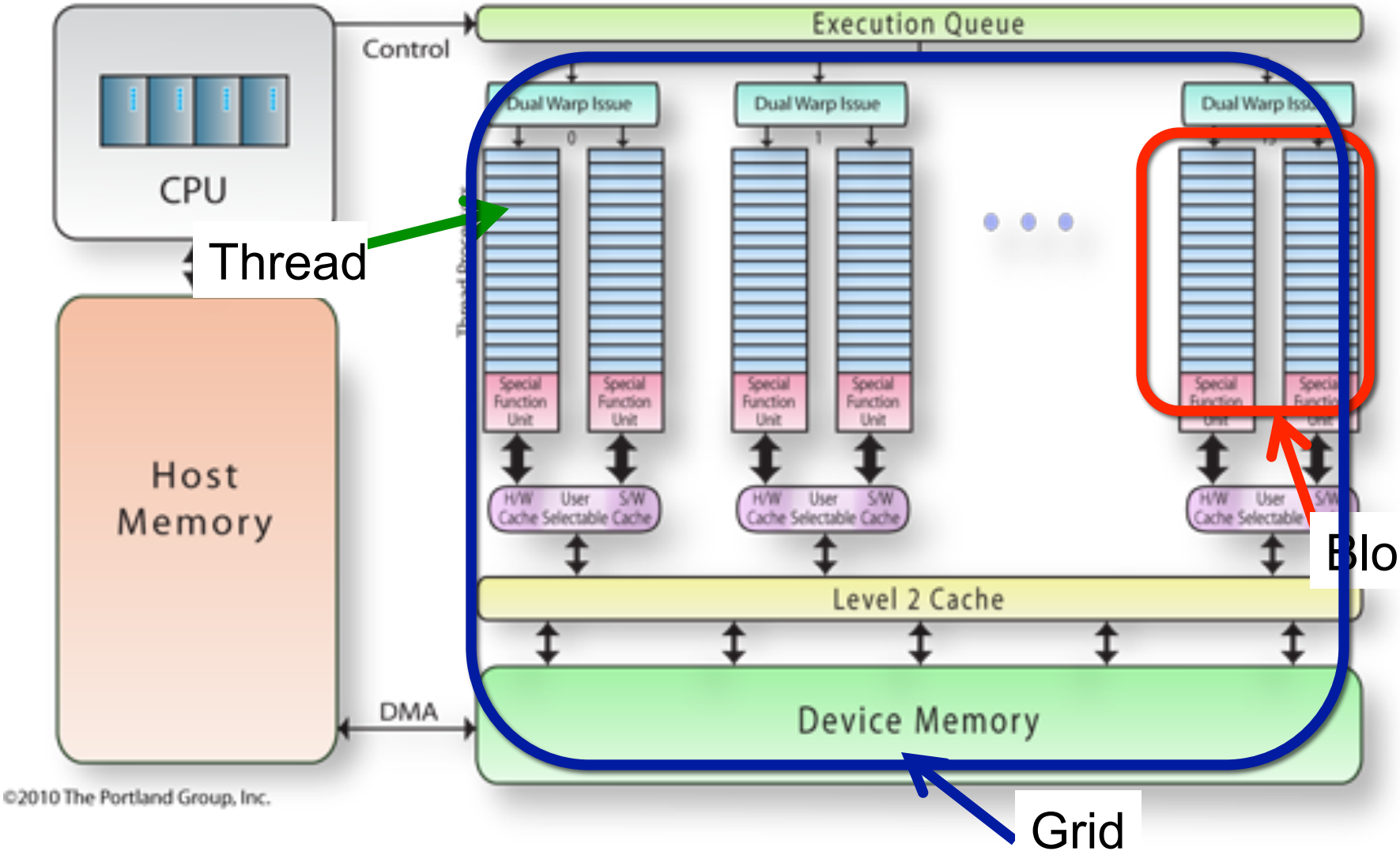
CUDA Model of Parallelism

- CUDA virtualizes the physical hardware
  - A block is a virtualized streaming multiprocessor
    - threads, shared memory
  - A thread is a virtualized scalar processor
    - registers, PC, state
- Threads are scheduled onto physical hardware without pre-emption
  - threads/blocks launch & run to completion
  - blocks must be independent

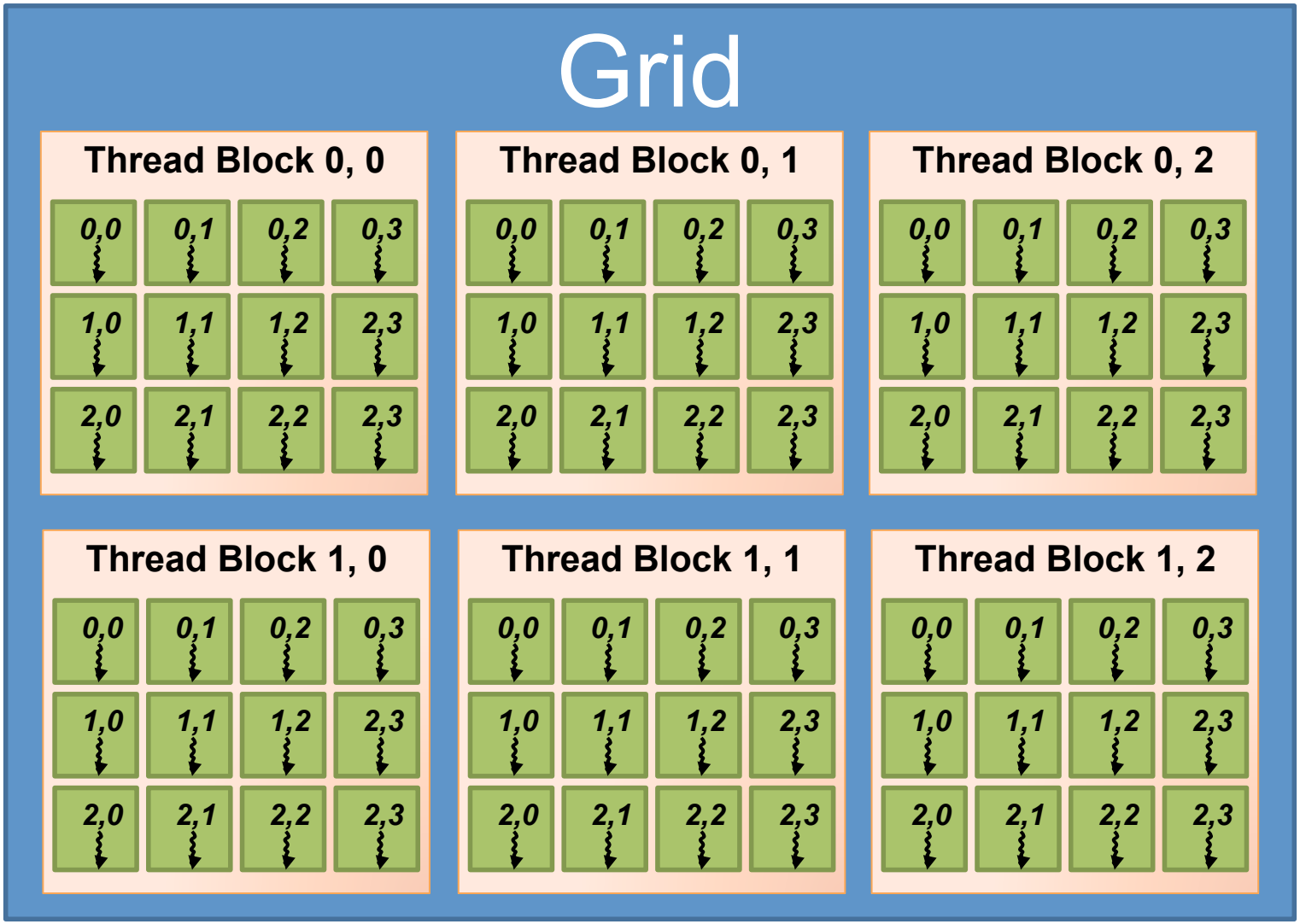
CUDA Model of Parallelism



Hierarchy of threads



Grids, Thread Blocks and Threads

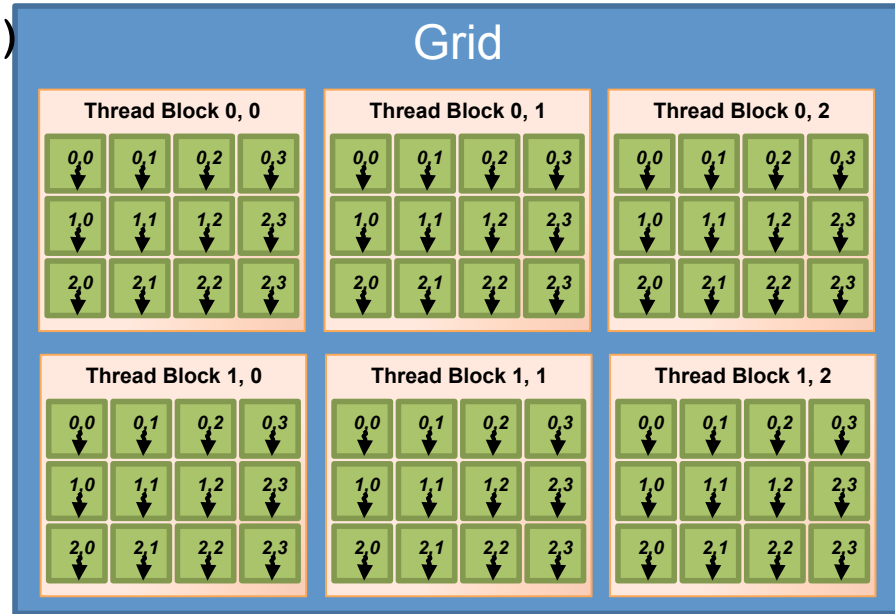


Kernels and grids

- Launch kernel (12 x 6 = 72 instances)

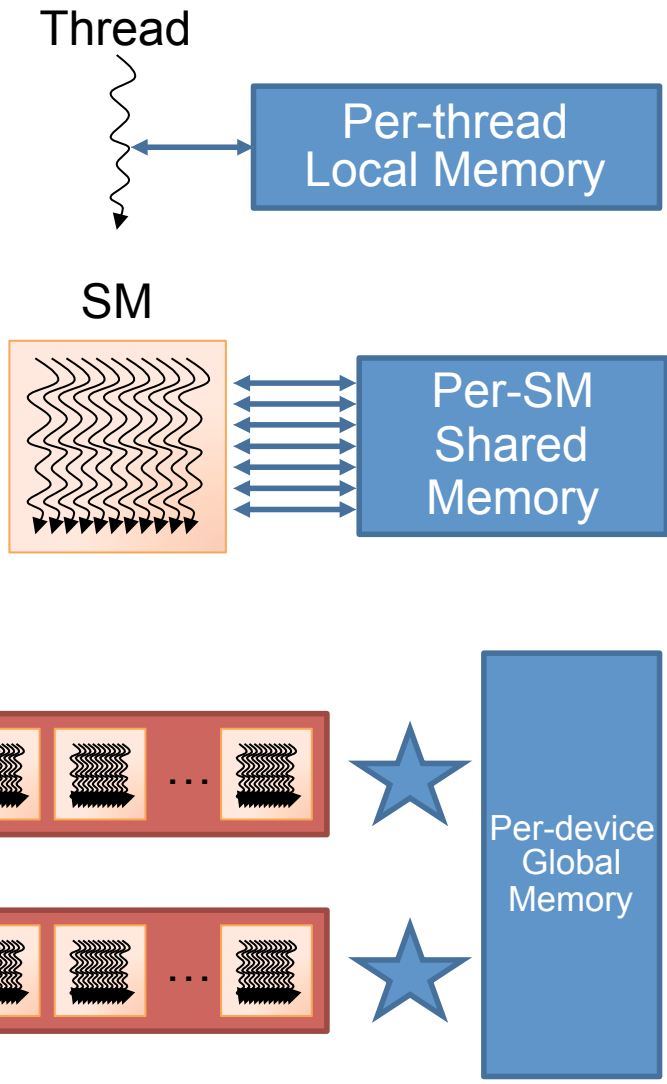
```
myKernel<<<numBlocks,threadsPerBlock>>>(…);
```

  - `dim3 threadsPerBlock(3,4);`
    - `threadsPerBlock.x = 3`
    - `threadsPerBlock.y = 4`
    - `Each thread:`
      `(threadIdx.x, threadIdx.y)`
  - `dim3 numBlocks(2,3);`
    - `blockDim.x = 2`
    - `blockDim.y=3`
    - `Each block :`
      `(blockIdx.x,blockIdx.y)`



Multiple Device Memory Scopes

- Per-thread *local memory*
  - Each thread has its own local memory.
  - Stores private data (in registers)
  - Accessible to a single thread only
- Per-SM *shared memory*
  - Each block has its own shared memory
  - Acts like an explicit **software cache**.
  - Stores data accessible to all threads in the same block.
- Device/*global memory*
  - GPU frame buffer
  - Accessible to any thread
- … and constant memory
  - Fast memory for read-only data



Memory spaces: Registers

Example:

```
__global__ void aKernel(float *C, float *A, float *B) {
    int tx = threadIdx.x; //local variable in registers
    float local_sum[4]; //small compile-time sized array
in registers
```

Registers:

- Thread-local scalars or small constant size arrays are stored as registers
- Implicit in the programming model
- Behavior is very similar with local variables
- Not persistent: kernel ends, data is lost

Memory spaces: global memory

Example:

```
__global__ void matmul_kernel(float *C, //C points to global memory
                              float *A, //A points to global memory
                              float *B) //B points to global memory
```

Global memory

- **Allocated by the host program using** cudaMalloc()
- **Initialized by the host program using** cudaMemcpy()**or previous kernels**
- Persistent = the values are retained between kernels
- Not coherent, writes by other threads might not be visible until kernel has finished

Memory spaces: Constant

Example

```
__constant__ float speed_of_light= 0.299792458; //scalars can be initialized directly
__constant__ float2 vertices[NUM_VERTICES]; //initialized by a host function
__global__ void cn_pnpoly(uint8_t* bitmap, float2* points, intn) {
...
for (intj=0; j<NUM_VERTICES; k = j++) {
float2 vj= vertices[j]; //index j does not depend on threadIdx
```

Constant memory:

- **Statically defined by the host program using** __constant__**qualifier**
- Defined as a global variable, visible only within the same translation unit
- **Initialized by the host program using** cudaMemcpyToSymbol()
- **Read-only to the GPU, cannot be accessed directly by the host**
- **Values are cached in a special cache optimized for broadcast access by multiple threads simultaneously, access should not depend on** threadIdx

Memory spaces: Shared

Example:

```
__global__ void matmul_kernel(float *C, float *A, float *B) {
__shared__ float sh_A[tile_size][tile_size]; //2D array in shared memory
for (k = 0; k < WIDTH; k += tile_size) {
__syncthreads(); //wait for all threads in the block
sA[ty][tx] = A[y*WIDTH + k + tx]; //fill shared memory with values
__syncthreads(); //wait again
```

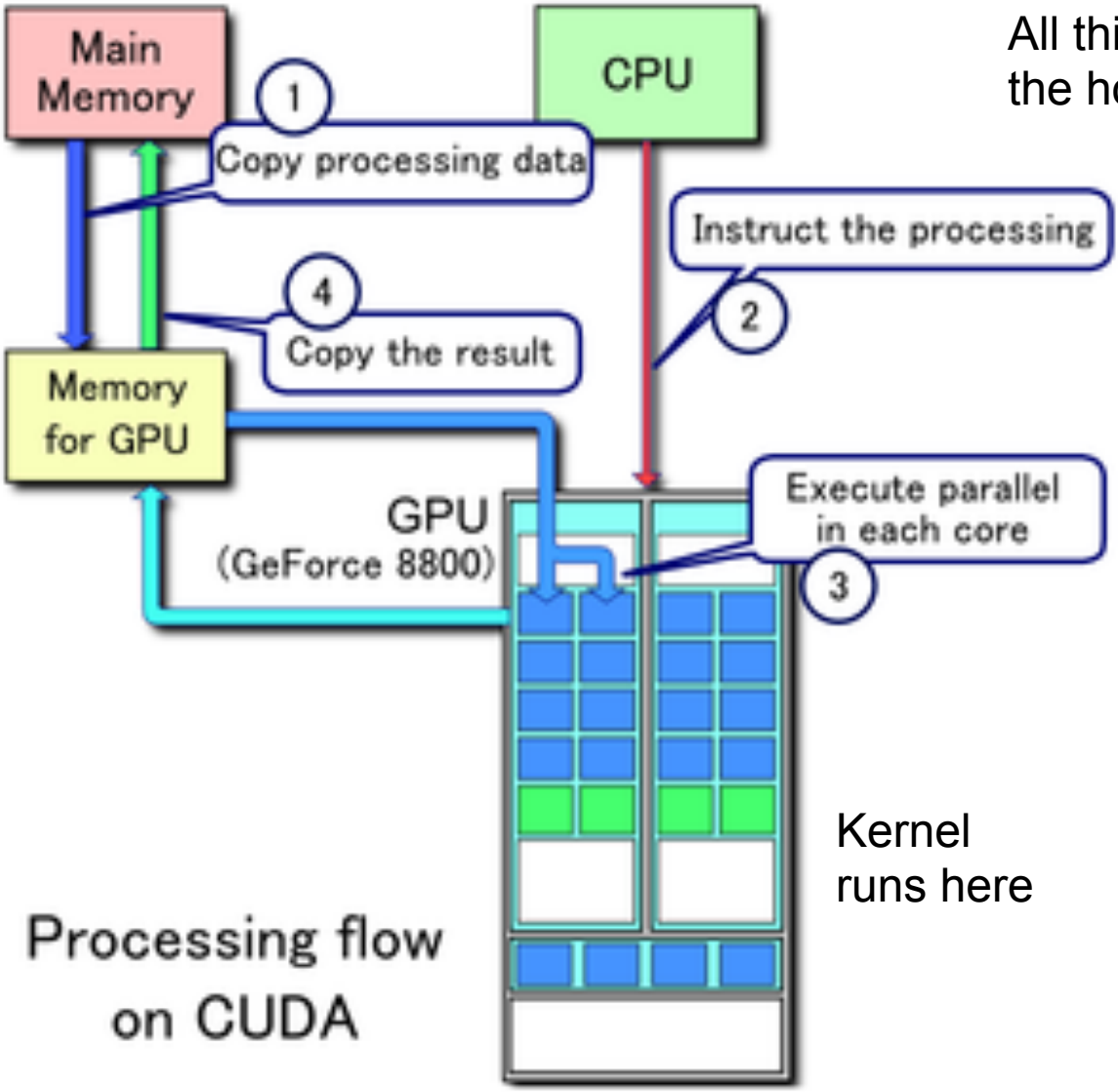
Shared memory

- Variables have to be declared using __shared__qualifier, size known at compile time
- In the scope of thread block, all threads in a thread block see the same piece of memory
- Not initialized, threads have to fill shared memory with meaningful values
- Not persistent, after the kernel has finished, values in shared memory are lost
- Not coherent, __syncthreads()is required to make writes visible to other threads within the thread block

Using CUDA

- Two parts of the code:
  - Device code = GPU code = kernel(s)
    - Sequential program
    - Write for 1 thread, execute for all
  - Host code = CPU code
    - Instantiate grid + run the kernel
    - Memory allocation, management, deallocation
    - C/C++/Java/Python/…
- Host-device communication
  - Explicit / implicit via PCI/e
  - Minimum: data input/output

Processing flow



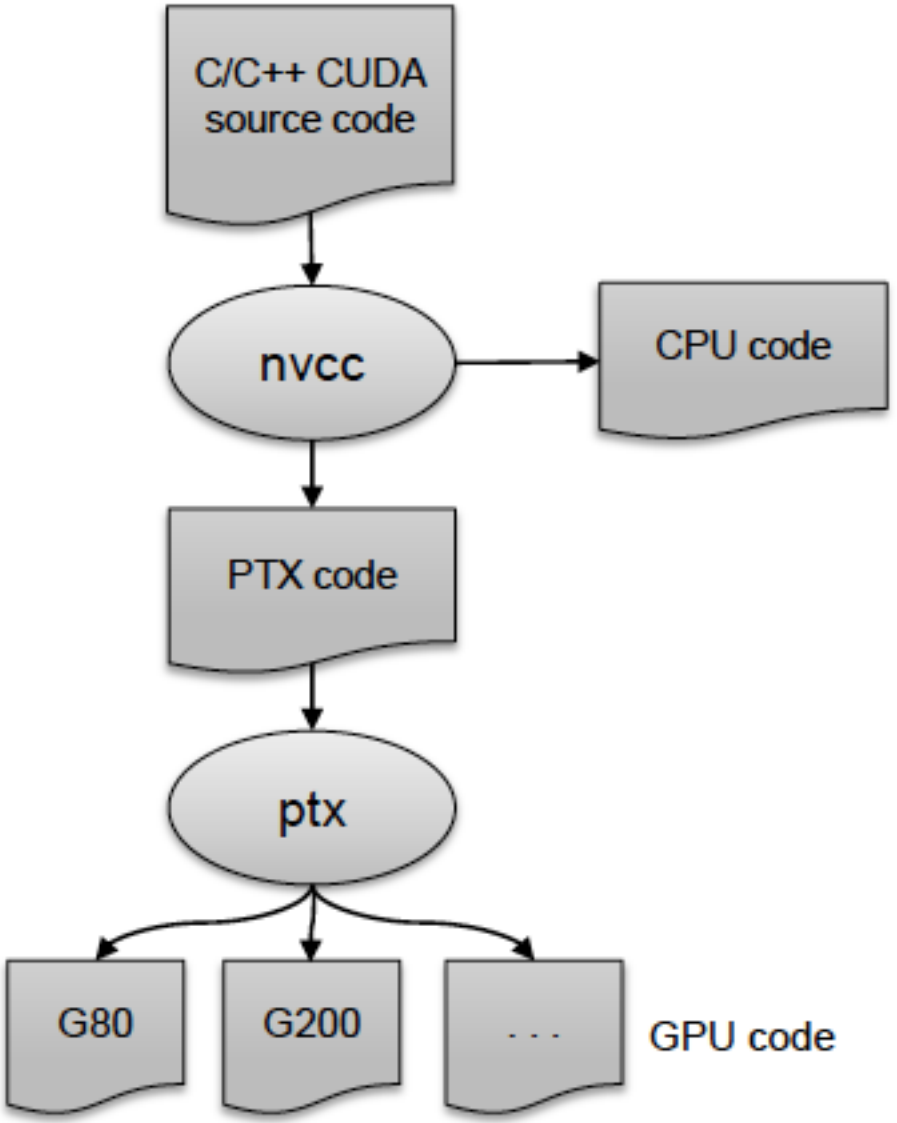
All this happens from the host code.

Kernel runs here

Image courtesy of Wikipedia

Compiling CUDA

- nvcc is a compiler driver
- Separates source code into:
  - device code (runs on GPU)
    - further processed by NVIDIA compi
  - host code (runs on CPU)
    - further processed by host compiler



CUDA: dummy example, host side

```
int n = 1024;
int nbytes = n * sizeof(int);
int* dataCPU = (int *)malloc(nbytes);
int* dataGPU;

cudaMalloc(&dataGPU, nbytes);
cudaMemset(dataGPU, 0, nbytes);

cudaMemcpy(dataGPU, dataCPU, nbytes,
                          cudaMemcpyHostToDevice);
myKernel<<<n/128,128>>>(n, dataGPU);
cudaMemcpy(dataCPU, dataGPU, nbytes,
                          cudaMemcpyDeviceToHost);
cudaFree(dataGPU);
free(dataCPU);
```

CUDA: Memory Allocation/Release

- All memory buffers – CPU **and** GPU must be allocated
- Host (CPU) manages device (GPU) memory:
  - `cudaMalloc(void **pointer, size_t nbytes)`
  - `cudaMemset(void *pointer, int val, size_t count)`
  - `cudaFree(void* pointer)`

CUDA: Data Copies

```
cudaMemcpy(void *dst, void *src,
           size_t nbytes,
           enum cudaMemcpyKind direction);
```

- blocks CPU thread until all bytes have been copied
- doesn't start copying until previous CUDA calls complete

- ```
  enum {
    cudaMemcpyHostToDevice,
    cudaMemcpyDeviceToHost,
    cudaMemcpyDeviceToDevice
    } cudaMemcpyKind
  ```

- Non-blocking copies are also available
  - `cudaMemcpyAsync`
  - DMA transfers, overlap computation and communication
  ```

CUDA: dummy example, kernel side

```
__global__ void myKernel(int n, int* data) {
    int whoAmI = … //determine my ID
    data[whoAmI] = my_device_func(n,data[whoAmI]);
}


__device__ int my_device_func(int x, int y) {
     return x*y;
}
```

CUDA: kernels and launch

- Function qualifiers:

```
__global__ void my_kernel() { }
__device__ float my_device_func() { }
```

- Execution configuration:

```
dim3 myGridSize(100,50); // 5000 thread blocks
dim3 myBlockSize(4,8,8); // 256 threads per block
my_kernel <<< gridDim, blockDim >>> (...); // Launch kernel
```

- Built-in variables and functions valid in device code:

```
dim3 gridDim;   // Grid   dimension
dim3 blockDim;  // Block  dimension
dim3 blockIdx;  // Block  index
dim3 threadIdx; // Thread index
```

EXAMPLE: VECTOR-ADD

Programming many-cores

= parallel programming:

- Choose/design algorithm
- Parallelize algorithm
  - Expose enough layers of parallelism
  - Minimize communication, synchronization, dependencies
  - Overlap computation and communication
- Implement parallel algorithm
  - Choose parallel programming model
  - (?) Choose many-core platform
- Tune/optimize application
  - Understand performance bottlenecks & expectations
  - Apply platform specific optimizations
  - (?) Apply application & data specific optimizations

First CUDA program

- Determine mapping of operations and data to threads
- Write kernel(s)
  - Sequential code
  - Written per-thread
- Determine block geometry
  - Threads per block, blocks per grid
  - Number of grids (>= number of kernels)
- Write host code
  - Memory initialization and copying to device
  - Kernel(s) launch(es)
  - Results copying to host
- Optimize the kernels

Vector add: sequential

```
void vector_add(int size, float* a, float* b, float* c) {
    for(int i=0; i<size; i++) {
        c[i] = a[i] + b[i];
    }
}
```

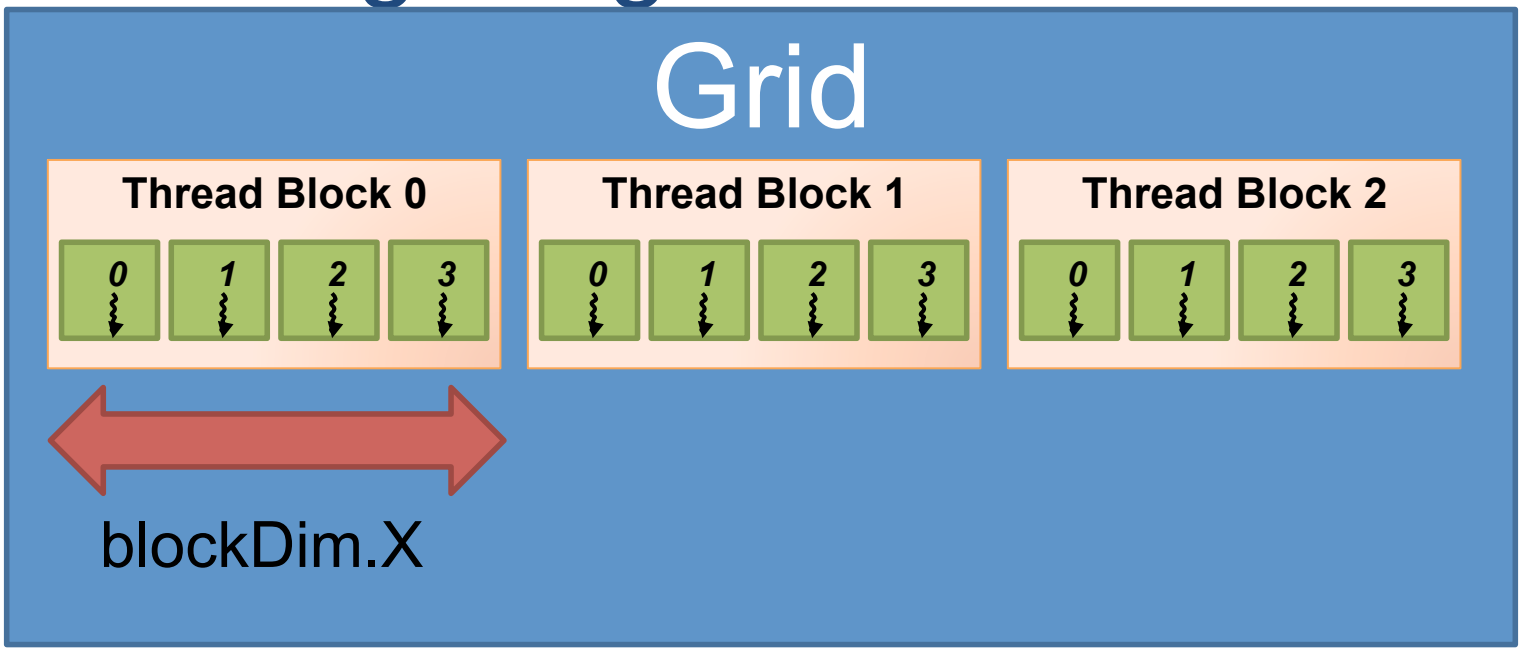
How do we parallelize this?

- What does each thread compute?
  - One addition per thread
  - Each thread deals with *different* elements
  - How do we know which element?
    - Compute a mapping of the grid to the data
      - Any mapping will do!

Vector add: Kernel

```
// compute vector sum c = a + b
// each thread performs one pair-wise addition
__global__ void vector_add(float* A, float* B, float* C) {
    int i = ?
    C[i] = A[i] + B[i];
}
```

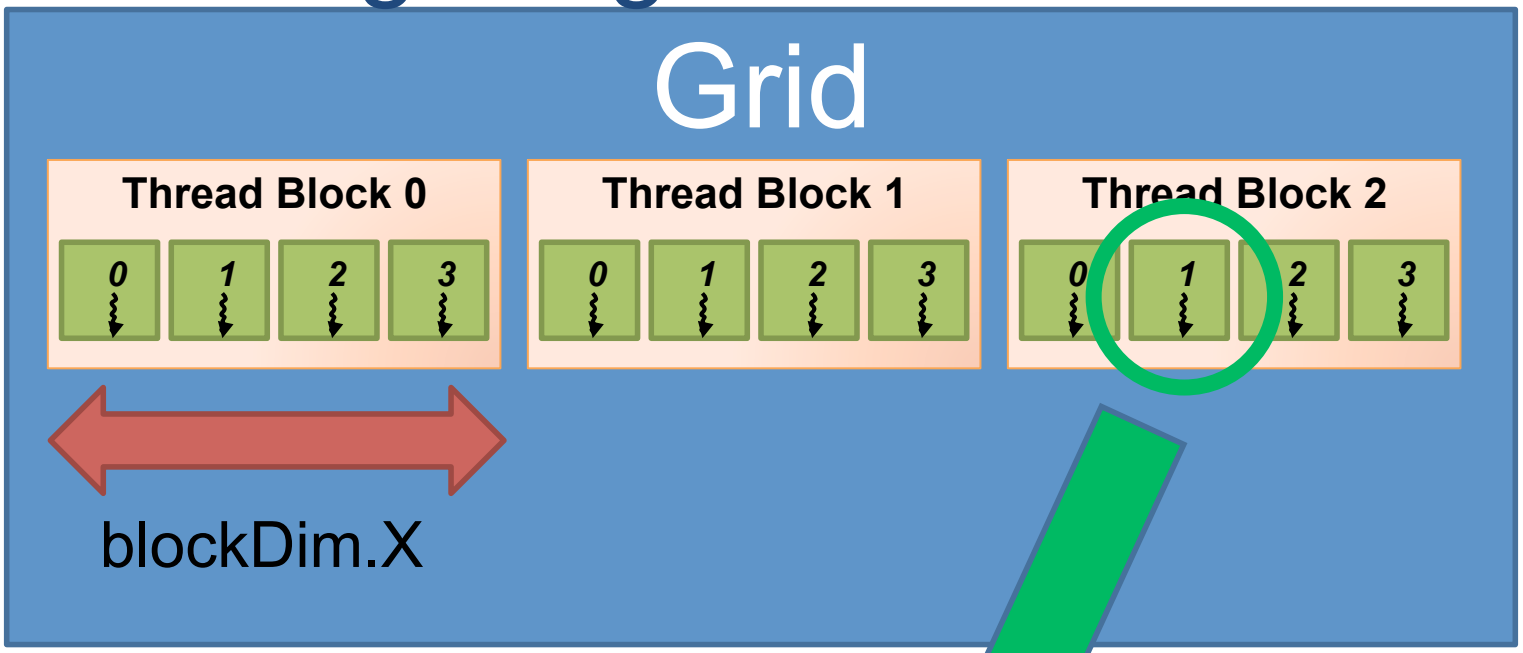
Calculating the global thread index



"global" thread index:

```
blockDim.x * blockIdx.x + threadIdx.x;
```

Calculating the global thread index



"global" thread index:

```
blockDim.x * blockIdx.x + threadIdx.x;
```

```
4          *          2          +          1          = 9
```

Vector add: Kernel

```
// compute vector sum c = a + b
// each thread performs one pair-wise addition
__global__ void vector_add(float* A, float* B, float* C) {
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    C[i] = A[i] + B[i];
}
```

Done with the kernel!

Vector add: Launch kernel

```
// compute vector sum c = a + b
// each thread performs one pair-wise addition
__global__ void vector_add(float* A, float* B, float* C) {
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    C[i] = A[i] + B[i];
}
```

GPU code

```
int main() {
  // initialization code here ...
  N = 5120;
  // launch N/256 blocks of 256 threads each
  vector_add<<< N/256, 256 >>>(deviceA, deviceB, deviceC);
  // cleanup code here ...
}
```

Host code

(in the same file)

Vector add: Launch kernel

```
// compute vector sum c = a + b
// each thread performs one pair-wise addition
__global__ void vector_add(float* A, float* B, float* C) {
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    C[i] = A[i] + B[i];
}
```

GPU code

```
int main() {
  // initialization code here ...
  N = 5000;  // <- what happens?
  // launch N/256 blocks of 256 threads each
  vector_add<<< N/256, 256 >>>(deviceA, deviceB, deviceC);
  // cleanup code here ...
}
```

Host code

(in the same file)

Vector add: Launch kernel

```
// compute vector sum c = a + b
// each thread performs one pair-wise addition
__global__ void vector_add(float* A, float* B, float* C) {
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    if (i<N) C[i] = A[i] + B[i];
}
```

GPU code

```
int main() {
  // initialization code here ...
  N = 5000;  // <- what happens?
  // launch N/256 blocks of 256 threads each
  vector_add<<< N/256+1, 256 >>>(deviceA, deviceB, deviceC);
  // cleanup code here ...
}
```

Host code

(in the same file)

Vector add: Host

```
int main(int argc, char** argv) {
  float *hostA, *deviceA, *hostB, *deviceB, *hostC,
*deviceC;
  int size = N * sizeof(float);

  // allocate host memory
  hostA = malloc(size);
  hostB = malloc(size);
  hostC = malloc(size);

  // initialize A, B arrays here...

  // allocate device memory
  cudaMalloc(&deviceA, size);
  cudaMalloc(&deviceB, size);
  cudaMalloc(&deviceC, size);
```

Vector add: Host

```
 // transfer the data from the host to the device
 cudaMemcpy(deviceA, hostA, size,
cudaMemcpyHostToDevice);
 cudaMemcpy(deviceB, hostB, size,
cudaMemcpyHostToDevice);

 // launch N/256 blocks of 256 threads each
 vector_add<<<N/256, 256>>>(deviceA, deviceB,
deviceC);

 // transfer the result back from the GPU to the
host
 cudaMemcpy(hostC, deviceC, size,
cudaMemcpyDeviceToHost);
}
```

Done with the host code!

Vector add: OpenACC

```
void vector_add(int size, float* a, float* b, fl
#pragma acc kernels,
   copyin(a[0:n],b[0:n]), copyout(result[0:n])
    for(int i=0; i<size; i++) {
        c[i] = a[i] + b[i];
    }
}
```

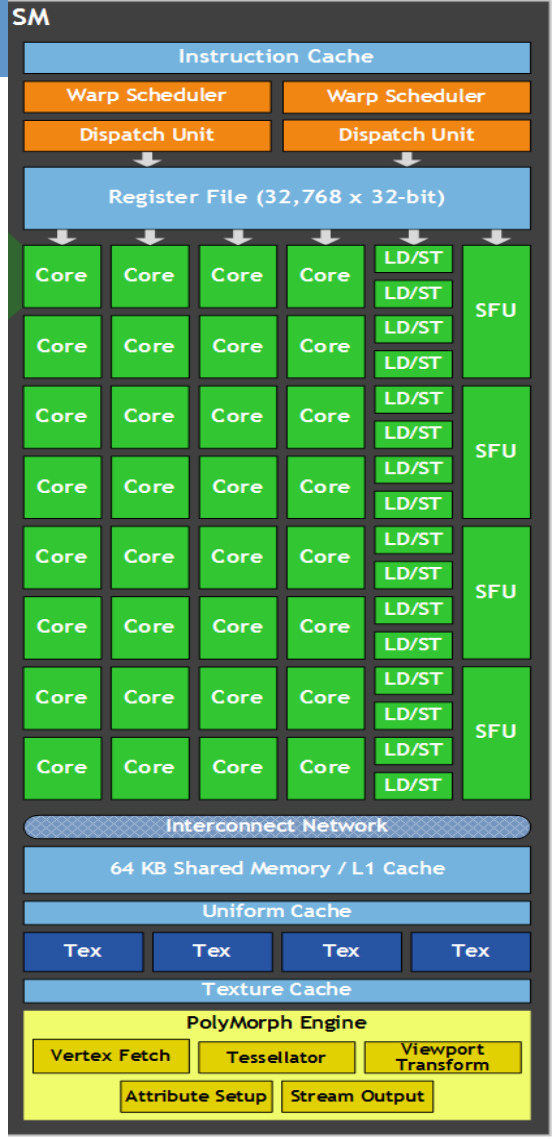
ADVANCED CONCEPTS

Thread Scheduling

- Order of threads within a block is undefined!
  - Threads are grouped in warps (32 threads/warp)
    - AMD calls it “a wavefront” (64 threads/wavefront)
- Order in which thread blocks are mapped and scheduled is undefined!
  - Blocks run to completion on one SM without preemption
  - Can run in any order
    - Any possible interleaving of blocks should be valid
  - Can run concurrently OR sequentially

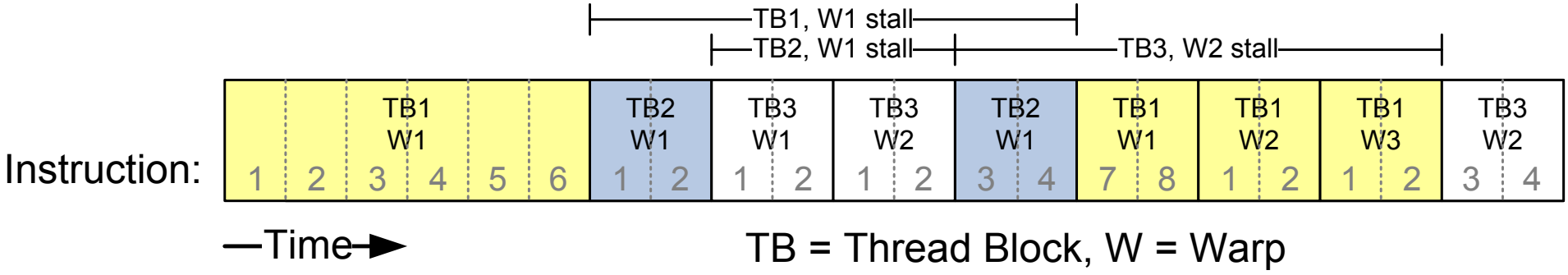
Warps = 32 threads

- Threads are scheduled in warps
  - AMD calls them “wavefronts”
- One warp => on one SM
  - Same SM till completion
- Scheduling
  - GigaThread Unit : schedules blocks per SM’s
  - Inside SM: warp scheduler(s) + instruction dispatcher
    - Replace warps that are stalled by warps waiting to compute
    - Very fast context switching



Thread Scheduling

- SMs implement zero-overhead warp scheduling
  - A warp is a group of 32 threads that runs concurrently on an SM
  - At any time, the number of warps concurrently executed by an SM is limited by its number of cores.
  - Warps whose next instruction has its inputs ready for consumption are eligible for execution
  - Eligible Warps are selected for execution on a prioritized scheduling policy
  - All threads in a warp execute the same instruction when selected



Stalling warps

- What happens if all warps are stalled?
  - No instruction issued → performance lost
- Most common reason for stalling?
  - Waiting on global memory
- If your code reads global memory every couple of instructions
  - You should try to maximize **occupancy**

CUDA: THREAD DIVERGENCE

Thread divergence

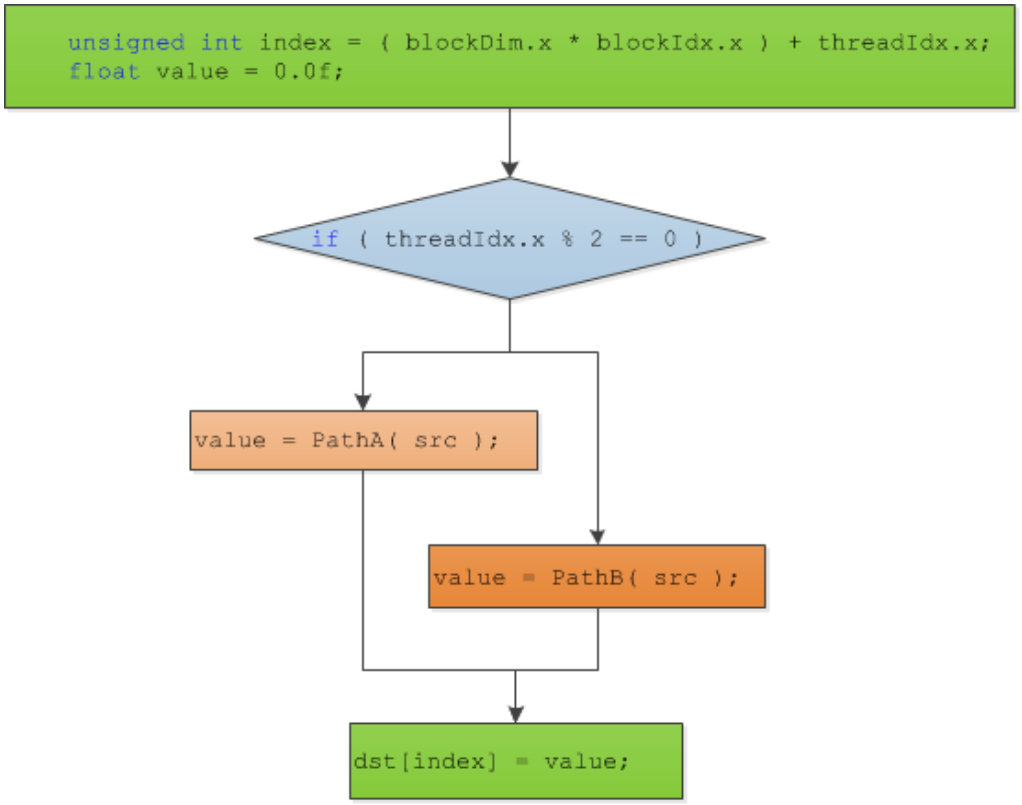
“I heard GPU branching is expensive. Is this true?”

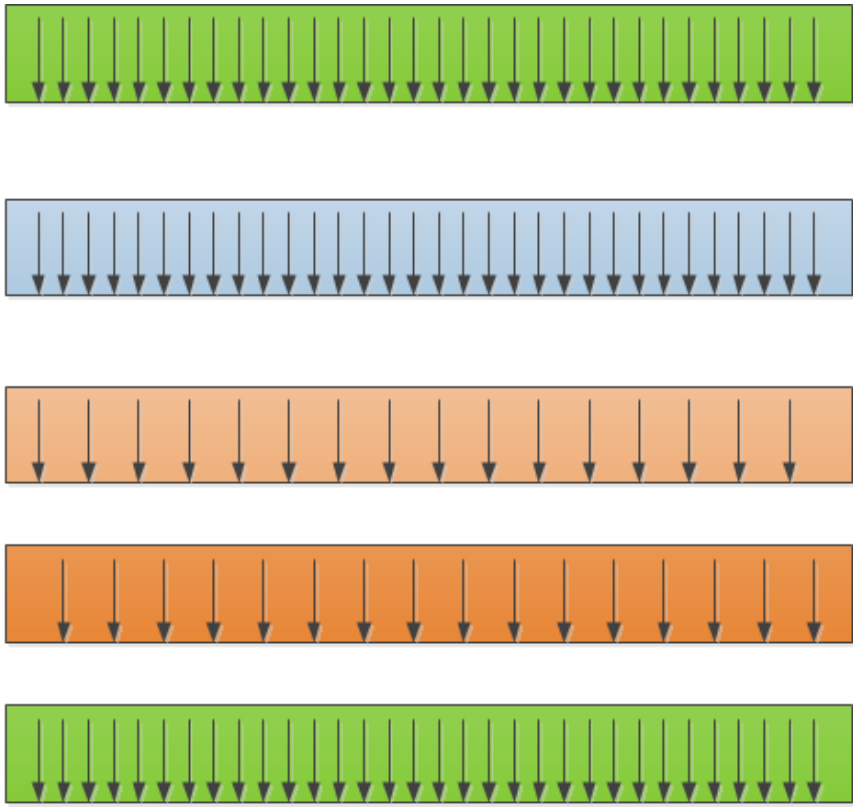
```
__global__ void Divergence(float* dst,float* src )
{
    float value = 0.0f;

    if ( threadIdx.x % 2 == 0 )
// active threads : 50%
        value = src[0] + 5.0f;
    else
// active threads : 50%
        value = src[0] – 5.0f;

    dst[index] = value;
}
```

Execution

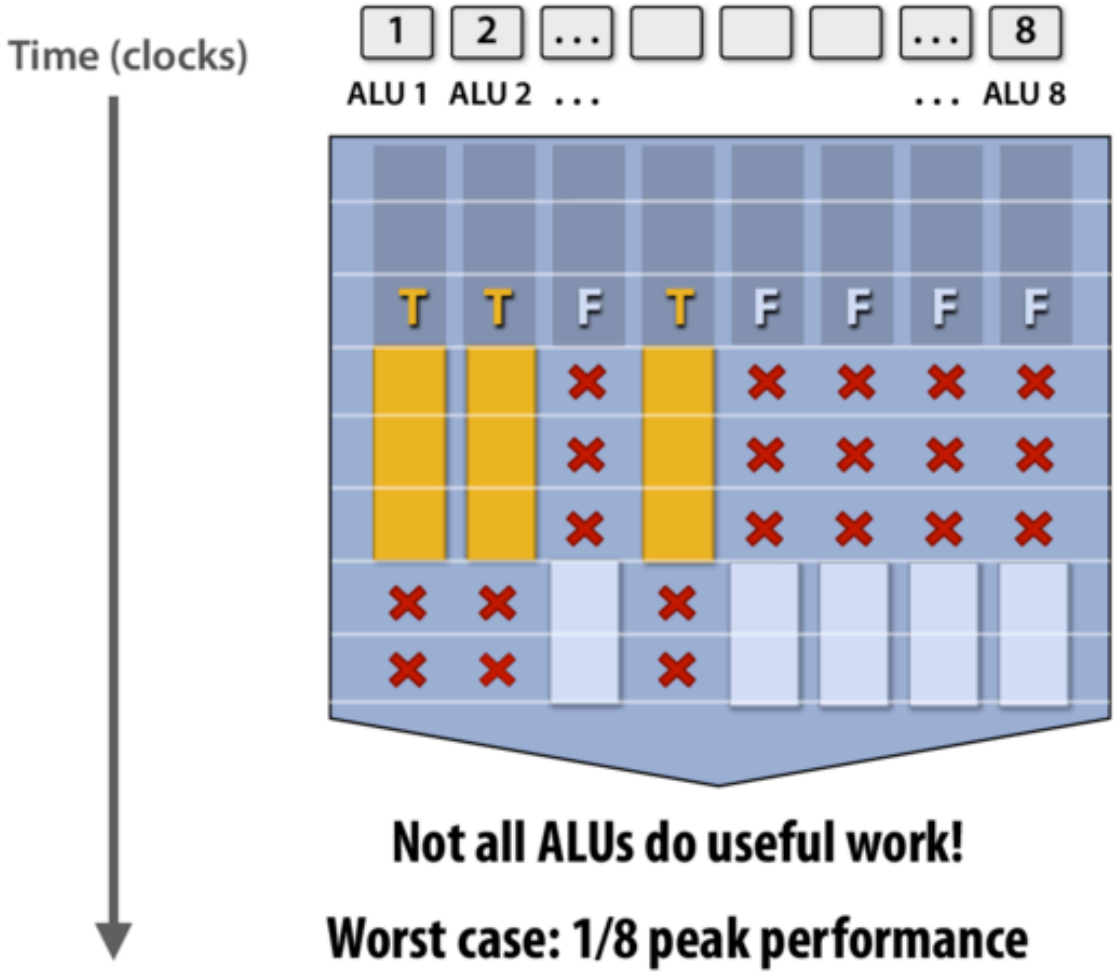




Worst case performance loss:

50% compared with the non divergent case.

Another example



(assume logic below is to be executed for each element in input array 'A', producing output into the array 'result')

```
<unconditional code>

float x = A[i];

if (x > 0) {
   float tmp = exp(x,5.f);
   tmp *= kMyConst1;
   x = tmp + kMyConst2;
} else {
   float tmp = kMyConst1;
   x = 2.f * tmp;
}

<resume unconditional code>

result[i] = x;
```

Performance penalty?

- Depends on the amount of divergence
  - Worst case: 1/32 performance
    - When each thread does something different
- Depends on whether branching is data- or ID- dependent
  - If ID – consider grouping threads differently
  - If data – consider sorting
- Non-diverging warps => NO performance penalty
  - In this case, branches are not expensive …

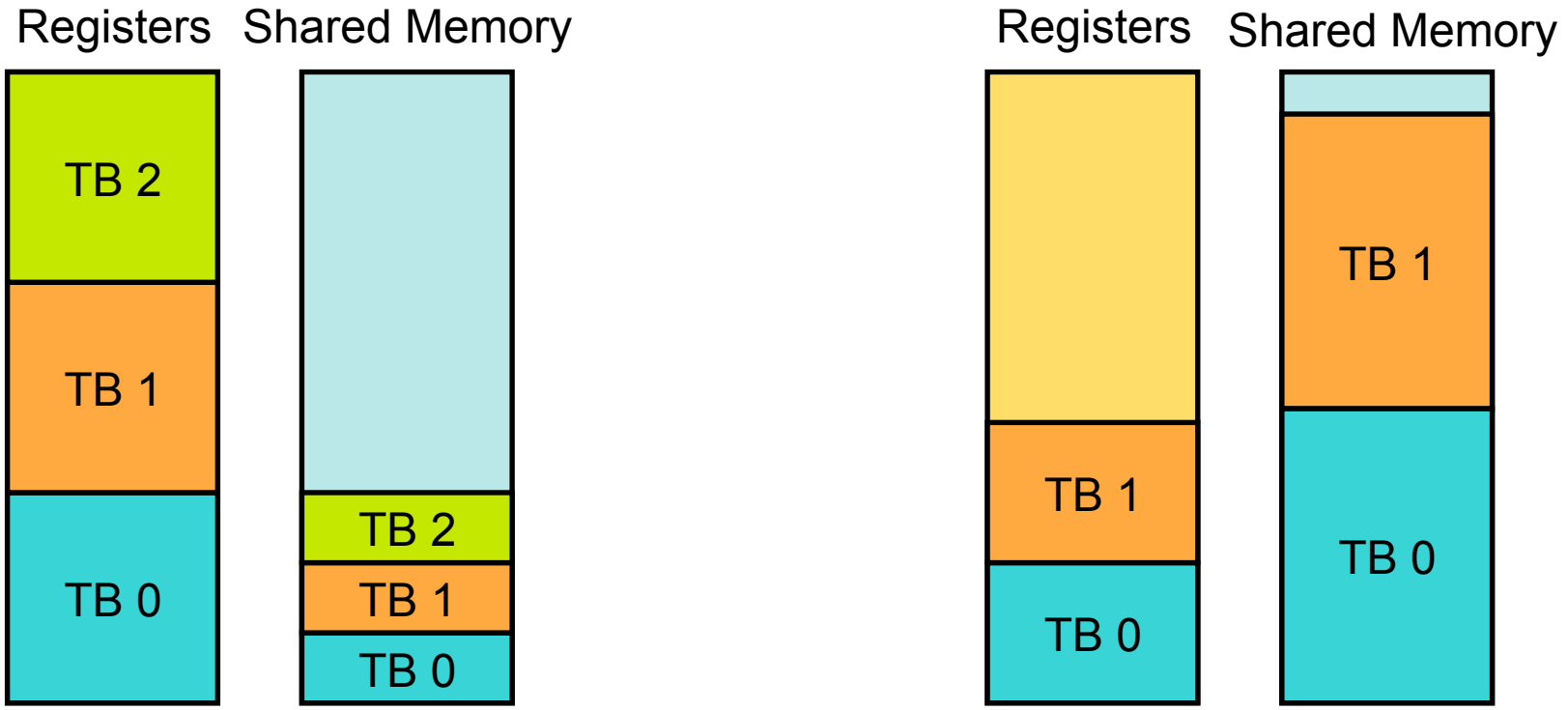
CUDA: OCCUPANCY

Occupancy

Occupancy = Active Warps / Maximum Active Warps

- Remember: resources are allocated for the entire block!
- Resources are finite
  - Utilizing too many resources per thread may limit the occupancy
- Potential occupancy limiters:
  - Register usage
  - Shared memory usage
  - Block size

Resource Limits



- Pool of registers and shared memory per SM
  - Each thread block grabs registers & shared memory
  - If one or the other is fully utilized => no more thread blocks

How do you know what you're using?

- Use compiler flags to get register and shared memory usage
  - "`nvcc -Xptxas -v`"
- Use the NVIDIA Profiler
- Plug those numbers into CUDA Occupancy Calculator

- Maximize occupancy for improved performance
  - Empirical rule! Don't overuse!

CUDA_Occupancy_calculator.xlsm - Microsoft Excel

Security Warning Macros have been disabled. Options...

MyRegCount | 25

1.) Select Compute Capability (click):	1.3	(Help)	
2.) Enter your resource usage:			
Threads Per Block	128	(Help)	
Registers Per Thread	25		
Shared Memory Per Block (bytes)	640		
(Don't edit anything below this line)			
3.) GPU Occupancy Data is displayed here and in the graphs:			
Active Threads per Multiprocessor	512	(Help)	
Active Warps per Multiprocessor	16		
Active Thread Blocks per Multiprocessor	4		
Occupancy of each Multiprocessor	50%		

Physical Limits for GPU Compute Capability:	1.3
Threads per Warp	32
Warps per Multiprocessor	32
Threads per Multiprocessor	1024
Thread Blocks per Multiprocessor	8
Total # of 32-bit registers per Multiprocessor	16384
Register allocation unit size	512
Register allocation granularity	block
Shared Memory per Multiprocessor (bytes)	16384
Shared Memory Allocation unit size	512
Warp allocation granularity (for register allocation)	2

Allocation Per Thread Block	
Warps	4
Registers	3584
Shared Memory	1024

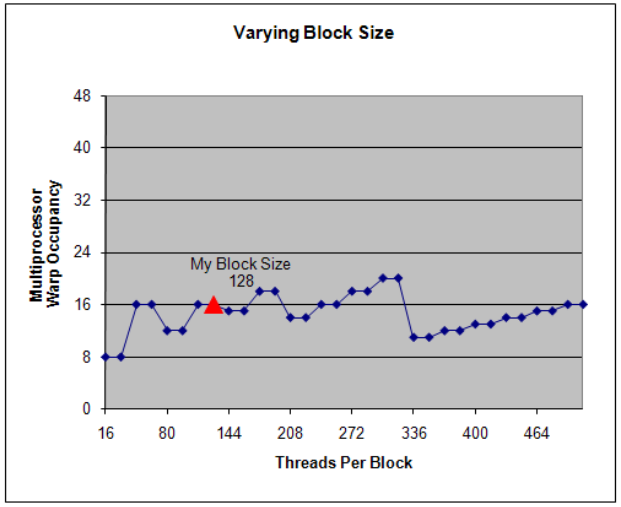
These data are used in computing the occupancy data in blue

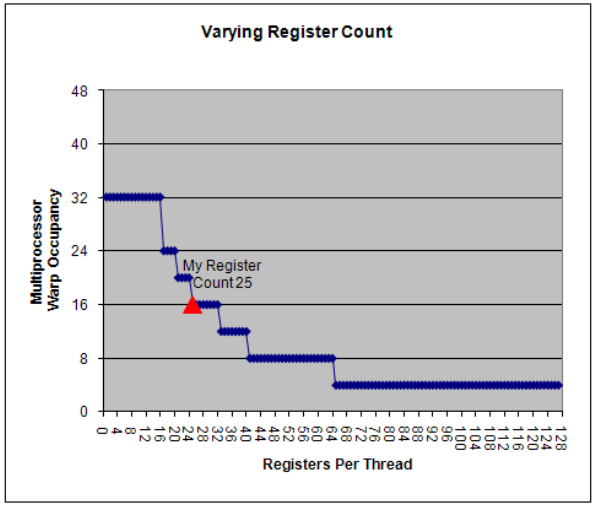
Maximum Thread Blocks Per Multiprocessor	Blocks
Limited by Max Warps / Blocks per Multiprocessor	8
Limited by Registers per Multiprocessor	4
Limited by Shared Memory per Multiprocessor	16

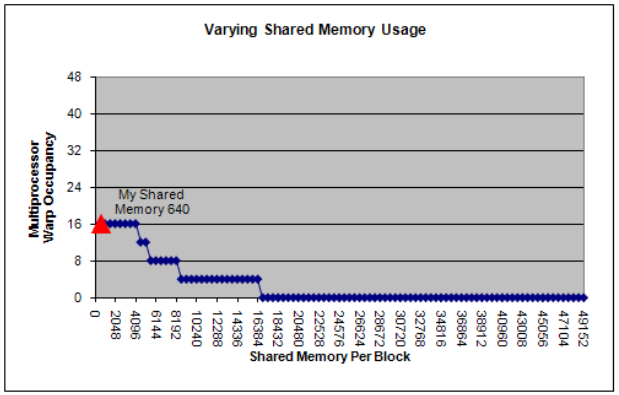
Thread Block Limit Per Multiprocessor highlighted RED

CUDA Occupancy Calculator	
Version:	2.0
Copyright and License	

The other data points represent the range of possible block sizes, register counts, and shared memory allocation.







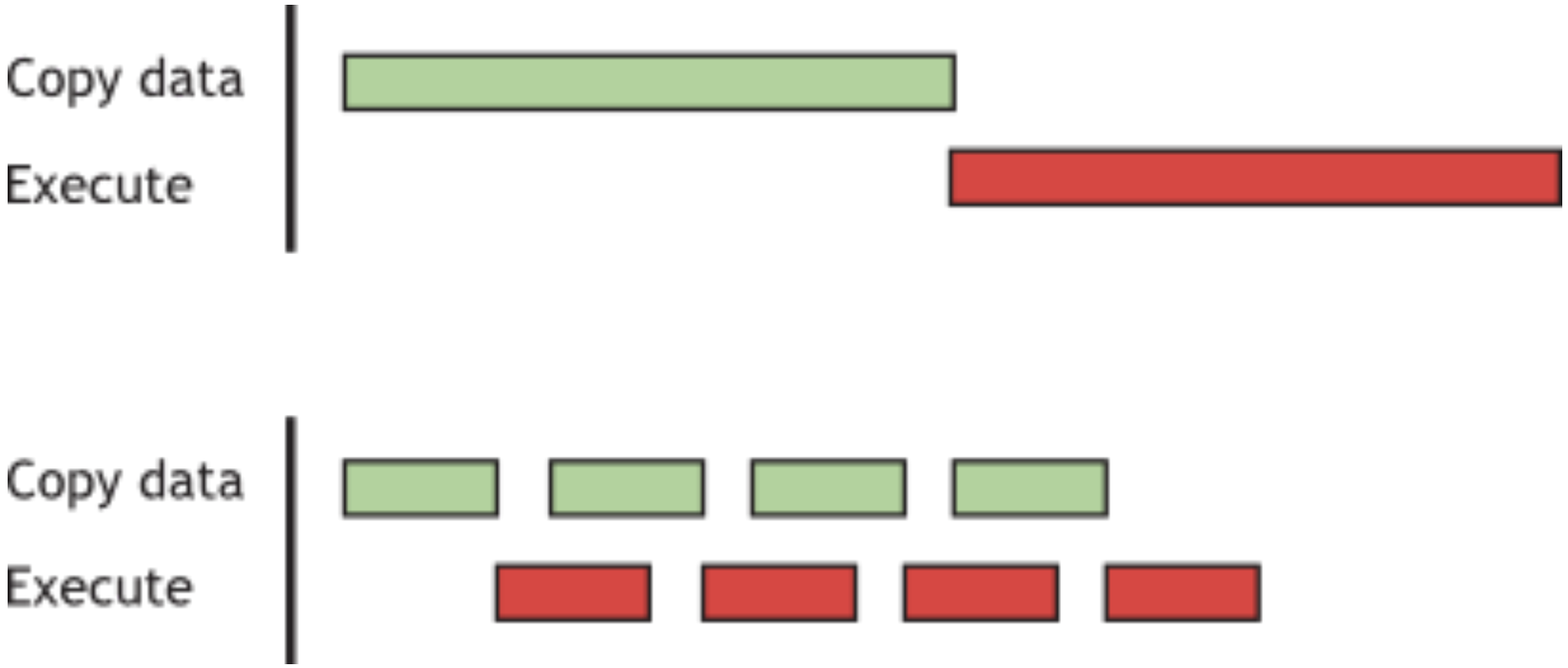
Calculator | Help | GPU Data | Copyright & License

Ready

CUDA: STREAMS

Overlap Computation and Communication

- Main idea: while executing a kernel, bring data in for the next kernel:



What are streams?

- **Stream** = a sequence of operations that execute on the device in the order in which they are issued by the host code.

- Same stream: In-Order execution
- Different streams: Out-of-Order execution

- Default stream = Synchronizing stream
  - No operation in the default stream can begin until all previously issued operations in any stream on the device have completed.
  - An operation in the default stream must complete before any other operation in any stream on the device can begin.

Default stream: example

```
cudaMemcpy(d_a, a, numBytes,cudaMemcpyHostToDevice);
 increment<<<1,N>>>(d_a);
 CpuFunction(b);
cudaMemcpy(a, d_a, numBytes,cudaMemcpyDeviceToHost);
```

- All operations happen in the same stream
- Device (GPU)
  - Synchronous execution
    - all operations execute (in order), one after the previous has finished
  - Unaware of `CpuFunction()`
- Host (CPU)
  - Launches `increment` and regains control
  - *May* execute `CpuFunction` *before* increment has finished
  - Final copy starts *after* both `increment` and `CpuFunction()` have finished

Non-default streams

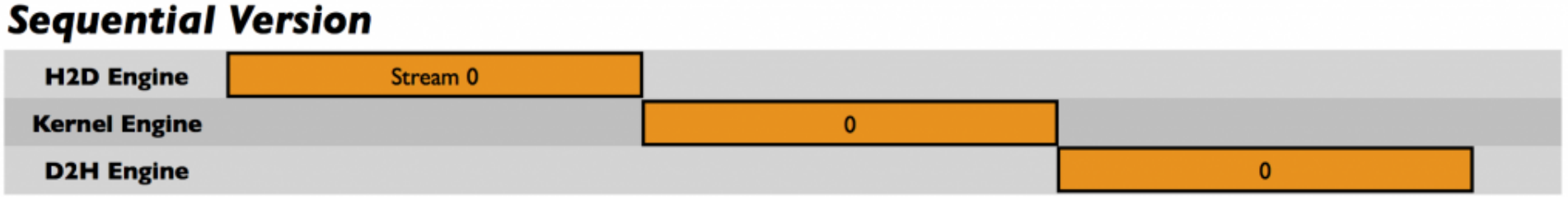
- Enable asynchronous execution and overlaps
  - Require special creation/deletion of streams
    - `cudaStreamCreate(&stream1)`
    - `cudaStreamDestroy(stream1)`
  - Special memory operations
    - `cudaMemcpyAsync(deviceMem, hostMem, size, cudaMemcpyHostToDevice, stream1)`
  - Special kernel parameter (the 4th one)
    - `increment<<<1, N, 0, stream1>>>(d_a)`
- Synchronization
  - All streams
    - `cudaDeviceSynchronize()`
  - Specific stream:
    - cudaStreamSyncrhonize(stream1)

Computation vs. communication

```
//Single stream, numBytes = 16M, numElements = 4M
cudaMemcpy(d_a, a, numBytes, cudaMemcpyHostToDevice);
kernel<<blocks,threads>>(d_a, firstElement);
cudaMemcpy(a, d_a, numBytes, cudaMemcpyDeviceToHost);
```

C1060 (pre-Fermi): 12.9ms

C2050 (Fermi): 9.9ms

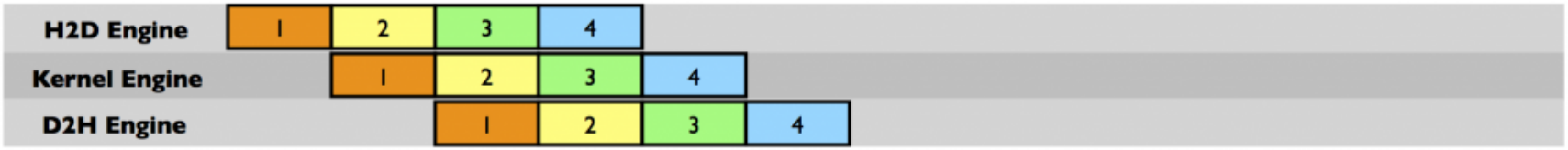


Computation-communication overlap[1]*

```
for (int i = 0; i < nStreams; ++i) {
  int offset = i * streamSize;
  cudaMemcpyAsync(&d_a[offset], &a[offset], streamBytes,
stream[i]);
  kernel<<blocks,threads,0,stream[i]>>(d_a, offset);
  cudaMemcpyAsync(&a[offset], &d_a[offset], streamBytes,
stream[i]);
}
```

C1060 (pre-Fermi): 13.63 ms (worse than sequential)

C2050 (Fermi): 5.73 ms (better than sequential)



https://github.com/parallel-forall/code-samples/blob/master/series/cuda-cpp/overlap-data-transfers/async

Computation-communication overlap[2]*

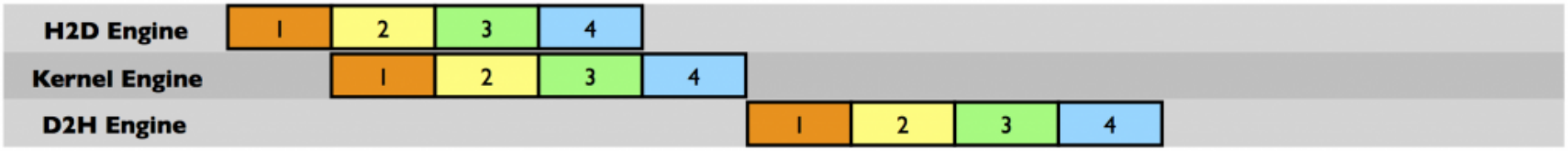
```
for (int i = 0; i < nStreams; ++i) offset[i]=i * streamSize;
for (int i = 0; i < nStreams; ++i)
   cudaMemcpyAsync(&d_a[offset[i]], &a[offset[i]], streamBytes,
      cudaMemcpyHostToDevice, stream[i]);

for (int i = 0; i < nStreams; ++i)
      kernel<<blocks,threads,0,stream[i]>>(d_a, offset);

for (int i = 0; i < nStreams; ++i)
   cudaMemcpyAsync(&a[offset], &d_a[offset], streamBytes,
      cudaMemcpyDeviceToHost, stream[i]);
```

C1060 (pre-Fermi): 8.84 ms (better than sequential)

C2050 (Fermi): 7.59 ms (better than sequential, worse than v1)



SHARING & SYNCHRONIZATION

Global synchronization

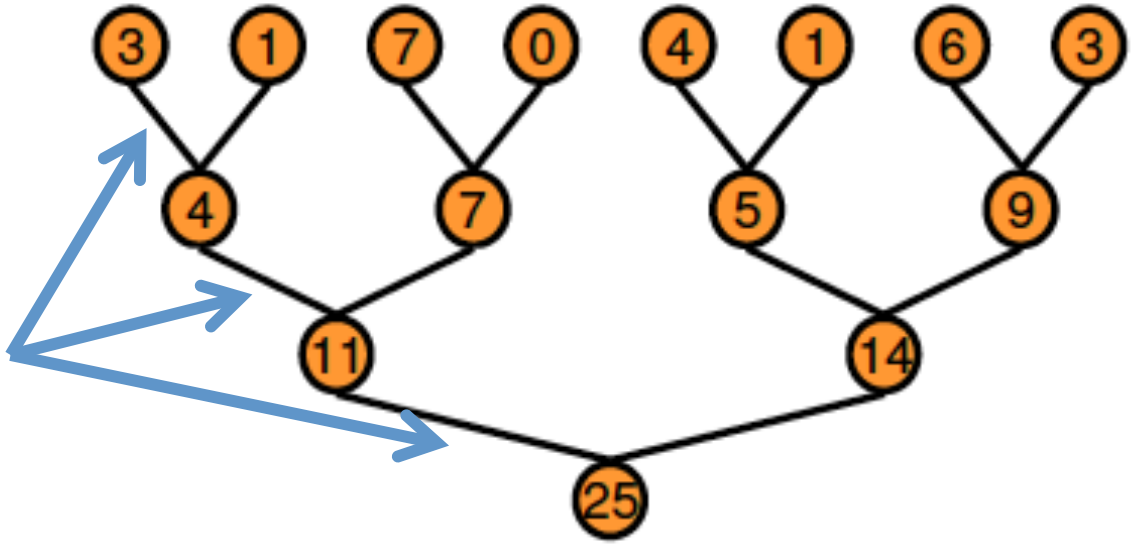
- We launch many more blocks than physical SM's.
- Each block might/should have more threads than the SM's cores

```
__global__ void my_kernel() {
    step1; // compute some values in a global array
    // wait for *all* threads to finish
    __my_global_barrier();
    step2; // use the array
}

int main() {
  dim3 blockSize(32, 32);
  dim3 gridSize(100, 100, 100);
  my_kernel<<<gridDim, blockDim>>>();
}
```

An example: parallel reduction

- Given an array with data, “reduce” it to a single value
  - The sum of all elements
  - The min/max of all elements
- Sequentially: O(n)
- In parallel?
  - Tree-based algo.
    - O(log n)
  - Requires a barrier after each step

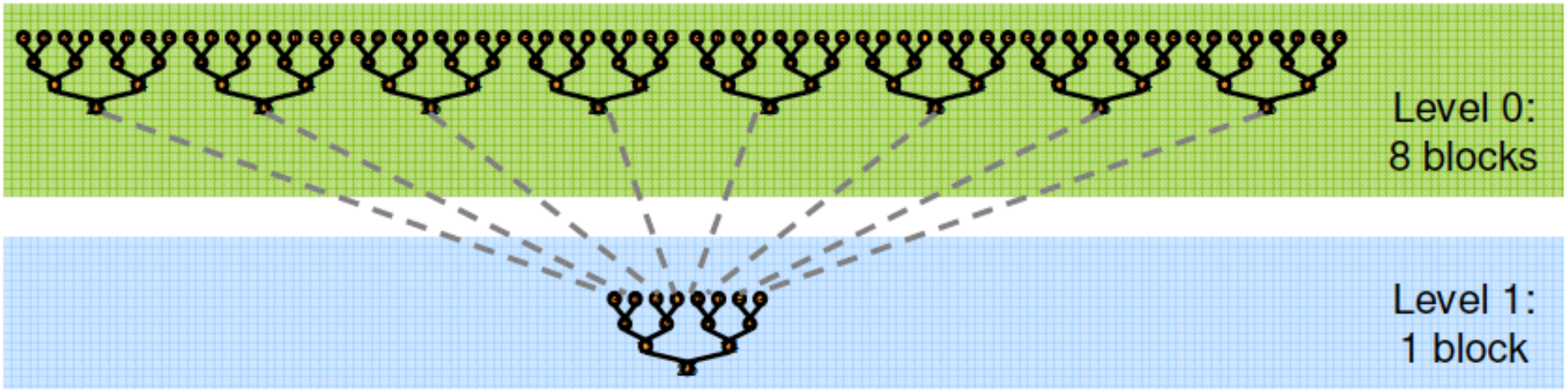


Parallel reduction in CUDA*

- One element per thread
- We need to use multiple blocks
  - Large arrays
  - Good GPU utilization
- We need global synchronization
  - Synchronization inside blocks is possible.
  - Synchronization between blocks is not possible!
- Solution: decompose into multiple kernels
  - Kernel launch serves as a global synchronization point
  - Kernel launch has negligible HW overhead, low SW overhead

*http://developer.download.nvidia.com/compute/cuda/1.1-Beta/x86_website/projects/reduction/doc/reduction.pdf

Parallel reduction in CUDA*



- Other optimizations
  - Use shared memory
  - Increase granularity
  - Avoid branching
  - Improve data access patterns

*http://developer.download.nvidia.com/compute/cuda/1.1-Beta/x86_website/projects/reduction/doc/reduction.pdf

Memory consistency

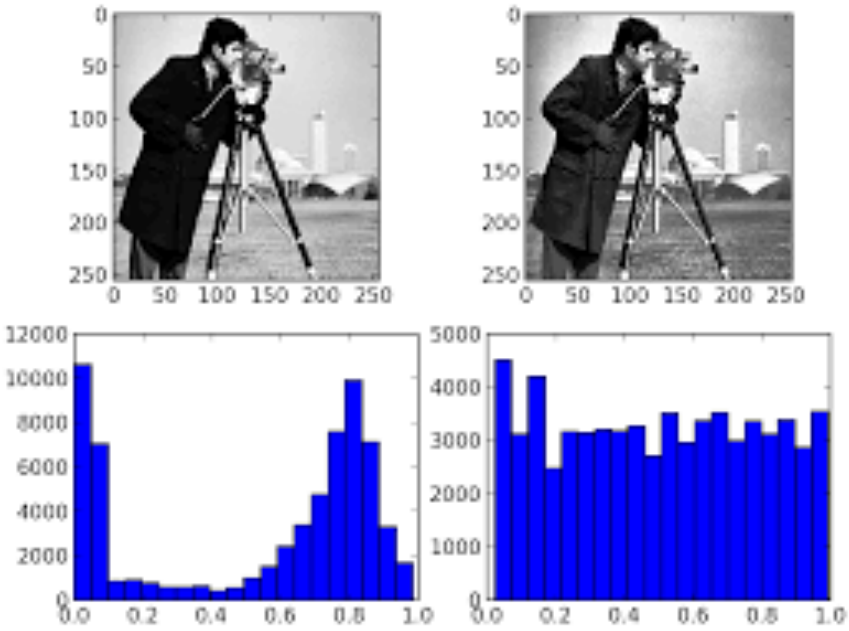
- Device (global) memory is not serially consistent
  - No ordering guarantees in shared/global memory Rd/Wr
- Share data between streaming multiprocessors
  - Potential write hazards!
- Use **atomics** to avoid data races for global (and shared) memory variables!
- Evolution:
  - Fermi has reasonable atomics for both shared and global memory
  - Kepler increases *global memory atomics* performance vs. Fermi
  - Maxwell uses native support for shared memory atomics
    - Much faster than Fermi and Kepler

Atomics

- Guarantee that only a single thread has access to a piece of memory during an operation
  - Ordering is still arbitrary
- Different types of atomic instructions
  - Add, Sub, Exch, Min, Max, Inc, Dec, CAS, And, Or, Xor
- Both for device memory and shared memory
- Much more expensive than load + operation + store

An example: image histogram

- The histogram of an image: the distribution of the pixels in the image.
  - In practice: count the pixels of each color
  - Useful image feature detection for image recognition.



An example: image histogram

```
// Determine frequency of colors in a picture.
// Colors have already been converted into integers
// between 0 and 255.
// Each thread looks at one pixel,
// and increments a counter

__global__ void histogram(int* colors, int* buckets)
{
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    int c = colors[i];
    buckets[c] += 1;  // incorrect!
}
```

An example: image histogram

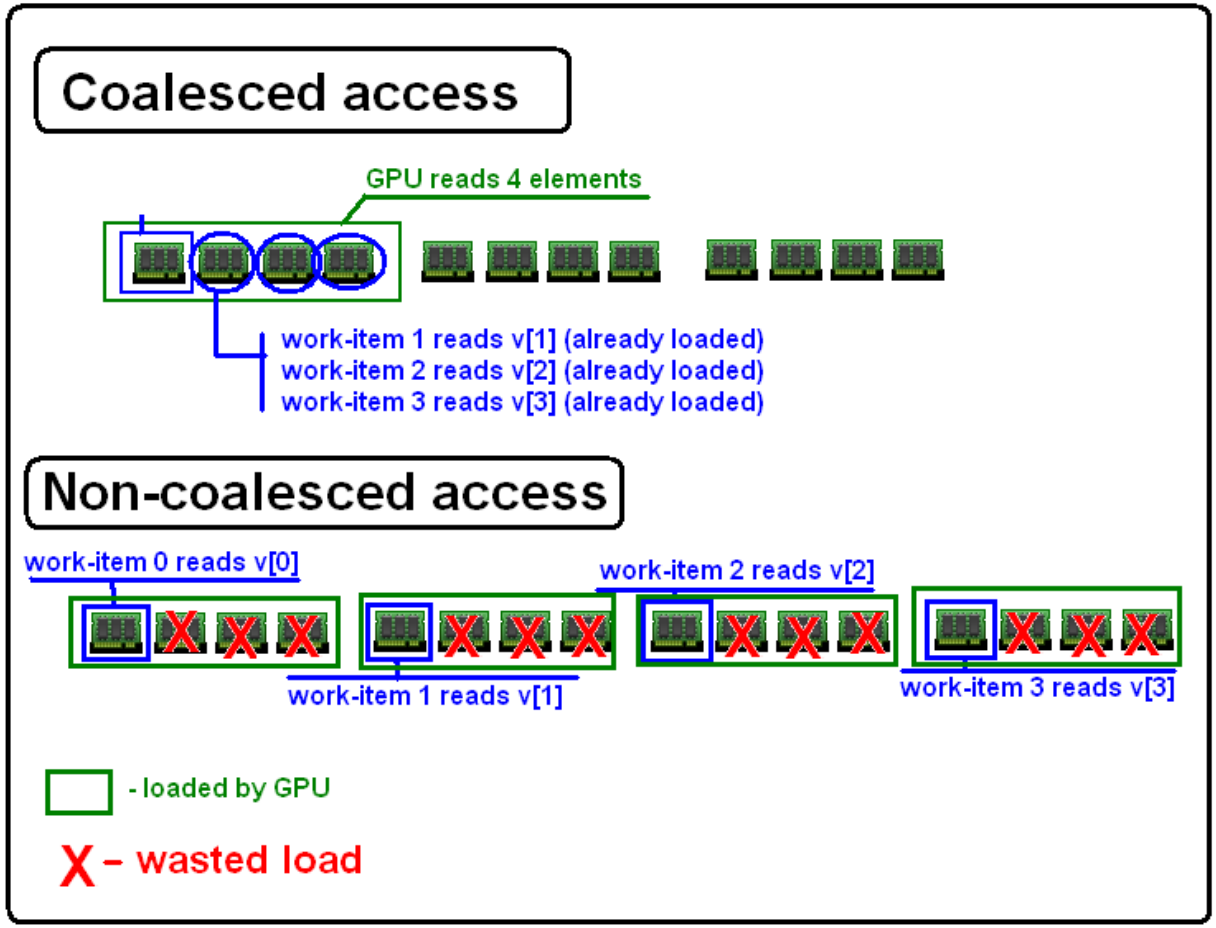
```
// Determine frequency of colors in a picture.
// Colors have already been converted into integers
// between 0 and 255.
// Each thread looks at one pixel,
// and increments a counter atomically

__global__ void histogram(int* colors, int* buckets)
{
    int i = threadIdx.x + blockDim.x * blockIdx.x;
    int c = colors[i];
    atomicAdd(&buckets[c], 1);
}
```

CUDA: MEMORY COALESCING

Memory Coalescing

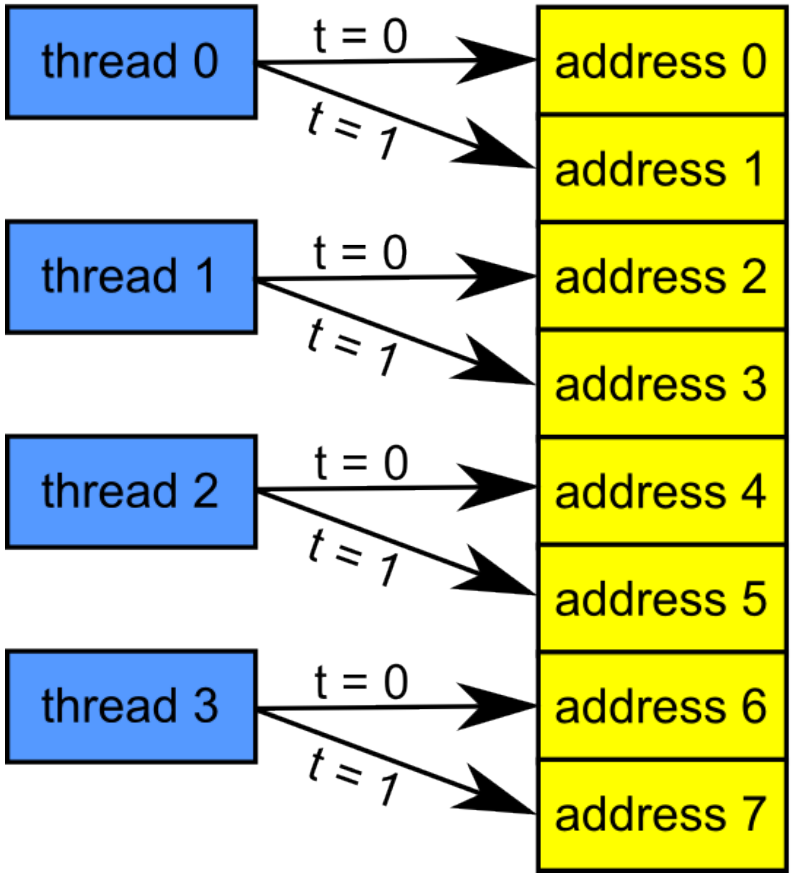
- **Memory coalescing** refers to combining multiple **memory** accesses into a single transaction



Caching vs. Coalescing

traditional multi-core

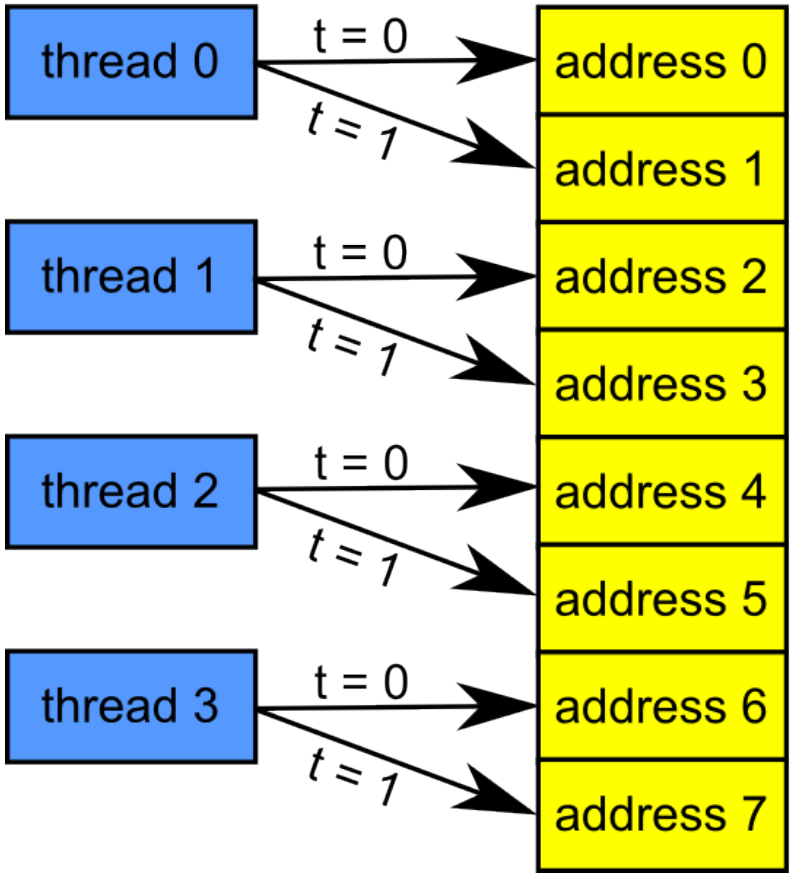
optimal memory access pattern



Caching

Caching vs. Coalescing

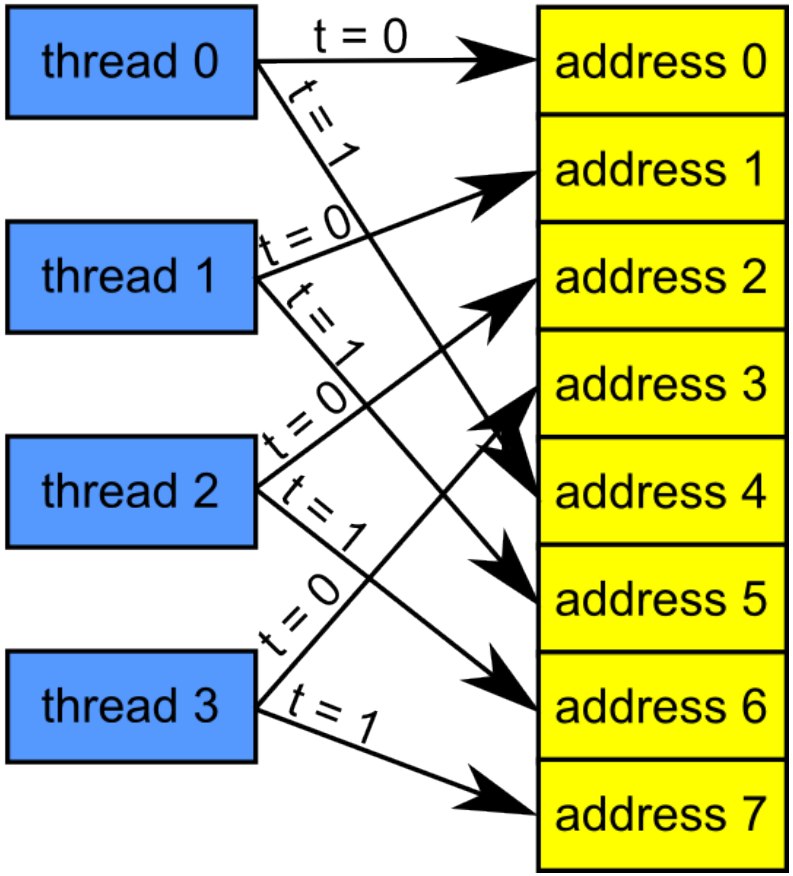
traditional multi-core
optimal memory access pattern



Caching

vs.

many-core GPU
optimal memory access pattern



Coalescin

Consider the stride of your accesses

	Input[0]	Input[1]	Input[2]	Input[3]	Input[4]	Input[5]	Input[6]	Input[7]
Stride1	T0	T1	T2	T3	T4	T5	T6	T7
Stride2		T0		T1		T2		T3
“random”			T7		T1			T4

```
__global__ void foo(int* input, float3* input2) {
  int i = blockDim.x * blockIdx.x + threadIdx.x;
  // Stride 1, full bandwidth used!
  int a = input[i];
  // Stride 2, 50% of the bandwidth is wasted
  int b = input[2*i+1];
  // “Random” stride – ?? up to 7/8 bandwidth wasted
  int c = input[f(i)];
}
```

Example: Array of Structures (AoS)

```
Struct AoS{
    int key;
    int value;
    int flag;
};

record *d_AoS_data;
cudaMalloc((void**)&d_AoS_data, ...);

kernel {
 threadID = blockDim.x * blockIdx.x + threadIdx.x;
 // …
 d_AoS_data[threadID].value += i; // wastes bandwidth!
 // …
}
```

Example: Structure of Arrays (SoA)

```
Struct SoA {
    int* keys;
    int* values;
    int* flags;
};

SoA d_SoA_data;
cudaMalloc((void**)&d_SoA_data.keys, ...);
cudaMalloc((void**)&d_SoA_data.values, ...);
cudaMalloc((void**)&d_SoA_data.flags, ...);

kernel {
  threadID = blockDim.x * blockIdx.x + threadIdx.x;
…
  d_SoA_data.values[threadID] += i; // full
 bandwidth!
… }
```

Memory Coalescing*

- Group memory accesses in as few memory transactions as possible.
  - 128-byte
- Stride 1 access patterns are preferred!
  - Other patterns can still get benefits
- Structure of arrays is often better than array of structures
- Unpredictable/ irregular access patterns
  - Case-by-case performance impact
- No coalescing => performance loss ~10x or more !
  - Caching might improve this impact …

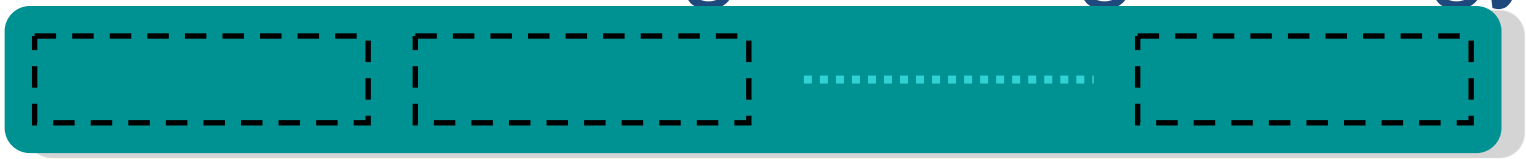
*http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/#axzz3nJ0kBsWe

CUDA: USING SHARED MEMORY

Using shared memory

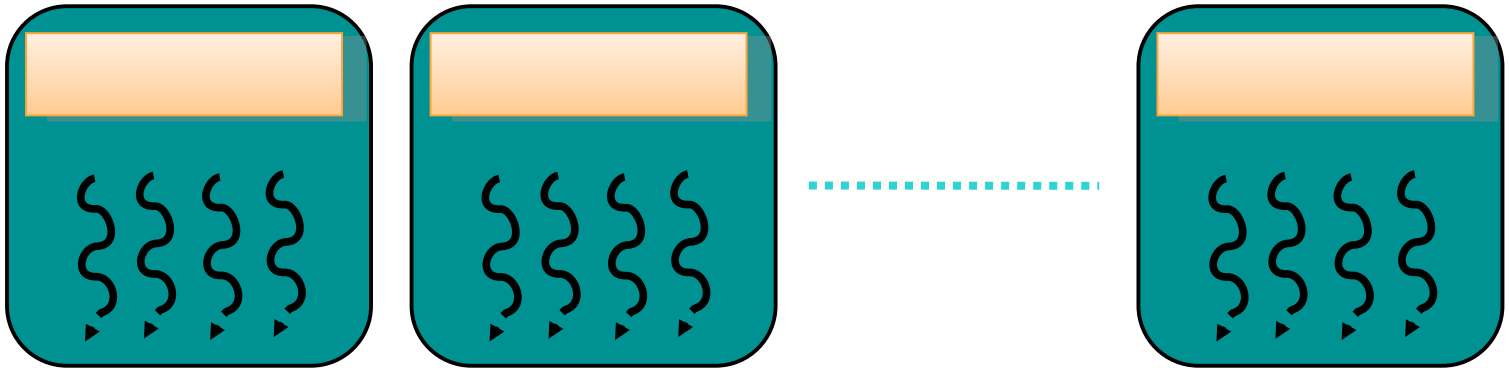
- Equivalent with providing software caching
  - Explicit: Load data to be re-used in shared memory
  - Use it for computation
  - Explicit: Store results back to global memory
- All threads in a block share memory
  - Load/Store: using all threads
  - Barrier: __syncthreads
    - Guard against using uninitialized data – not all threads have finished loading data to shared memory
    - Guard against corrupting live data – not all threads have finished computing

A Common Programming Strategy



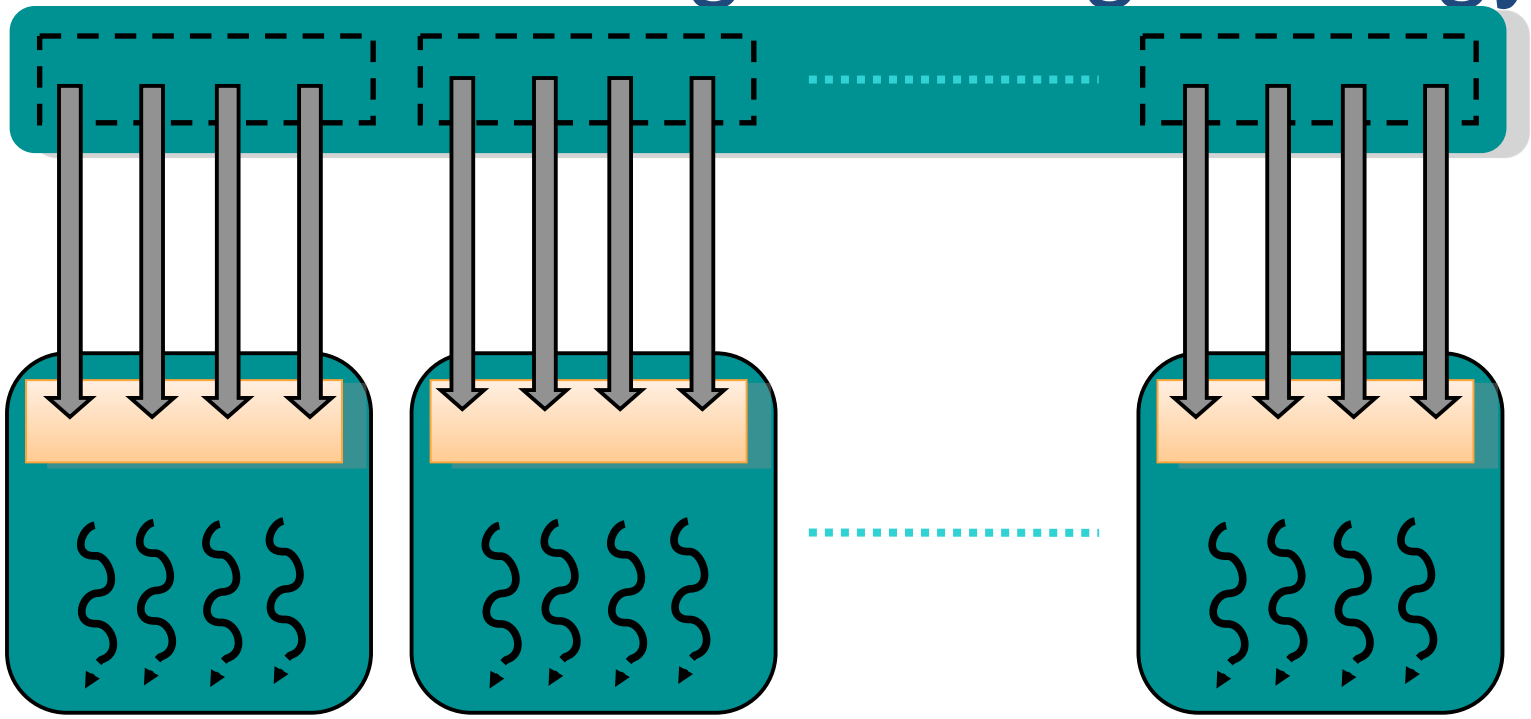
- Partition data into subsets that fit into shared memory

A Common Programming Strategy



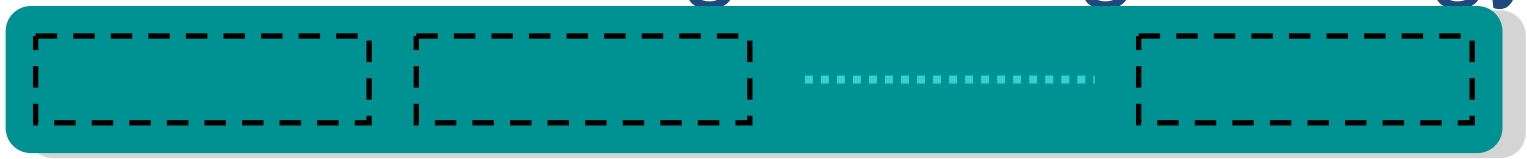
- Handle each data subset with one thread block

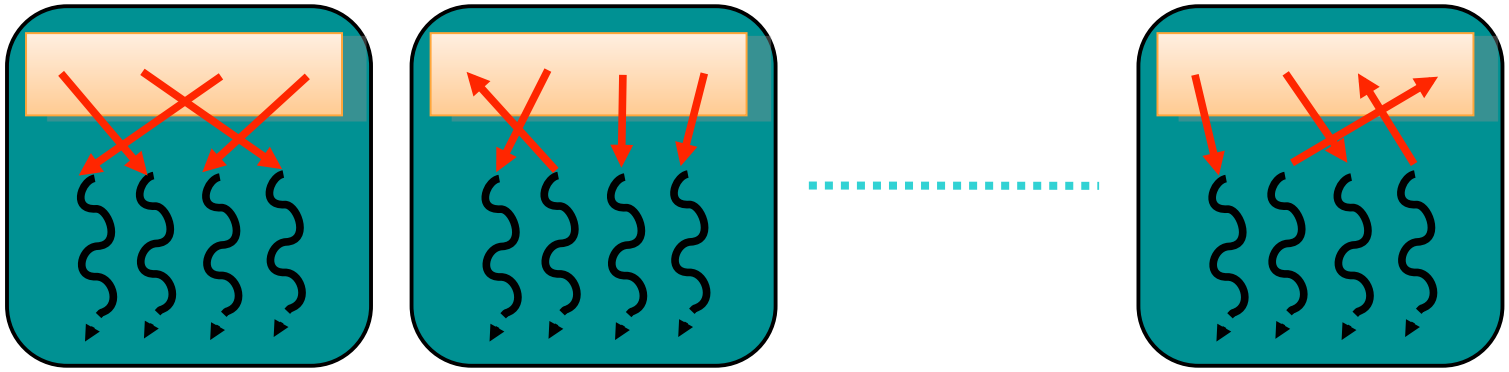
A Common Programming Strategy



- Load the subset from device memory to shared memory, using multiple threads to exploit memory-level parallelism

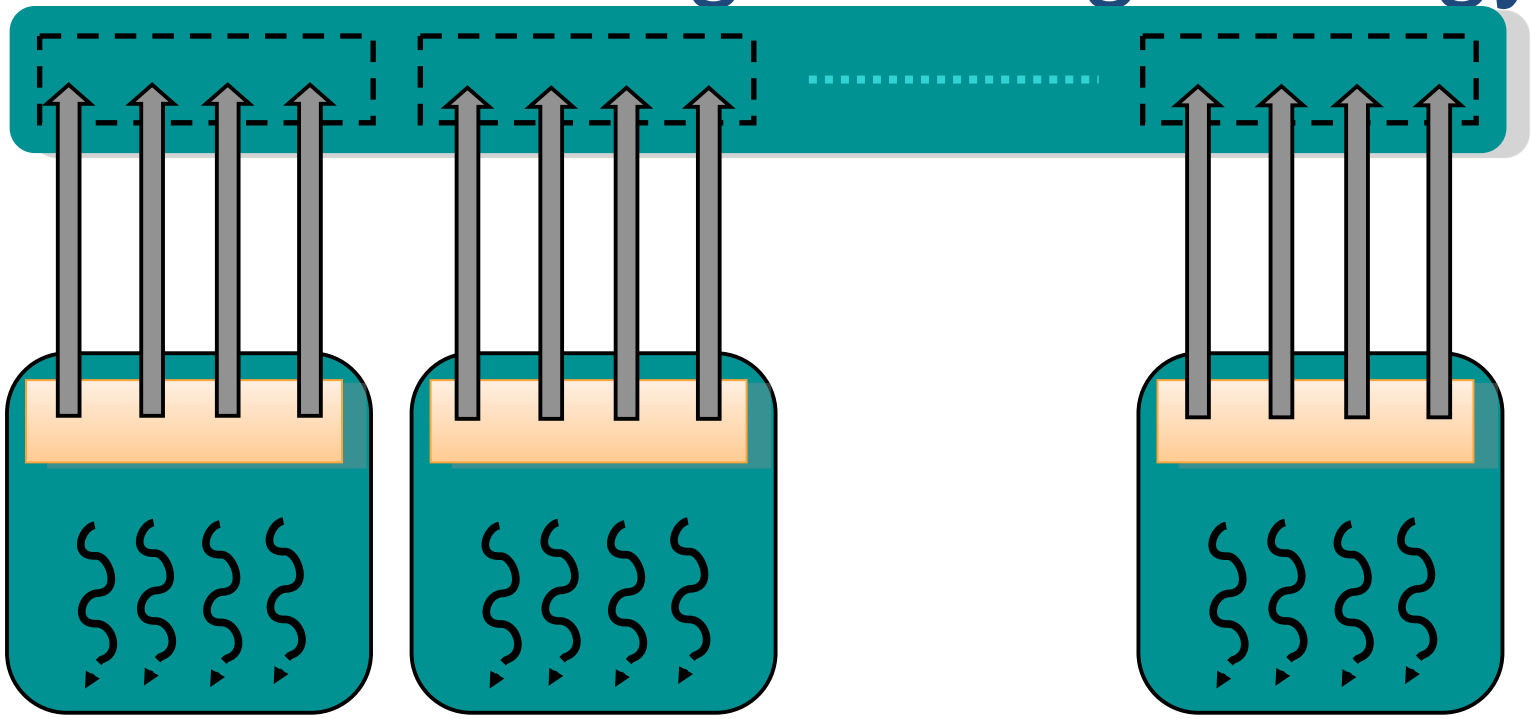
A Common Programming Strategy





- Perform the computation on the subset from shared memory

A Common Programming Strategy



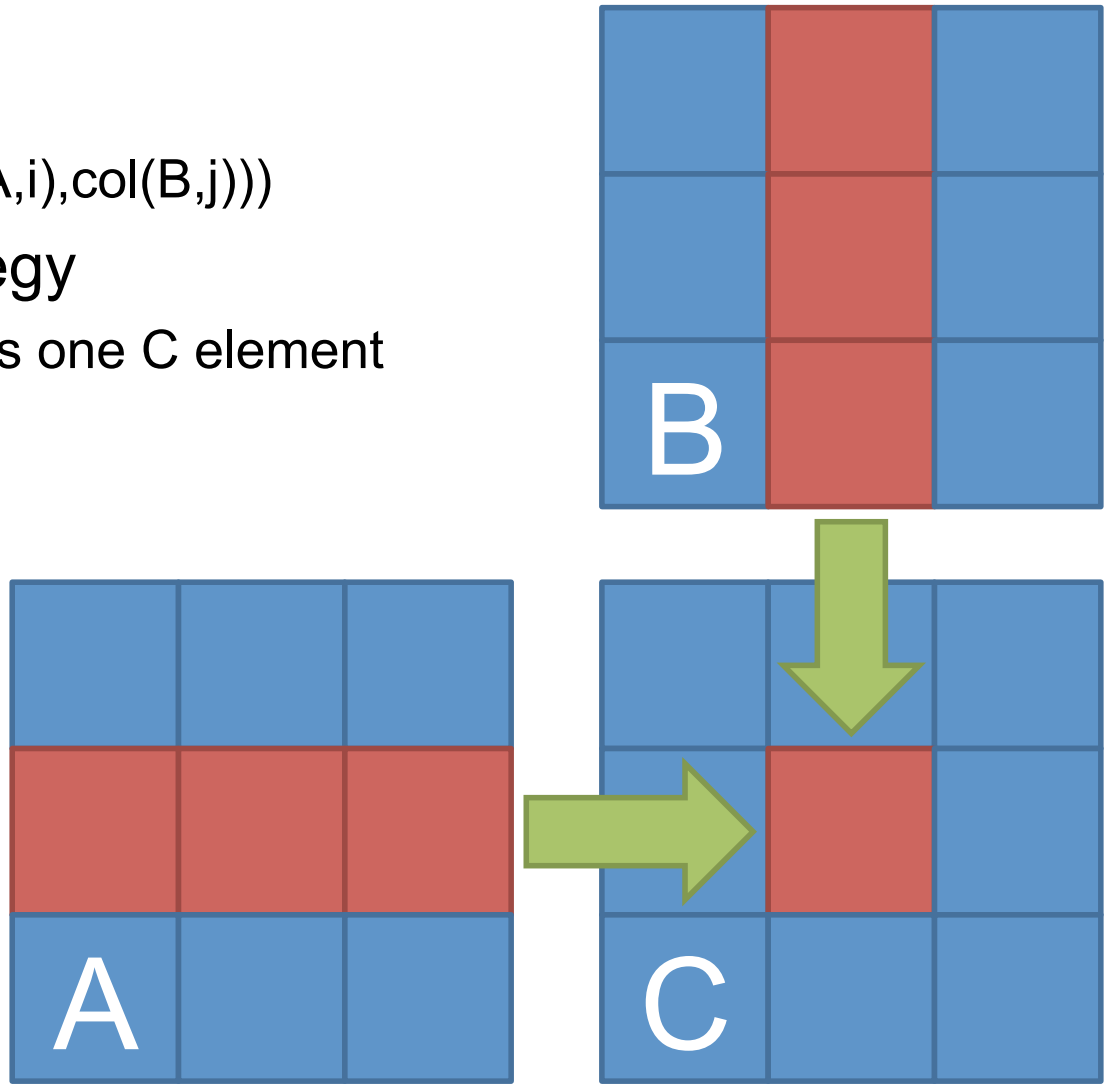
- Copy the result from shared memory back to device memory

Caches vs. Shared Memory

- Since Fermi, NVIDIA GPUs feature BOTH hardware **L1 caches** and **shared memory** per SM
  - They share the same space
    - ¾ Cache + ¼ Shared Memory OR
    - ¼ Cache + ¾ Shared Memory
- L1 Cache
  - Hardware caching enabled
    - The HW decides what goes in or out and when
- Shared memory
  - Software manages what goes in/out
  - Allows more complex access patterns to be cached

Example: Matrix multiplication

- C = A * B
  - C(i,j) = sum(dot(row(A,i),col(B,j)))
- Parallelization strategy
  - Each thread computes one C element
  - 2D kernel

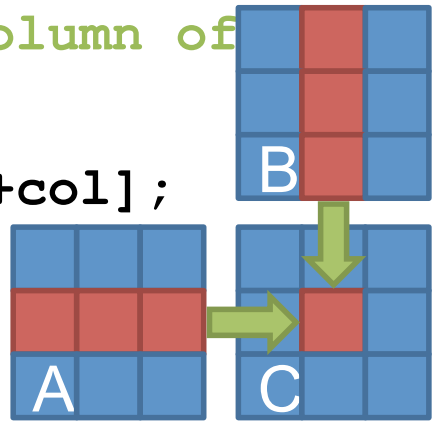


Matrix multiplication implementation

```
__global__ void mat_mul(float *a, float *b,
                        float *c, int width)
{
  // calc row & column index of output element
  int row = blockIdx.y*blockDim.y + threadIdx.y;
  int col = blockIdx.x*blockDim.x + threadIdx.x;

 float result = 0;

  // do dot product between row of a and column of
  for(int k = 0; k < width; k++) {
      result += a[row*width+k] * b[k*width+col];
 }
  c[row*width+col] = result;
}
```

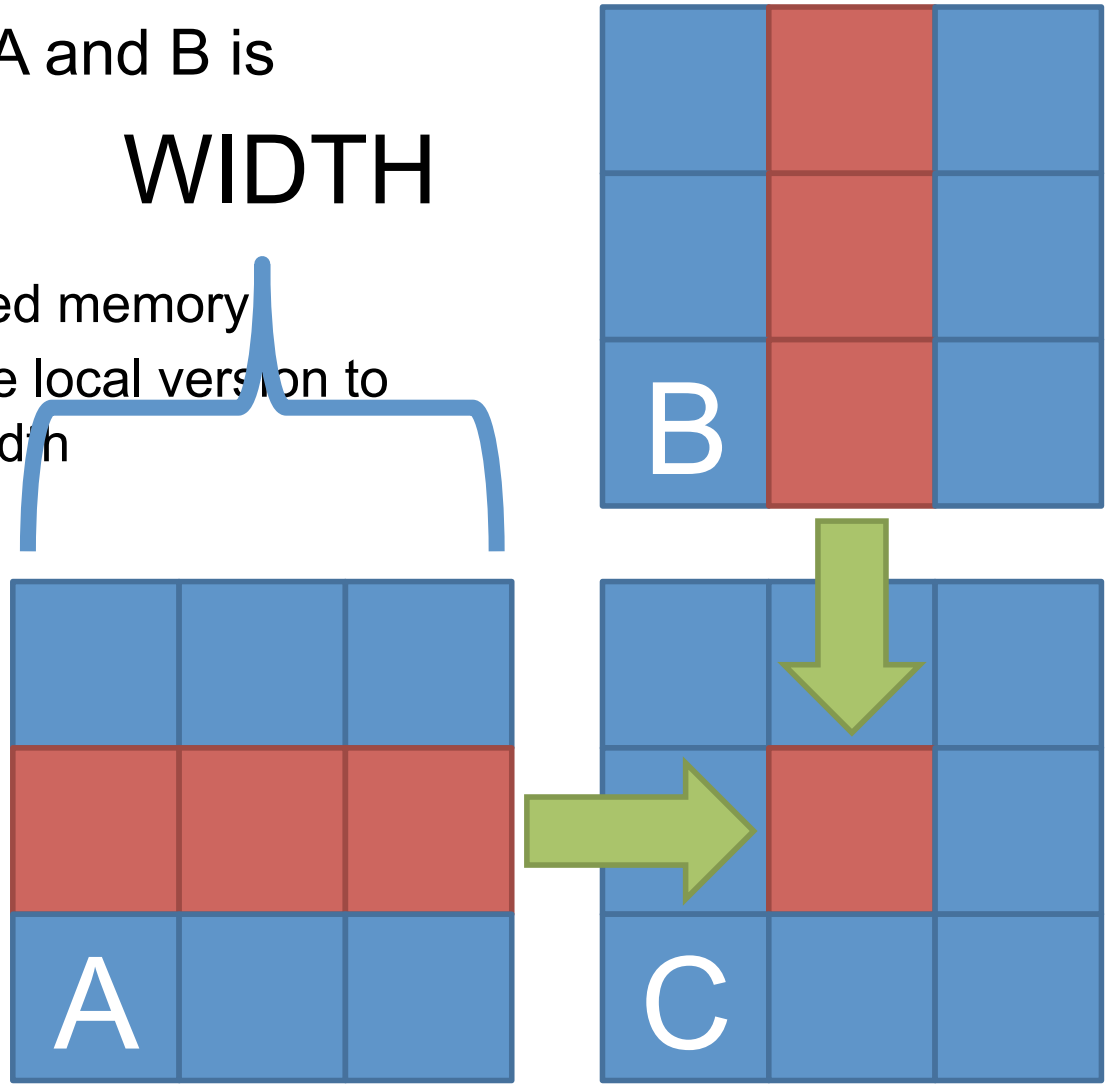


Matrix multiplication performance

Loads per dot product term	2 (a and b) = 8 bytes
FLOPS	2 (multiply and add)
AI	2 / 8 = 0.25
Performance GTX 580	1581 GFLOPs
Memory bandwidth GTX 580	192 GB/s
Attainable performance	192 * 0.25 = 48 GFLOPS
Maximum efficiency	3.0 % of theoretical peak

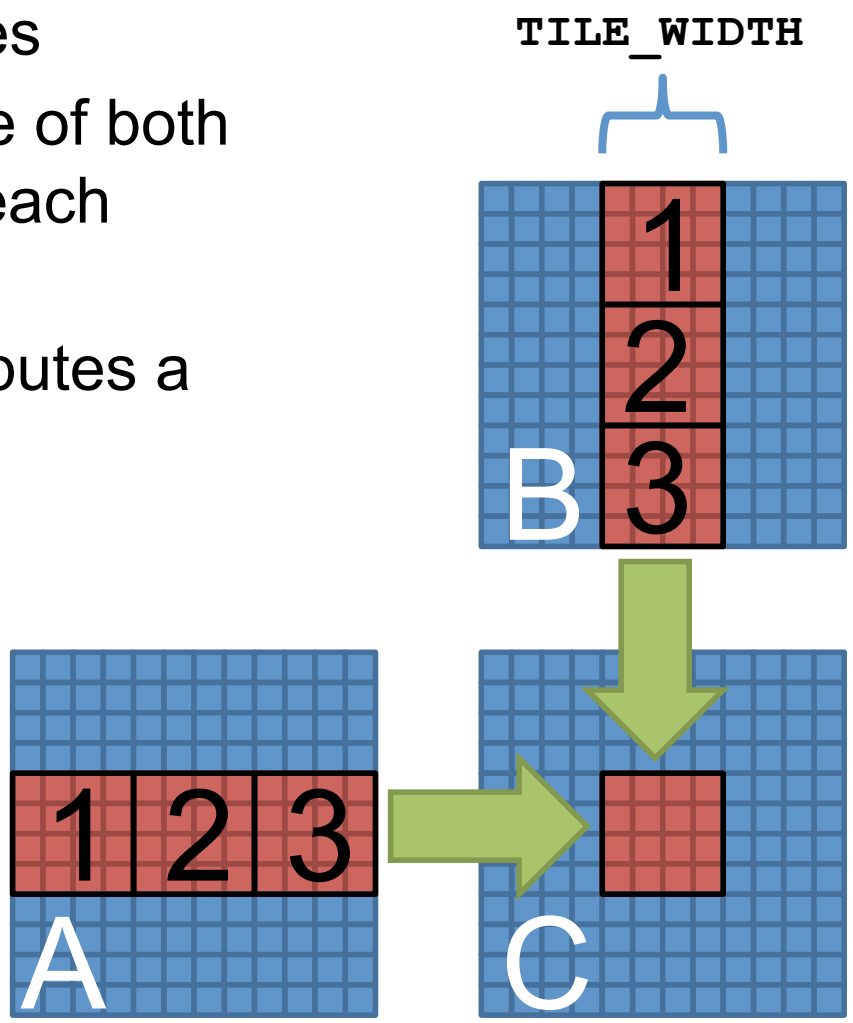
Data reuse

- Each input element in A and B is read WIDTH times
- IDEA:
  - Load elements into shared memory
  - Have several threads use local version to improve memory bandwidth



Using shared memory

- Partition kernel loop into phases
- In each thread block, load a tile of both matrices into shared memory each phase
- Each phase, each thread computes a partial result



Matrix multiply with shared memory

```
__global__ void mat_mul(float *a, float *b,
                        float *c, int width) {
  // shorthand
  int tx = threadIdx.x, ty = threadIdx.y;
  int bx = blockIdx.x,  by = blockIdx.y;

  // allocate tiles in shared memory
  __shared__ float s_a[TILE_WIDTH][TILE_WIDTH];
  __shared__ float s_b[TILE_WIDTH][TILE_WIDTH];

  // calculate the row & column index from A,B
  int row = by*blockDim.y + ty;
  int col = bx*blockDim.x + tx;

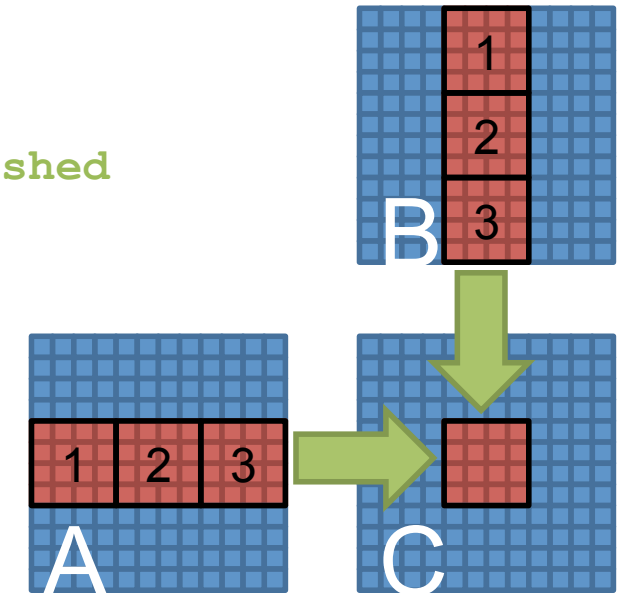
 float result = 0;
```

Matrix multiply with shared memory

```
  // loop over input tiles in phases, p = crt. phase
  for(int p = 0; p < width/TILE_WIDTH; p++) {
    // collaboratively load tiles into shared memory
    s_a[ty][tx] = a[row*width + (p*TILE_WIDTH + tx)];
    s_b[ty][tx] = b[(p*TILE_WIDTH + ty)*width + col];
// barrier: ALL writes to shared memory finished
    __syncthreads();

    // dot product between row of s_a and col of s_b
    for(int k = 0; k < TILE_WIDTH; k++) {
      result += s_a[ty][k] * s_b[k][tx];
    }
// barrier: ALL reads of shared memory finished
    __syncthreads();
  }

  c[row*width+col] = result;
}
```



Use of Barriers in mat_mul

- Two barriers per phase:
  - __syncthreads after all data is loaded into shared memory
  - __syncthreads after all data is read from shared memory
    - Second __syncthreads in phase p guards the load in phase p+1

- Formally, __synchthreads is a barrier for shared memory for a block of threads:

"void __syncthreads();

waits until all threads in the thread block have reached this point **and** all global and shared memory accesses made by these threads prior to __syncthreads() are visible to all threads in the block."

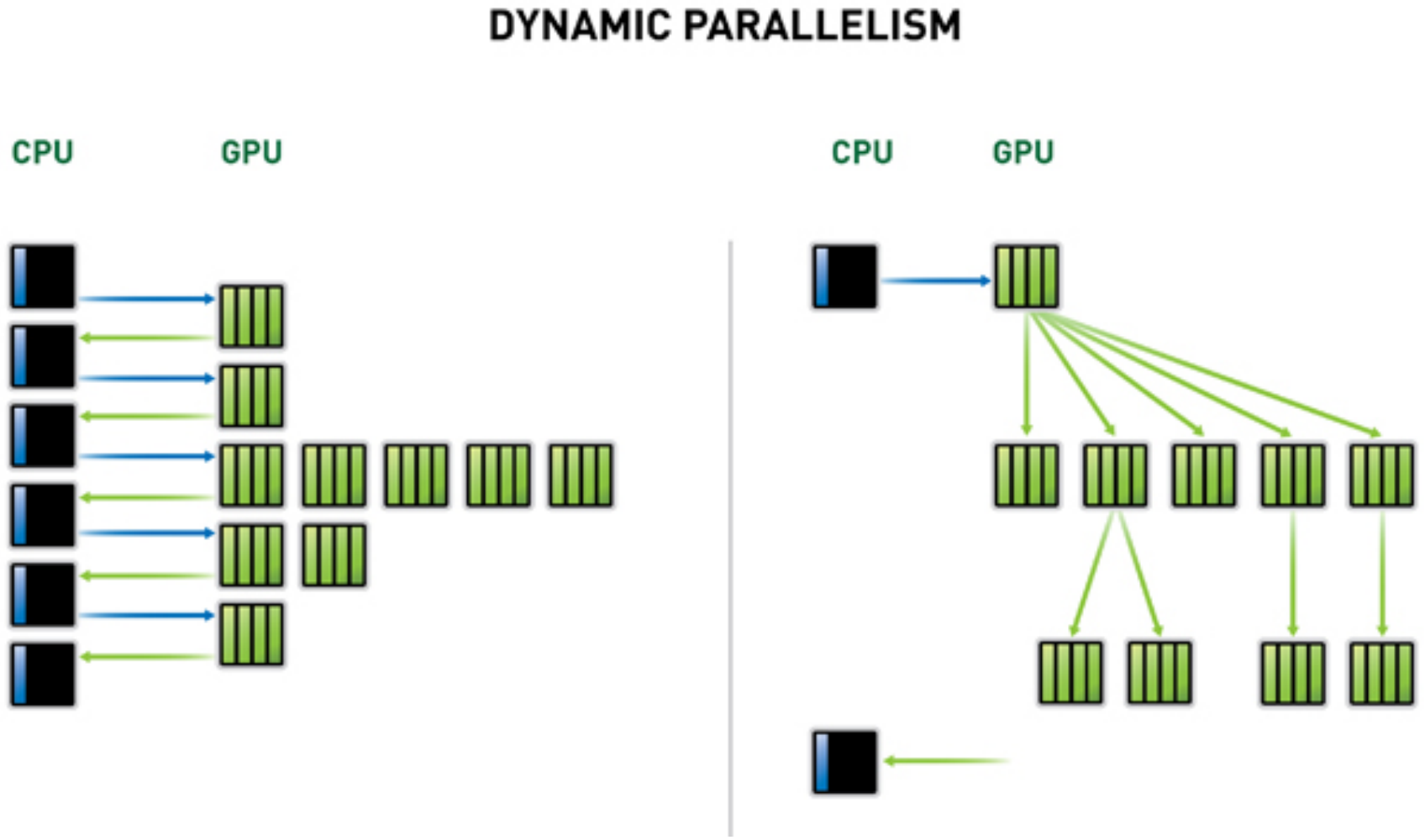
Matrix multiplication performance

	Original	shared memory
Global loads	$2N^3$ * 4 bytes	($2N^3$ / TILE_WIDTH) * 4 bytes
Total ops	$2N^3$	$2N^3$
AI	0.25	0.25 * TILE_WIDTH

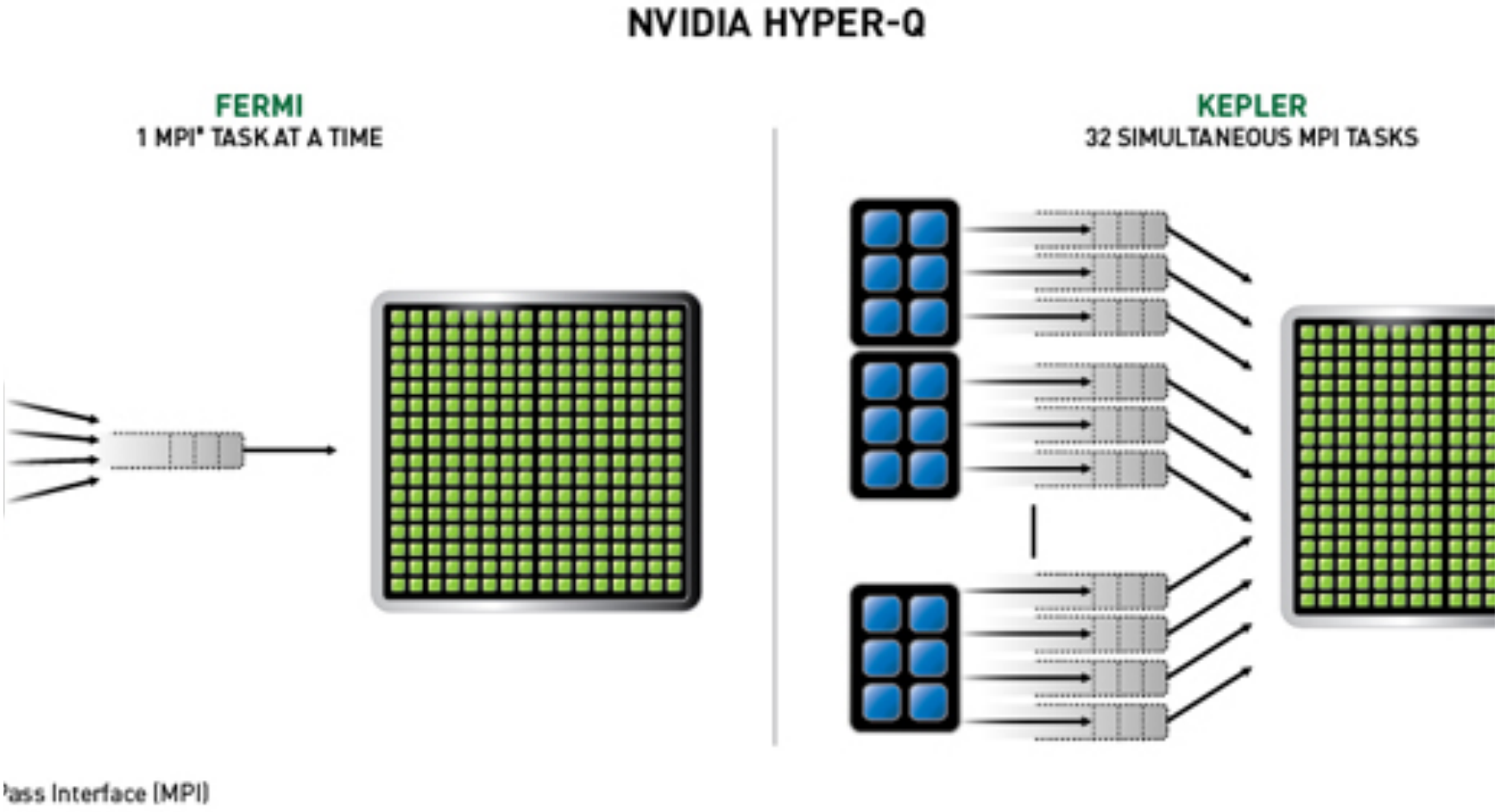
Performance GTX 580	1581 GFLOPs
Memory bandwidth GTX 580	192 GB/s
AI needed for peak	1581 / 192 = **8.23**
TILE_WIDTH required to achieve peak	0.25 * TILE_WIDTH = **8.23**, TILE_WIDTH = 32.9

CUDA: MANY OTHER FEATURES

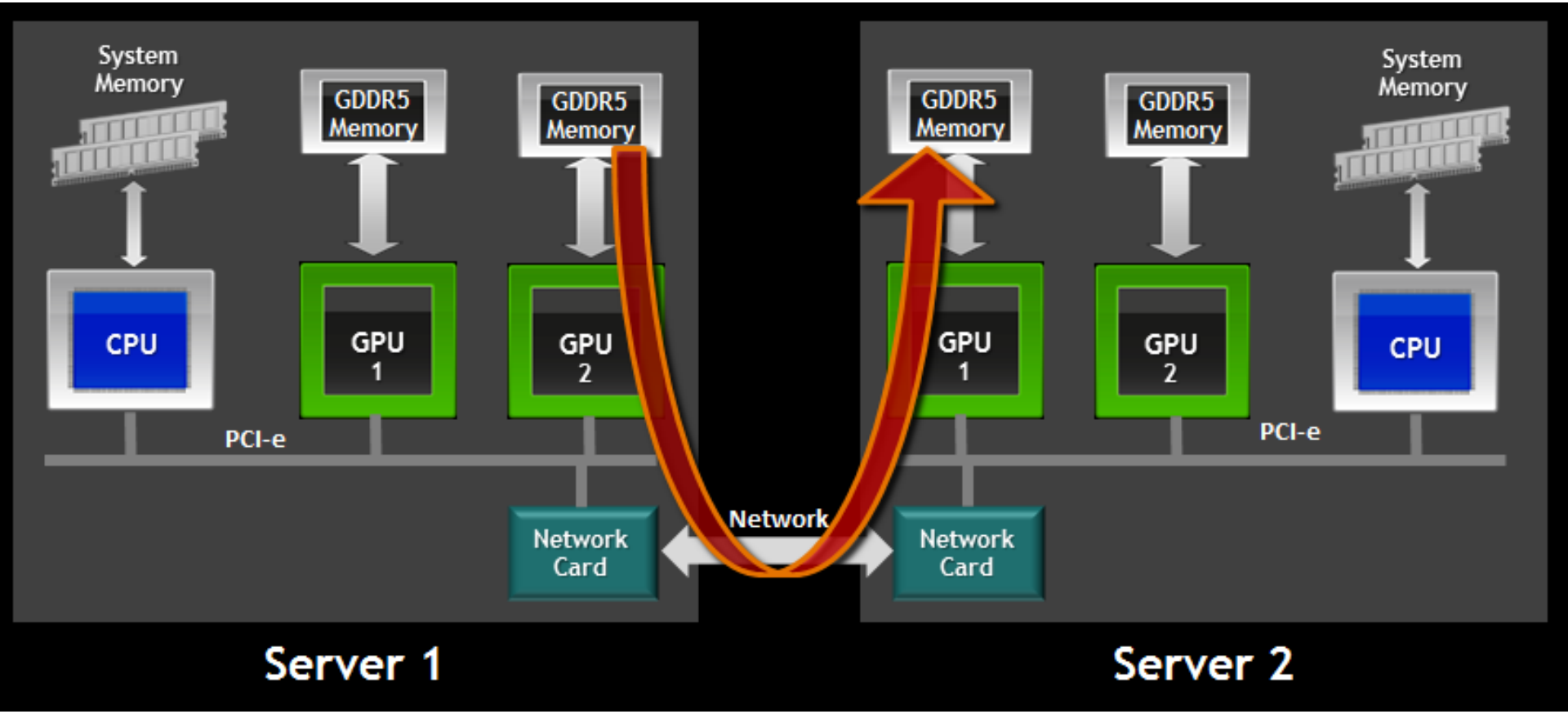
Kepler: Dynamic parallelism



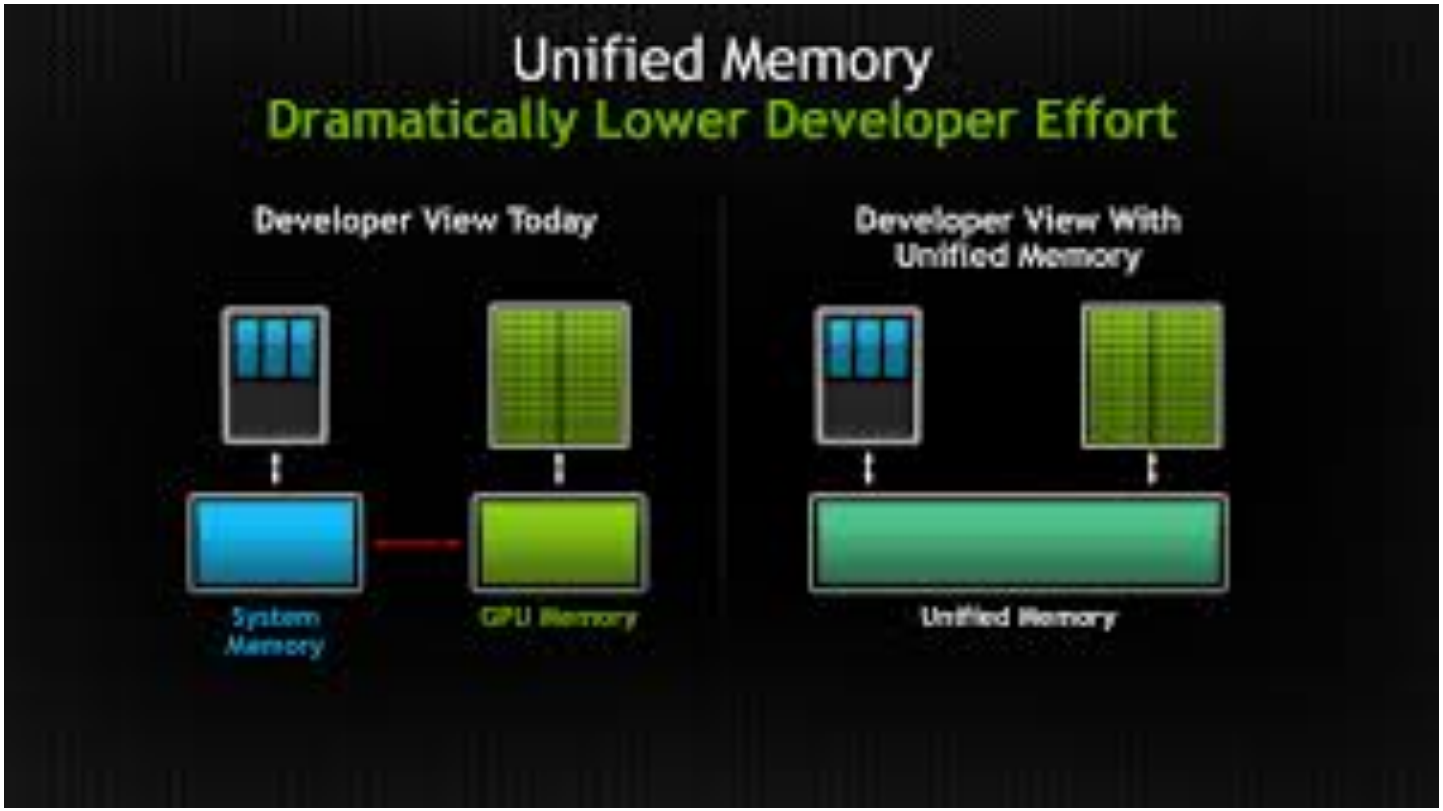
Kepler: Hyper-Q



Kepler GK110: GPU Direct



Unified memory



SUMMARY

Take home message

- GPUs are massively parallel architectures with limited flexibility, but very high throughput
- Pro's:
  - Much higher compute capabilities
  - Higher bandwidth
- Con's
  - Limited on-card memory
  - Low-bandwidth communication with host
- Debate-able
  - Programmability & productivity

Open research questions

- Shall we port all applications on GPUs?
  - If yes – can we automate the process?
  - If not – can we decide how to select?
- Shall we use GPUs in large-scale systems?
- Shall we use heterogeneous CPU+GPU systems?
- Can we improve the GPU design …
  - For HPC?
  - For other application domains?

Questions? Comments? Suggestions?

- A.L.Varbanescu@uva.nl

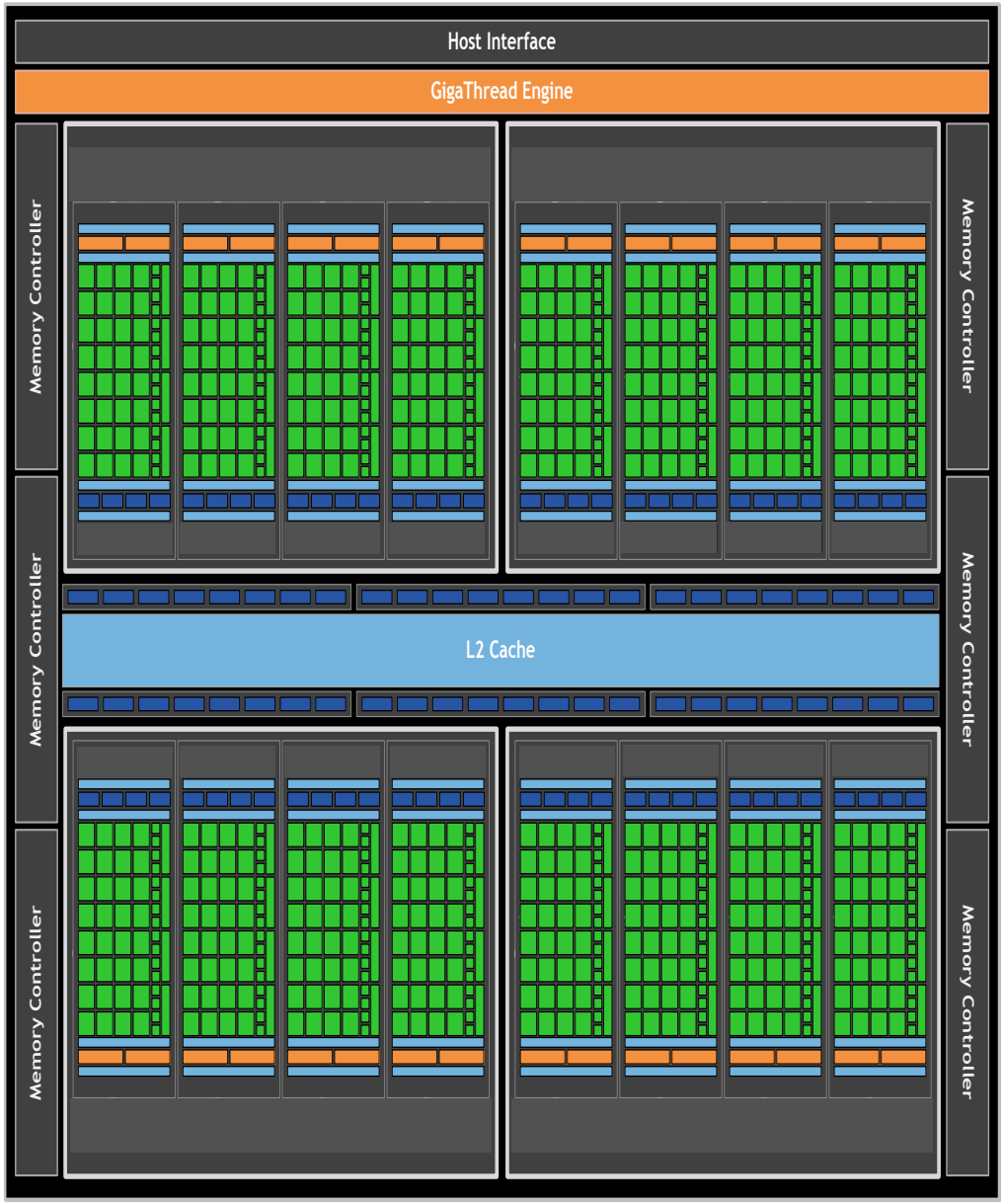
… also if you want to work on GPU-related projects OR in a team that works on heterogeneous computing.

… All you have to do is ask ☺

Back-up slides

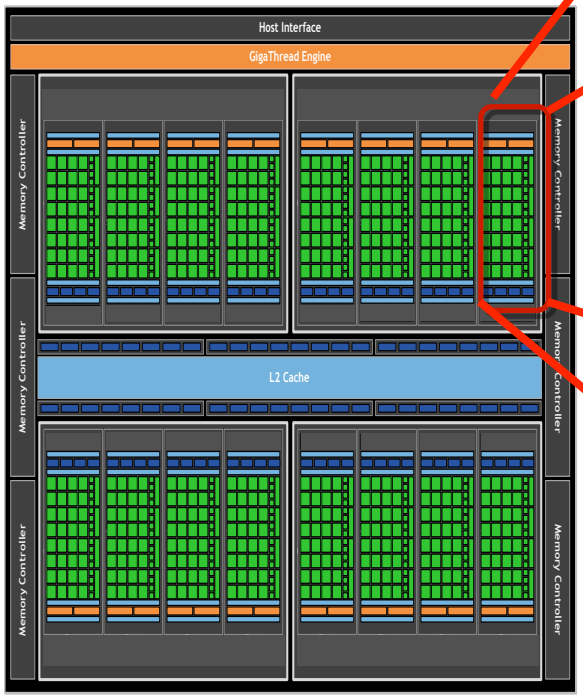
Fermi

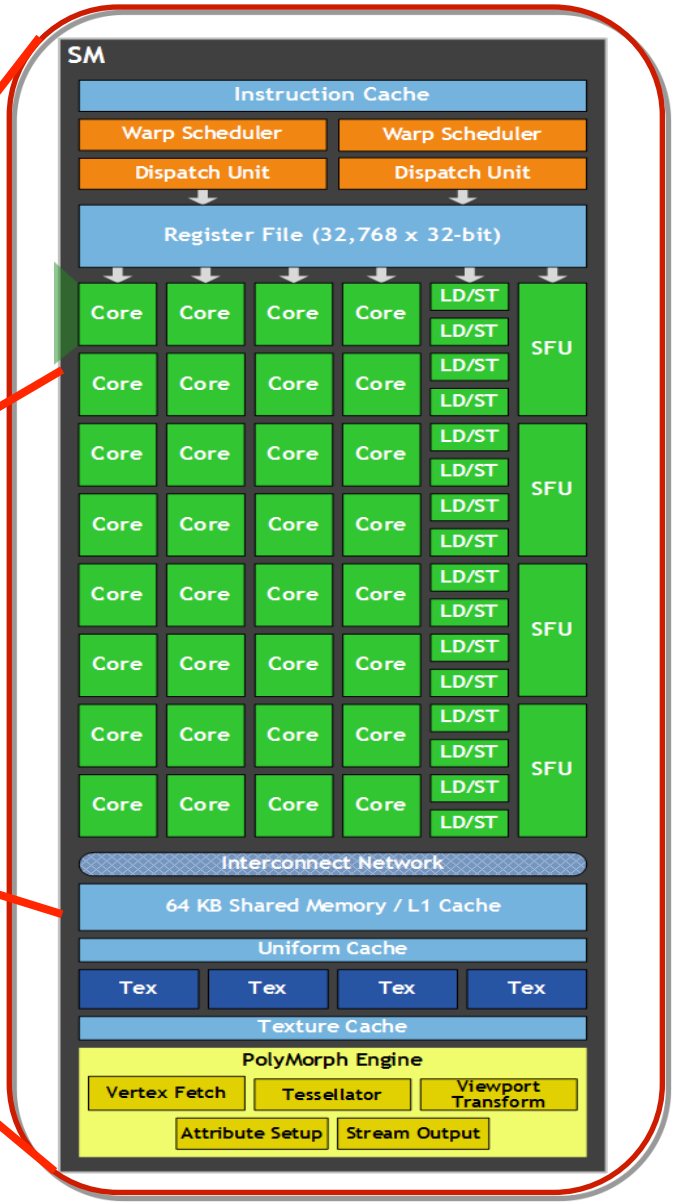
- Consumer: GTX 480, 580
- HPC: Tesla C2050
  - More memory, ECC
  - 1.0 Tlop SP
  - 515 GFlop SP
- 16 streaming multiprocessors (SM)
  - GTX 580: 16
  - GTX 480: 15
  - C2050: 14
- SMs are independent
- 768 KB L2 cache



Fermi Streaming Multiprocessor (SM)

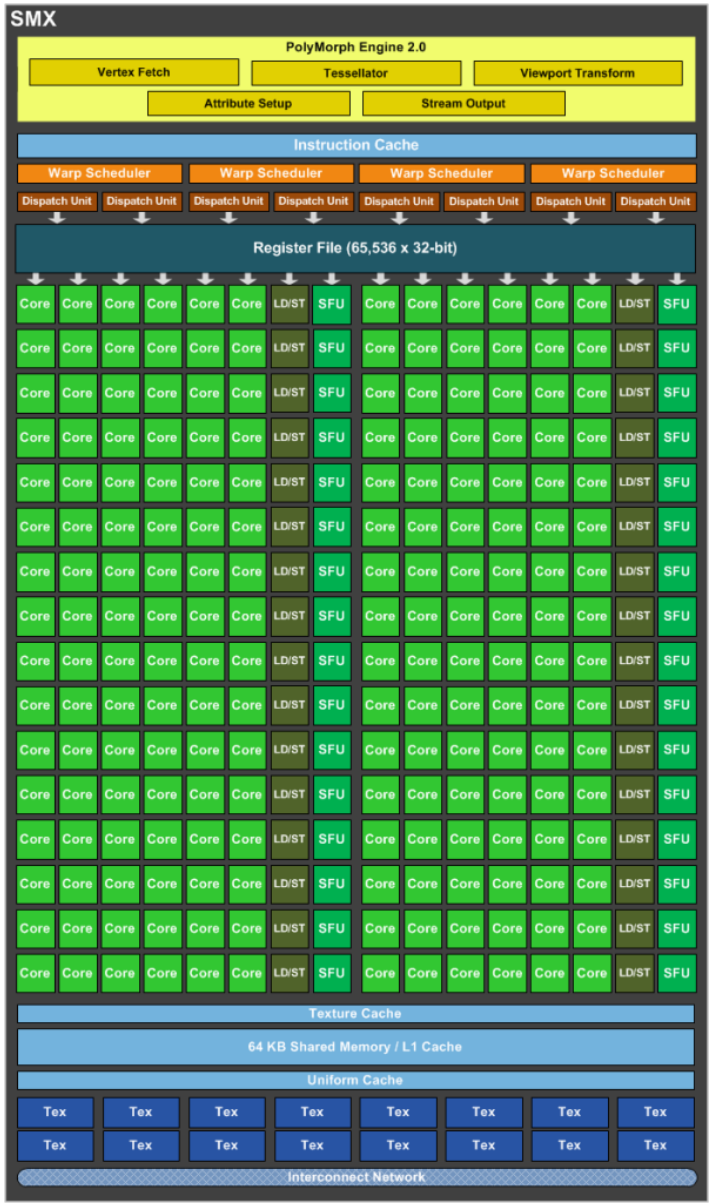
- 32 cores per SM (512 cores total)
- 64KB configurable
  L1 cache / shared memory
- 32,768 32-bit registers





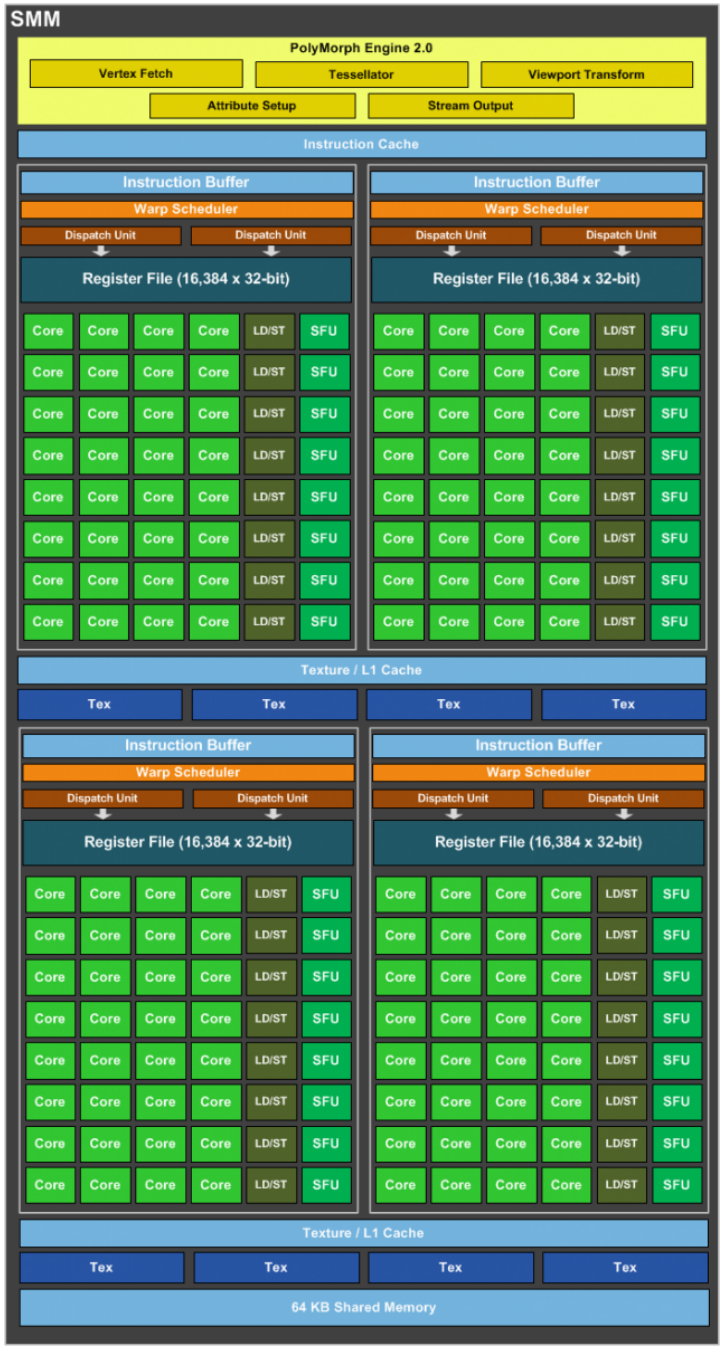
Kepler: SMX

- Consumer:
  - GTX680, GTX780, GTX-Titan
- HPC
  - Tesla K10..K40, K80
- SMX features
  - 192 CUDA cores
    - 32 in Fermi
  - 32 Special Function Units (SFU)
    - 4 for Fermi
  - 32 Load/Store units (LD/ST)
    - 16 for Fermi
- 3x Perf/Watt improvement
- 4x more texture memory



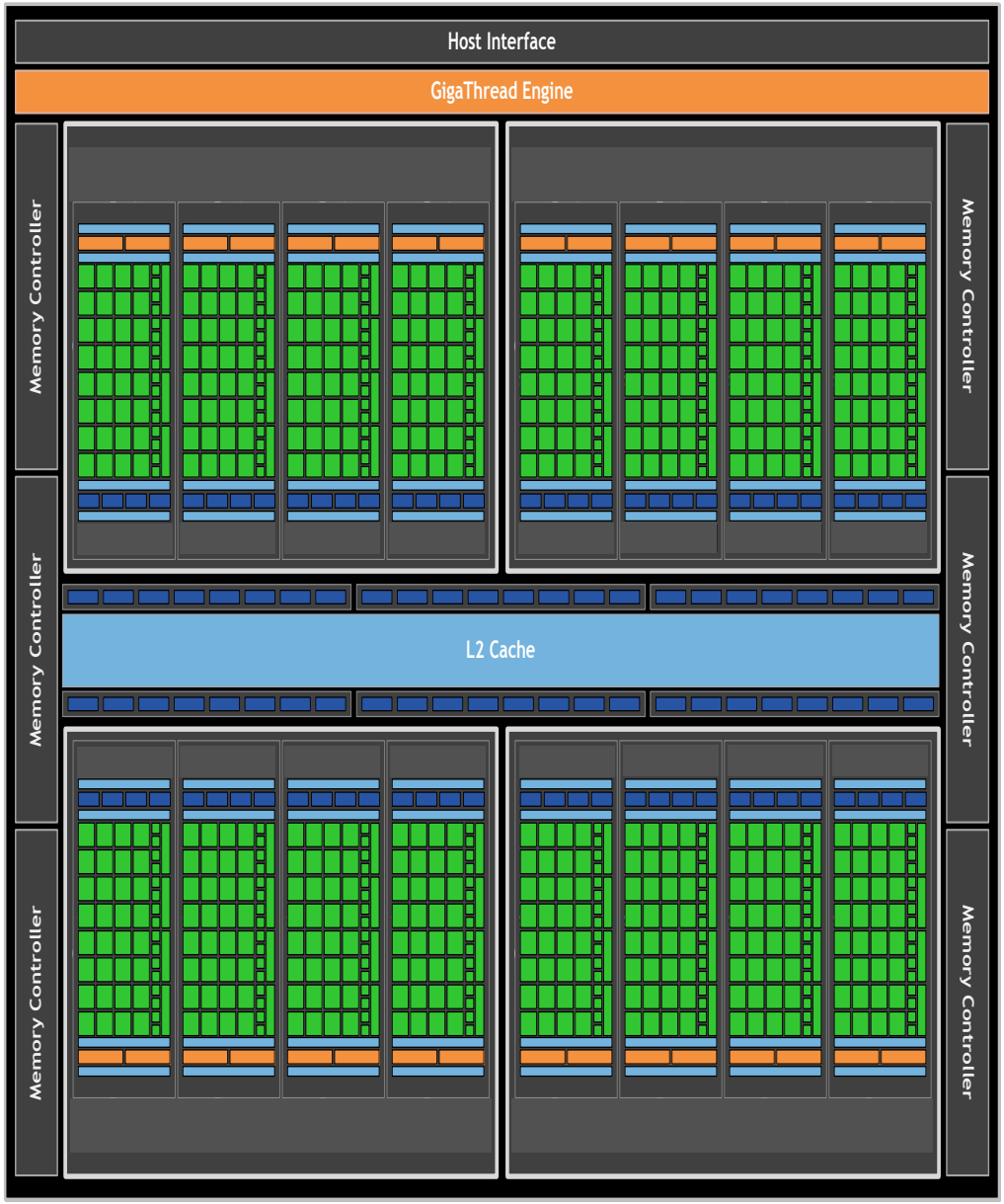
Maxwell: SMM

- Consumer:
  - GTX 970, GTX 980, …
- HPC:
  - Tesla M40
- SMM Features:
  - 4 subblocks of 32 cores
  - Dedicated L1/LM per 64 cores
  - Dispatch/decode/registers per 32 cores
- L2 cache: 2MB (~3x vs. Kepler)
- 40 texture units
- Lower power consumption



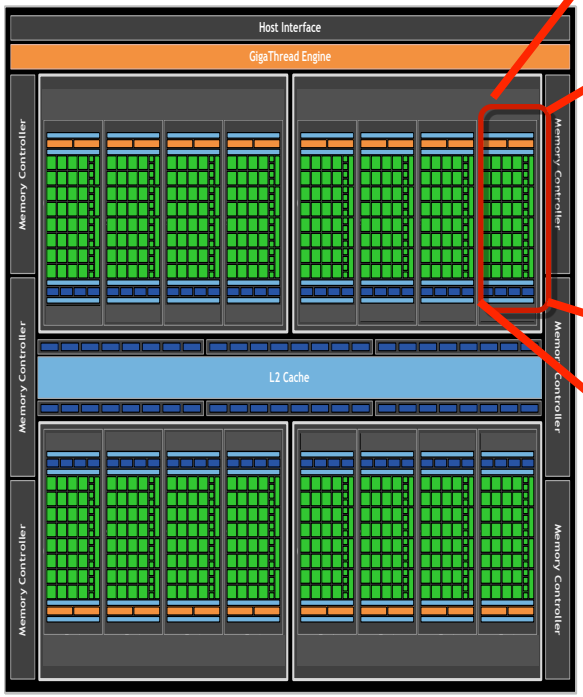
Fermi

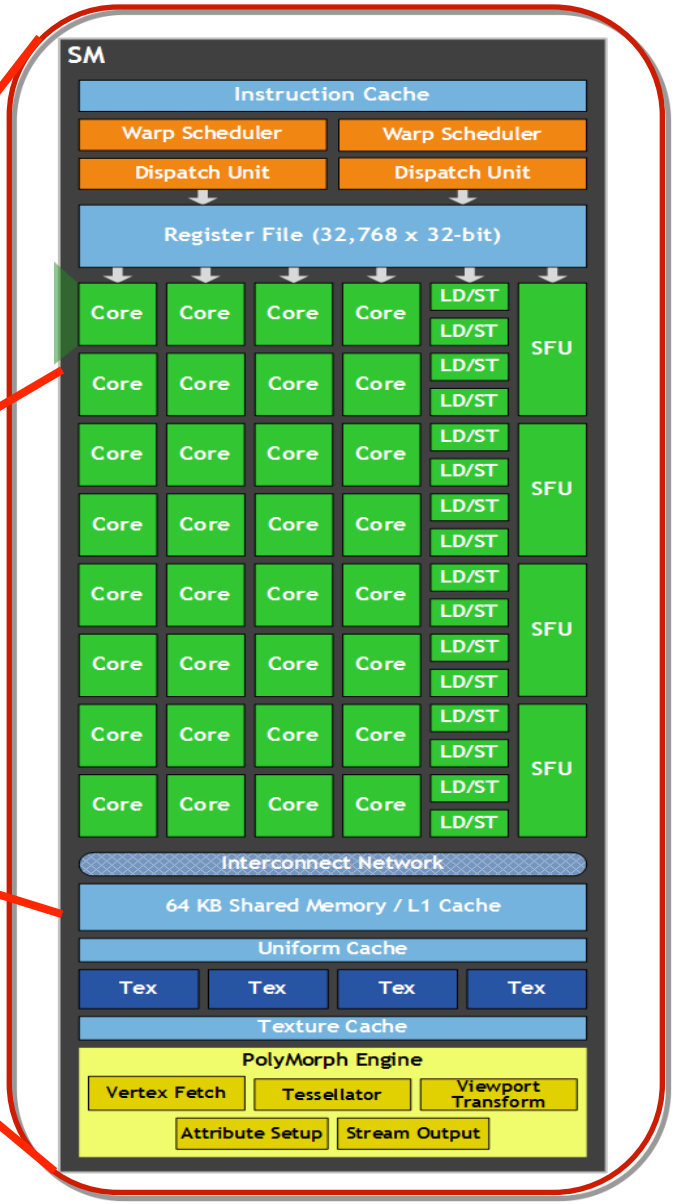
- Consumer: GTX 480, 580
- HPC: Tesla C2050
  - More memory, ECC
  - 1.0 Tlop SP
  - 515 GFlop SP
- 16 streaming multiprocessors (SM)
  - GTX 580: 16
  - GTX 480: 15
  - C2050: 14
- SMs are independent
- 768 KB L2 cache



Fermi Streaming Multiprocessor (SM)

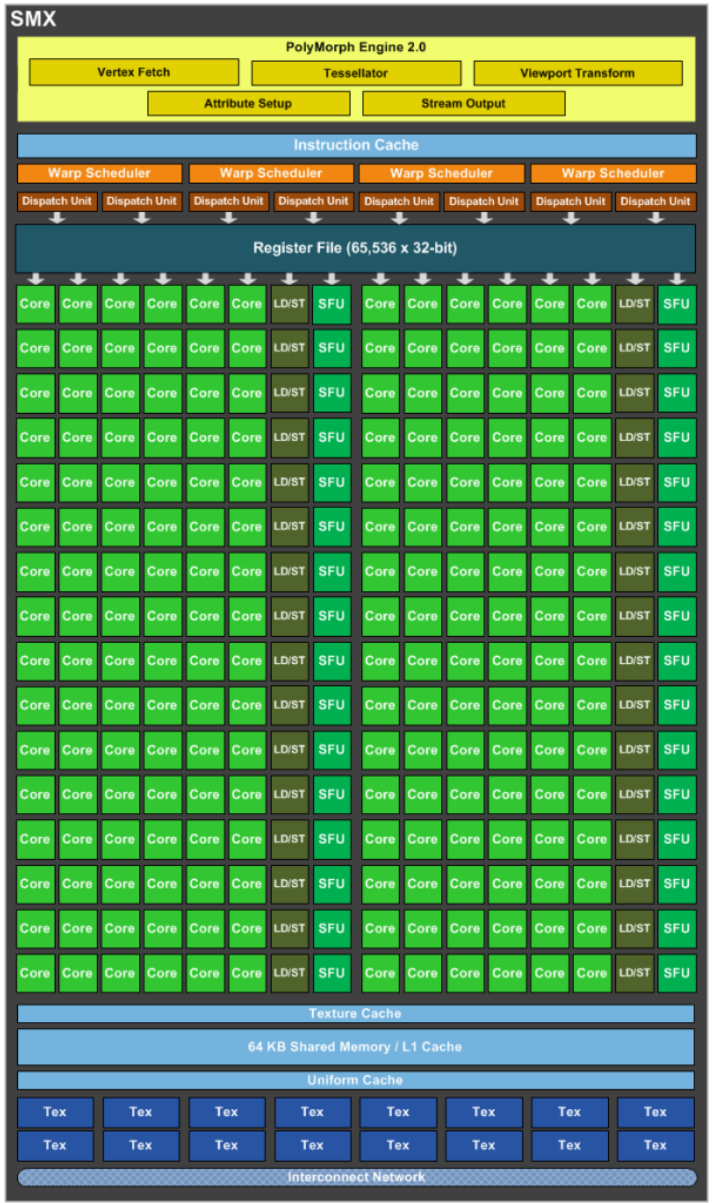
- 32 cores per SM (512 cores total)
- 64KB configurable
  L1 cache / shared memory
- 32,768 32-bit registers





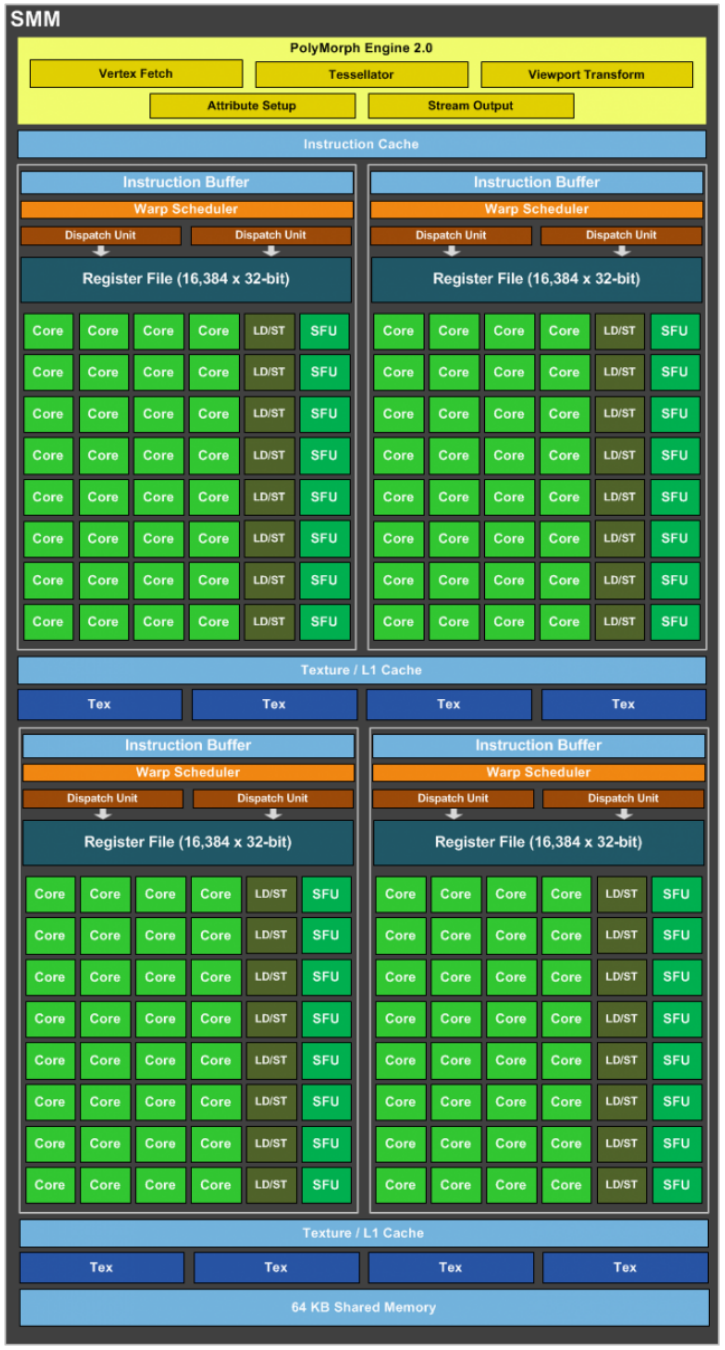
Kepler: SMX

- Consumer:
  - GTX680, GTX780, GTX-Titan
- HPC
  - Tesla K10..K40, K80
- SMX features
  - 192 CUDA cores
    - 32 in Fermi
  - 32 Special Function Units (SFU)
    - 4 for Fermi
  - 32 Load/Store units (LD/ST)
    - 16 for Fermi
- 3x Perf/Watt improvement
- 4x more texture memory



Maxwell: SMM

- Consumer:
  - GTX 970, GTX 980, …
- HPC:
  - Tesla M40
- SMM Features:
  - 4 subblocks of 32 cores
  - Dedicated L1/LM per 64 cores
  - Dispatch/decode/registers per 32 cores
- L2 cache: 2MB (~3x vs. Kepler)
- 40 texture units
- Lower power consumption



Pascal: SMP

- 64 single-precision (FP32) CUDA Cores.
  - Maxwell = 128
  - Kepler = 192
- Focus on DP
- Energy efficiency

Inside an NVIDIA SM (Volta)

- Volta:
  - Tensor cores (8 / SM)
  - Enhanced L1 cache
  - L0, L1 instruction cache