COSC 407
Intro to Parallel Computing

Topic 14: Scheduling, Warps and memory

Outline

Previously:

- Kernel Launch Configuration: nD grids/blocks
- CUDA limits
- Thread Cooperation
- Running Example: Matrix Multiplication

Today:

- Tiling (Improving Performance of Matrix Multiplication)
- CUDA Scalability
- Thread Scheduling on the H/W: Thread Lifecycle
- zero-overhead and latency tolerance
- GPU limits
- CUDA Memories Types (and Performance)

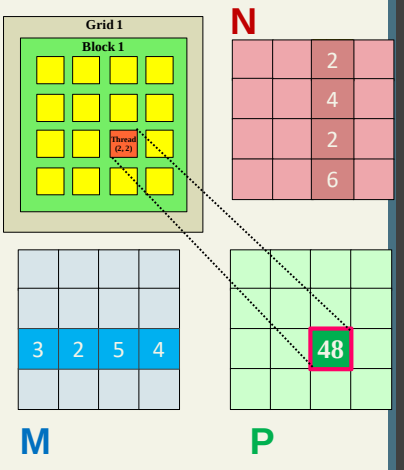
Parallel Code: Using *One* Block

Basic Idea

- Only **ONE** block used to compute the output matrix P
- Each thread computes one element of P as follows:
  - Loads a row of matrix M
  - Loads a column of matrix N
  - Perform one multiply and addition for each pair of M and N elements
  - Compute and stores the result on an off-chip memory (DRAM)

Limitation:

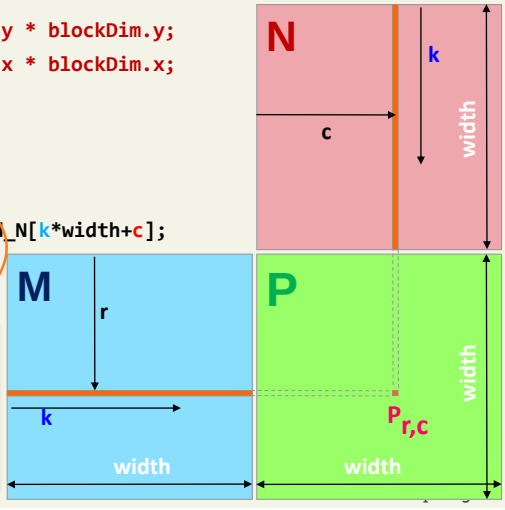
- Size of P is limited to 32x32
  - i.e. the number of threads allowed in a thread block.



Parallel: Kernel – *One* Block

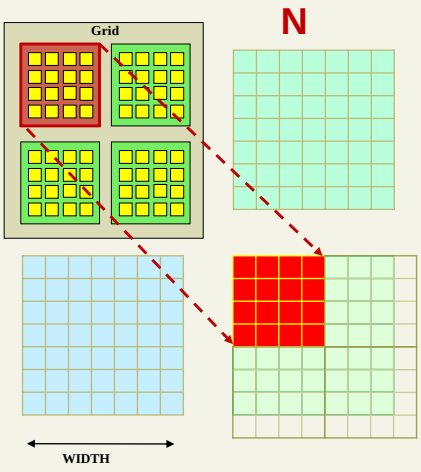
```
// Matrix multiplication kernel – each thread computes one P element
__global__ void MatrixMulKernel(float* d_M, float* d_N, float* d_P, int width){
    //find index of P_{r,c} element
    int r = threadIdx.y + blockIdx.y * blockDim.y;
    int c = threadIdx.x + blockIdx.x * blockDim.x;
    //compute P's element
    if(r<width && c<width){
      float value = 0;
      for (int k=0; k<width; k++)
        value += d_M[r*width+k] * d_N[k*width+c];
      d_P[r*width+c] = value;
  }
}
```

Also ok to use
`int r = threadIdx.y;`
`int c = threadIdx.x;`
But it is better to use the general formula (here, we use only one block, and thus blockIdx = 0). ***WHY BETTER?***



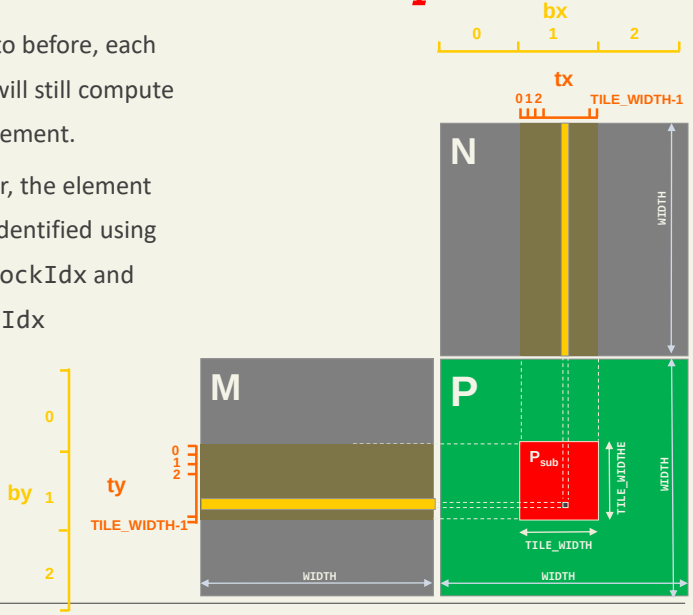
Using *Multiple* Blocks

- We saw that using only **one block has a serious limitation**: size of matrix limited by 1024.
- Also, you are not fully using your GPU
- **Solution**: use multiple blocks
  - We shall apply the method explained previously



Parallel Code: Multiple Blocks

- Similar to before, each thread will still compute one P element.
- However, the element will be identified using both `blockIdx` and `threadIdx`

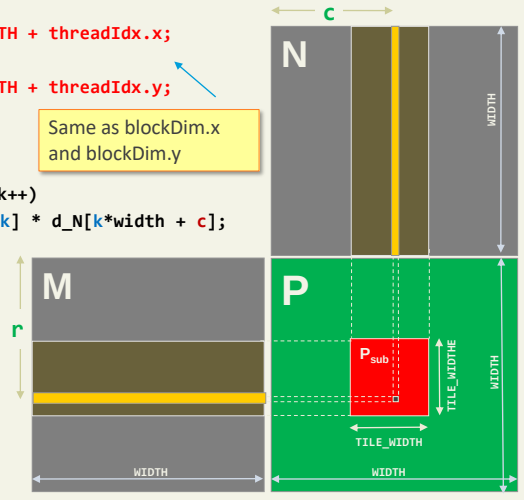


Parallel: Kernel - *Multiple* Blocks

```
__global__ void MatrixMul(float* d_M,float* d_N,float* d_P,int width) {
   //find index of P_r,c element
   int r = blockIdx.x * TILE_WIDTH + threadIdx.x;
   //TILE_WIDTH = TILE_HEIGHT
   int c = blockIdx.y * TILE_WIDTH + threadIdx.y;
   //compute P_r,c element
   if(r<width && c<width){
     float value = 0;
     for (int k = 0; k < width; k++)
         value += d_M[r*width + k] * d_N[k*width + c];
     d_P[r*width + c] = value;
   }
}
```

Same as blockDim.x and blockDim.y

The kernel is the same as before.. The difference is in the host code!



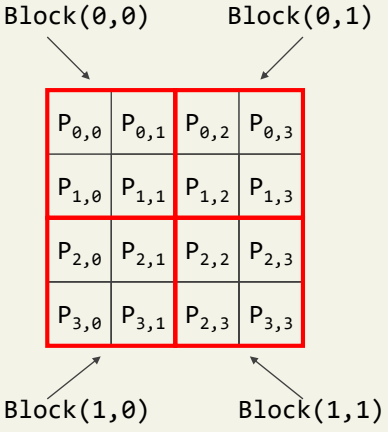
Parallel : Host - *Multiple* Blocks

- The host code will be the same as before except that we need to ***setup the kernel launch configuration***

```
//block dimensions - how many threads per block (our choice but must be <1024)
int TILE_WIDTH = TILE_HEIGHT = n;          // e.g. n = 32
dim3 blockSize(TILE_WIDTH, TILE_HEIGHT);
//grid dimensions (how many blocks are required to cover the whole matrix P)
int nblocks_x = 1+(WIDTH-1)/TILE_WIDTH;    // WIDTH = # elements along x
int nblocks_y = 1+(WIDTH-1)/TILE_HEIGHT;   // HEIGHT = WIDTH
dim3 gridSize(nblocks_x, nblocks_y);
//launch the kernel
MatrixMul<<<gridSize, blockSize>>>(d_M, d_N, d_P, WIDTH);
```

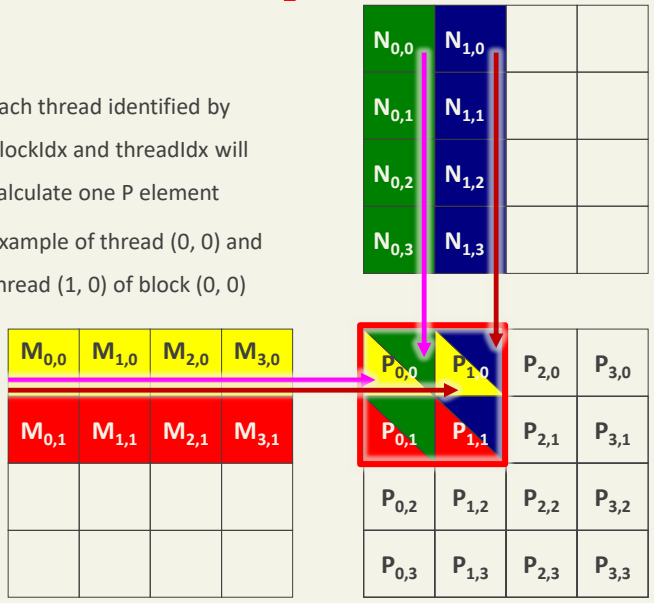
Parallel : *Multiple* Blocks - Details

- To illustrate how the algorithm works, assume
  - TILE_WIDTH = blockDim.x = blockDim.y = 2
  - Each block has 4 threads
- If WIDTH = 4, how many blocks?
  - WIDTH / TILE_WIDTH = 2
  - Use 2* 2 = 4 blocks
- How to identify element $P_{x,y}$?
  - x = TILE_WIDTH * bx + tx
  - y = TILE_WIDTH * by + ty



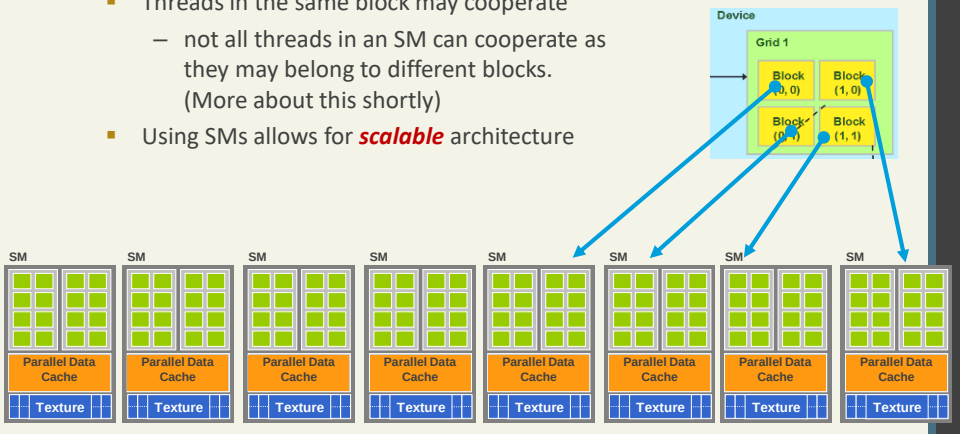
Parallel : *Multiple* blocks - Details

- Each thread identified by blockIdx and threadIdx will calculate one P element
- Example of thread (0, 0) and thread (1, 0) of block (0, 0)



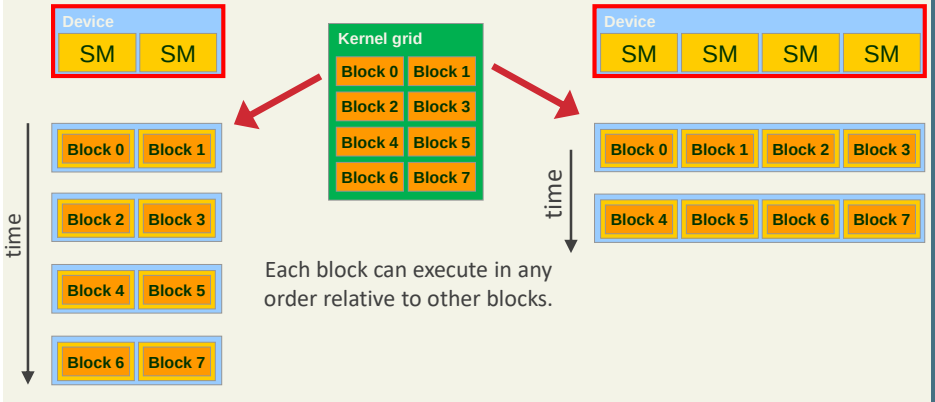
Transparent Scalability

- The GPU is responsible for assigning thread blocks to SMs
  - A block must be assigned to exactly one SM.
  - An SM can run more than one thread block
- Threads in the same block may cooperate
  - not all threads in an SM can cooperate as they may belong to different blocks. (More about this shortly)
- Using SMs allows for ***scalable*** architecture



Transparent Scalability

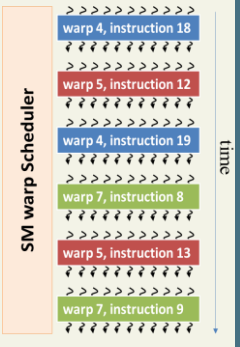
- Hardware is free to assigns blocks to any processor at any time
  - A kernel scales across any number of parallel processors



Each block can execute in any order relative to other blocks.

Warps

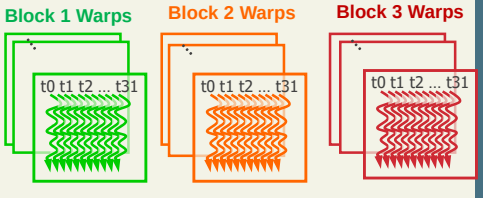
1. Blocks are assigned to SMs (as explained before)
2. Each SM splits threads in its blocks into **Warps**.
   - Groups of threads known as warps in SIMT fashion (execute same instruction)
   - Warps are the **scheduling units of SM**
   - **Thread IDs** within a warp are consecutive and increasing:
     - Warp 0 starts with Thread ID 0
   - Size of the warp is implementation specific
     - Generally **# of threads in a warp (32) = # of SPs** in SM
   - The warp scheduler of SM decides which of the warp gets prioritized during issuance of instructions.
- **DO NOT rely on any ordering between warps**
  - If there are any dependencies between threads, you must synchronize them to get correct results (more on this later).



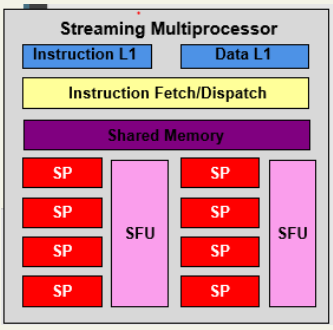
- Warps are not part of the CUDA specification, but
  - Can help optimize the performance in particular devices (discussed later)

Thread Scheduling

- **Each block** is executed as subsequent Warps
  - All threads in a single warp execute in parallel
  - A warp in an SM runs in parallel with Warps in other SMs



- **Question**: Consider a GPU with warp = 32 threads
  - if 3 blocks are assigned to an SM and each block has 256 threads, **how many Warps are there in an SM**?
  - Each Block is divided into 256 / 32 = 8 Warps
  - There are 8 * 3 = 24 Warps



Thread Life Cycle on the HW

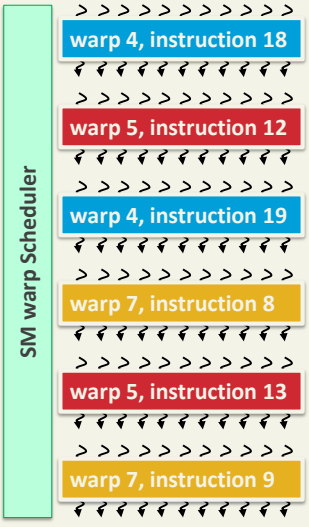
The complete story

1. The ***Grid*** is Launched
2. Blocks are assigned to SMs in arbitrary order
   - Each block is assigned to one SM.
     Each SM is assigned zero or more blocks.
   - There are limits on the number of blocks/threads the SM can track simultaneously. This is taken care of by the GPU.
3. Each block is divided into Warps whose execution is interleaved.
4. Warps are executed by the SM (each SP executes one thread).
   - Threads in a warp run simultaneously.
   - All threads in a warp execute the same instruction when selected

Zero-Overhead and Latency Tolerance

With many warps. those which are ready for consumption are eligible for execution (scheduling priority).

- **Latency hiding**:
  - While a warp is waiting for result from a **long-latency operation** (e.g. global memory access ~500 cylces, floating-point arithmetic, etc), the SM will pick **another warp** that's ready to execute to:
    - **avoid idle time**
    - make full use of the hardware despite long latency operations.
- **Zero-overhead thread scheduling**
  - *Having zero idle time* is referred to as **zero-overhead thread scheduling** in processor designs.



GPU Limits

- CUDA (the software) has limits (as discussed before).
- **GPU** also has limits on how many blocks and threads it can ***simultaneously track*** *(and schedule)*.
  - Hardware resources are required for SMs to maintain the thread, block IDs, and track their execution status.
- For example:
  - **G80** (16 SMs)
    - Each **SM** can track up to **8 blocks** or **768 threads** at a time
      - 3 blocks x 256 threads, or
      - 6 blocks x 128 threads, or …. etc
    - Max number threads at a time = 16 SMs x 768 threads = 12,288 threads
  - **G200** (30 SMs)
    - Each **SM** can process up to **8 blocks** or **1024 threads** at a time
    - Max threads: 30 SMs x 1024 threads = 30,720 threads
- If we assign to the SM more than its max amount of blocks (as per CUDA limits), they will be **scheduled for later execution**.

Benefits?

- Why is it good to know this stuff? (i.e. warps, GPU limits, etc.)
  - One benefit is to allow for **full utilization of each SM** on the GPU
  - Will discuss more on how these concepts are used when **improving performance** in the "CUDA Best Practices"

Granularity

An Example

Consider G80

- **CUDA Limits**: 512 threads per block, 2^{16} x 2^{16} blocks per grid.
  - These are the limits for CUDA 1.0 supported by G80
- **GPU Limits**: 8 blocks or 768 threads per SM

- Assume we have thousands of threads to run. To fully utilize each SM on G80, should we use 8X8, 16X16 or 32X32 threads per block?
  - For **32X32**, we have 1024 threads per Block. Not even one can fit into an SM!
  - For **8X8**, we have 64 threads per Block. Since each SM can take work with only 8 blocks at a time, this means 64x8=512 threads will go into each SM. But since SM needs 768 threads for full utilization, → **66% full - *underutilized (fewer warps to schedule)***
  - For **16X16**, we have 256 threads per Block. Since each SM can take up to 768 threads, it can take up to 3 Blocks and **achieve full capacity and a lot of warps to schedule**.

Utilizing the Hardware

Key things that need to be considered:

- Have a number of blocks >= the number of SMs
  - Want to utilize all SMs
- Have a reasonable number of threads per block
  - Fully utilize each SM

- **Occupancy**
  - Occupancy is defined as the ratio of active warps on an SM to the maximum number of active warps supported by the SM
  - Occupancy varies over time as warps begin and end, and can be different for each SM
  - Low occupancy results in poor instruction issue efficiency; not enough eligible warps to hide latency between dependent instructions

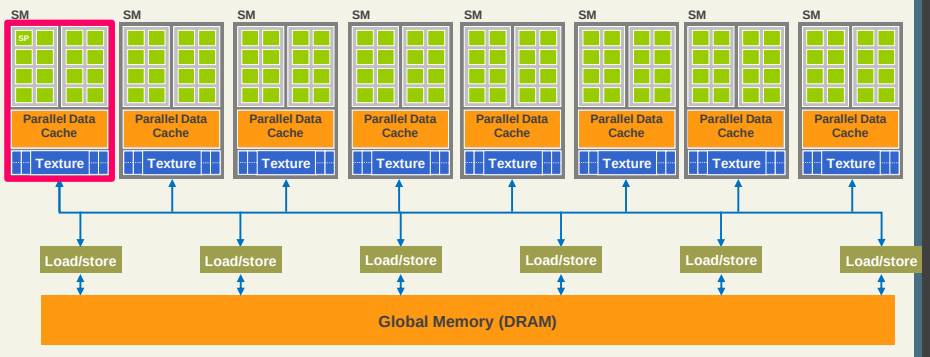
https://docs.nvidia.com/gameworks/content/developertools/desktop/analysis/report/cudaexperiments/kernellevel/achievedoccupancy.htm

Device Memories

Two types

1. ***Programmable***
   - Can control which data to put in that memory
   - Includes
     - Registers
     - Shared memory
     - Local memory
     - Constant memory
     - Global memory
2. ***Non-programmable***
   - *Cannot* control which data is put in that memory
   - Includes
     - L1 Cache memory
     - L2 Cache memory

Remember: GPU Design



Remember:

- A block is assigned to exactly one SM. An SM may run many blocks concurrently.
- All SMs can access the global memory
- We have different memories (as discussed before)

Device Memories

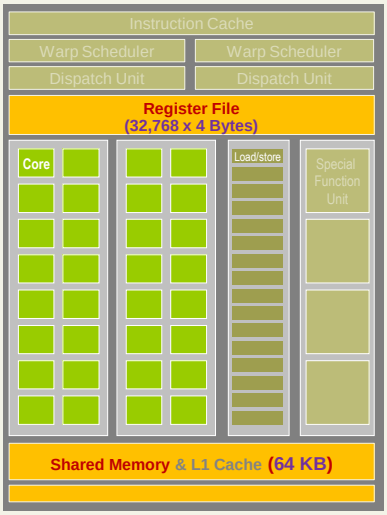
Registers

- **Private to each thread**
- ***Partitioned among threads* in a block**
- More threads means less registers per thread.
  - In diagram: if # of active threads = 1024, then each thread gets 32,768/1024 = **32 registers**

Shared memory

- **Shared by threads in the same block**
- ***Partitioned among blocks***
  - Remember that several blocks may be assigned to the same SM at the same time.
  - More blocks means less shared memory per block.
- There is **64KB** on-chip configurable memory, which is **partitioned between shared memory and L1 cache**.

?



SINGLE SM on Fermi Architecture

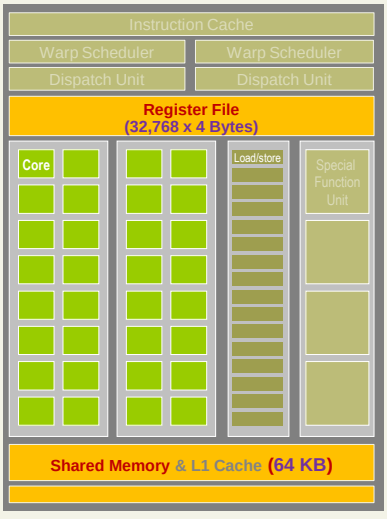
Device Memories

Global Memory (*not shown in figure*)

- **Shared by all threads in the grid**

Question: How this structure helps implementing zero-overhead thread scheduling?

- Zero-overhead (negligible context switching because each thread has its own registers and each block has its own shared memory
- GPU implements ***latency-hiding***
  - Many warps to run, and long latencies from reading global memory are hidden by warp-scheduling
  - The long waiting time of warp instructions is hidden by executing instructions from other warps

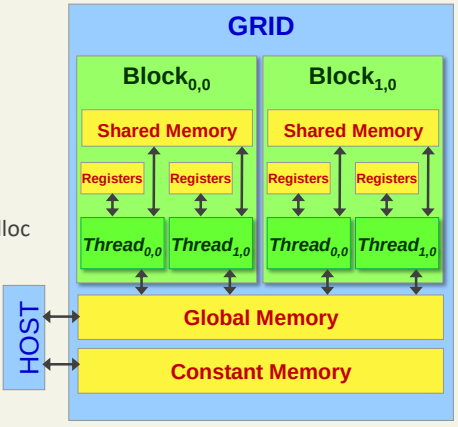


SINGLE SM on Fermi Architecture

Device Memory Model

Device can:

- Read/Write registers
  - ~1 cycle.
- Read/Write shared memory
  - ~5 cycles
- Read/Write global memory
  - ~500 cycles
  - Allocated by *host* using cudaMalloc
- Read only constant memory
  - ~5 cycles with caching
  - a **static global memory** area which is **cache**d.



Host can transfer data per-grid to/from global/constant memory.

Constant Memory

- Like global memory (DRAM) but has a ***dedicated on-chip cache*** for improved performance.
- **Initialized** in **host** code
  - ***Host*** can ***read/write***
  - ***Kernel*** can ***read-only***
- Has limited size (64 KB)
- **Which data is stored in constant memory?**
  - variables declared as `__constant__`
  - `__global__` **function parameters** are passed to the device **via** constant memory (limit is 4 KB)

Parallel Memory Sharing

Registers / Local Memory:

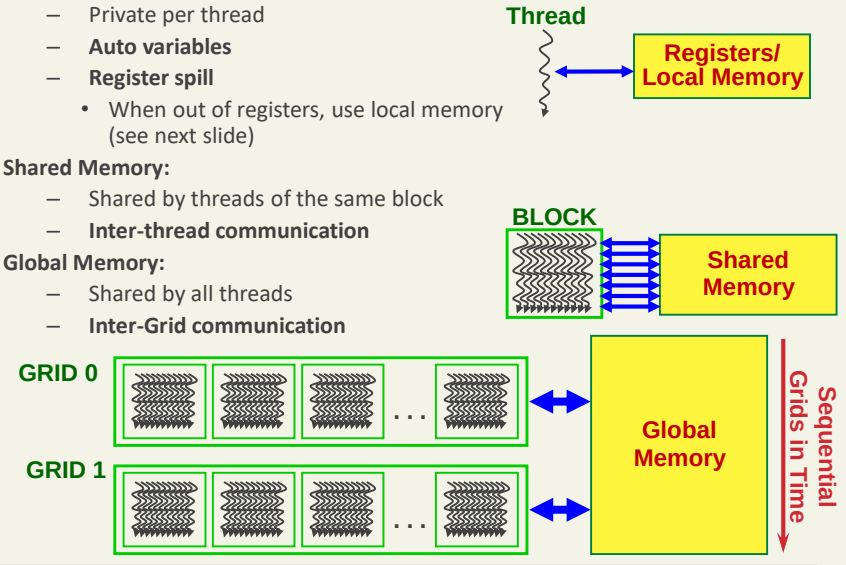
- Private per thread
- **Auto variables**
- **Register spill**
  - When out of registers, use local memory (see next slide)

Shared Memory:

- Shared by threads of the same block
- **Inter-thread communication**

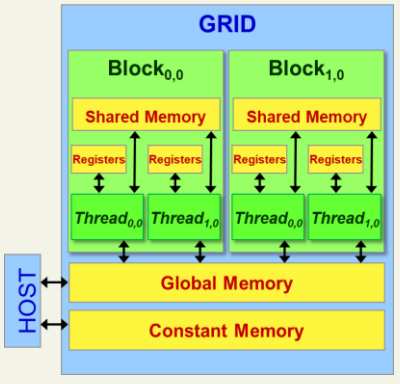
Global Memory:

- Shared by all threads
- **Inter-Grid communication**



Where is the Local Memory?

- Does ***not*** refer to a new physical memory
  - It is on the **global memory**
    - Data is put by the compiler
  - ***Local*** because **each thread** has its own **private area.**
  - Unlike global memory, it is **cached** (L1)



When is the Local Memory Used?

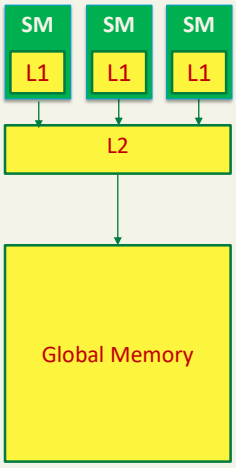
- When we run out of registers (Called *register spilling*)
  - Remember: there is a limit on # of registers per thread
- When declaring arrays inside kernels
  - Some arrays are still stored in registers if small and the compiler can resolve indexing (***registers aren't indexable***)

L1 and L2 Cache

- Cache is ***non-programmable*** small memory
- Cache memories (L1 and L2) help multiple threads that access the ***same memory segment*** so that they do not need to all go to the DRAM

Aside: L2 is coherent. L1 is not coherent.

- "*Not coherent*" means that if two SMs are working on the same global memory location, it is no guaranteed that one SM will immediately see the changes made by the other SM



Summary

Today:

- Tiling
- CUDA Scalability
- Thread Scheduling on the H/W: Thread Lifecycle
- zero-overhead and latency tolerance
- GPU limits
- CUDA Memories Types (and Performance)

Next:

- CUDA Memories Types (and Performance)
- Memory Access Challenges
- Thread Performance
- More Example: Improving Performance of Matrix Multiplication