Assignment #3

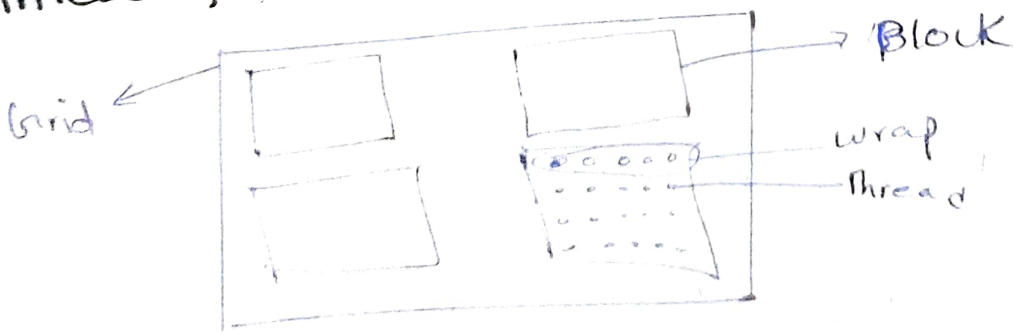
Abdullah Latif
SP23BCS-005

(a)

In GPU computing, these terms define the hierarchy of execution.

1. **Thread**: The smallest unit of execution, running a single sequence of instruction. Each thread has its own Rigisters and a unique ID to calculate which piece of Data it should work on.
2. **Block**: A group of threads that cooperate and share data through shared memory.
3) **Grid**: A collection all thread blocks launched in sinyle kernel execution
4) **Wrap**: A group of 32 consective threads in a block executed simultaneously by a streaming multiprocessor.

Thread, block, Grid, Wrap.



Thread Organization in a Block

Thread in a block can be organized in 1D, 2D or 3D, accessed via threadIdx

(b)

Memory Hierarchy in GPUs

Registers. Fast, on-chip, private to each thread

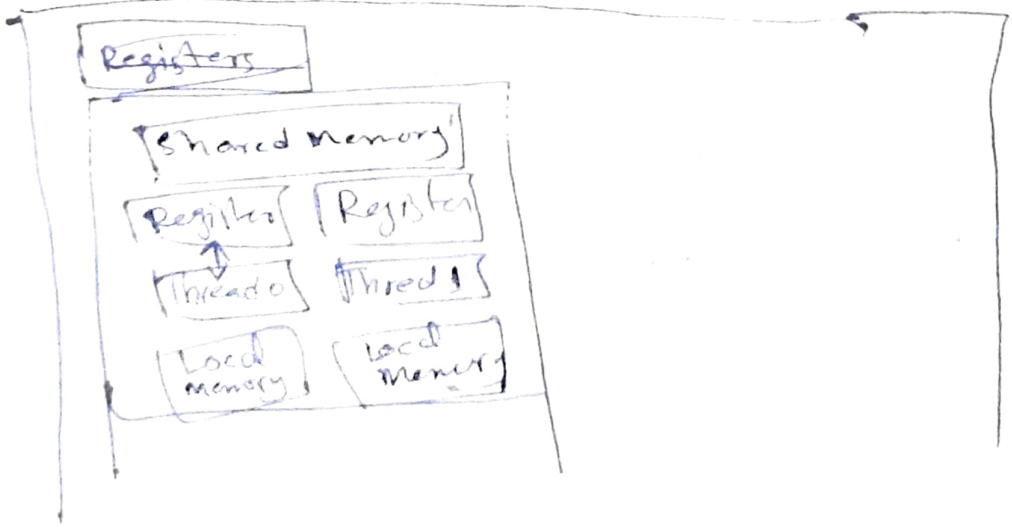
Shared memory: Very fast on-chip, shared by threads in a block.

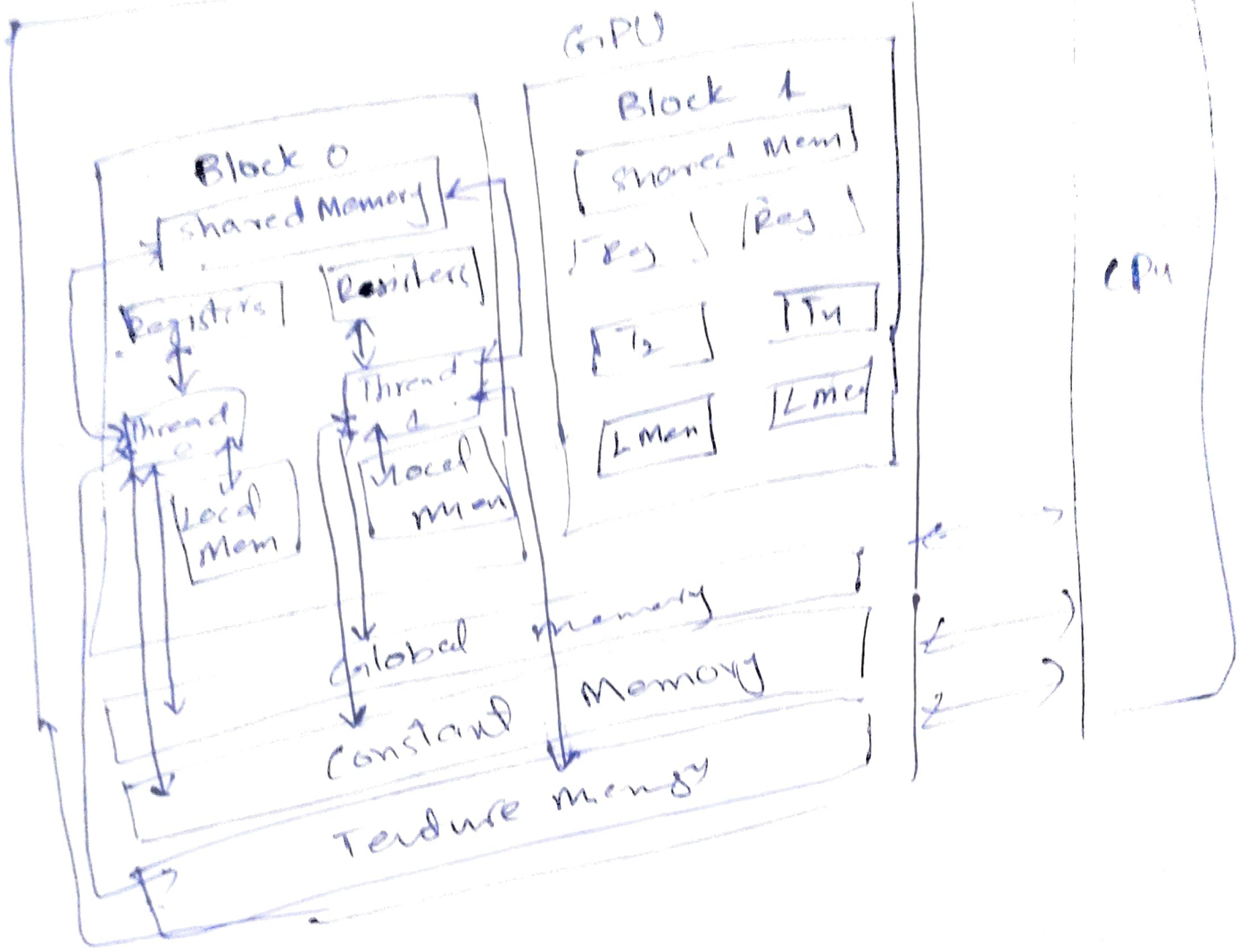
Global Memory: Largest capacity, slow, accesible by all threads.

Constant / Texture Memory

Read only, cached accessible by all threads.

Accels: Thread use registers for local veriables, shared memory for block wide communication and global memory for large datasets.




GPU
Block 0
Shared Memory
Registers
Registers
Thread 0
Thread 1
Local Mem
Local Mem
Block 1
Shared Mem
Reg
Reg
Th
Th
L Mem
L Mem
Global Memory
Constant Memory
Texture Memory
CPU

(C)

Wrap Switching and Latency Hiding

Fast wrap switching is enabled by having dedicated hardware and zero-overhead context switching.

Context-switching

GPU maintain multiple wraps on each SM.

* when one wrap stall.
  → The SM instantly switch to another wrap.

Why Fast
→ state already stored in M.
No costly OS.

Why effective.
→ keep ALUs busy
→ Hides long mem Latency.
→ Increase throughput.

(d)

Branch Instruction.

* GPU use SIMT
* If thread wrap take diff branch
  * Execution become serialized
  * This is called the wrap divergence

Performance Impact.

* Reduce parallel efficiency
* Best to keep wrap thread same path following.

(e)

No of wraps in 16x16.

T per blo = 256

W size = 32

256/32 = 8 wraps

8 wraps.

(f)

Num of Block on a SM with 1536Ts.

T per block = 256

Max T = 1536

1536/256 = 6 Blocks.

(9)

Maximi Occupancy

8x8 block = 8 x 64 = 512 33%.

16x16 Block = 6 x 256 = 1536 100%.

24x24 Block = 2 x 576 = 1152 75%

conclusion.

The 16x16 would be most efficient.