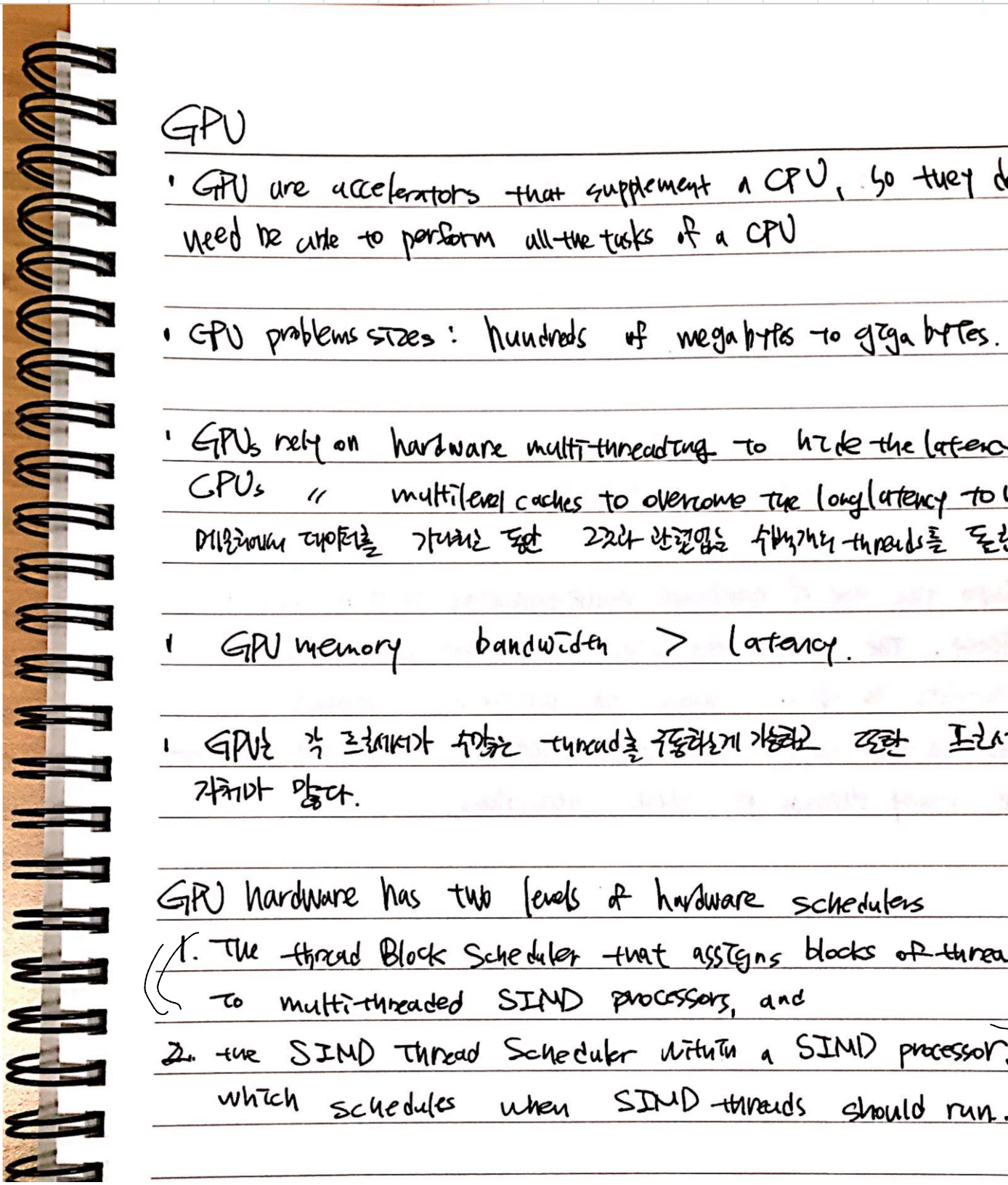
컴퓨터구조03

2018년 3월 19일 월요일 오후 12:35

컴퓨터구조 03



GPU

• GPU are accelerators that supplement a CPU, so they d need be able to perform all the tasks of a CPU

• GPU problems sizes : hundreds of mega bytes to giga bytes.

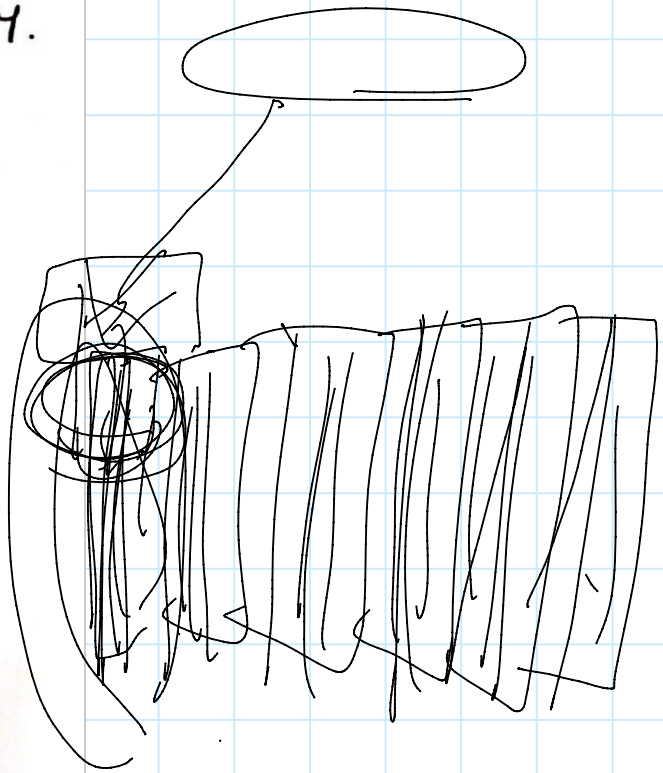
• GPUs rely on hardware multi-threading to hide the latenc GPUs " multilevel caches to overcome the long latency to 대용량의 데이터를 가져올 때 그것과 관련없는 수백개의 threads를 돌

• GPU memory bandwidth > latency.

• GPU는 각 프레임마다 수많은 thread를 동적으로 생성하고 종료 프로 가져야 한다.

GPU hardware has two levels of hardware schedulers

1. The thread Block Scheduler that assigns blocks of threa to multi-threaded SIMD processors, and
2. the SIMD Thread Scheduler within a SIMD processor, which schedules when SIMD threads should run.



Local memory! Shared by SIMD Lanes within multithreaded SIMD processor. But this memory is not shared between SIMD processors.

GPU memory! off chip DRAM shared by the whole GPU and all th

- 전체 application을 모두 담아야 large caches ✗
  smaller streaming caches and rely on extensive multi-threading of threads of SIMD instructions. to hide the long latency to DRAM. O

- working set이 hundreds of megabytes 일 수 있으므로 현재에도 last level cache에 안담김.

- Given the use of hardware multithreading to hide DRAM latency. the chip area used for cache in system processors is spend instead on computing resources and on the large number of registers to hold the state of the many threads of SIMD instructions.

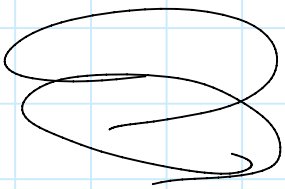
multi-thread

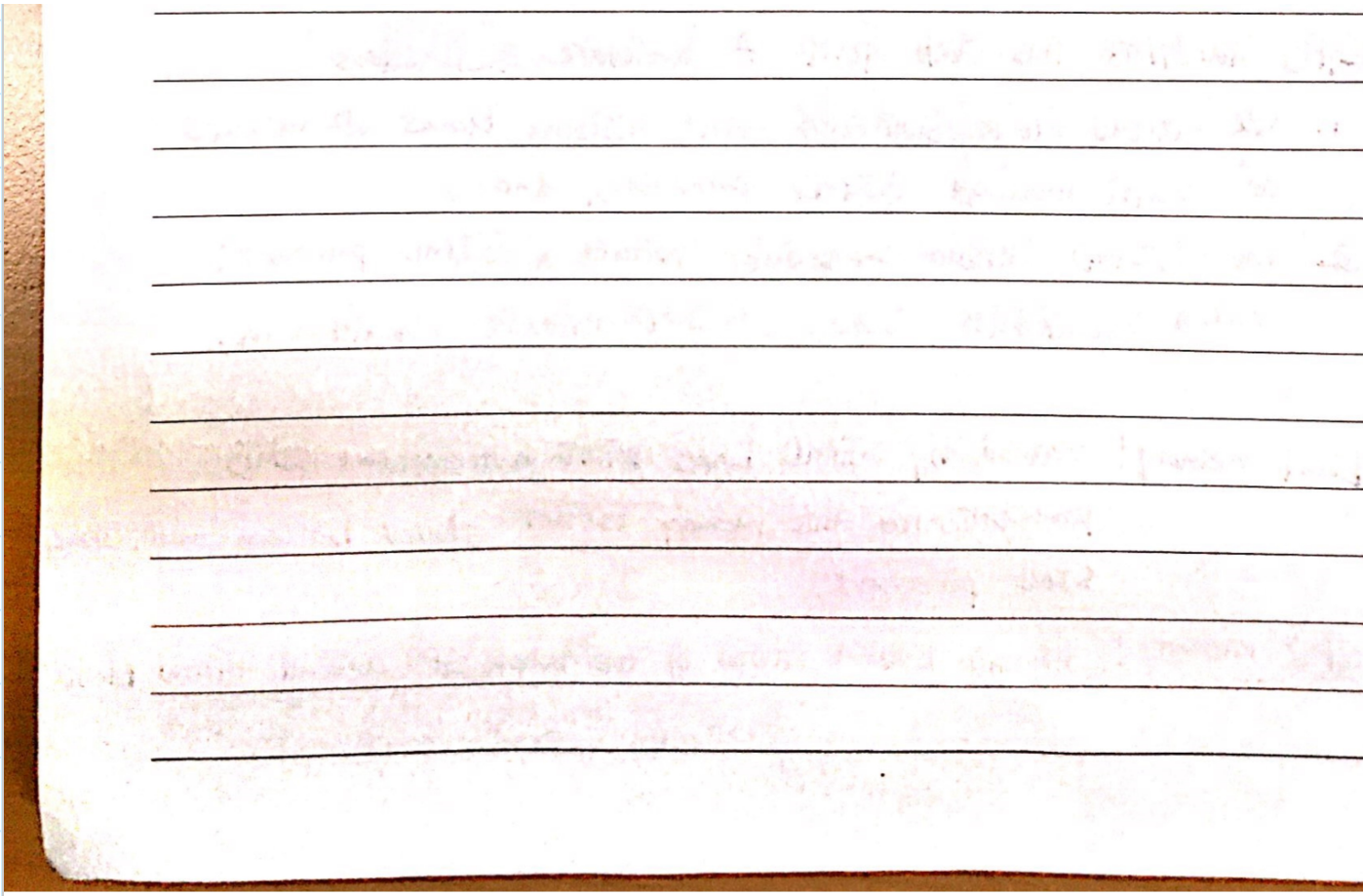
read blocks

학교

개발원

Scanner





Feature	Multicore with SIMD	
SIMD processors	4 to 8	
SIMD lanes/processor	2 to 4	
Multithreading hardware support for SIMD threads	2 to 4	
Largest cache size	8 MiB	
Size of memory address	64-bit	
Size of main memory	8 GiB to 256 GiB	4
Memory protection at level of page	Yes	
Demand paging	Yes	
Cache coherent	Yes	

FIGURE 6.11 Similarities and differences between multicore with Multimedia extensions and recent GPUs.

Scanner

GPU
8 to 16
8 to 16
16 to 32
0.75 MiB
64-bit
GiB to 6 GiB
Yes
No
No

SIMD

Type	More descriptive name	Closest old term outside of GPUs	Official CUDA/ NVIDIA GPU term	Book definition
Program abstractions	Vectorizable Loop	Vectorizable Loop	Grid	A vectorizable loop, executed on the GPU, up of one or more Thread Blocks (bodies o vectorized loop) that can execute in paralle
	Body of Vectorized Loop	Body of a (Strip-Mined) Vectorized Loop	Thread Block	A vectorized loop executed on a multithrea SIMD Processor, made up of one or more of SIMD instructions. They can communica Local Memory.
	Sequence of SIMD Lane Operations	One iteration of a Scalar Loop	CUDA Thread	A vertical cut of a thread of SIMD instructi corresponding to one element executed by SIMD Lane. Result is stored depending on and predicate register.
Machine object	A Thread of SIMD Instructions	Thread of Vector Instructions	Warp	A traditional thread, but it contains just SI instructions that are executed on a multith SIMD Processor. Results stored depending per-element mask.
	SIMD Instruction	Vector Instruction	PTX Instruction	A single SIMD instruction executed across Lanes.
Processing hardware	Multithreaded SIMD Processor	(Multithreaded) Vector Processor	Streaming Multiprocessor	A multithreaded SIMD Processor executes threads of SIMD instructions, independent other SIMD Processors.
	Thread Block Scheduler	Scalar Processor	Giga Thread Engine	Assigns multiple Thread Blocks (bodies of vectorized loop) to multithreaded SIMD Processors.
	SIMD Thread Scheduler	Thread scheduler in a Multithreaded CPU	Warp Scheduler	Hardware unit that schedules and issues t of SIMD instructions when they are ready t execute; includes a scoreboard to track SI Thread execution.
	SIMD Lane	Vector lane	Thread Processor	A SIMD Lane executes the operations in a of SIMD instructions on a single element. stored depending on mask.
Memory hardware	GPU Memory	Main Memory	Global Memory	DRAM memory accessible by all multithrea SIMD Processors in a GPU.
	Local Memory	Local Memory	Shared Memory	Fast local SRAM for one multithreaded SIM Processor, unavailable to other SIMD Proc
	SIMD Lane Registers	Vector Lane Registers	Thread Processor Registers	Registers in a single SIMD Lane allocated a full thread block (body of vectorized loop

FIGURE 6.12 Quick guide to GPU terms. We use the first column for hardware terms. Fou cluster these 12 terms. From top to bottom: Program Abstractions, Machine Objects, Processing H and Memory Hardware.

made of el.
ded threads ate via
ons one mask
MD hreaded g on a
SIMD
of
hreads to MD
thread Results
aded
MD essors.
l across o).

ur groups
Hardware,