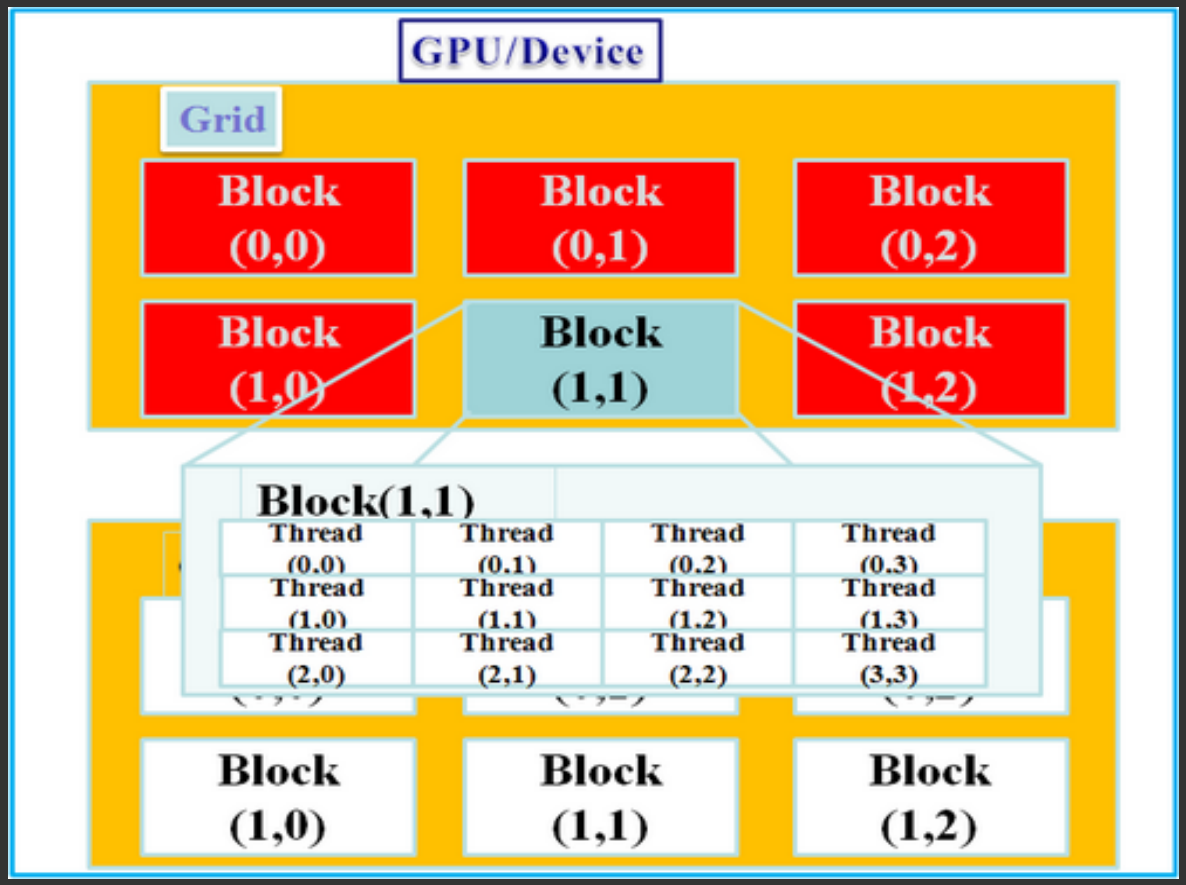
CUDA

Ethan Chen, William Chiu, and
Wenchuan Weng

Overview

- GPUs massively parallel processors earlier than CPUs
- CUDA
  - Compute Unified Device Architecture
  - Programming model introduced by NVIDIA in 2007
  - Minimal extension of C and C++ language
- Divides execution
  - Little or no parallelism on CPU *(host)*
  - Data parallelism on GPU *(device)*
- Extremely lightweight threads
  - Zero overhead for thread scheduling
  - Negligible overhead for running or creating thread

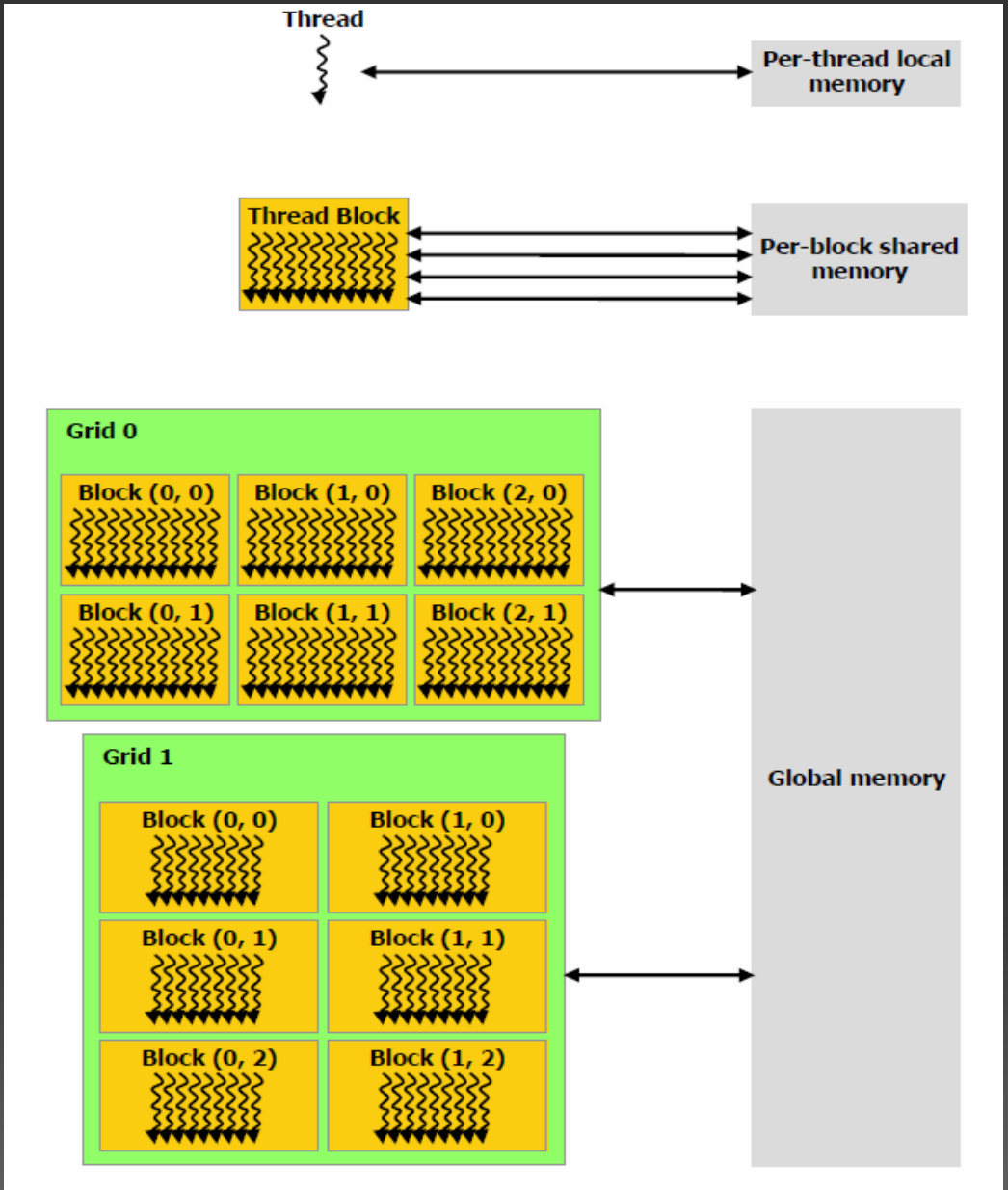
Hierarchy



- Threads indexed by *threadIdx*
- Thread-blocks indexed by *blockIdx*
- Grids

Source: http://inha.inwebcard.kr/sub.php?tname=1243339800

Memory Model



- Memory Model
  - Private local memory
  - Shared thread-block memory
  - Global kernel memory
- Synchronization
  - Barrier among thread blocks
  - Global among kernels

Source: http://cyberaide.googlecode.com/svn/trunk/papers/thesis-pangborn/images/

Demo

Map Reduce
Parallel Prefix

Questions?

References

NVIDIA, "CUDA C Programming Guide"

Mark Harris, "Parallel Prefix Sum (Scan) with CUDA"

Mark Harris, "Optimizing Parallel Reduction in CUDA"

http://inha.inwebcard.kr/sub.php?tname=1243339800

http://cyberaide.googlecode.com/svn/trunk/papers/thesis-pangborn/images/