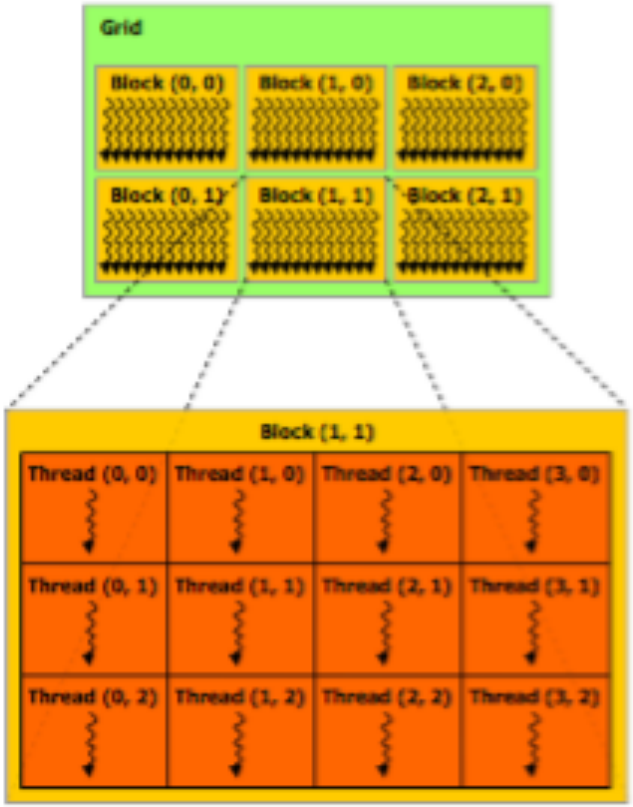
Intro to CUDA Programming
CS 540 – High Performance Computing

CUDA Thread Organization

- CUDA Thread Concepts you MUST know:
  - Grids
  - Blocks
- Grids made of blocks
- Blocks made of threads
- threadIdx – threadId
  - built-in variable that stores the id of each thread
- blockIdx – blockId
- blockDim – block dimensions
- gridDim – grid dimensions

CUDA Grids and Blocks



NVIDIA: **http://docs.nvidia.com/cuda/cuda-c-programming-guide/graphics/grid-of-thread-blocks.png**

Grids and Blocks cont.

- Maximum number of threads per block:
  - 512 (older GPU's)
  - 1024
- Maximum number of blocks per grid is 65535
- If you exceed these limits , the GPU return useless data or terminate executing
- When invoking a kernel, threads are divided into a **Grid** of **Blocks**
  - Kernel<<<**block_count**, **thread_per_block**>>>(...);

References

- CUDA by Example – Jason Sanders et.al.
- http://courses.cms.caltech.edu/
- https://en.wikipedia.org/wiki/Thread_block
- https://developer.nvidia.com/cuda-education
- https://en.wikipedia.org/wiki/Context_switch
- CME 214 Introduction to parallel computing using MPI, openMP and CUDA – Eric Darve, Stanford University