

The background of the slide is a light gray gradient. It is decorated with numerous realistic water droplets of various sizes. Some droplets are large and prominent, while others are small and subtle. They are scattered across the slide, with a higher concentration in the top-left and bottom-right corners. Each droplet has a highlight and a shadow, giving it a three-dimensional appearance.

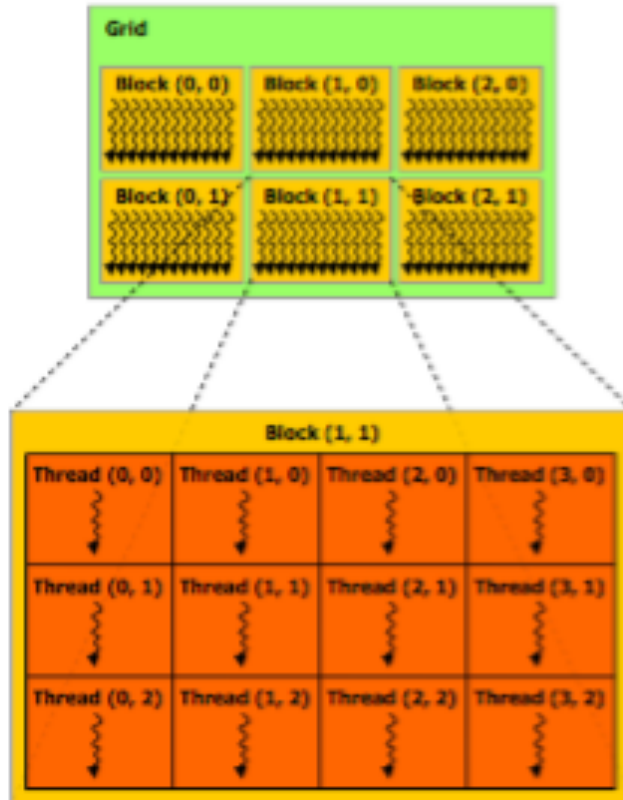
# Intro to CUDA Programming

## 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 – threadIdx
  - built-in variable that stores the id of each thread
- blockIdx – blockIdx
- 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://en.wikipedia.org/wiki/Thread_block)
- <https://developer.nvidia.com/cuda-education>
- [https://en.wikipedia.org/wiki/Context\\_switch](https://en.wikipedia.org/wiki/Context_switch)
- CME 214 Introduction to parallel computing using MPI, openMP and CUDA – Eric Darve, Stanford University