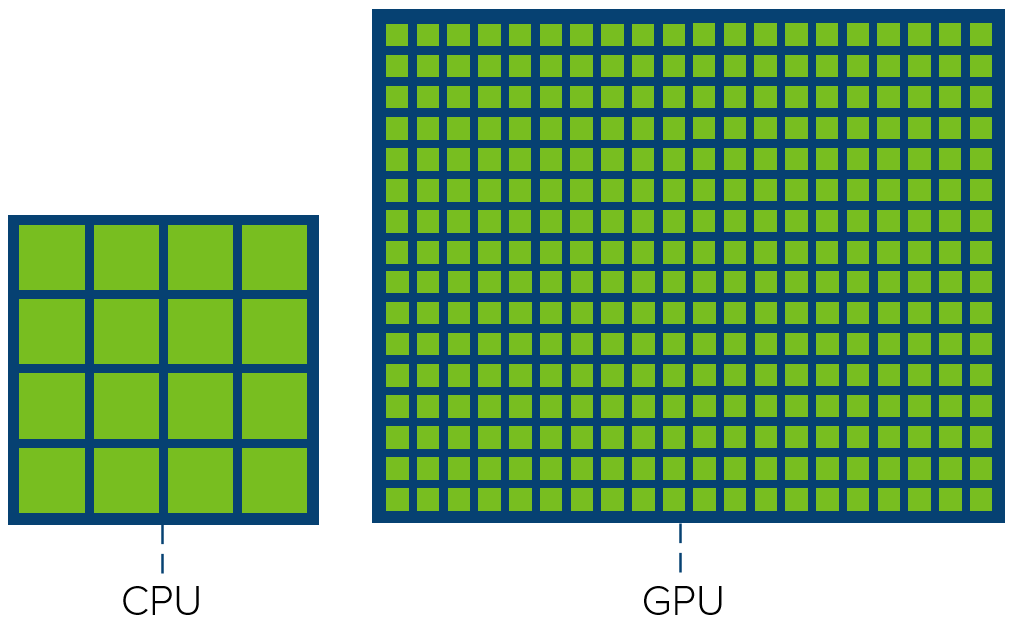
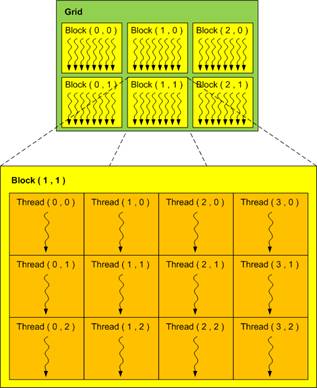
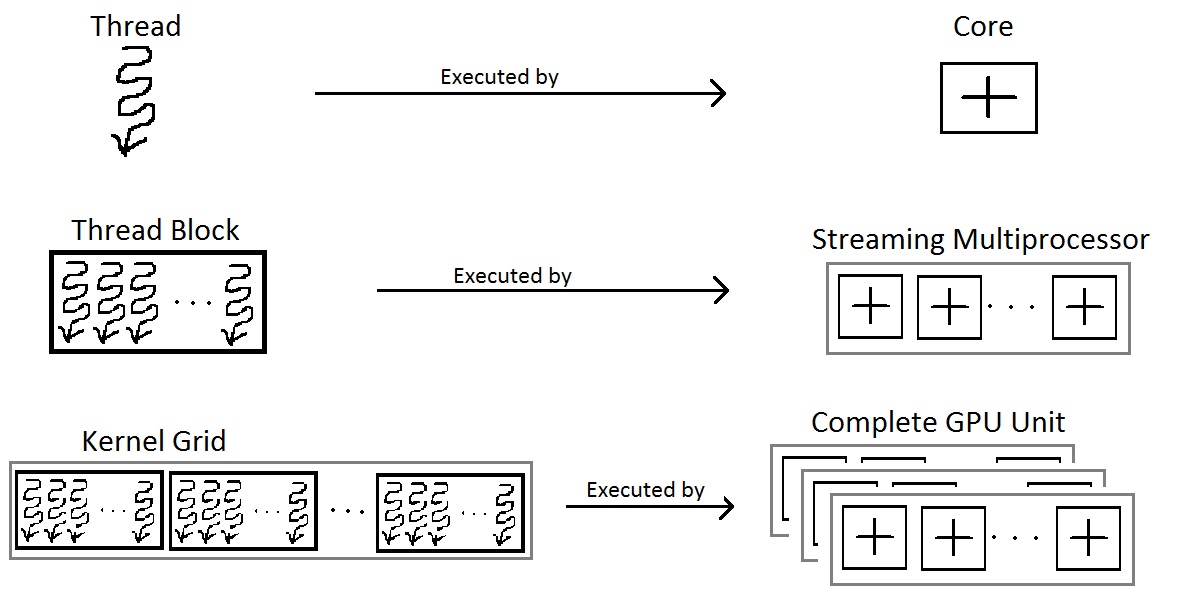
GPU architecture:

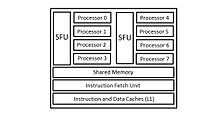




Although we have stated the hierarchy of threads, we should note that, threads, thread blocks and grid are essentially a programmer’s perspective. In order to get a complete gist of thread block, it is critical to know it from a hardware perspective. The hardware groups threads that execute the same instruction into warps. Several warps constitute a thread block. Several thread blocks are assigned to a Streaming Multiprocessor (SM). Several SM constitute the whole GPU unit (which executes the whole Kernel Grid).[



Each architecture in GPU (say [Kepler](https://en.wikipedia.org/wiki/Kepler_(microarchitecture)) or [Fermi](https://en.wikipedia.org/wiki/Fermi_(microarchitecture))) consists of several SM or Streaming Multiprocessors. These are general purpose processors with a low clock rate target and a small cache. An SM is able to execute several thread blocks in parallel. As soon as one of its thread blocks has completed execution, it takes up the serially next thread block. In general, SMs support [instruction-level parallelism](https://en.wikipedia.org/wiki/Instruction-level_parallelism) but not branch prediction.[[7]](https://en.wikipedia.org/wiki/Thread_block_(CUDA_programming)#cite_note-:1-7)

[](https://en.wikipedia.org/wiki/File:Streaming-Multiprocessor.jpg)

An illustration of a streaming multiprocessor and its resources.[[8]](https://en.wikipedia.org/wiki/Thread_block_(CUDA_programming)#cite_note-8)

To achieve this purpose, an SM contains the following:[[7]](https://en.wikipedia.org/wiki/Thread_block_(CUDA_programming)#cite_note-:1-7)

* Execution cores. (single precision floating-point units, double precision floating-point units, special function units (SFUs)).
* Caches:

1. L1 [cache](https://en.wikipedia.org/wiki/Cache_(computing)). (for reducing memory access latency).
2. [Shared memory](https://en.wikipedia.org/wiki/Shared_memory). (for shared data between threads).
3. Constant cache (for broadcasting of reads from a read-only memory).
4. [Texture cache](https://en.wikipedia.org/wiki/Texture_cache). (for aggregating bandwidth from texture memory).