WARPS

WARPS

- Warps are the basic unit of execution on the GPU.
- The GPU is effectively a collection of SIMD vector processors.
- Each group of threads, or warps, is executed together.
- In the ideal case, only one fetch from memory for the current instruction and a broadcast of that instruction to the entire set of SPs in the warp.
- This is much more efficient than the CPU model, which fetches independent execution streams to support task-level parallelism.
- In the CPU model, for every core you are running an independent task, you can conceptually divide the memory.

WARPS

- With GPU programming, It’s vector architecture and expects you to write code that runs on thousands of threads.
- This approach allows you to check things, such as whether memory copying to/from the GPU is working correctly, before introducing parallelism into the application.
- Warps on the GPU are currently 32 elements, although nVidia reserves the right to change this in the future.