HOST

INPUT ASSEMBLER

THREAD EXECUTION MANAGER

Registers
Registers hold local variables. Just like L1 Cache, they are shared among all threads in a block.

A **warp** is a group of **32 threads** that:
- Execute in lockstep on the GPU using the SIMD model.
- Share the same program counter (PC) but operate on unique data.
- Are managed by the GPU scheduler for efficient parallel execution.
- Threads in a warp can share data through **shared memory**, but care must be taken to avoid bank conflicts.

Warp Divergence
- **What is it?**
  - Divergence occurs when threads in the same warp follow different execution paths, typically due to conditional branching (`if`, `switch`, etc.).
- **Impact**:
  - Diverged paths are executed sequentially, with threads not on the current path being masked (inactive), reducing efficiency.
- **Mitigation**:
  - Minimize divergence by structuring code so threads in a warp follow similar execution paths.

Block
A **block** is a collection of threads that execute on a **Streaming Multiprocessor (SM)**:
- Threads in a block can synchronize using `__syncthreads()`.

Block Grouping
A block consists of multiple threads (up to 1024 threads on modern GPUs).
- Threads in a block are uniquely identified by a **thread index** (`threadIdx.x`, `threadIdx.y`, `threadIdx.z`).
- I can't have a block spanning over multiple SMs, but I can have more blocks running on the same SM.

Shared Mem. and Registers
Threads in a block share a common **shared memory space**.
- Shared memory is faster than global memory and is used for inter-thread communication within a block.
- Resources like **registers** and **shared memory** are divided among all threads in a block.
  - Larger blocks require more resources, which can limit the number of active blocks per SM.

Shared Memory
Fast on-chip memory to hold frequently used data. Can be used to exchange data between the corse of the same SM

BUILDING BLOCK

SM
(Streaming Multiprocessor/CUDA Core)

REGISTERS

~ 8 TB/s

BLOCK

WARP

thread

~1.7 TB/s

L1 CACHE (small shared memory)

~ XXX GB/s

L2 CACHE

~200 MB/s

BOARD MEMORY

TEXTURE MEMORY

GLOBAL MEMORY

CONSTANT MEMORY

Texture and surface memory
Content managed by special hardware that permits fast implementation of some filtering/interpolation operatotions

Global memory
Main part of the off-chip memory. High capacity but relatively slow. The only part accessible by the host through CUDA functions

Constant memory
Can only store constants. It is cached, and allows broadcasting of a single value to all threads in a warp (less appealing on newer GPUs that have a cache anyway)