SITCON'21

GPU MEMORY HIERARCHY

- Shared memory

- Per blocks / MP

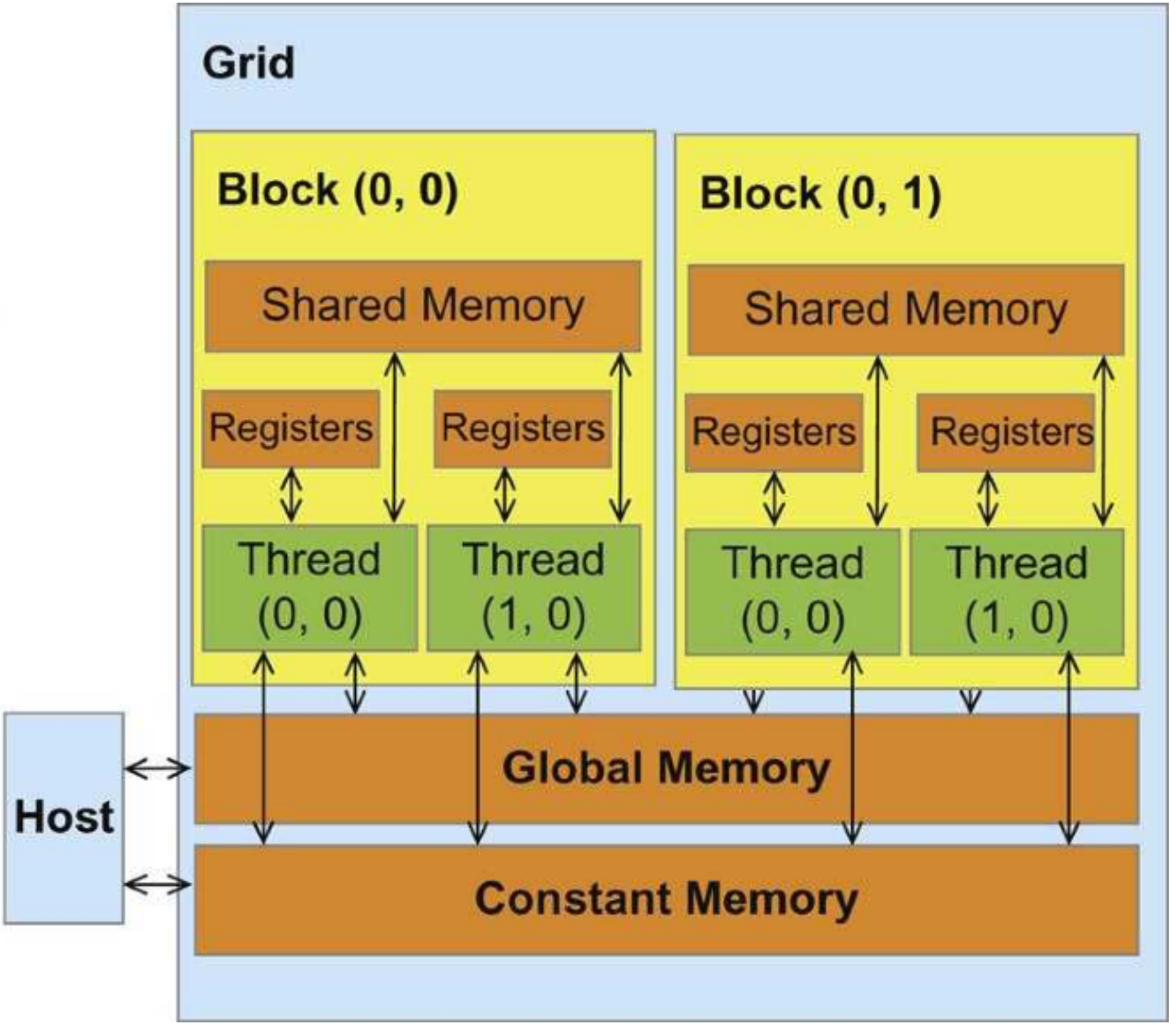
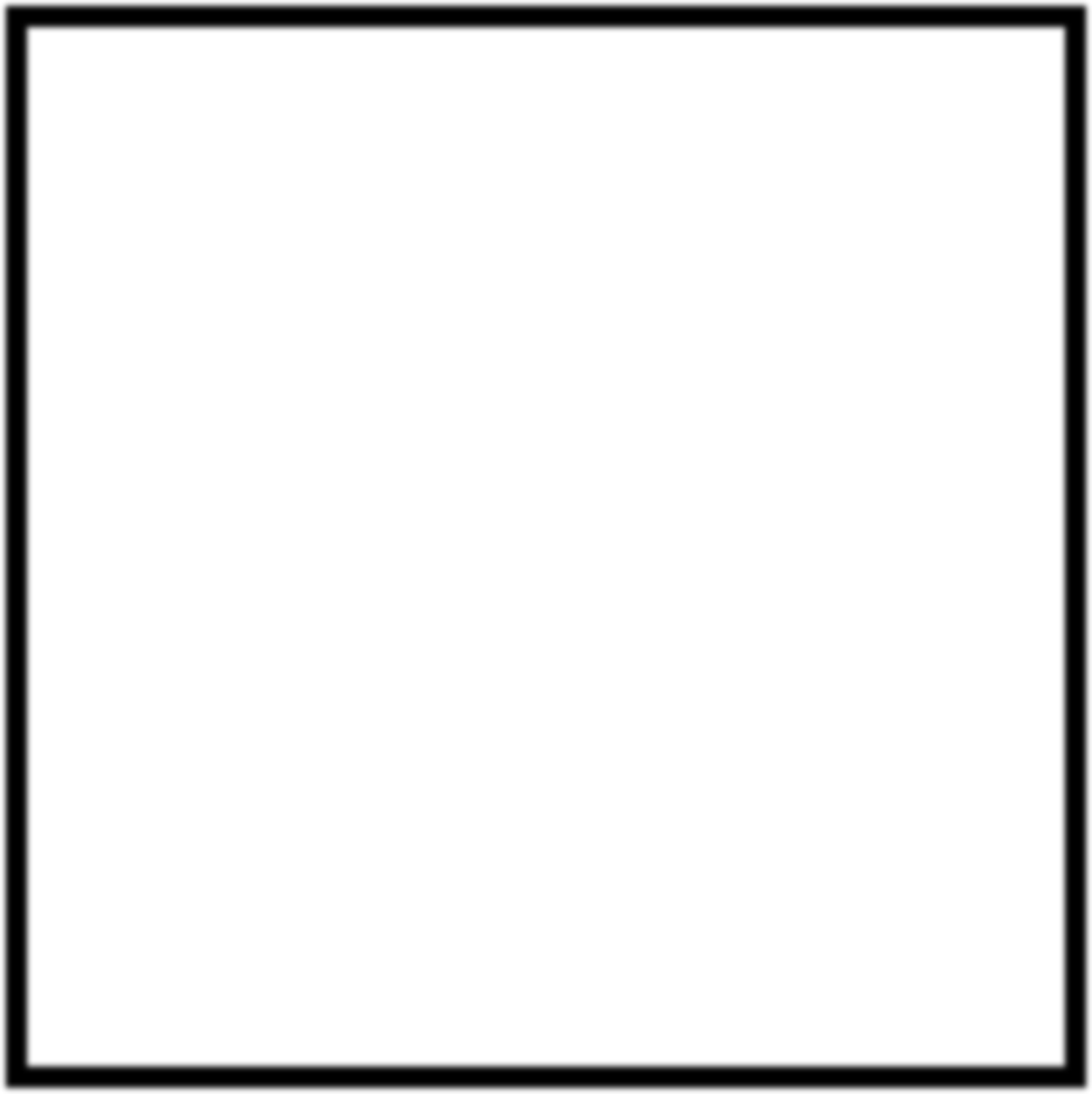
L1 cache

- Per grid / GPU device

- Global memory

- Constant / Texture memory

L2 cache


Grid
Block (0, 0)
Block (0, 1)
Shared Memory
Shared Memory
Registers
Registers
Registers
Registers
Thread (0, 0)
Thread (1, 0)
Thread (0, 0)
Thread (1, 0)
Host
Global Memory
Constant Memory