CUDA访存模型

可编程内存

- Registers
- Local memory
- Shared memory (on-chip)
- Constant memory (with dedicated cache)
- Texture memory (with dedicated cache)
- Global memory
  - L2 cache[, L1 cache]

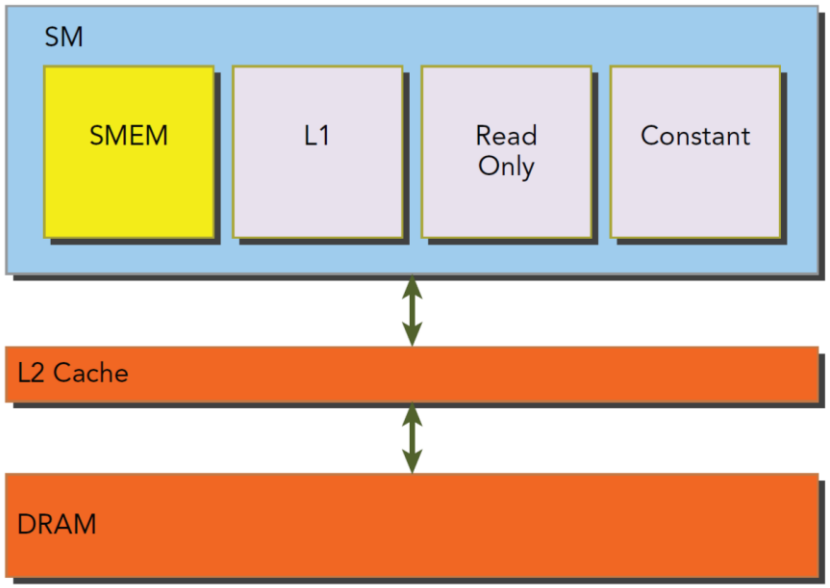

FIGURE 5-1

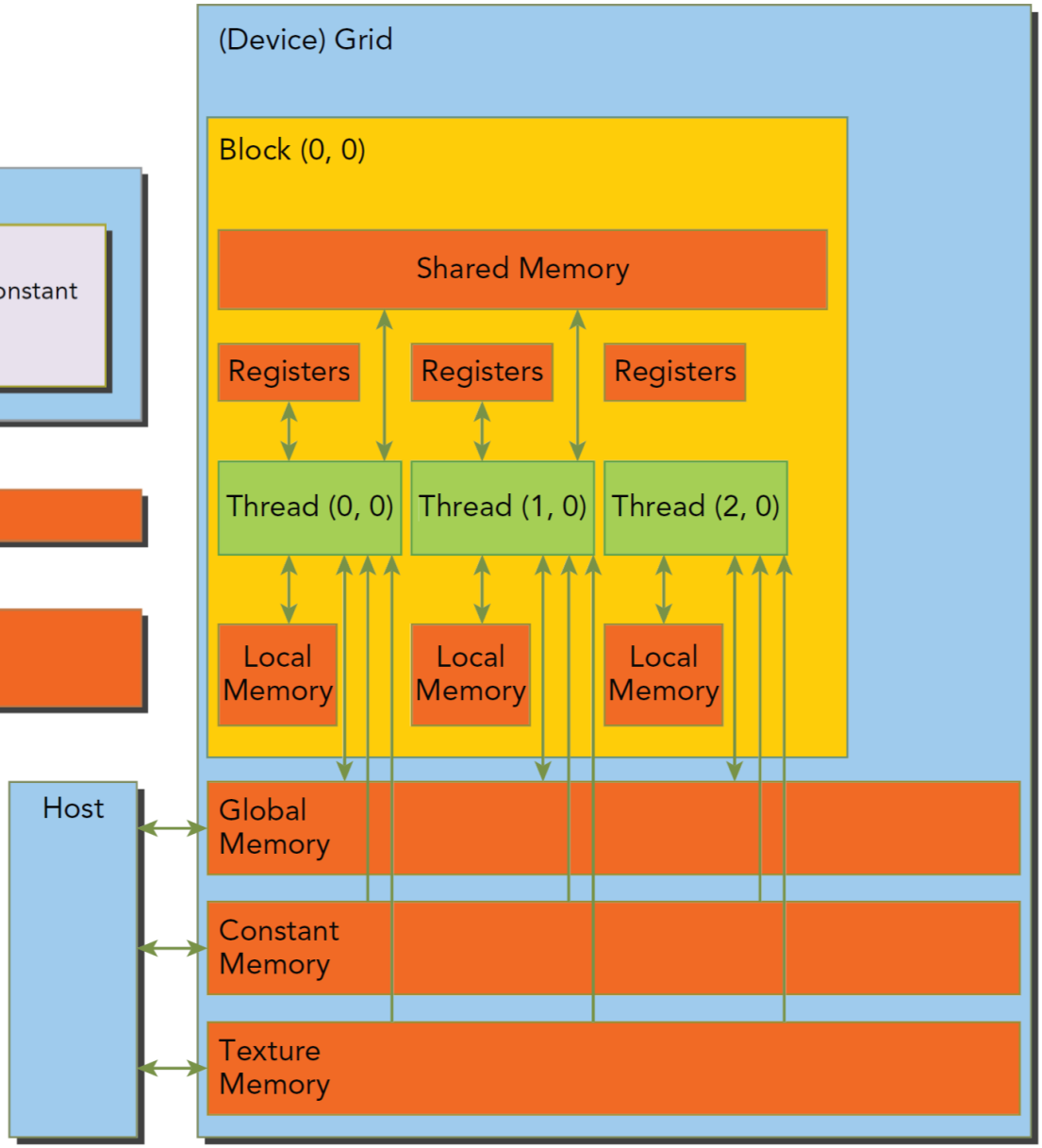

FIGURE 4-2

Registers

- 在kernel中声明的，无限定符的
  - 变量
  - 编译时可确定大小的数组
- 特点：
  - 线程私有
  - 分配给活跃warp
  - Spilling

Local memory

- 在kernel中声明的，无限定符的
  - 编译时不确定大小的数组
  - 大的结构或数组
  - Spilled registers
- 特点：
  - 线程私有
  - 与global memory类似
  - Warp shuffle支持

Global memory

- 声明
  - Host code: cudaMalloc, cudaFree
  - Device code: 限定符__device__
- 特点:
  - 最大，延迟最高
  - L2 (32-byte line), L1 cache (128-byte line)
  - 写操作不使用L1 cache

Global memory

- Pinned memory
  - 更快读写，批量读写
  - 申请开销大，可用内存变小可能影响CPU性能
- Zero-copy memory
  - 映射到device地址空间的pinned memory
- Unified virtual addressing
  - 统一地址空间
- Unified memory
  - 数据自动迁移，指针统一

Global memory

- 访存模式
  - 由device内存事务支持
    - 读操作32-byte/128-byte
    - 写操作32-byte，每次可写1/2/4个段
  - 对齐访问，连续访问
  - Prefer Structure of Arrays
- 优化
  - 目标：更多对齐和连续访问，更多并发访存
  - Unrolling，修改执行参数，对角化block坐标

Shared memory

- 在kernel内声明，限定符__shared__
- 特点：
  - On-chip，低延迟、高带宽
  - 分配给block
  - Block内线程间通信，__syncthreads();
  - 与L1 cache共享硬件

Shared memory

- 访存模式
  - 一个warp的请求由1-32个事务完成
  - 访问同一个word由multicast实现
  - 划分为32个bank
    - 按列进行读操作时padding，避免冲突
  - 同步：barrier, memory fence和volatile限定符

Constant memory

- 全局声明，限定符__constant__
- 特点：
  - Read-only
  - 专用cache
  - Broadcast，访问相同位置时性能最好

Texture memory

- 访问
  - __ldg函数
  - 参数限定符const __restrict__
- 特点：
  - Read-only
  - 专用cache，专用硬件支持
  - 访问分散位置时性能较好