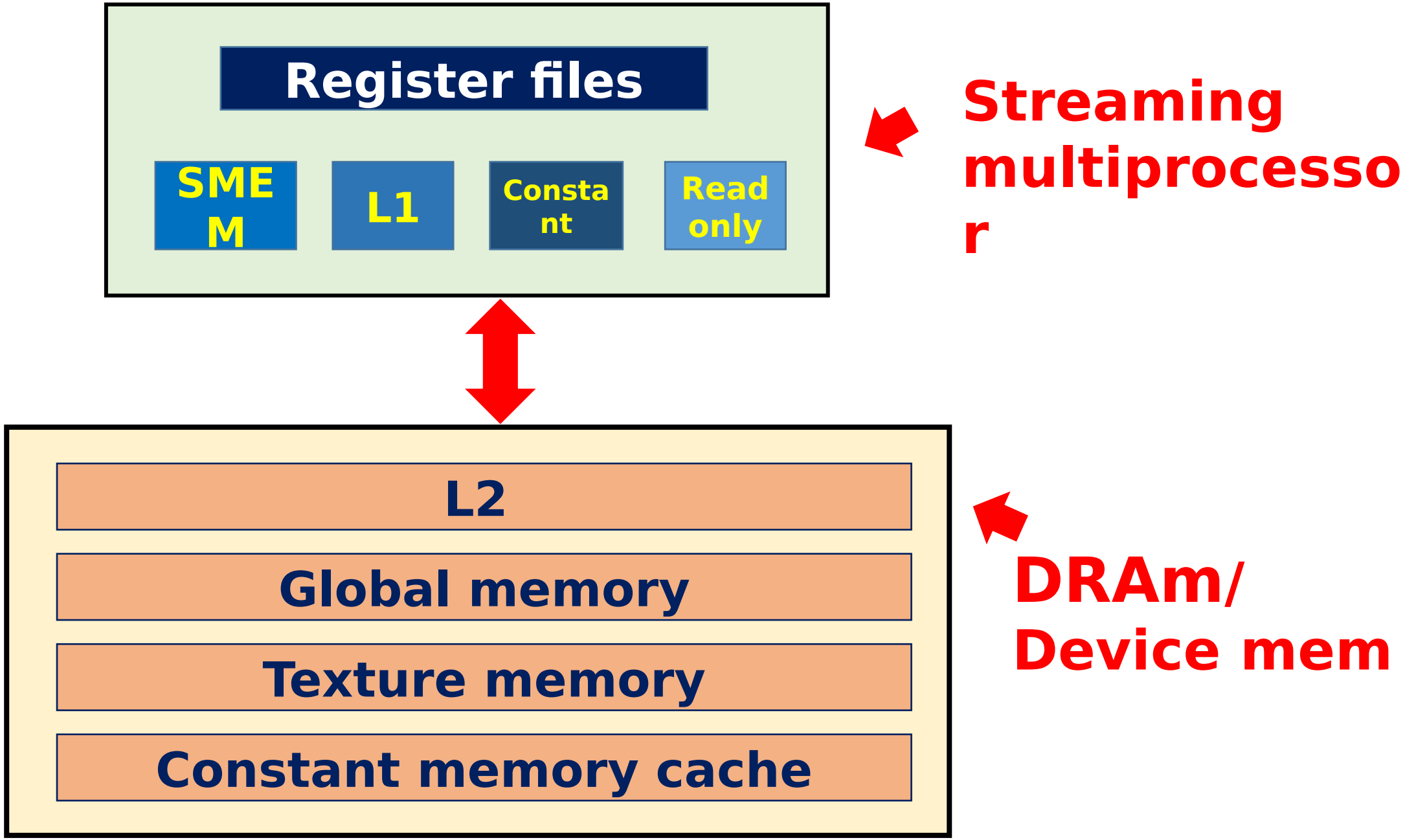
CUDA memory types


Register files
SME M
L1
Consta nt
Read only
Streaming multiprocesso r
L2
Global memory
Texture memory
Constant memory cache
DRAm/ Device mem

Registers

* Fastest memory space in the GPU

* Use to hold frequently accessed thread-private variables, and arrays if the indices are constant or can be determine at compile time.

* share their lifetime with the kernel

- On Fermi GPUs one thread can have maximum of 63 registers. But all other microarchitectures allowed to have maximum of 255 registers per thread.

Register spills

If a kernel uses more registers than the hardware limit, the excess registers will spill over to local memory. This register spilling can have adverse performance consequences.

Launch bounds

__launch_bounds__(maxThreadsPerBlock,

minBlocksPerMultiprocessor)

Ex : __global__ void __launch_bounds__(48)
register_usage_test(int * results, int size)

-maxrregcount=32

Local memory

- Store variables which are eligible for registers but cannot fit into the register space
  - local arrays with indices which cannot resolve at compiler time.
  - Large local structures

- Not an on-chip memory, allocates in DRAM so have high latency memory access

Shared memory

- Shared memory is on chip memory which partition among thread blocks.

- __shared__

- The L1 cache and shared memory for an SM use the same on-chip memory

- Constant memory
- Texture Memory
- Global Memory
- GPU Caches