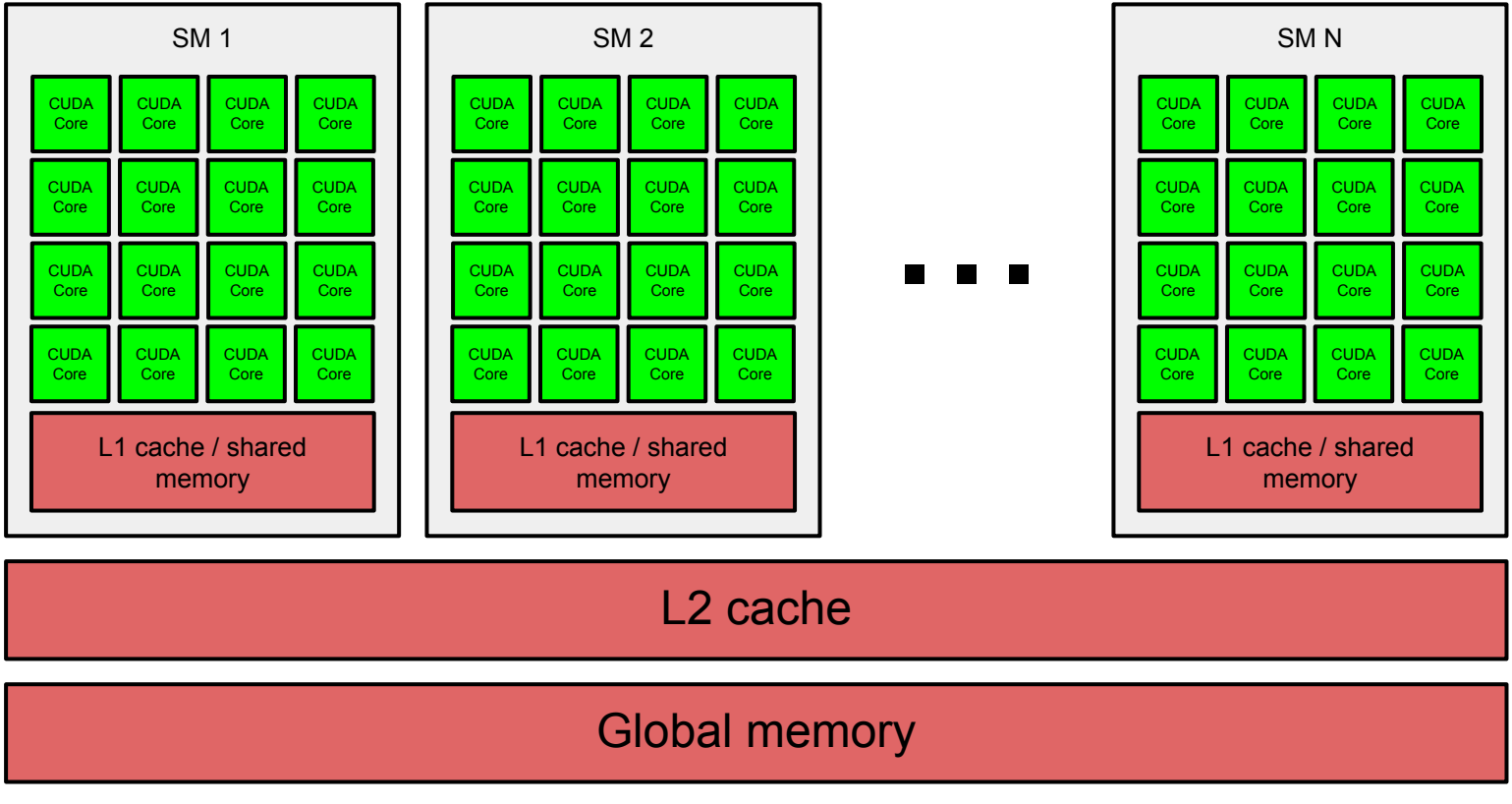
SM 1
CUDA Core
L1 cache / shared memory
SM 2
SM N
L2 cache
Global memory