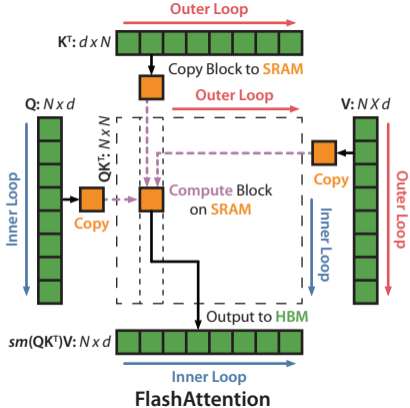


Memory Hierarchy with Bandwidth & Memory Size

SRAM: 19 TB/s (20 MB)

HBM: 1.5 TB/s (40 GB)

DRAM: 12.8 GB/s (>1 TB)



FlashAttention

