Memory Hierarchy - RAM, VRAM, Cache

Registers
Fastest • Smallest

↓

L1 Cache
~1ns • KB size

↓

L2 Cache
~4ns • MB size

↓

L3 Cache
~10ns • MB size

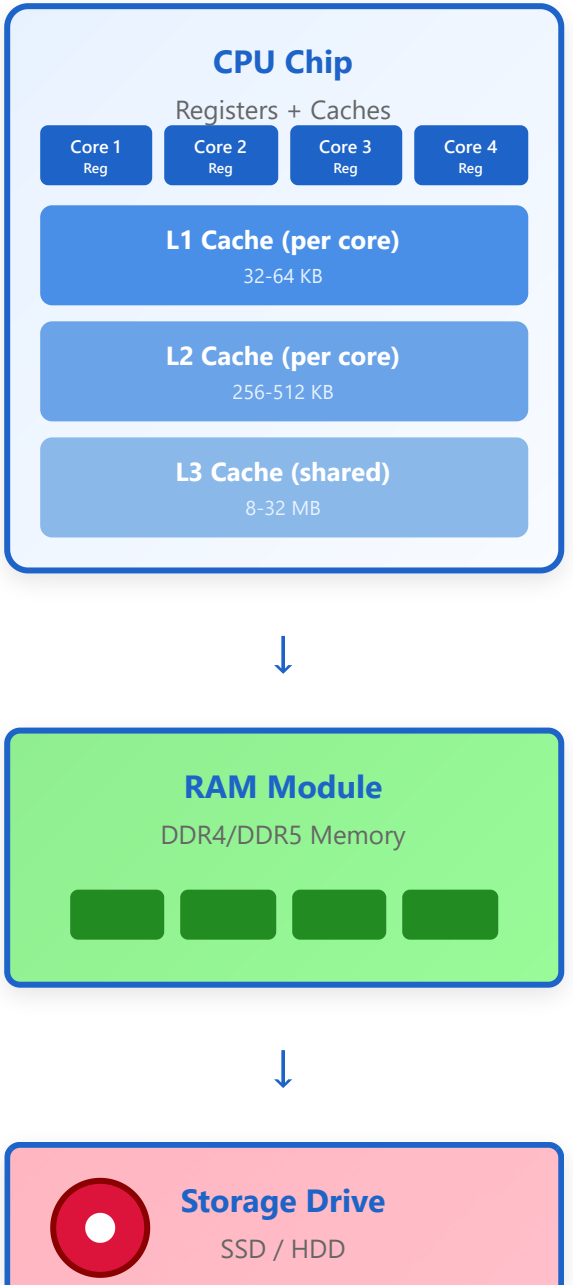
↓

RAM
~100ns • 16-512GB

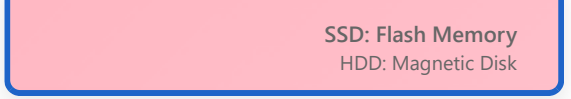
↓

Storage
~100µs • TB size

Physical Hardware Components



Trade-off: **Faster = Smaller + Higher Cost**



RAM

16-512GB

~20GB/s bandwidth

System memory

VRAM (GPU)

8-80GB

~1-2TB/s bandwidth

HBM2/HBM3

Cache

L1: **KB**

L2: **MB**

L3: **MB**

ML Bottleneck: RAM ↔ VRAM *Data transfer overhead*