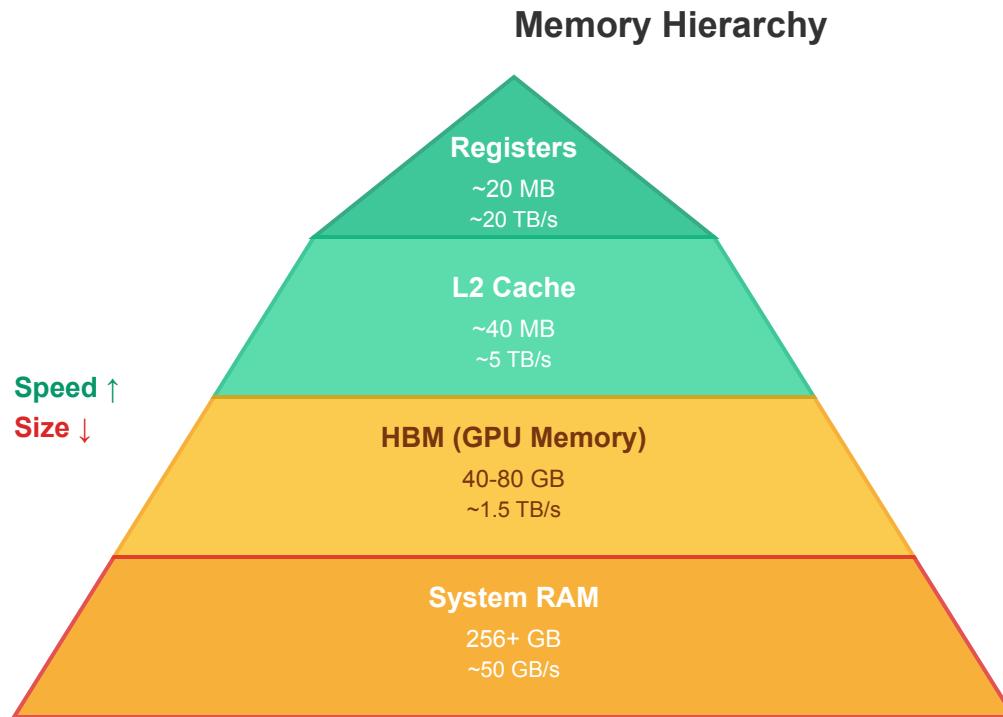


GPU Memory Hierarchy and Compute vs Memory Bound Operations



Latency:
1 cycle →
1000 cycles

Compute-Bound vs Memory-Bound Operations

Memory-Bound Region

- Limited by bandwidth
- Element-wise operations
- Small matrix multiplications
- Batch normalization

Compute-Bound Region

- Limited by FLOPs
- Large matrix multiplications
 - Convolutions
 - Attention mechanisms

Arithmetic Intensity (FLOPs/byte) →

~50 FLOPs/byte

$$\text{A100: } 312 \text{ TFLOPS} \div 1.5 \text{ TB/s} = 208 \text{ FLOPs/byte}$$

Key Insight for Leaders:

Small operations are memory-limited • Large operations are compute-limited • Optimization strategy depends on bottleneck • Batch size affects which regime you're in