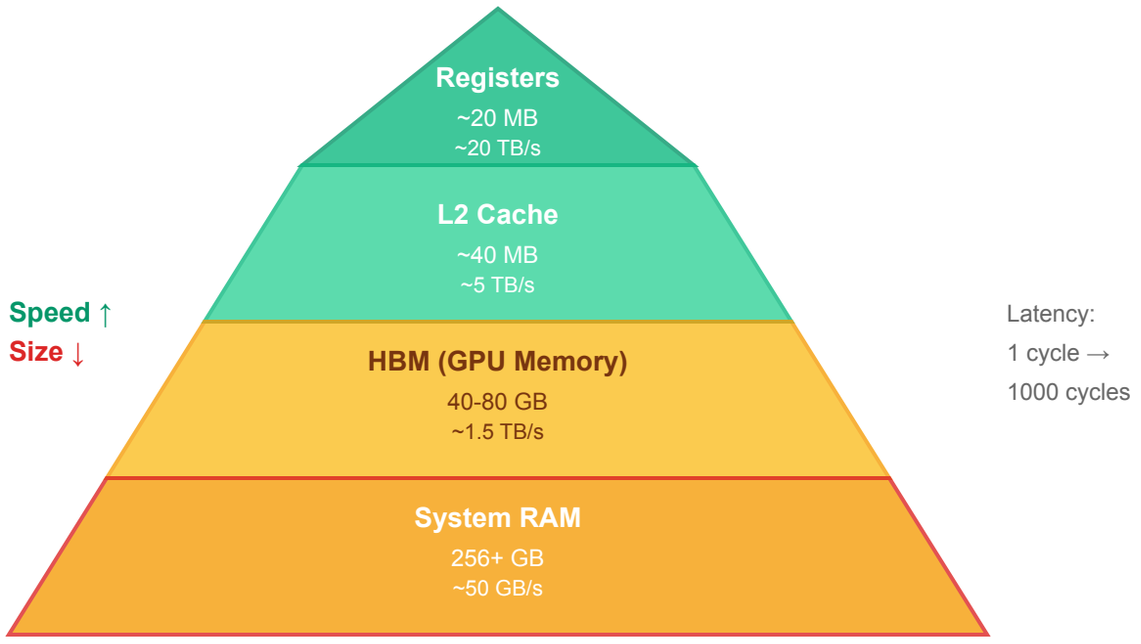
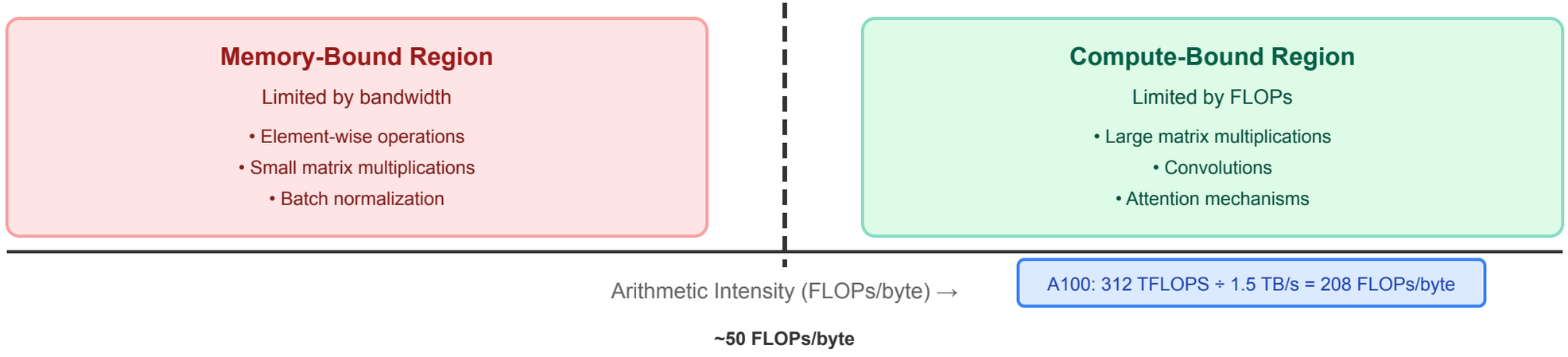
GPU Memory Hierarchy and Compute vs Memory Bound Operations

Memory Hierarchy



Compute-Bound vs Memory-Bound Operations



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