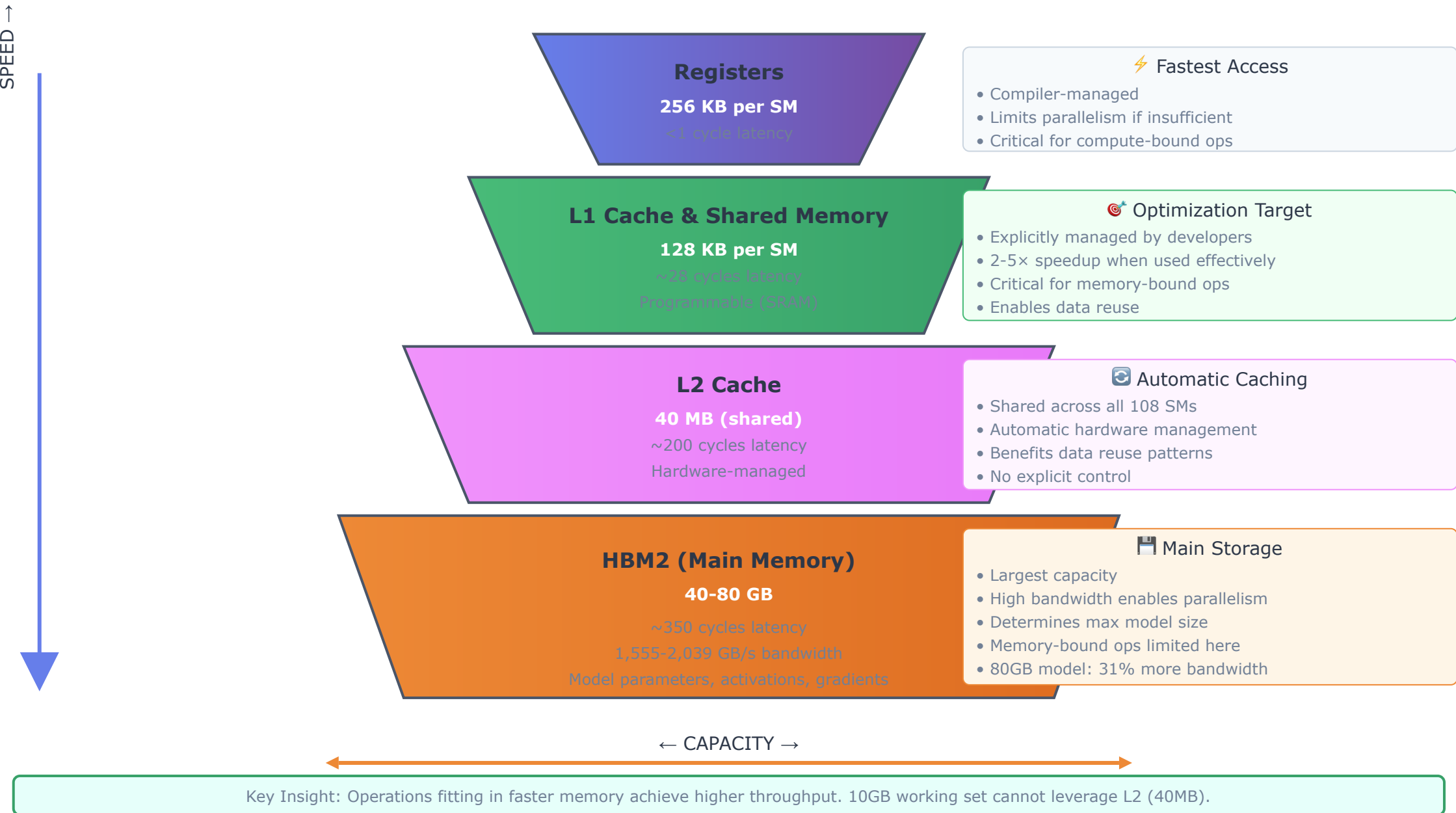
GPU Memory Hierarchy: A100 Architecture

Speed vs Capacity Trade-off



Key Insight: Operations fitting in faster memory achieve higher throughput. 10GB working set cannot leverage L2 (40MB).