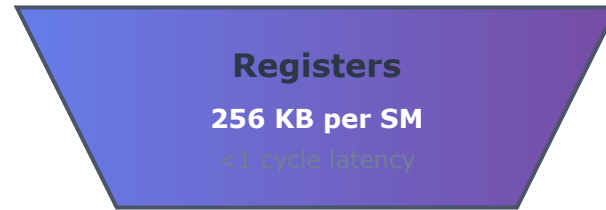


GPU Memory Hierarchy: A100 Architecture

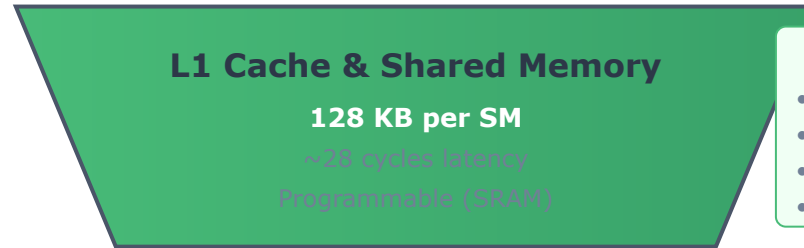
Speed vs Capacity Trade-off

SPEED ↑



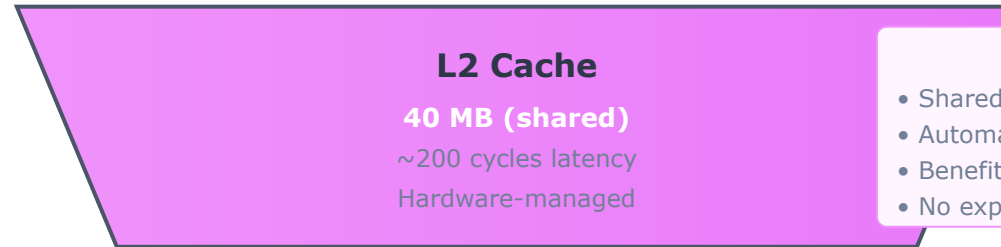
⚡ Fastest Access

- Compiler-managed
- Limits parallelism if insufficient
- Critical for compute-bound ops



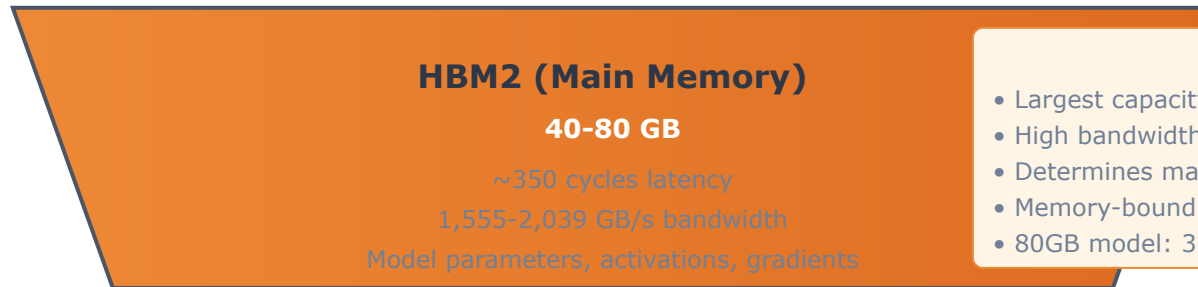
🎯 Optimization Target

- Explicitly managed by developers
- 2-5× speedup when used effectively
- Critical for memory-bound ops
- Enables data reuse



🔄 Automatic Caching

- Shared across all 108 SMs
- Automatic hardware management
- Benefits data reuse patterns
- No explicit control



💾 Main Storage

- Largest capacity
- High bandwidth enables parallelism
- Determines max model size
- Memory-bound ops limited here
- 80GB model: 31% more bandwidth

← CAPACITY →

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