Caching Optimization

Request →

L1: Query Cache
✓ Exact query match
TTL: 1h • Hit: 40%

Miss ↓

L2: Embedding Cache
Vector embeddings
TTL: 24h • Hit: 70%

Miss ↓

L3: Document Cache
Retrieved documents
TTL: 7d • Hit: 85%
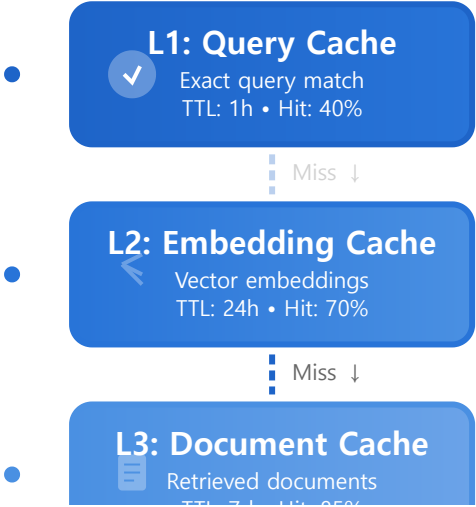
✓ Cache Hit Path

L1: ~20ms

L2: ~50ms

L3: ~80ms

Miss: ~200ms

Overall Latency: 200ms → 20ms (90% reduction)
Cost savings: 60% | DB load: -80%

⚙ Redis Configuration

Memory: 16GB with LRU eviction

Persistence: RDB + AOF for durability

Cluster: 3 nodes with replication

📊 Performance Impact

Latency reduction: **200ms → 20ms**

Cost savings: **60% reduction**

DB load: **-80%**