

Memory Management

Why Memory Management Matters

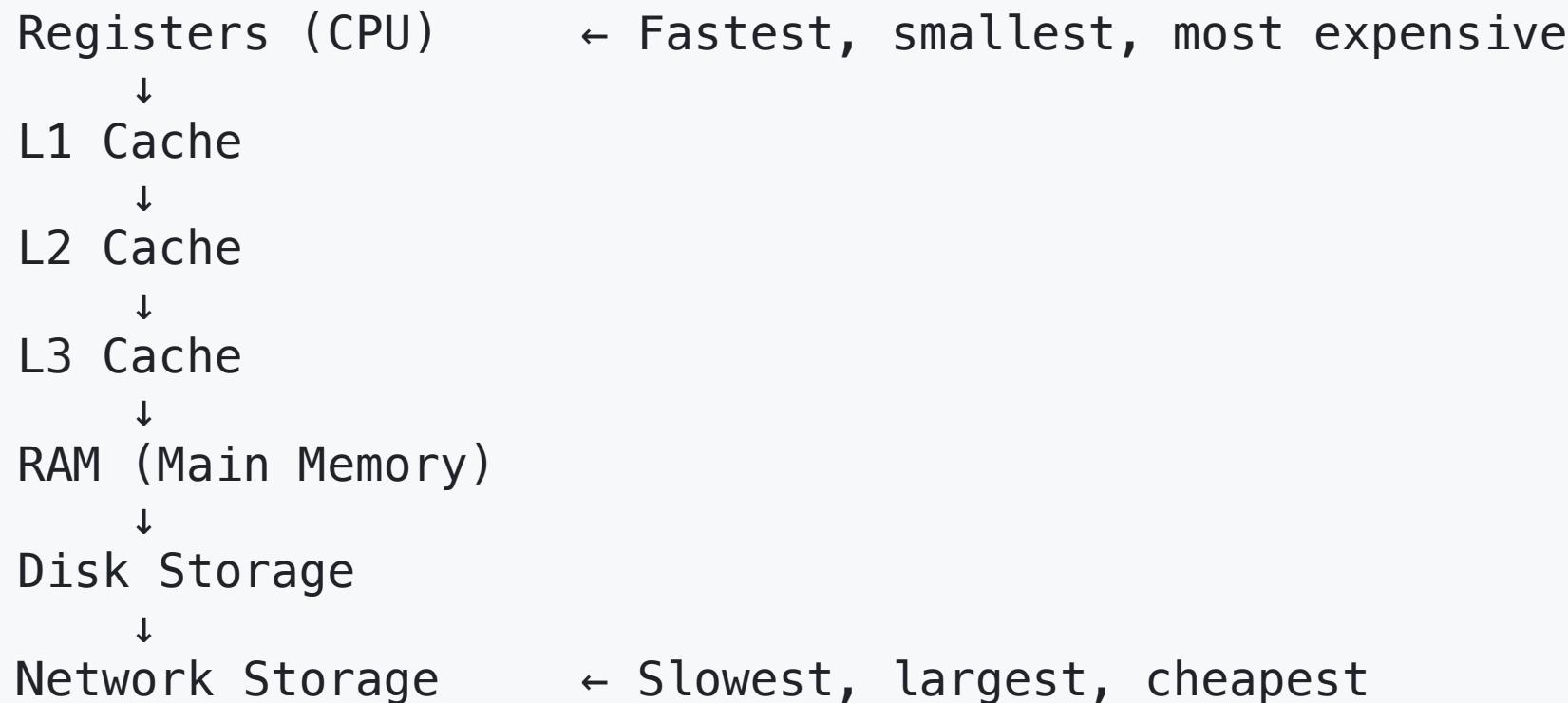
Real-world impact:

- 1. Performance:** Efficient memory use = faster programs
- 2. Stability:** Memory errors cause crashes and security vulnerabilities
- 3. Scalability:** Poor memory management limits how big your system can grow
- 4. Cost:** Memory is expensive at scale (servers, cloud computing)

Part 1: Memory Hierarchy

Memory Hierarchy

Modern computers organize memory by speed and cost:



Key Principle: Maximize use of faster memory, minimize slower memory access

Access Times Comparison

Level	Typical Size	Access Time	Relative Speed
Registers	Bytes	< 1 ns	1x
L1 Cache	KB	1-2 ns	~4x slower
L2 Cache	MB	3-10 ns	~10x slower
RAM	GB	50-100 ns	~100x slower
SSD	TB	50-150 µs	~100,000x slower
HDD	TB	1-10 ms	~10,000,000x slower

Takeaway: Memory location dramatically affects performance!

Active Learning: Quick Poll

Question: If we need to sum 1 million integers, which is faster?

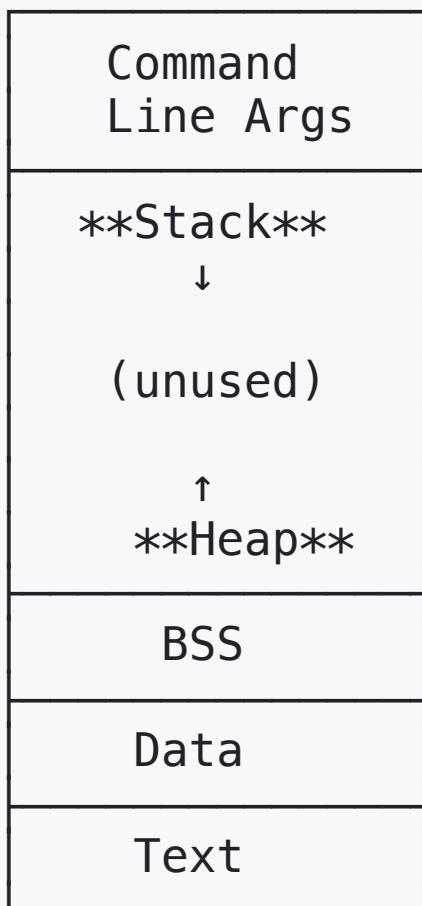
- A) Array stored contiguously in memory (sequential access)
- B) Linked list with nodes scattered in memory (random access)

Think about why one might be faster

Part 2: Program Memory Layout

Memory Layout of a Running Program

High Address



- ← Arguments, environment variables
- ← Grows downward (local variables)
- ← Grows upward (dynamic allocation)
- ← Uninitialized global/static data
- ← Initialized global/static data
- ← Program code (read-only)

Low Address

Stack Memory

Characteristics:

- Automatic allocation when entering function
- Automatic deallocation when exiting function
- Very fast (just pointer adjustment)
- Limited size (typically 1-8 MB)
- LIFO (Last In, First Out) structure

Example (C):

```
void function() {  
    int x = 10;           // Allocated on stack  
    char buffer[100];    // Allocated on stack  
    // Automatically freed when function returns  
}
```

Heap Memory

Characteristics:

- Explicit allocation (malloc/new)
- Explicit deallocation (free/delete)
- Slower than stack
- Large size (limited by available RAM)
- Manual management or garbage collection
- Can become fragmented

Example (C):

```
void function() {  
    int* ptr = (int*)malloc(sizeof(int) * 100); // Heap  
    // Use the memory...  
    free(ptr); // MUST explicitly free!  
}
```

Stack vs Heap Summary

Stack ✓

- **Fast** allocation/deallocation
- **Automatic** cleanup
- **No** fragmentation
- **Simple** management

Heap ✓

- **Large** size available
- **Flexible** lifetime
- Can return allocated memory

Heap ✗

- **Slower** allocation
- **Manual** management needed
- **Fragmentation** possible
- Memory leak risks

Stack ✗

- **Limited** size
- **Fixed** lifetime (function scope)
- Can't return pointers to local variables

Active Learning: Stack or Heap?

For each scenario, decide: **Stack or Heap?**

1. Small integer counter in a function
2. 10 MB image buffer
3. Function return value (int)
4. Data structure needed after function returns
5. Array of size known at compile time (100 elements)
6. Array of size determined at runtime

Part 3: Memory Management in C/C++

Manual Memory Management

In C/C++, you are responsible for memory:

Allocation:

```
int* ptr = (int*)malloc(sizeof(int) * 10);      // C style  
int* ptr = new int[10];                          // C++ style
```

Deallocation:

```
free(ptr);          // C style  
delete[] ptr;     // C++ style
```

Critical Rule: Every `malloc` needs a `free`, every `new` needs a `delete`

Common Memory Errors (1/2)

1. Memory Leak

```
void leak_example() {
    int* ptr = malloc(100 * sizeof(int));
    // Use ptr...
    // Forgot to free(ptr)!
} // Memory is lost forever
```

2. Use After Free (Dangling Pointer)

```
int* ptr = malloc(sizeof(int));
*ptr = 42;
free(ptr);
*ptr = 10; // ERROR: Undefined behavior!
```

Common Memory Errors (2/2)

3. Double Free

```
int* ptr = malloc(sizeof(int));
free(ptr);
free(ptr); // ERROR: Crash or corruption!
```

4. Buffer Overflow

```
char* buf = malloc(10);
strcpy(buf, "This string is way too long"); // ERROR: Writes past end
```

5. Stack Overflow

```
void infinite_recursion() {
    infinite_recursion(); // ERROR: Stack exhausted
}
```

Tools to Find Memory Errors

Valgrind (Linux/Mac):

```
valgrind --leak-check=full ./myprogram
```

AddressSanitizer (GCC/Clang):

```
gcc -fsanitize=address myprogram.c  
./myprogram
```

Benefits:

- Detects leaks, use-after-free, buffer overflows
- Provides detailed error reports with line numbers
- Essential for debugging memory issues

C++ Smart Pointers To Avoid Memory Errors

- `std::unique_ptr`
- Uses RAII pattern
 - When `unique_ptr` goes out of scope, memory is freed
- The code that has the `unique_ptr` is responsible for it
- Can't copy
- Can move (transfer ownership)

```
#include <memory>

{
    // Creation
    std::unique_ptr<int> ptr1 = std::make_unique<int>(42);

    // Usage - like a raw pointer
    *ptr1 = 100;
    std::cout << *ptr1 << std::endl; // 100

    // Cannot copy
    // std::unique_ptr<int> ptr2 = ptr1; // X Compile error!

    // Can move (transfer ownership)
    std::unique_ptr<int> ptr2 = std::move(ptr1);
    // Now ptr1 is nullptr, ptr2 owns the memory
}

// Now memory is deleted.
```

Part 4: Memory Management in Java

Automatic Memory Management

Java uses **garbage collection** - no explicit `free`:

```
void method() {  
    // Objects allocated on heap  
    String s = new String("Hello");  
    int[] arr = new int[1000];  
  
    // No need to explicitly free  
    // Garbage collector reclaims when unreachable  
}
```

Tradeoff: Convenience vs control (and some runtime overhead)

Java Memory Regions

Heap (Generational):

1. Young Generation

- Eden Space (new objects)
- Survivor Spaces S0, S1 (survived minor GC)

2. Old Generation (long-lived objects)

Non-Heap:

3. Metaspace (class metadata)

4. Stack (method frames, local primitives)

Key Idea: Most objects die young, so optimize GC for that

Garbage Collection Process

How it works:

- 1. Mark** - Starting from roots, mark all reachable objects
 - Roots: local variables, static variables, JNI references
- 2. Sweep** - Reclaim memory from unreachable objects
- 3. Compact (optional)** - Move objects together to reduce fragmentation

Types: Minor GC (young gen), Major/Full GC (old gen)

Java Memory Best Practices

Do:

- Reuse objects when possible (e.g., `StringBuilder`)
- Null out references when done (helps GC)
- Use primitives instead of wrappers when possible
- Close resources (`try-with-resources`)

Don't:

- Create unnecessary objects in loops
- Hold onto large objects longer than needed
- Ignore `OutOfMemoryError`
- Assume GC runs at specific times



Active Learning: Optimize This Code

```
public String concatenate(List<String> strings) {  
    String result = "";  
    for (String s : strings) {  
        result = result + s; // What's the problem?  
    }  
    return result;  
}
```

Question: If `strings` has 10,000 elements, how many `String` objects are created?

Discuss: How would you fix this?

Part 5: Memory Optimization

Optimization Strategy 1: Algorithm Choice

Different algorithms have different memory needs:

Algorithm	Time	Space	Notes
Bubble Sort	$O(n^2)$	$O(1)$	In-place
Merge Sort	$O(n \log n)$	$O(n)$	Needs temp array
Quick Sort	$O(n \log n)$	$O(\log n)$	Recursion stack
Hash Table	$O(1)$ avg	$O(n)$	Extra buckets

Tradeoff: Often must choose between time and space efficiency

Optimization Strategy 2: Memory Pooling

Problem: Many small allocations are slow

Solution: Pre-allocate a pool

```
// Inefficient - 1000 allocations
for (int i = 0; i < 1000; i++) {
    int* temp = malloc(sizeof(int));
    process(temp);
    free(temp);
}

// Efficient - 1 allocation
int* pool = malloc(1000 * sizeof(int));
for (int i = 0; i < 1000; i++) {
    process(&pool[i]);
}
free(pool);
```

Optimization Strategy 3: Cache-Friendly Code

Memory is loaded in cache lines (smallest unit of data xfer) (~64 bytes)

Cache-Unfriendly ✗

```
// Column-major access
for (int col = 0; col < N; col++) {
    for (int row = 0; row < N; row++) {
        sum += matrix[row][col];
    }
}
```

Jumps around memory

Impact: Can be 10x faster for large matrices!

Cache-Friendly ✓

```
// Row-major access
for (int row = 0; row < N; row++) {
    for (int col = 0; col < N; col++) {
        sum += matrix[row][col];
    }
}
```

Sequential access

Optimization Strategy 4: Data Structure Alignment

Problem: CPUs prefer aligned memory access

```
struct Bad {
    char c;          // 1 byte
    int i;           // 4 bytes - needs alignment
    char c2;         // 1 byte
    // Total: 12 bytes (with padding)
};

struct Good {
    int i;           // 4 bytes
    char c;          // 1 byte
    char c2;         // 1 byte
    // Total: 8 bytes (better packed)
};
```

Tip: Order struct fields largest to smallest

Measuring Memory Usage

C/C++:

```
#include <sys/resource.h>
struct rusage usage;
getrusage(RUSAGE_SELF, &usage);
printf("Memory: %ld KB\n", usage.ru_maxrss);
```

Java:

```
Runtime runtime = Runtime.getRuntime();
long memory = runtime.totalMemory() - runtime.freeMemory();
System.out.println("Memory: " + memory + " bytes");
```

Rule: Always measure before optimizing!

Summary and Best Practices

Key Takeaways

Core Concepts:

1. Memory hierarchy: registers → cache → RAM → disk
2. Program memory: heap vs stack
 - i. Stack: Fast, automatic, limited
 - ii. Heap: Flexible, manual, large
3. C/C++: Manual management
4. Java: Garbage collection

Key Takeaways

Optimization:

1. Choose algorithms considering space-time tradeoffs
2. Use memory pools, cache-friendly access patterns
3. Measure before optimizing

Common Pitfalls to Avoid

- ✗ Memory leaks (forgetting to free)
- ✗ Use after free (dangling pointers)
- ✗ Double free (freeing twice)
- ✗ Buffer overflows (writing past bounds)
- ✗ Stack overflows (deep recursion)
- ✗ Creating unnecessary objects
- ✗ Premature optimization
- ✗ Ignoring memory profiler warnings

