Computer Architecture (CS F342)

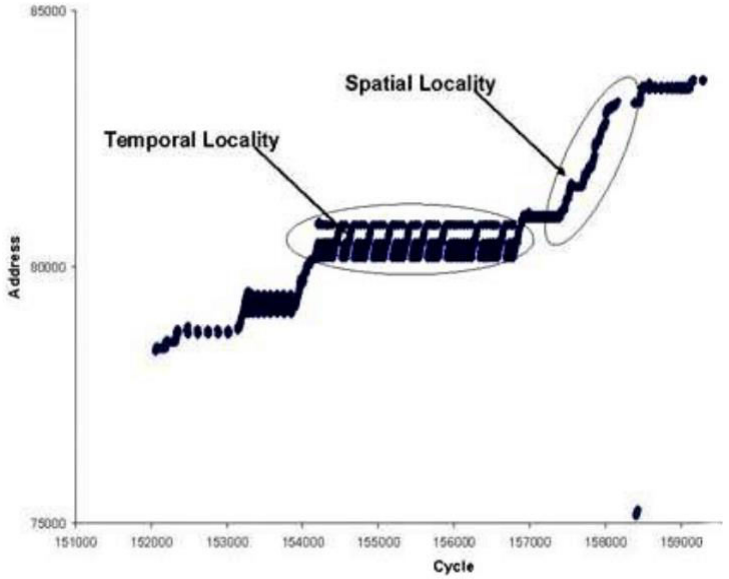
Memory/Storage Hierarchy

&

Fast Storage Unit: Cache Memory Architecture & Organization

Why do we need to study memory/storage hierarchy?

- CPU is a component in the computer systems
- Others components: Memory and I/O systems
- Programs exhibit principle of locality
  - Temporal
  - Spatial
  - A rule of thumb (the 90/10 rule): 90% of the execution of programs spends in only 10% of the code



Behavior of a program

Matrix multiplication

- Data stored in row-major order
- Data of A, B & C can be used in near future
- Data neighboring to previously accessed data
- Instructions are also to be used in near future
- Principle of locality

```
for (i=0; i<l; i++)
    for (j=0; j<m; j++)
        for (k=0; k<n; k++)
            A[i][j] += B[i][k] x C[k][j];
```

Why do we need to study memory/storage hierarchy?

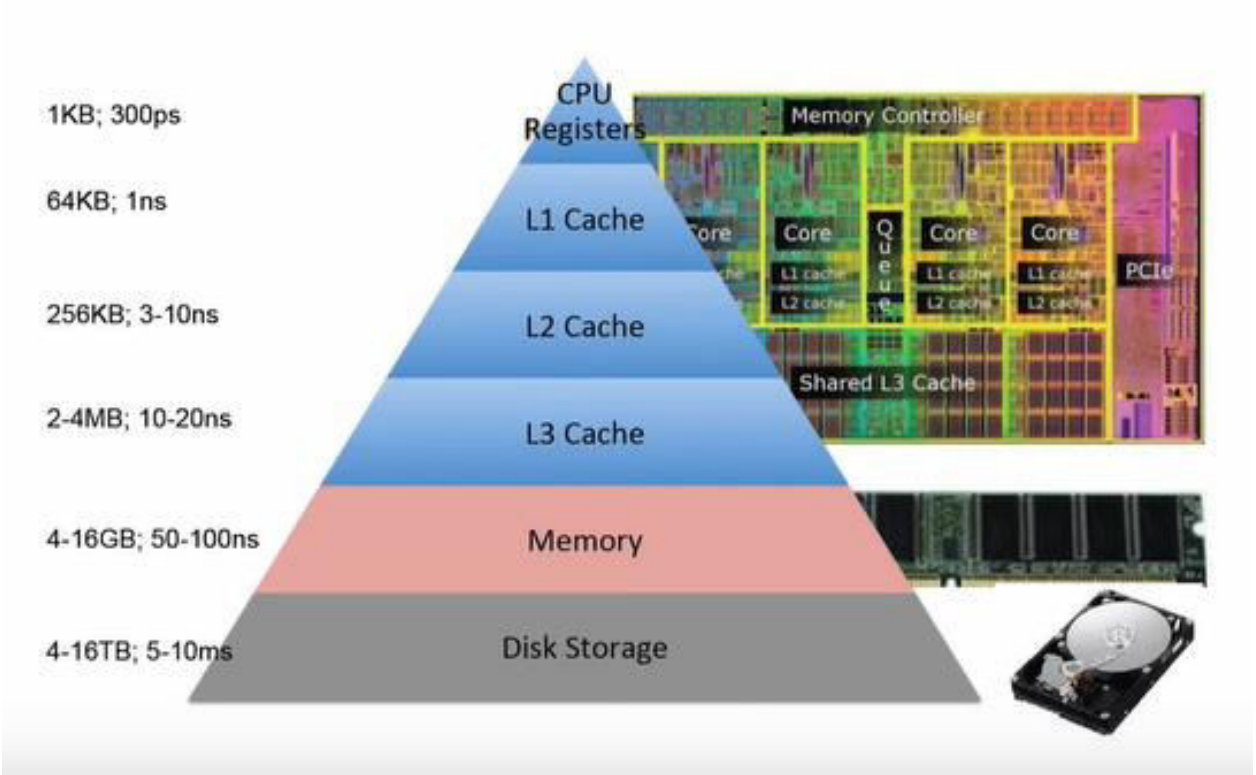
- Principle of locality can be found in most of the programs

- To exploit such principle we need memory hierarchy
  - Keep the repeatedly accessed data & instruction near to CPU
  - Not the entire code

- Memory hierarchy
  - Faster but smaller memory closer to CPU
  - Slower but larger memory faraway from the CPU

Memory hierarchy

Access **time** and **space** increase



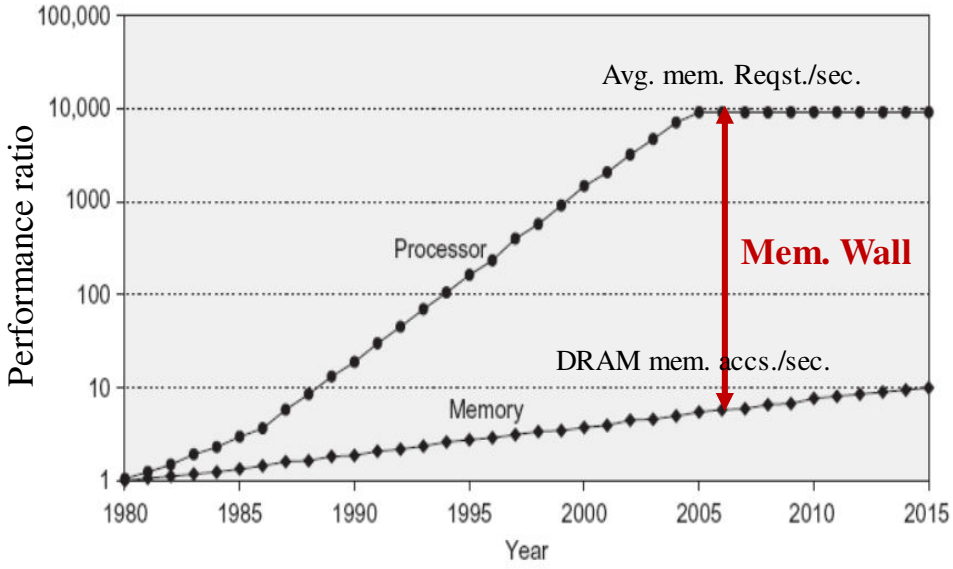
Cost increase

What makes improvement in the storage access time?

- Static Random Access Memory (SRAM) Technology
  - Registers, L1, L2 & L3 cache
- Dynamic Random Access Memory (DRAM) Technology
  - Main memory
- Magnetic Technology
  - Hard disk takes longer access time because of mechanical components

Do we really need memory hierarchy?

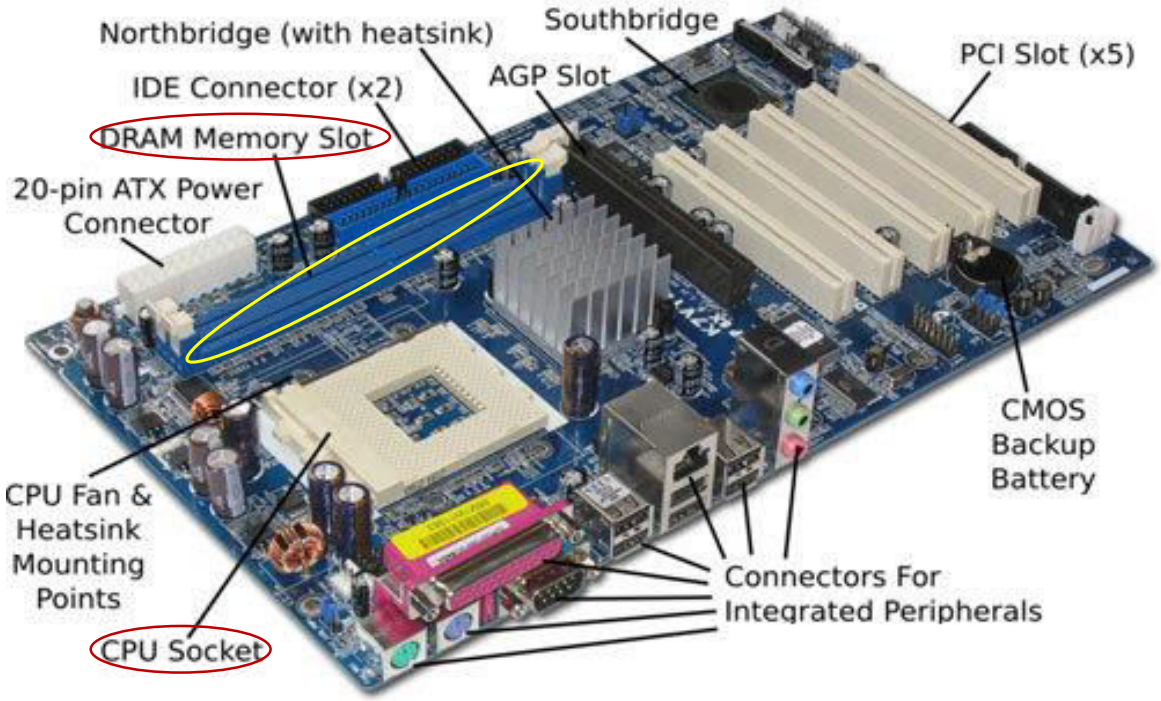
- Memory Wall Problem:
  - Significant increase in processor performance over the years
  - Not significant increase in main-memory performance over the years



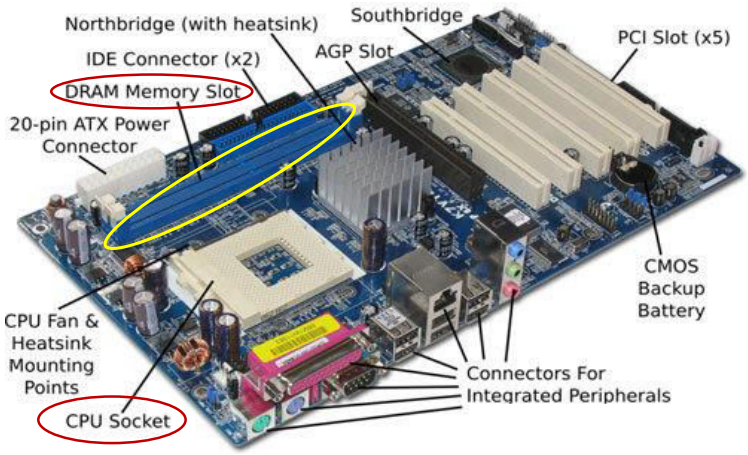
Necessity of memory hierarchy

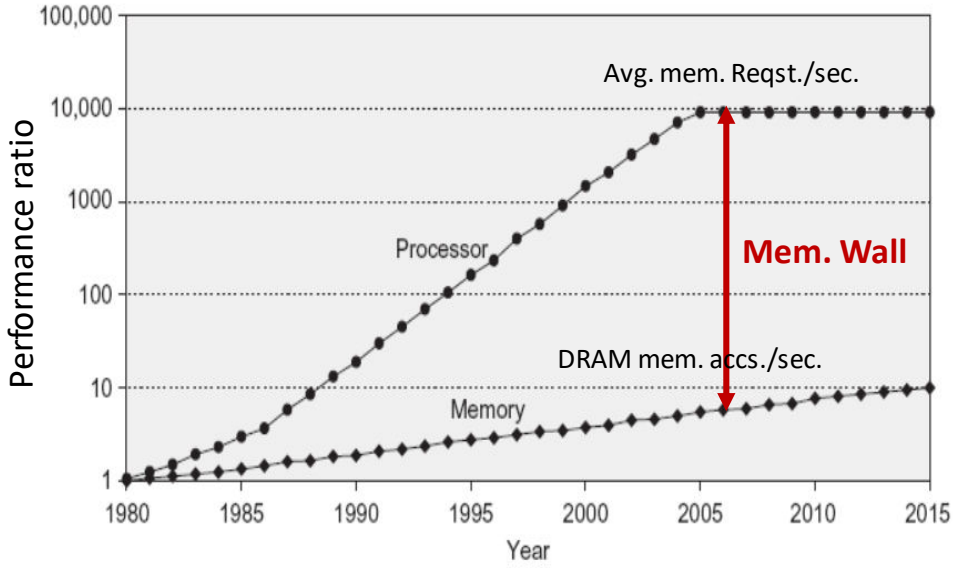
- From the previous graph
  - The gap is increasing
- The previous graph did not include the multiprocessors
  - The aggregated peak bandwidth requirement increases with no. cores/processors
- How does one deal with this increasing gap?
  - Need memory hierarchy: multi-levels of cache hierarchy

Why do we need cache memory?



Memory Wall Problem and Necessitv of Cache Memorv

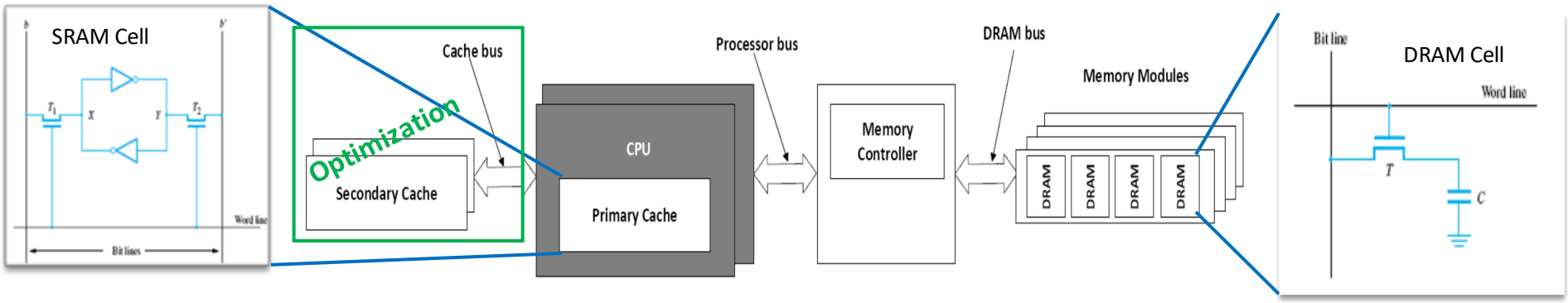




- Performance difference between processor (CPU) and memory by technology and memory is placed far away (*nm* scale) from CPU (off-chip)
- There is a gap in CPU's request (rate) for the memory accesses and the service (rate) for those request by memory
-
- Cache memory technique can speed up the performance of the memory accesses time

Cache Memory Architecture

- Cache memory is a high-speed storage unit
- **What makes it faster as compare to the memory in question?**
  - 1) Position 2) SRAM-based memory technology
- **What would be the size and characteristic of the Cache memory?**
  - Smaller in size and bit access time must be faster



Program's Behavior and Characteristics of Cache Memory

Which memory request should be keep in the cache?

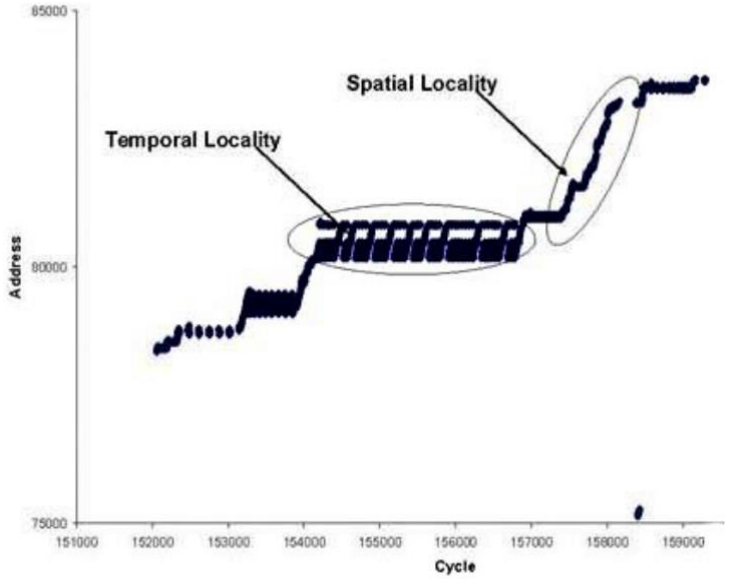
- Program's behavior: programs tend to reuse data and instructions they have used recently.
- Spatial locality and Temporal locality

How many such instructions and/or data one can keep in the cache?

- More than one or a block of instructions and/or data

How does one decide the block size?

- Processor's cache is managed by its own set of heuristics or rules
- For example, in Intel i7, block size of the primary cache is 64 bytes



Behavior of a program

Characteristics of Cache Memory

The following items are embedded in the cache:

Organization

The logical arrangement of storage unit/data

Content-management heuristics

Decide the best possible items for caching and evict out the candidate to make room for more important data not yet cached

Consistency-management heuristics

Ensure that the instructions and data that the program expects to receive are the ones the program does, indeed, receive

Consistency with 1) self, 2) main memory 3) other caches

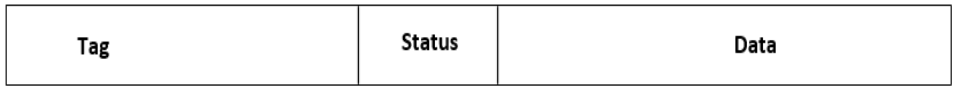
Cache Organization: Blocks, Tags and Set

A cache stores chunks of data (called cache blocks or cache lines) that come from the memory.

A cache is typically much smaller than the memory:

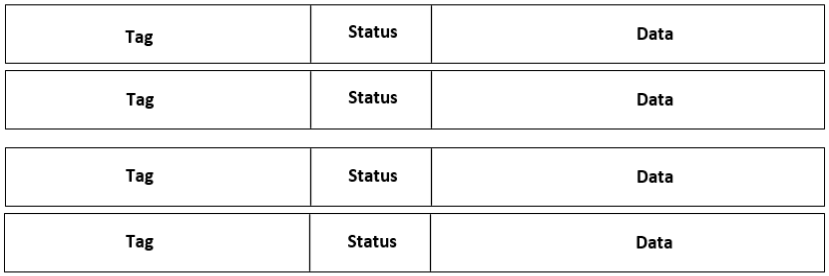
How does CPU know whether any particular datum is present in the cache or not?

Cache tags fulfil this necessity.



What if the cache contains more than one block?

Cache can have the set of choices for the incoming blocks



Optimization Techniques to Improve the Memory Access Time (or Miss Rate)

How does following items affect the cache miss rate?:

- Cache size (higher or smaller)
- Cache block size (higher or smaller)
- Set or associativity size (higher or smaller)
- Cache hierarchies
  - Issues in hierarchy management

Summary

- Necessity of Memory Hierarchy
- Necessity of Cache memory
- Characteristic of Cache memory
- Program's behavior
- Elements of Cache Organization
- Elements of Cache Optimization