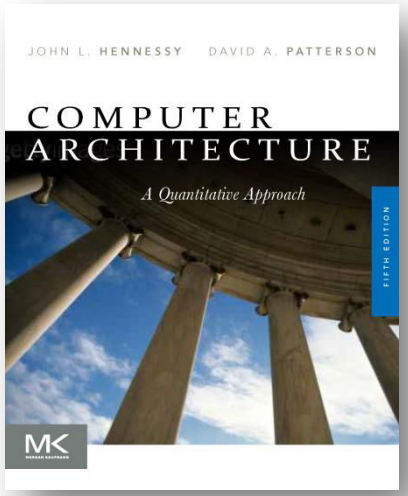


Chapter 2

Memory Hierarchy Design

Section 2.1 Only

Introduction

- Programmers want unlimited amounts of memory with low latency
- Fast memory technology is more expensive per bit than slower memory
- Solution: organize memory system into a hierarchy
  - Entire addressable memory space available in largest, slowest memory
  - Incrementally smaller and faster memories, each containing a subset of the memory below it, proceed in steps up toward the processor
- Temporal and spatial locality insures that nearly all references can be found in smaller memories
  - Gives the allusion of a large, fast memory being presented to the processor

PRINCIPLE OF LOCALITY

- Temporal locality (locality in time): if an item is referenced, it will tend to be referenced again soon.
- Spatial locality (locality in space): if an item is referenced, items whose addresses are close by will tend to be referenced soon.

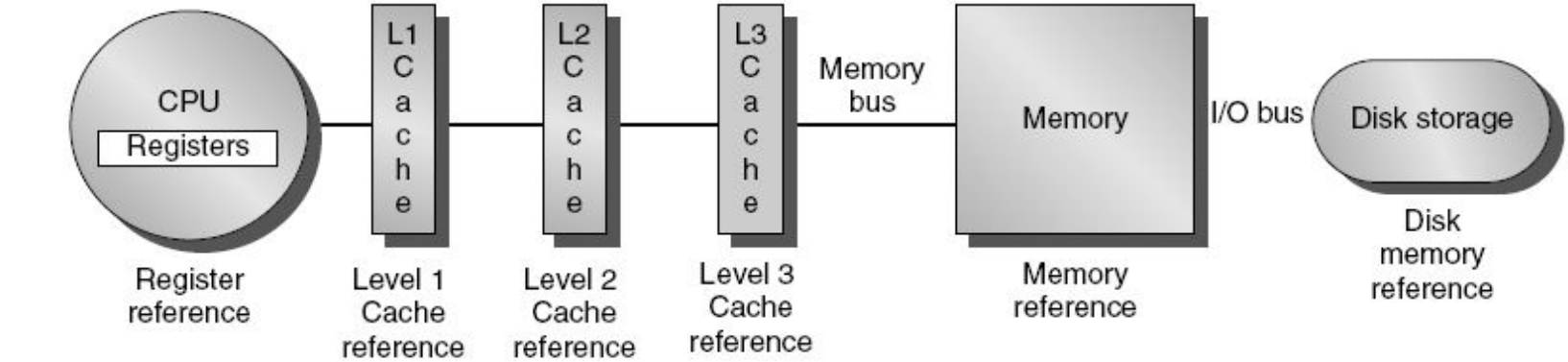
```
// Multiply the two matrices together
  for ( ty = 0 ; ty < BLOCK_SIZE ; ty++ ){
    for ( tx = 0 ; tx < BLOCK_SIZE ; tx++ ){
      Csub = 0.0 ;
      for (k = 0; k < BLOCK_SIZE; ++k ){
         Asub = As[ty][k ] ;
         Bsub = Bs[k ][tx] ;
         Csub += Asub * Bsub ;
      }
      c = wB * BLOCK_SIZE * by + BLOCK_SIZE * bx;
      C[c + wB * ty + tx] += Csub;
    }// for tx ;
  }// for ty
```

for-loop is temporal locality

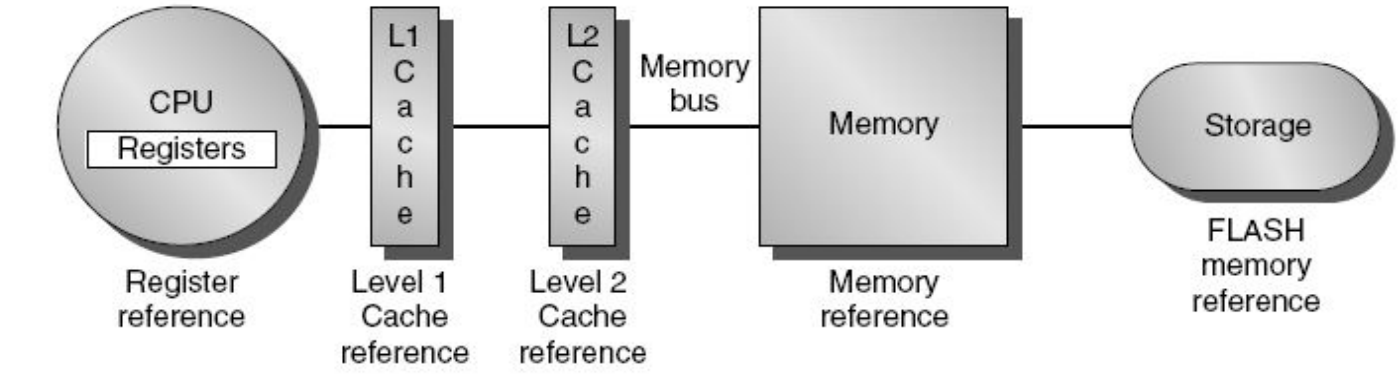
array is spatial locality

Observation: temporal locality means that we don't put all program into memory whereas spatial locality means that we don't put all data into memory, hence we have "Memory Hierarchy"

Memory Hierarchy

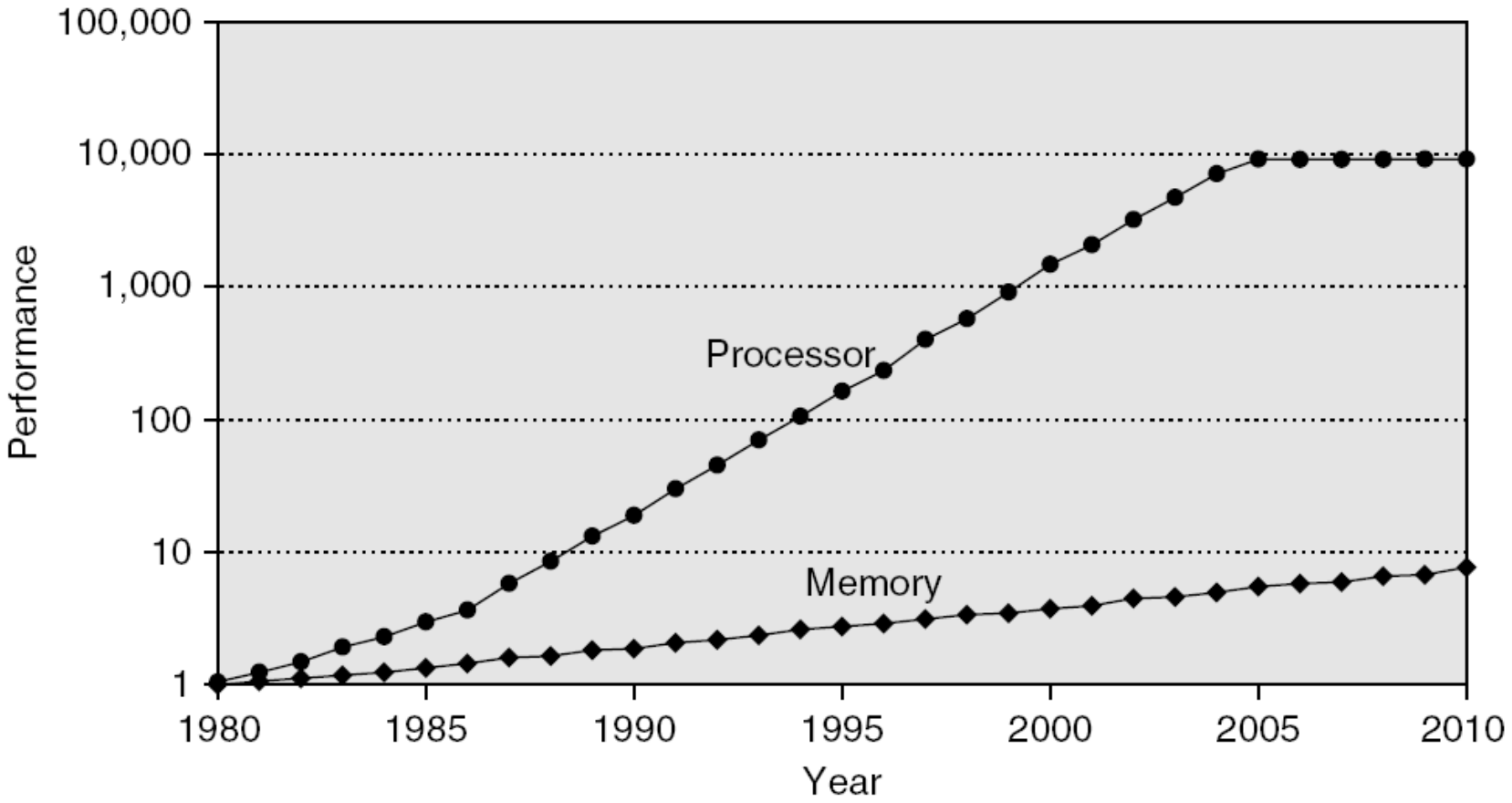


(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Memory Performance Gap



Memory Hierarchy Design

- Memory hierarchy design becomes more crucial with recent multi-core processors:
  - Aggregate peak bandwidth grows with # cores:
    - Intel Core i7 can generate two references per core per clock
    - Four cores and 3.2 GHz clock
      - 25.6 billion 64-bit data references/second +
      - 12.8 billion 128-bit instruction references
      - = 409.6 GB/s!
    - DRAM bandwidth is only 6% of this (25 GB/s)
    - Requires:
      - Multi-port, pipelined caches
      - Two levels of cache per core
      - Shared third-level cache on chip

Performance and Power

- High-end microprocessors have >10 MB on-chip cache
  - Consumes large amount of area and power budget

Q & A