Computer Organization & Architecture

Chapter 8 – Memory Hierarchy

Zhang Yang 张杨

cszyang@scut.edu.cn

Autumn 2025

Contents of this lecture

- 8.5 Memory Hierarchy
  - Why have a memory hierarchy?
  - Figure of memory hierarchy
  - Fundamental idea of memory hierarchy
  - How does memory hierarchy work?
  - How is the hierarchy managed?
  - Four questions for memory hierarchy designer
  - Summary

Why Have a Memory Hierarchy? (1)

- Ideal Memory
  - Zero access time (latency)
  - Infinite capacity
  - Zero cost
  - Infinite bandwidth (to support multiple accesses in parallel)

Why Have a Memory Hierarchy? (2)

- The problem is
  - Ideal memory's requirements oppose each other.
    - Bigger is slower.
      - Bigger → Takes longer to determine the location.
    - Faster is more expensive.
      - Memory technology: SRAM vs. DRAM
    - Higher bandwidth is more expensive.
      - Need more banks, more ports, higher frequency, or faster technology.

Why Have a Memory Hierarchy? (3)

- Memory Technology

Memory Technology	Typical Access Time	$ per GB in 2012
SRAM	0.5 – 2.5 ns	$500 - $1000
DRAM	50 – 70 ns	$10 - $20
Flash Memory	5,000-50,000ns	$0.75-$1.00
Magnetic Disk	5,000,000 – 20,000,000 ns (5 – 20 ms)	$0.05 - $0.10

Why Have a Memory Hierarchy? (4)

- Storage Trends
  - SRAM

metric	1980	1985	1990	1995	2000	2005	2010	*2010:1980*
$/MB	19,200	2,900	320	256	100	75	*60*	*320*
access (ns)	300	150	35	15	3	2	*1.5*	*200*

  - DRAM

metric	1980	1985	1990	1995	2000	2005	2010	*2010:1980*
$/MB	8,000	880	100	30	1	0.1	0.06	*130,000*
access (ns)	375	200	100	70	60	50	40	*9*
typical size(MB)	0.064	0.256	4	16	64	2,000	8,000	*125,000*

Why Have a Memory Hierarchy? (5)

- Storage Trends (ctd.)
  - Disk

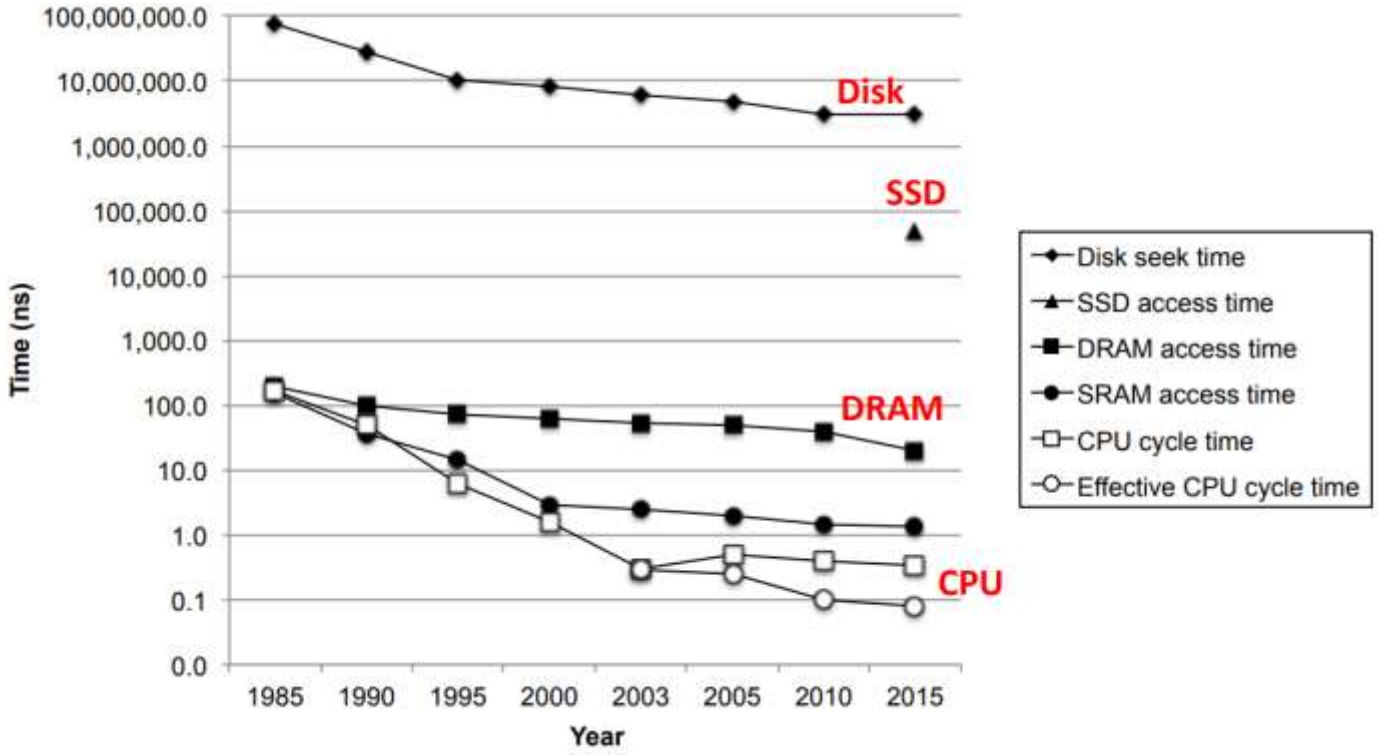
metric	1980	1985	1990	1995	2000	2005	2010	*2010:1980*
$/MB	500	100	8	0.30	0.01	0.005	0.0003	*1,600,000*
access (ms)	87	75	28	10	8	*4*	*3*	*29*
typical size(MB)	1	10	160	1,000	20,000	160,000	1,500,000	*1,500,000*

  - CPU Clock Rates

	1980	1985	1990	1995	2000	2005	*2005:1980*
processor	8080	286	386	Pentium	P-III	P-4	
clock rate(MHz)	1	6	20	150	750	3,000	3,000
cycle time(ns)	1,000	166	50	6	1.3	0.3	3,333

Why Have a Memory Hierarchy? (6)

- The CPU-Memory Gap
  - The gap widens between DRAM, disk, and CPU speeds.

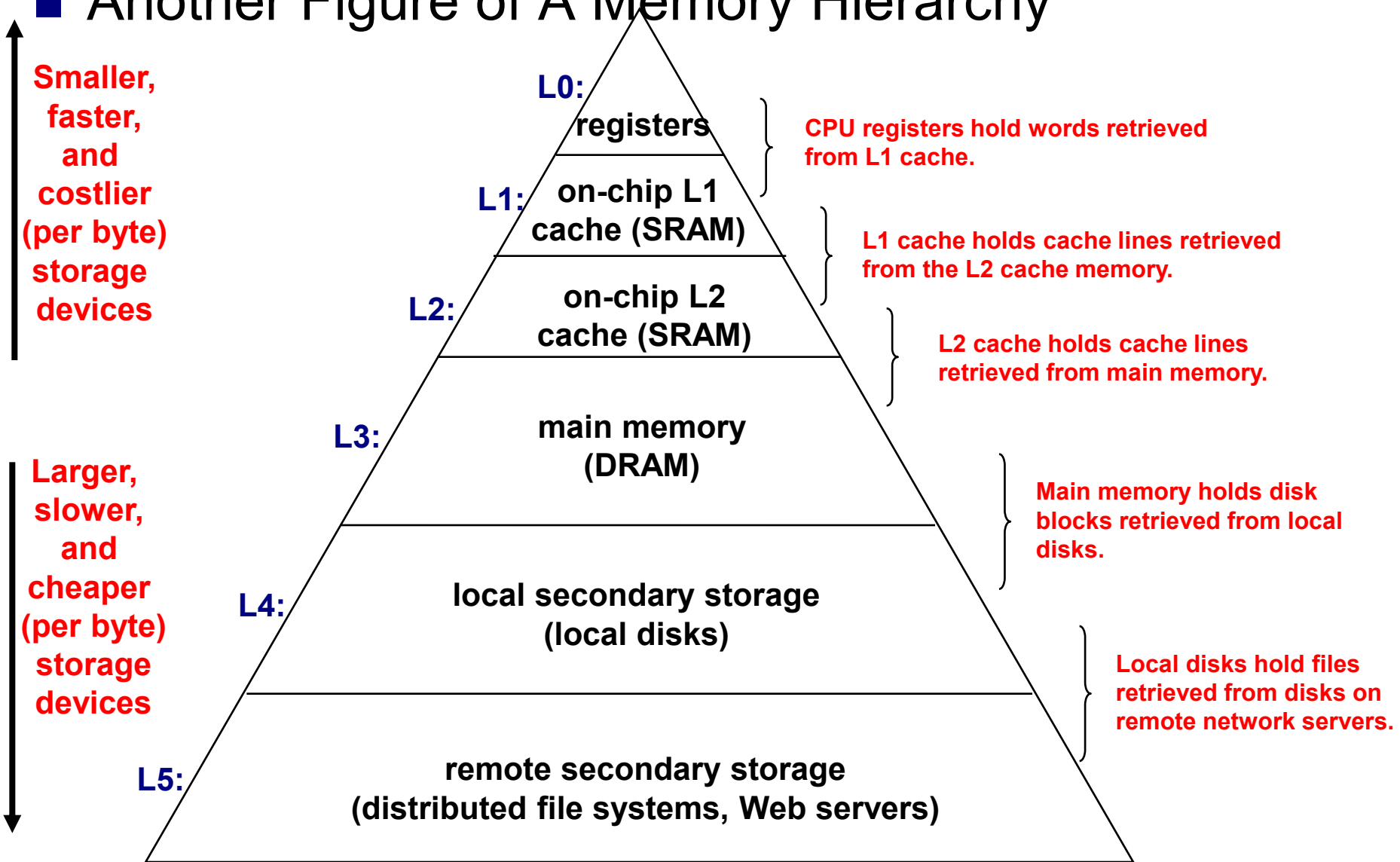


Why Have a Memory Hierarchy? (7)

- Summary
  - We want both fast and large memory.
  - But we cannot achieve both with a single level of memory.
  - Idea: Have multiple levels of storage (progressively bigger and slower as the levels are farther from the processor) and ensure most of the data the processor needs is kept in the fast(er) level(s).

Figure of Memory Hierarchy

- Another Figure of A Memory Hierarchy

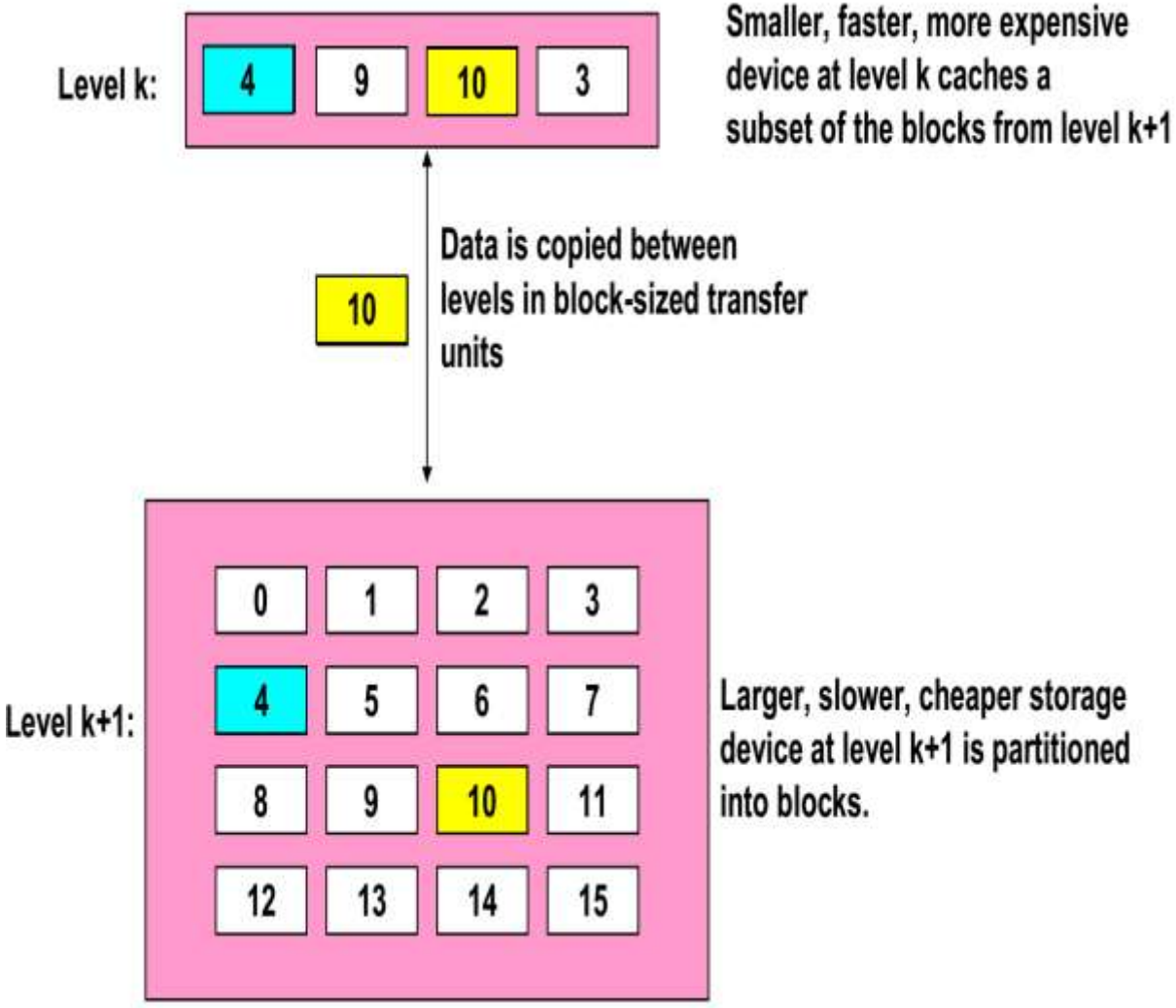


Fundamental Idea of The Memory Hierarchy (1)

- Cache
  - A smaller, faster storage device that acts as a staging area for a subset of the data in a larger, slower device.
  - For each k, the faster, smaller device at level k serves as a cache for the larger, slower device at level k+1.
  - Data is copied between two adjacent levels.
  - All data is stored at the lowest level.

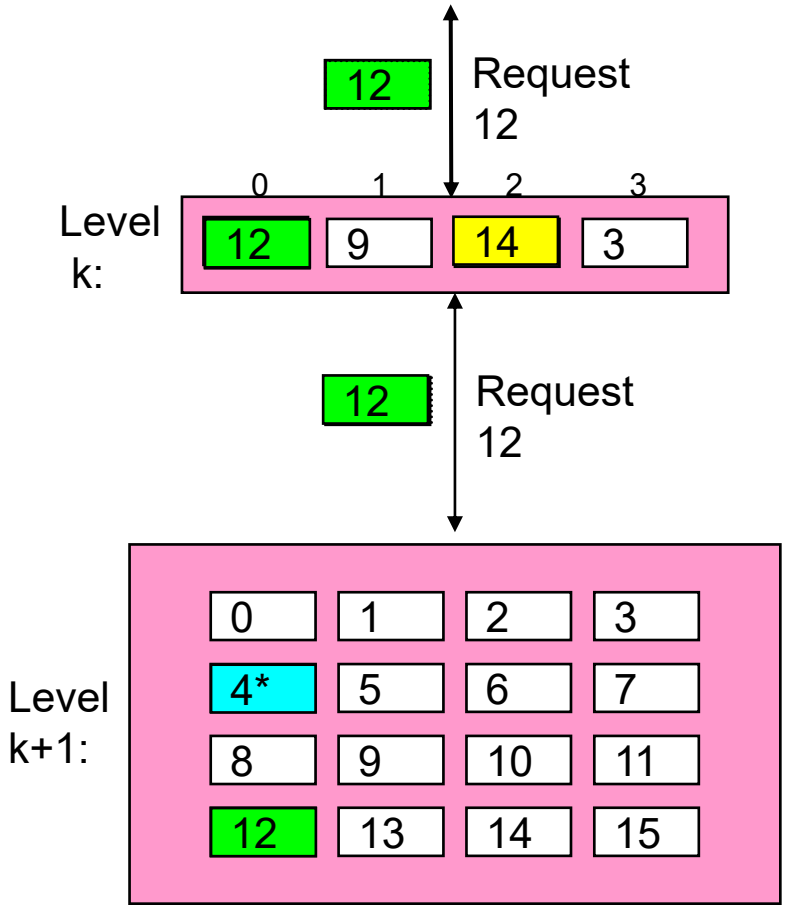
Fundamental Idea of The Memory Hierarchy (2)

- Cache
  - Example



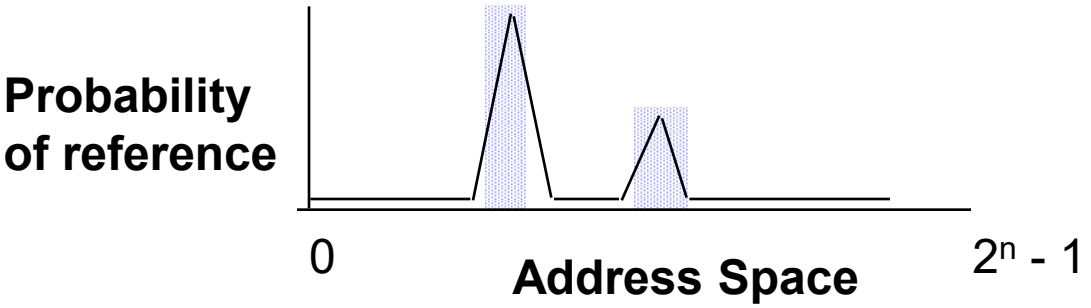
Fundamental Idea of The Memory Hierarchy (3)

- General Caching Concepts
  - Program needs object d, which is stored in some block b.
  - Hit
    - Program finds b in the cache at level k.
    - E.g., block 14.
  - Miss
    - b is not at level k, so level k cache must fetch it from level k+1.
    - E.g., block 12.



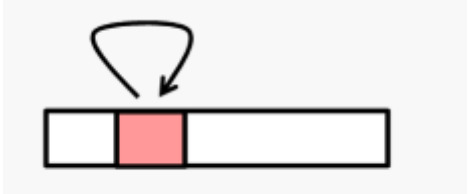
How does The Memory Hierarchy Work? (1)

- Fact: Large memories are slow, fast memories are small, How do we create a memory that gives the illusion of being large, cheap and fast (most of the time)?
- By taking advantage of
  - The Principle of Locality(局部性原理):Programs access a relatively small portion of the address space at any instant of time.

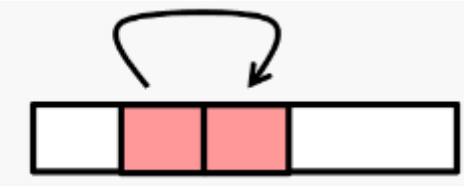


How does The Memory Hierarchy Work? (2)

- The Principle of Locality
  - Temporal Locality – locality in time (时间局部性)
    - If item was referenced recently, it will likely be referenced again soon.



  - Spatial locality – locality in space (空间局部性)
    - If an item is referenced, items near it might be referenced soon.
    - 80/20 rule – 80% of the time is spent executing 20% of the code



How does The Memory Hierarchy Work? (3)

- The Principle of Locality (ctd.)
  - Locality Example
    - Instructions
      - Reference instructions in sequence: Spatial Locality
      - Cycle through loop repeatedly: Temporal Locality
    - Data
      - Reference array elements in succession : Spatial Locality
      - Reference `sum` each iteration: Temporal Locality

```
sum = 0;
for (i = 0; i < n; i++)
sum += a[i];
return sum;
```

How does The Memory Hierarchy Work? (4)

- In the memory hierarchy
  - Temporal Locality (Locality in Time):
    - => Keep most recently accessed data items closer to the processor
  - Spatial Locality (Locality in Space):
    - => Move blocks consists of contiguous words to the upper levels

How does The Memory Hierarchy Work? (5)

- Summary
  - Programs tend to access the data at level k more often than they access the data at level k+1.
  - Thus, the storage at level k+1 can be slower, and thus larger and cheaper per bit.
  - Net effect: A large pool of memory that costs as much as the cheap storage near the bottom, but that serves data to programs at the rate of the fast storage near the top.

How is the Hierarchy Managed?

- Registers <-> Memory
  - By compiler.
- Cache <-> Main memory
  - By the hardware.
- Main Memory <-> Disks
  - By the hardware and operating system (virtual memory).
  - By the programmer (files).

Four Questions for Memory Hierarchy Designers

- Q1: Where can a block be placed in the upper level? *(Block placement)*
- Q2: How is a block found if it is in the upper level?
  (Block identification)
- Q3: Which block should be replaced on a miss?
  (Block replacement)
- Q4: What happens on a write?
  (Write strategy)

Summary

- 知识点: Memory Hierarchy
  - Why does a computer have memory hierarchy?
  - What is memory hierarchy? Figure
  - Principle: locality of reference
  - Fundamental idea of memory hierarchy
- 掌握程度
  - 理解使用存储层次结构的原因。
  - 会画出存储层次结构图。
  - 理解局部性原理（时间，空间）。
  - 了解存储层次结构的基本思想

Exercise

- 简答: Why are computer's memory sytems typically built as hierarchies?

- Solution:
  - The faster a memory technology is, the more it tends to cost per bit of storage. Using a memory hierarchy allows the computer to provide a large memory capacity, fast average access time, and low memory cost.

Exercises

- Solution: (ctd.)
  - The lower levels of the memory hierarchy, which contain the most storage, are implemented using slow but cheap memory technologies. The higher levels, which contain smaller amounts of storage, are implemented in fast but expensive memory technologies.
  - As data is referenced, it is moved into the higher levels of the hierarchy. If enough references are handled by the top levels of the hierarchy, the memory system gives an average access time similar to that of the fastest level of the hierarchy, with a cost per bit similar to that of the lowest level of the hierarchy.