PRINCIPLE OF LOCALITY

Programs access a small proportion of their address space at any time.

Temporal locality

Items accessed recently are likely to be accessed again soon.

E.g., instructions in a loop, induction variables

Spatial locality

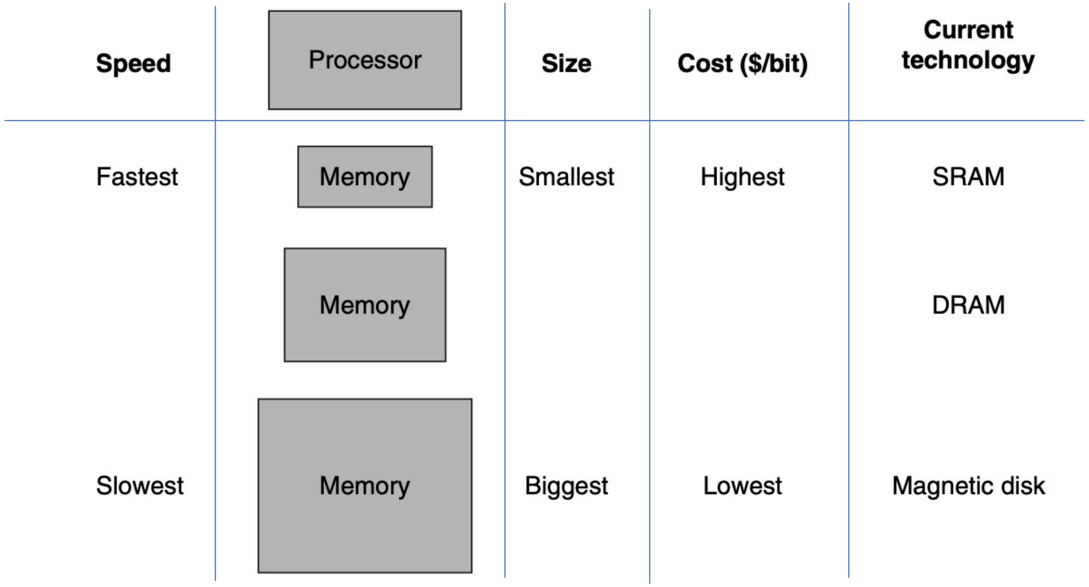
Items near those accessed recently are likely to be accessed soon.

E.g., sequential instruction access, array data

TAKING ADVANTAGE OF LOCALITY

Memory hierarchy

A structure that uses multiple levels of memories; as the distance from the processor increases, the size of the memories and the access time both increase.



Speed	Processor	**Size**	**Cost ($/bit)**	**Current technology**
Fastest	Memory	Smallest	Highest	SRAM
	Memory			DRAM
Slowest	Memory	Biggest	Lowest	Magnetic disk

- The goal is to present the user with as much memory as is available in the cheapest technology, while providing access at the speed offered by the fastest memory.

TAKING ADVANTAGE OF LOCALITY

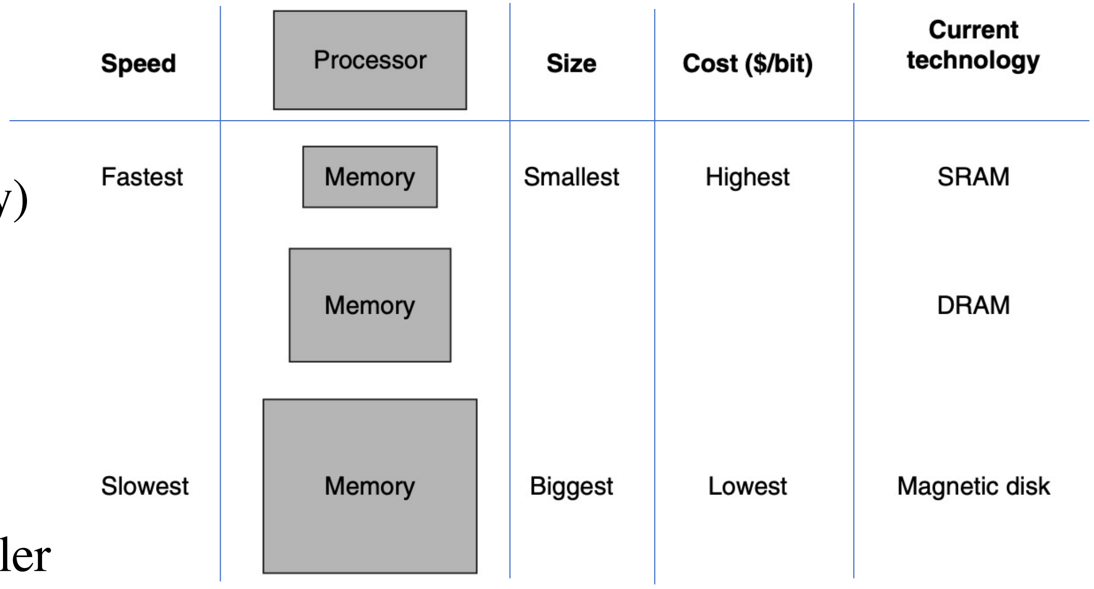
- Store everything on disk

- Copy recently accessed (and nearby) items from disk to smaller DRAM memory

  Main memory

- Copy more recently accessed (and nearby) items from DRAM to smaller SRAM memory

  Cache memory attached to CPU



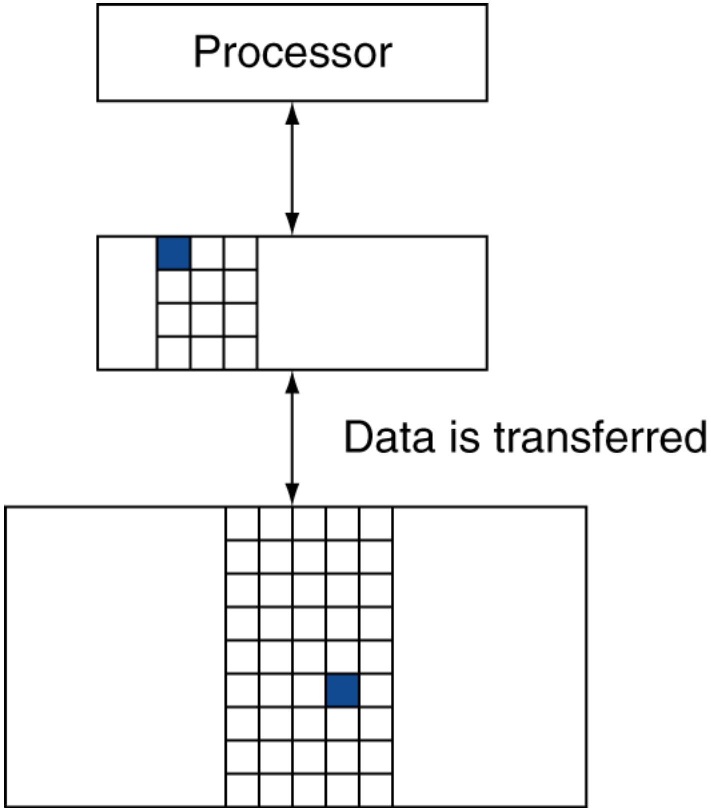
Speed	Processor	Size	Cost ($/bit)	Current technology
Fastest	Memory	Smallest	Highest	SRAM
	Memory			DRAM
Slowest	Memory	Biggest	Lowest	Magnetic disk

MEMORY HIERARCHY LEVELS

Block (aka line): unit of copying
May be multiple words

- If accessed data is present in upper level
  - Hit: access satisfied by upper level
    - Hit ratio: hits/accesses

- If accessed data is absent
  - Miss: block copied from lower level
  - Time taken: miss penalty
    - Miss ratio: misses/accesses
      - = 1 – hit ratio



MEMORY HIERARCHY LEVELS

Block (or line): The minimum unit of information that can be either present or not present in a cache.

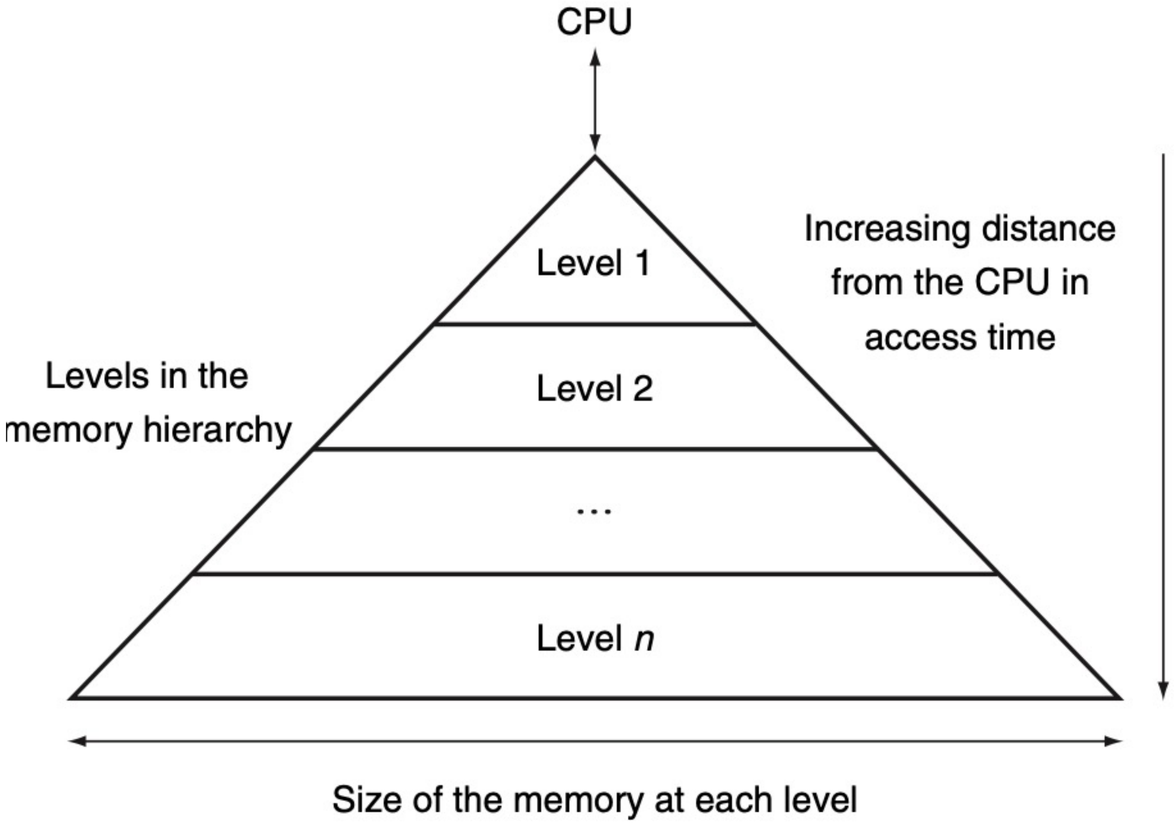
Hit rate: The fraction of memory accesses found in a level of the memory hierarchy.

Miss rate: The fraction of memory accesses not found in a level of the memory hierarchy.

Hit time**:** The time required to access a level of the memory hierarchy, including the time needed to determine whether the access is a hit or a miss.

Miss penalty: The time required to fetch a block into a level of the memory hierarchy from the lower level, including the time to access the block, transmit it from one level to the other, insert it in the level that experienced the miss, and then pass the block to the requestor

MEMORY HIERARCHY LEVELS


CPU
Level 1
Level 2
...
Level n
Levels in the memory hierarchy
Increasing distance from the CPU in access time
Size of the memory at each level

MEMORY HIERARCHY LEVELS

Memory hierarchies take advantage of temporal locality by keeping more recently accessed data items closer to the processor.

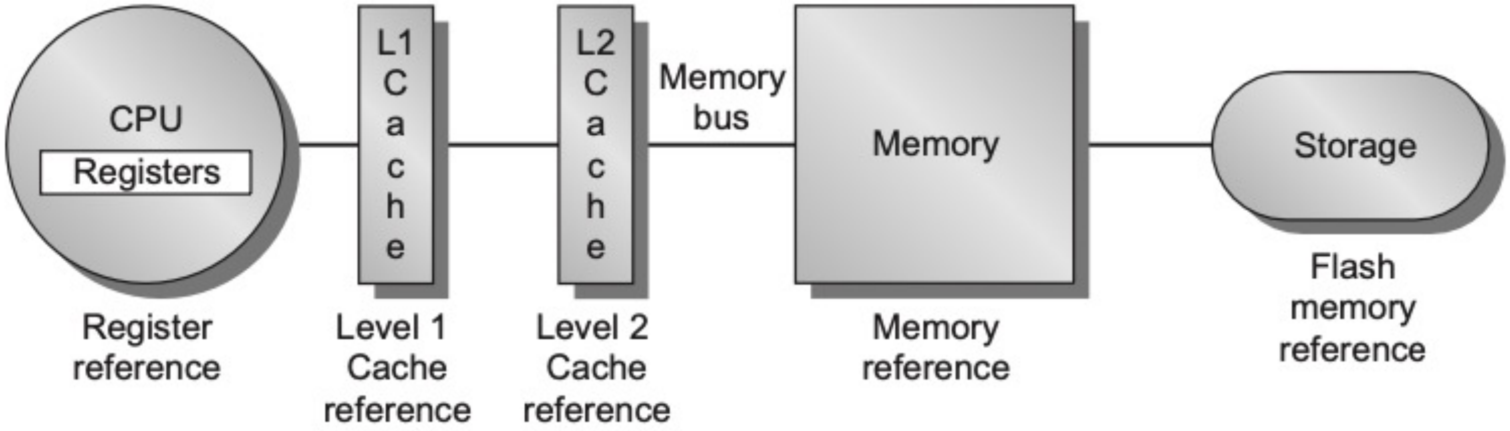
Memory hierarchies take advantage of spatial locality by moving blocks consisting of multiple contiguous words in memory to upper levels of the hierarchy.

Memory hierarchy uses smaller and faster memory technologies close to the processor. Thus, accesses that hit in the highest level of the hierarchy can be processed quickly. Accesses that miss go to lower levels of the hierarchy, which are larger but slower.

If the hit rate is high enough, the memory hierarchy has an effective access time close to that of the highest (and fastest) level and a size equal to that of the lowest (and largest) level.

In most systems, the memory is a true hierarchy, meaning that data cannot be present in level i unless it is also present in level $i + 1$.

MEMORY HIERARCHY LEVELS



CPU

Registers

L1 Cache

L2 Cache

Memory bus

Memory

Storage

Register reference

Level 1 Cache reference

Level 2 Cache reference

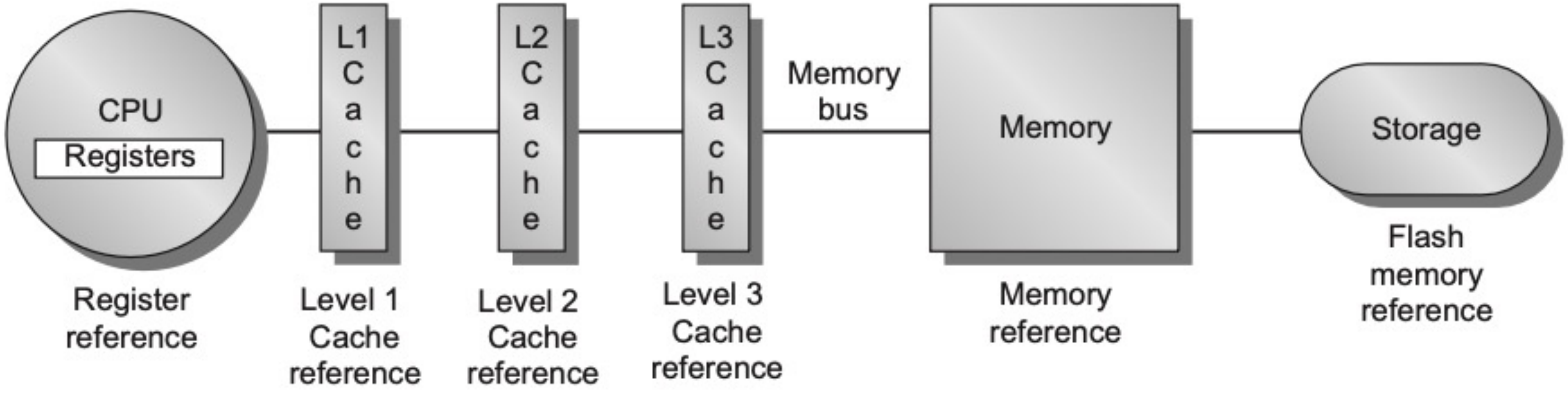
Memory reference

Flash memory reference

	Register reference	Level 1 Cache reference	Level 2 Cache reference	Memory reference	Flash memory reference
Size:	1000 bytes	64 KB	256 KB	1–2 GB	4–64 GB
Speed:	300 ps	1 ns	5-10 ns	50–100 ns	25–50 us

(A) Memory hierarchy for a personal mobile device

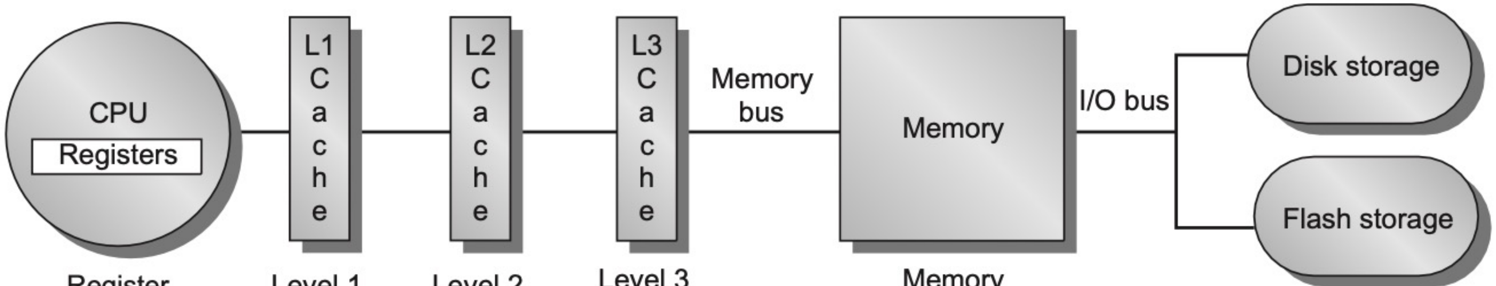
MEMORY HIERARCHY LEVELS



Laptop	Size:	1000 bytes	64 KB	256 KB	4-8 MB	4–16 GB	256 GB-1 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS
Desktop	Size:	2000 bytes	64 KB	256 KB	8-32 MB	8–64 GB	256 GB-2 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS

(B) Memory hierarchy for a laptop or a desktop

MEMORY HIERARCHY LEVELS



	Register reference	Level 1 Cache reference	Level 2 Cache reference	Level 3 Cache reference	Memory reference	Disk memory reference	Flash memory reference
Size:	4000 bytes	64 KB	256 KB	16-64 MB	32–256 GB	16–64 TB	1-16 TB
Speed:	200 ps	1 ns	3–10 ns	10 – 20 ns	50 –100 ns	5 –10 ms	100-200 us

(C)

Memory hierarchy for server

MEMORY TECHNOLOGIES

Four primary technologies used in memory hierarchies.

1. Main memory is implemented from DRAM (dynamic random access memory).

2. While levels closer to the processor (caches) use SRAM (static random access memory). DRAM is less costly per bit than SRAM, although it is substantially slower.

3. The flash memory is a non-volatile memory used as the secondary memory in Personal Mobile Devices.

4. The fourth technology, used to implement the largest and slowest level in the hierarchy in servers, is magnetic disk.

MEMORY TECHNOLOGIES

- Static RAM (SRAM)
  0.5ns – 2.5ns, $2000 – $5000 per GB
- Dynamic RAM (DRAM)
  50ns – 70ns, $20 – $75 per GB
- Flash Memory
  5000ns – 50,000ns, $5 – $10 per GB
- Magnetic disk
  5000,000ns – 20,000,000ns, $0.20 – $2 per GB

Ideal memory:

Access time of SRAM
Capacity and cost/GB of disk

MEMORY TECHNOLOGIES

Static RAM (SRAM)

IC’s that are memory arrays with (usually) a single access port that can provide either a read or a write.

Have a fixed access time to any datum, though the read and write access times may differ.

Don’t need to refresh and so the access time is very close to the cycle time.

Typically use six to eight transistors per bit to prevent the information from being disturbed when read. SRAM needs only minimal power to retain the charge in standby mode.

In the past, most PCs and server systems used separate SRAM chips for either their primary, secondary, or even tertiary caches. Today, all three levels of caches are integrated onto the processor chip, so the market for separate SRAM chips has nearly evaporated.

MEMORY TECHNOLOGIES

Dynamic RAM (DRAM)

In a SRAM, as long as power is applied, the value can be kept indefinitely.

In a dynamic RAM (DRAM), the value kept in a cell is stored as a charge in a capacitor.

A single transistor is then used to access this stored charge, either to read the value or to overwrite the charge stored there.

Because DRAMs use only a single transistor per bit of storage, they are much denser and cheaper per bit than SRAM.

As DRAMs store the charge on a capacitor, it cannot be kept indefinitely and must periodically be refreshed. That is why this memory structure is called dynamic, as opposed to the static storage in an SRAM cell.

MEMORY TECHNOLOGIES

Dynamic RAM (DRAM)

To refresh the cell, we merely read its contents and write it back.

The charge can be kept for several milliseconds. If every bit had to be read out of the DRAM and then written back individually, we would constantly be refreshing the DRAM, leaving no time for accessing it.

Fortunately, DRAMs use a two-level decoding structure, and this allows us to refresh an entire row (which shares a word line) with a read cycle followed immediately by a write cycle.