Microprocessors

Tuba Ayhan

MEF University

Memories

CH8

Basic Concepts

- k-bit addresses is capable of addressing up to 2^k memory locations.
- 16-bit addresses is capable of addressing up to 2^{16} = 64K (kilo) memory locations
- 32-bit addresses can utilize a memory that contains up to 2^{32} = 4G (giga) locations
- 64-bit addresses can access up to 2^{64} = 16E (exa) ≈ 16×10^{18} locations

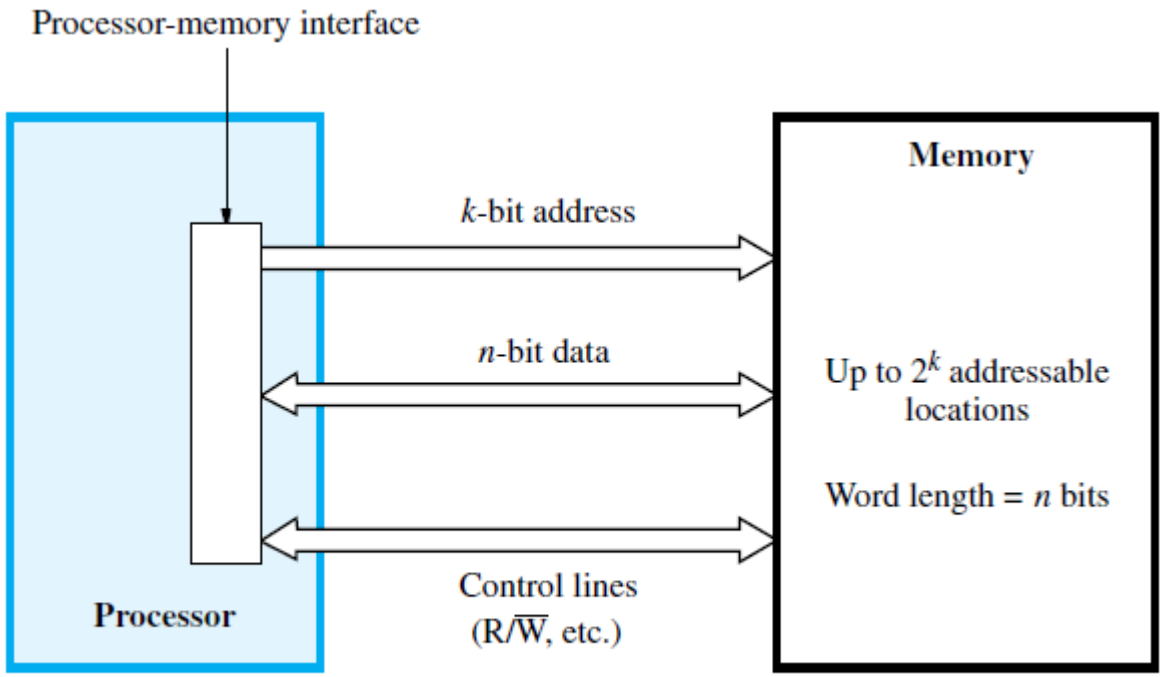


Figure 8.1 Connection of the memory to the processor.

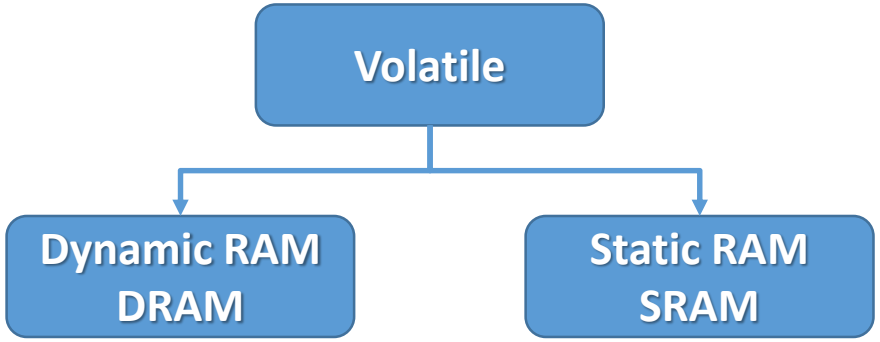
Basic Concepts

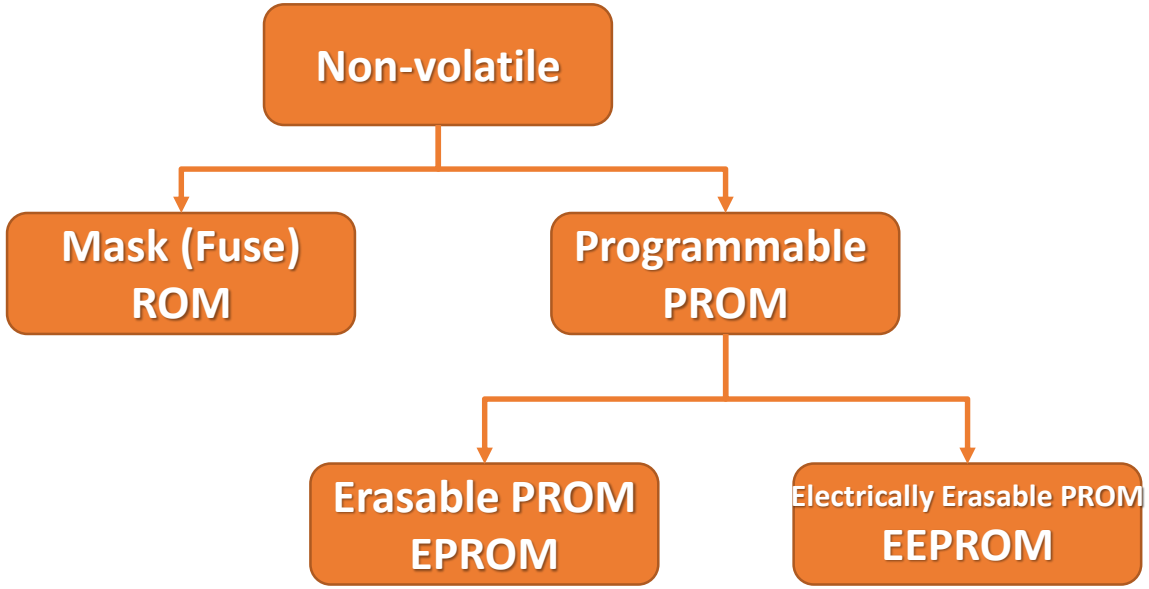
- memory access time: time that elapses between the initiation of an operation to transfer a word of data and the completion of that operation.
- memory cycle time: minimum time delay required between the initiation of two successive memory operations.
  - i.e., the time between two successive Read operations.
  - The cycle time is usually slightly longer than the access time, depending on the implementation details of the memory unit.
- random-access memory (RAM): the access time to any location is the same, independent of the location's address.
  - The processor of a computer can usually process instructions and data faster than they can be fetched from the main memory. Hence, the memory access time is the bottleneck in the system.

Basic Concepts

- Cache memory: a small, fast memory inserted between the larger, slower main memory and the processor.
  - It holds the currently active portions of a program and their data.
- Virtual memory: only the active portions of a program are stored in the main memory, and the remainder is stored on the much larger secondary storage device.
  - Sections of the program are transferred back and forth between the main memory and the secondary storage device.
  - As a result, the application program sees a memory that is much larger than the computer's physical main memory.
- Frequent data transfer between the main memory and the cache and between the main memory and the disk → Block Transfers
  - Data are always transferred in contiguous blocks involving tens, hundreds or thousands of words
  - Data transfers between the main memory and high-speed devices such as a graphic display or an Ethernet interface also involve large blocks of data.

Semiconductor memory types





- Flash memory
- Ferroelectric RAM (FRAM)
- Magnetoresistive RAM (MRAM)
- Resistive RAM (RRAM)
- Phase-Change RAM (PCRAM)
- Spin Torque Transfer RAM (STT)

Internal Organization of Memory Chips

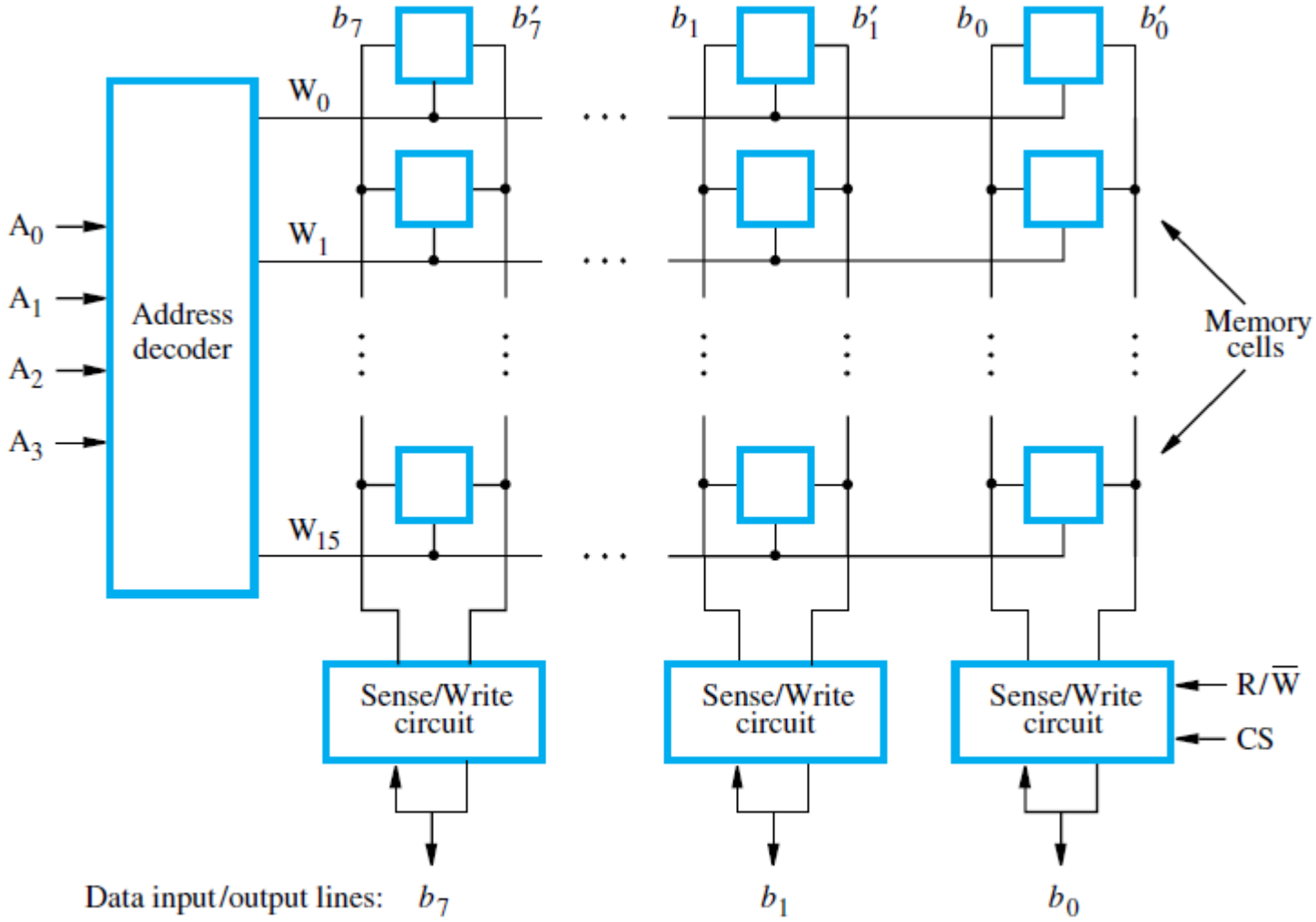


Figure 8.2 Organization of bit cells in a memory chip.

Internal Organization of Memory Chips

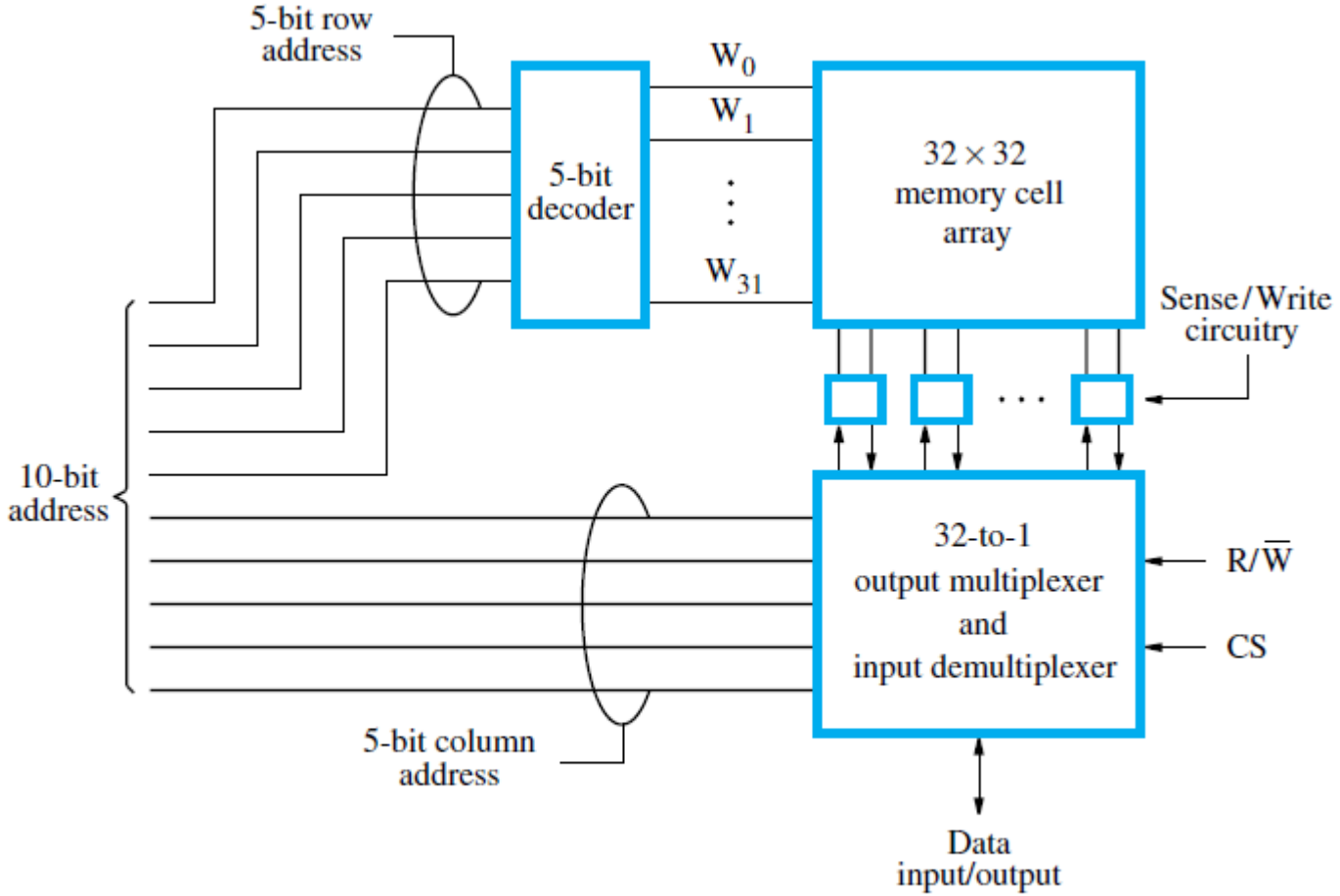


Figure 8.3 Organization of a 1K × 1 memory chip.

Static RAMs

- **Read Operation**
- The word line is activated: switches *T*1 and *T*2 are closed.
  - The cell is in state 1: the signal on bit line *b* is high and the signal on bit line *b* is low.
  - The cell is in state 0: the signal on bit line *b* is low and the signal on bit line *b* is high.
  - The Sense/Write circuit at the end of the two bit lines monitors their state and sets the corresponding output accordingly.
- **Write Operation**
- The Sense/Write circuit drives bit lines *b* and *b*.
- It places the appropriate value on bit line *b* and its complement on *b* and activates the word line. This forces the cell into the corresponding state, which the cell retains when the word line is deactivated.

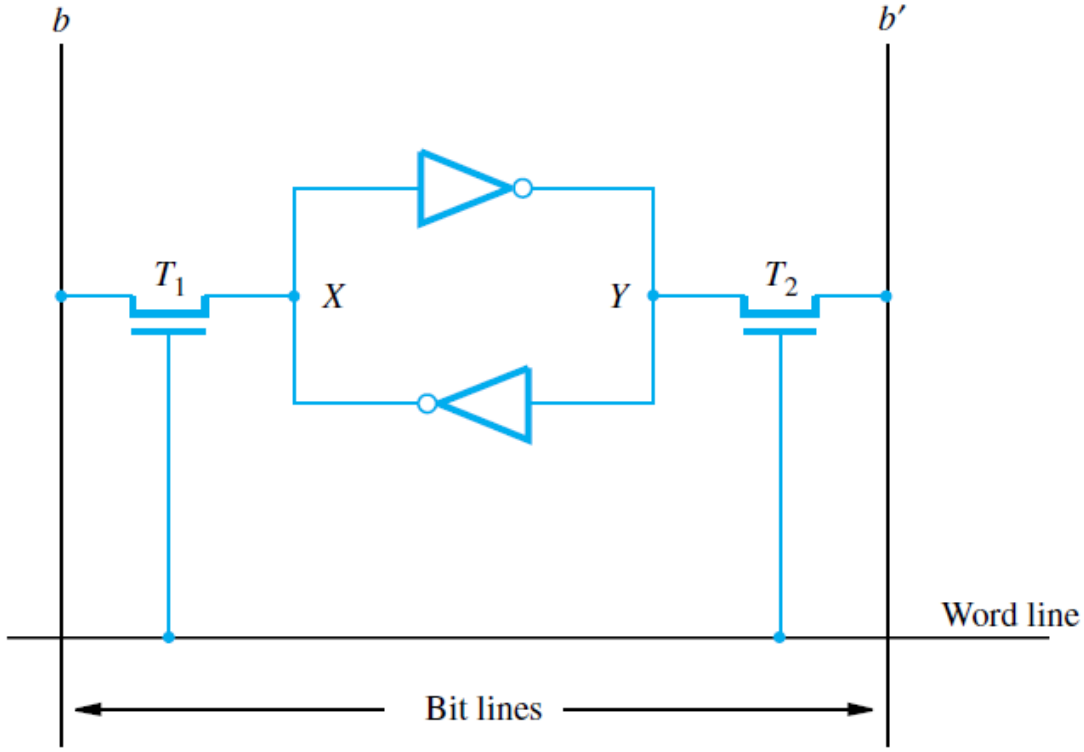


Figure 8.4 A static RAM cell.

Dynamic RAMs

- Information is stored in a dynamic memory cell in the form of a charge on a capacitor.
- This charge can be maintained for only tens of milliseconds. → **refresh is required!**
- Refresh: occurs when the contents of the cell are read or when new information is written into it.

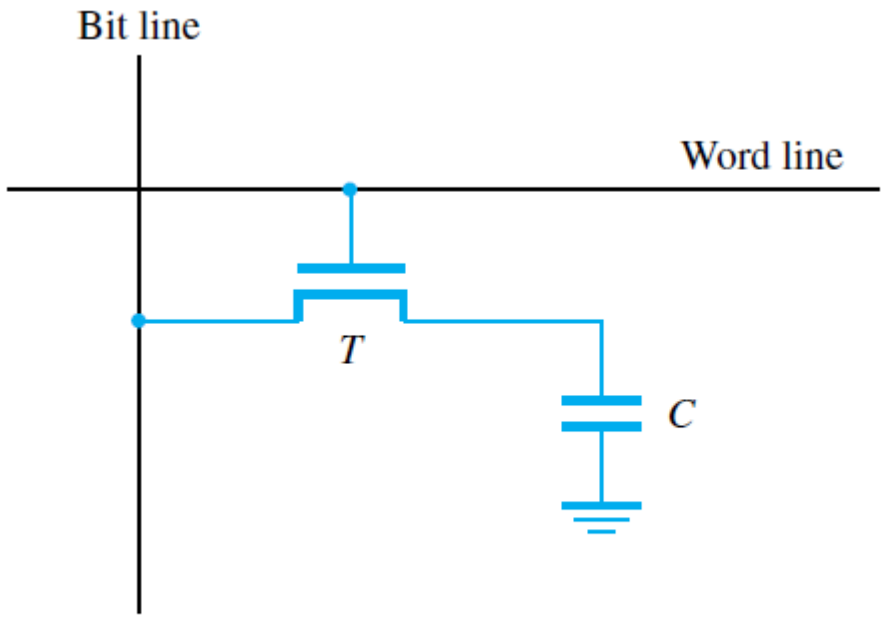


Figure 8.6 A single-transistor dynamic memory cell.

Dynamic RAMs

- The cells are organized in the form of a 16K × 16K array.
- The 16,384 cells in each row are divided into 2,048 groups of 8, forming 2,048 bytes of data. → 25-bit address is needed
  - The high-order 14 address bits are needed to select a row
  - The low-order 11 bits are needed to specify a byte in the selected row.
  - To reduce the number of pins needed for external connections, the row and column addresses are multiplexed on 14 pins.

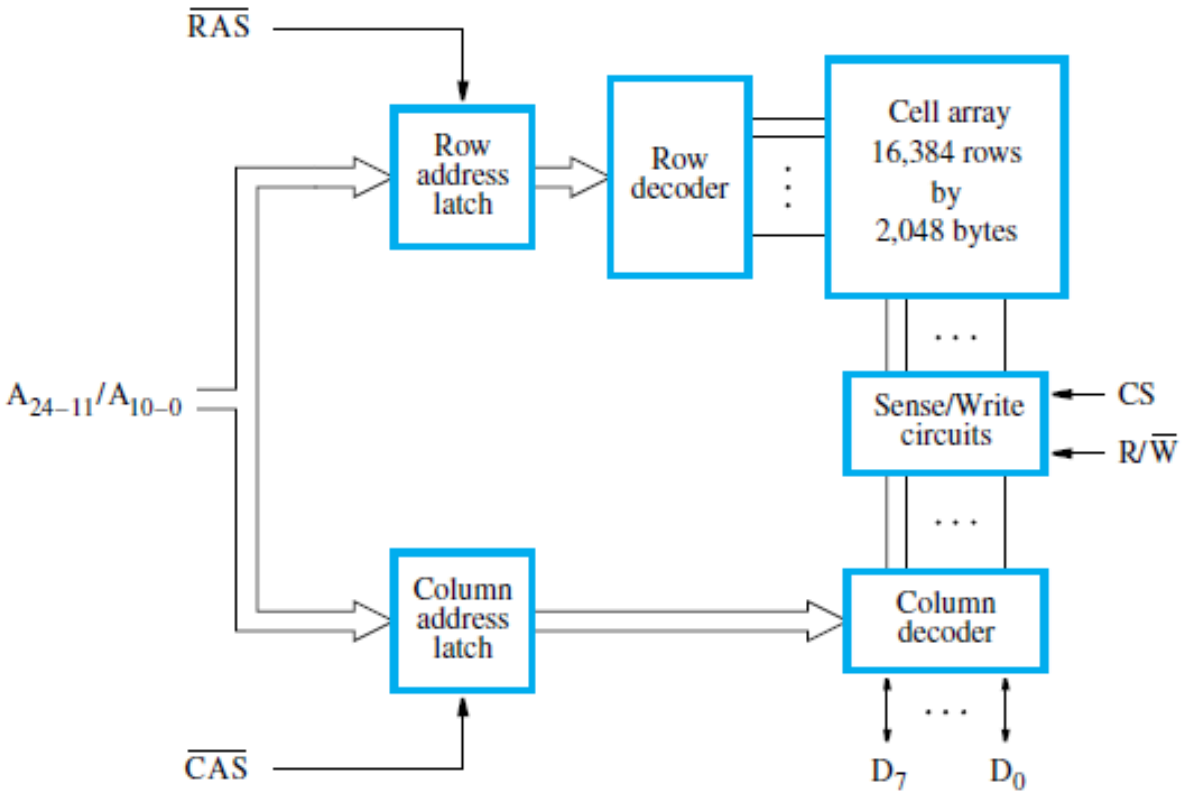


Figure 8.7 Internal organization of a 32M × 8 dynamic memory chip.

Dynamic RAMs

- During a Read or a Write:

1. The row address is applied first. It is loaded into the row address latch when the Row Address Strobe (RAS) is low.
2. The column address is applied to the address pins and loaded into the column address latch under control of Column Address Strobe (CAS).
3. The information in this latch is decoded and the appropriate group of 8 Sense/Write circuits is selected.

- Read operation (R/W = 1)
  - The output values of the selected circuits are transferred to the data lines, D7−0.
- Write operation (R/W = 0)
  - The information on theD7−0 lines is transferred to the selected circuits, then used to overwrite the contents of the selected cells in the corresponding 8 columns.

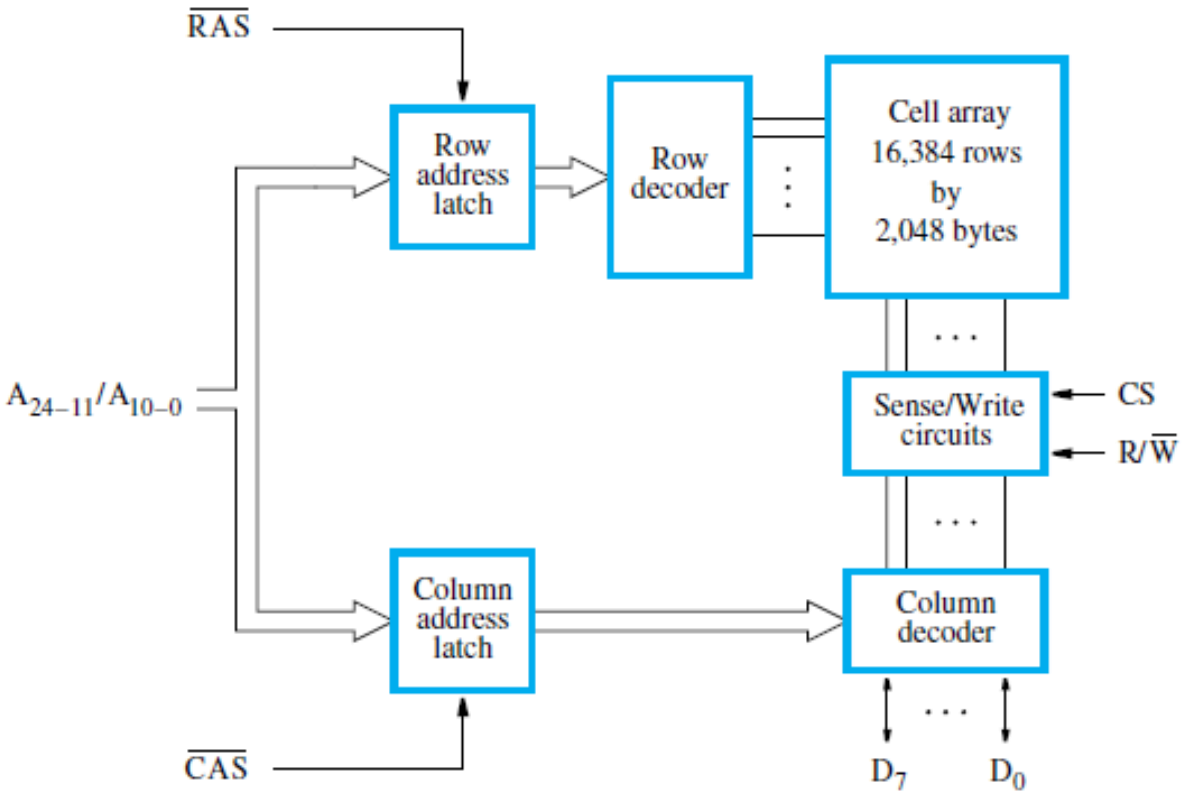


Figure 8.7 Internal organization of a 32M × 8 dynamic memory chip.

Dynamic RAMs - Fast Page Mode

- the contents of all 16,384 cells in the selected row are sensed, but only 8 bits are placed on the data lines, D7–0.
- How to access the other bytes in the same row without having to reselect the row?
- When a row address is applied, the contents of all cells in the selected row are loaded into the corresponding latches. Then, it is only necessary to apply different column addresses to place the different bytes on the data lines.
- All bytes in the selected row can be transferred in sequential order by applying a consecutive sequence of column addresses under the control of successive CAS signals.
- A block of data can be transferred at a much faster rate.
- The faster rate in the fast page mode makes dynamic RAMs well suited to **block transfers**.

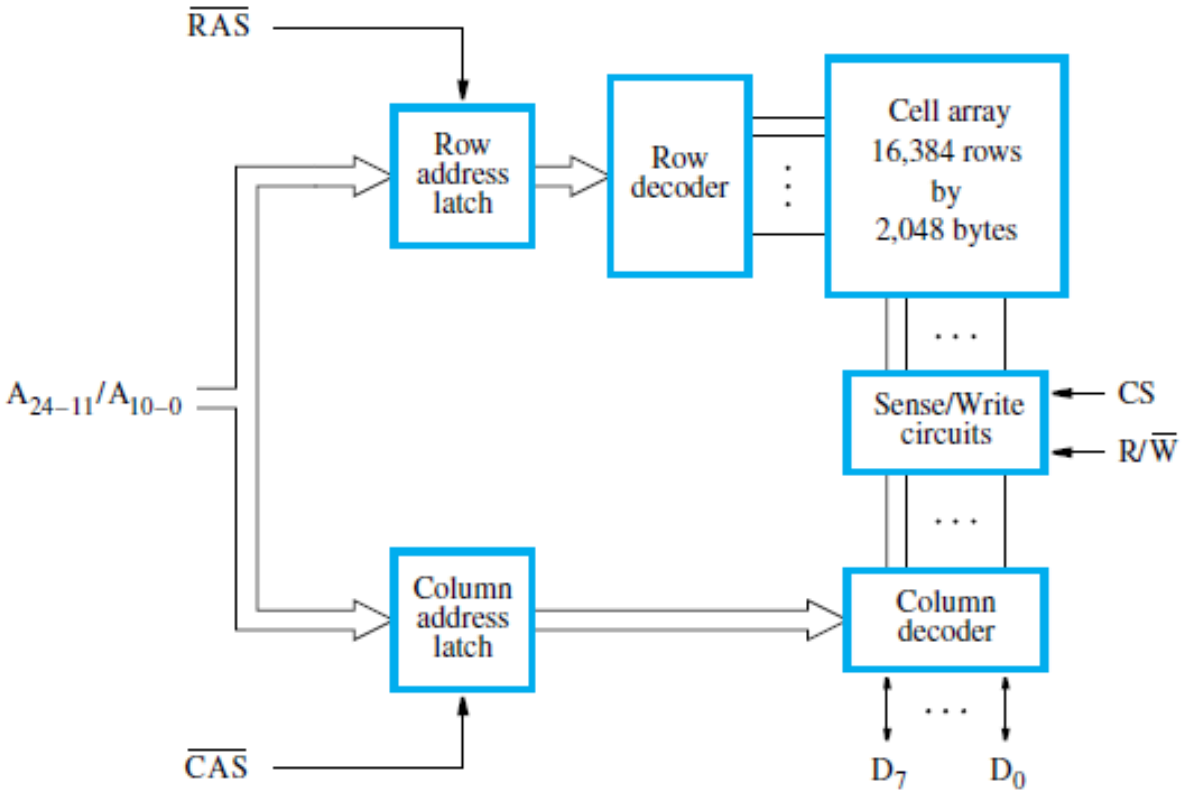


Figure 8.7 Internal organization of a 32M × 8 dynamic memory chip.

Synchronous DRAM - SDRAM

- **Main difference: «use of a clock signal»**
- have built-in refresh circuitry → the dynamic nature of these memory chips is almost invisible to the user.
- The address and data connections of an SDRAM may be buffered by means of registers.
- The Sense/Write amplifiers function as latches, as in asynchronous DRAMs.
- A Read operation causes the contents of all cells in the selected row to be loaded into these latches. The data in the latches of the selected column are transferred into the data register → the data output pins.
- Isolate external connections from the chip's internal circuitry: we can start a new access operation while data are being transferred to or from the registers.

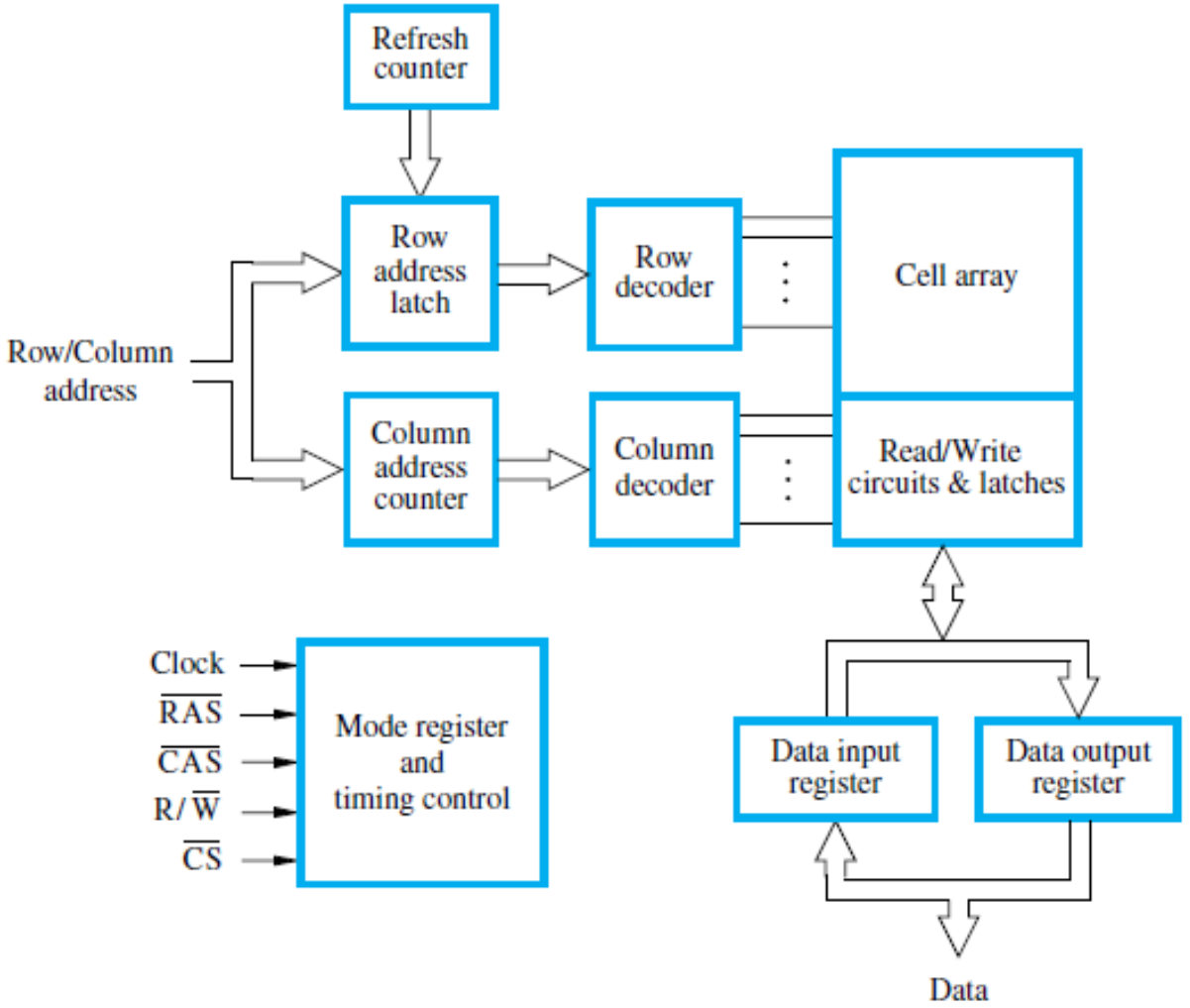


Figure 8.8 Synchronous DRAM.

Synchronous DRAM - SDRAM

- Example: a typical burst read of length 4.

1. The row address is latched under control of the RAS signal.
2. The memory typically takes 5 or 6 clock cycles (for simplicity, 2 in the figure) to activate the selected row.
3. The column address is latched under control of the CAS signal.
4. After a delay of one clock cycle, the first set of data bits is placed on the data lines.
5. The SDRAM automatically increments the column address to access the next three sets of bits in the selected row, which are placed on the data lines in the next 3 clock cycles.

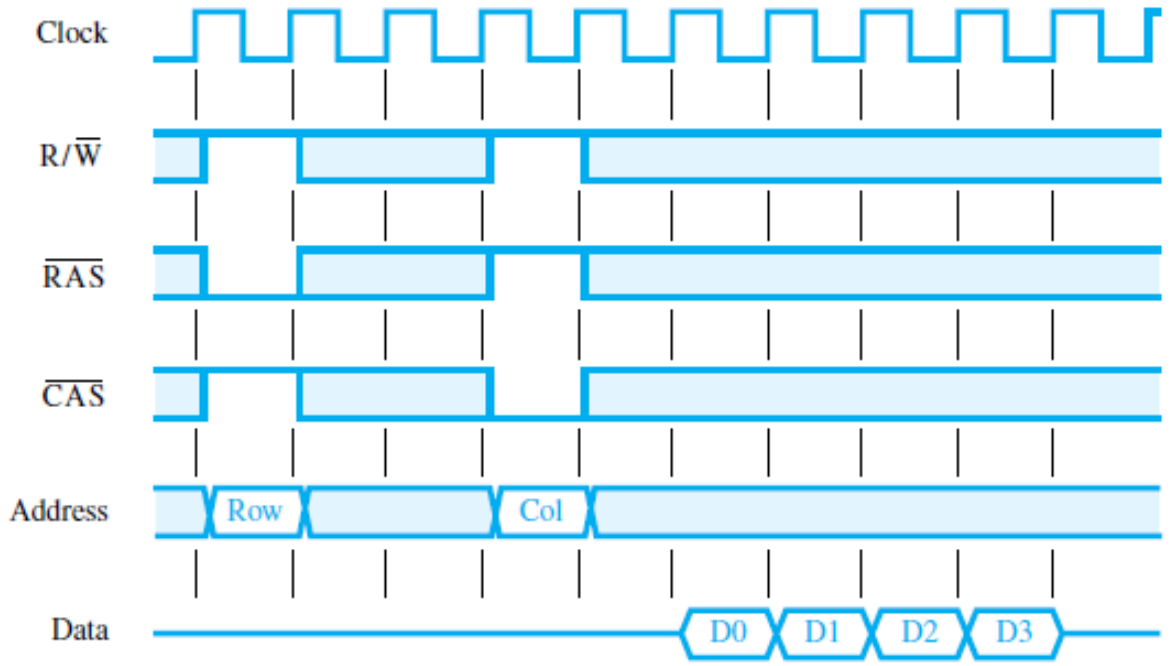


Figure 8.9 A burst read of length 4 in an SDRAM.

Latency and Bandwidth

- memory latency: is the amount of time it takes to transfer the first word of a block.
  - the access cycle begins with the assertion of the RAS signal → the latency is five clock cycles.
- Bandwidth: the number of bits or bytes that can be transferred in one second.
  - It depends on the speed of access to the stored data and on the number of bits that can be accessed in parallel.
  - The rate at which data can be transferred to or from the memory depends on the bandwidth of the system interconnections.
- The time between successive words of a block is much shorter than the time needed to transfer the first word.

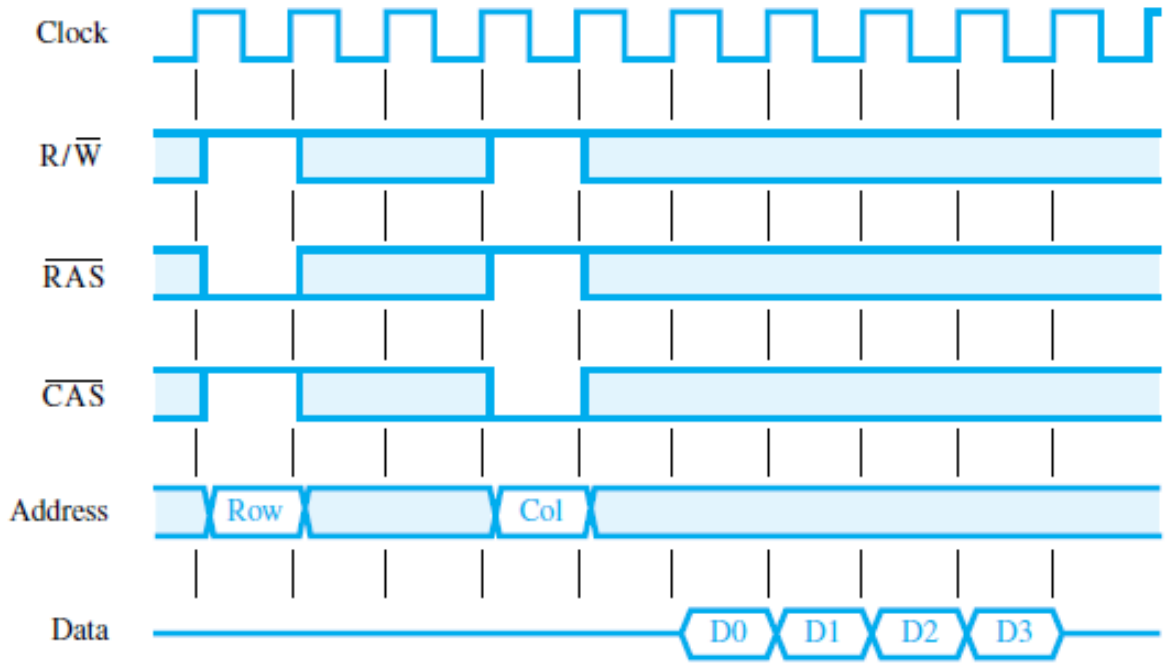


Figure 8.9 A burst read of length 4 in an SDRAM.

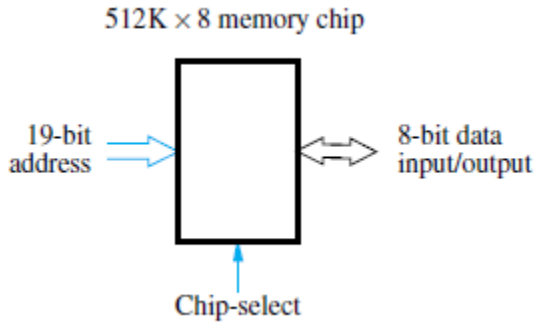
Double-Data-Rate SDRAM

- They are faster versions of SDRAMs.
  - faster circuits,
  - new organizational and operational features.
- Fact: A large number of bits are accessed at the same time inside the chip when a row address is applied.
- To make the best use of the available clock speed, data are transferred externally on both the rising and falling edges of the clock.
- Versions:
- DDR: earliest
- DDR2: 400MHz
- DDR3: up to 1067MHz
- DDR4: between 800 and 4266 MHz

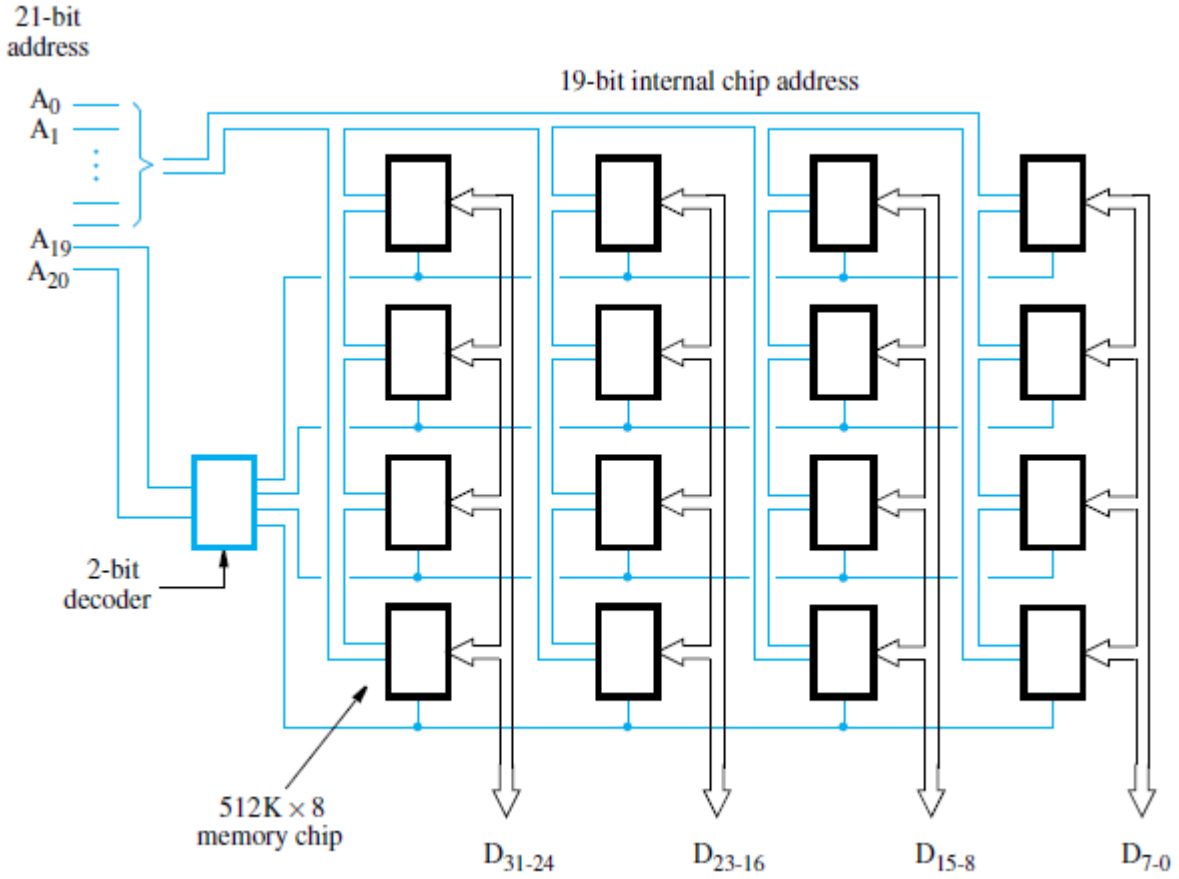
increased storage capacity,
lower power, and
faster clock speeds

Question

- Generate a 512K × 32 memory module using 512K × 8 static memory chips.
- Generate a 2M × 8 memory module using 512K × 8 static memory chips.
- Generate a 2M × 32 memory module using 512K × 8 static memory chips.



Structure of Larger Memories



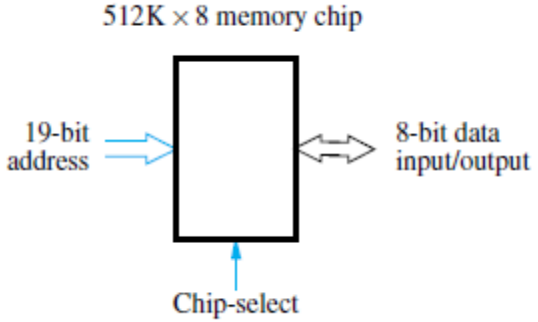


Figure 8.10 Organization of a 2M × 32 memory module using 512K × 8 static memory chips.

Read-only Memories

- ROM
- PROM: programmable ROM
  - Programming by inserting a fuse at point P
  - Before it is programmed, the memory contains all 0s.
  - This process is irreversible.
- EPROM: erasable PROM
  - The connection to ground at point P is made through a special transistor.
  - It can be turned on by injecting charge into it.
  - Erasure requires dissipating the charge trapped in the transistors that form the memory cells. → under UV light
- EEPROM: electrically erasable PROM
  - Flash Memory

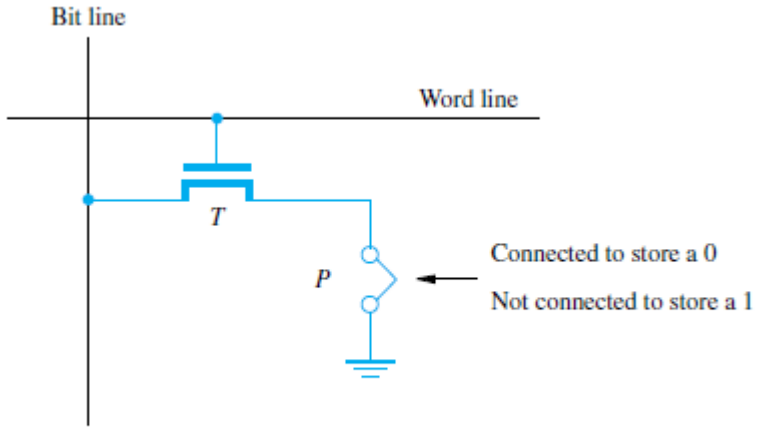
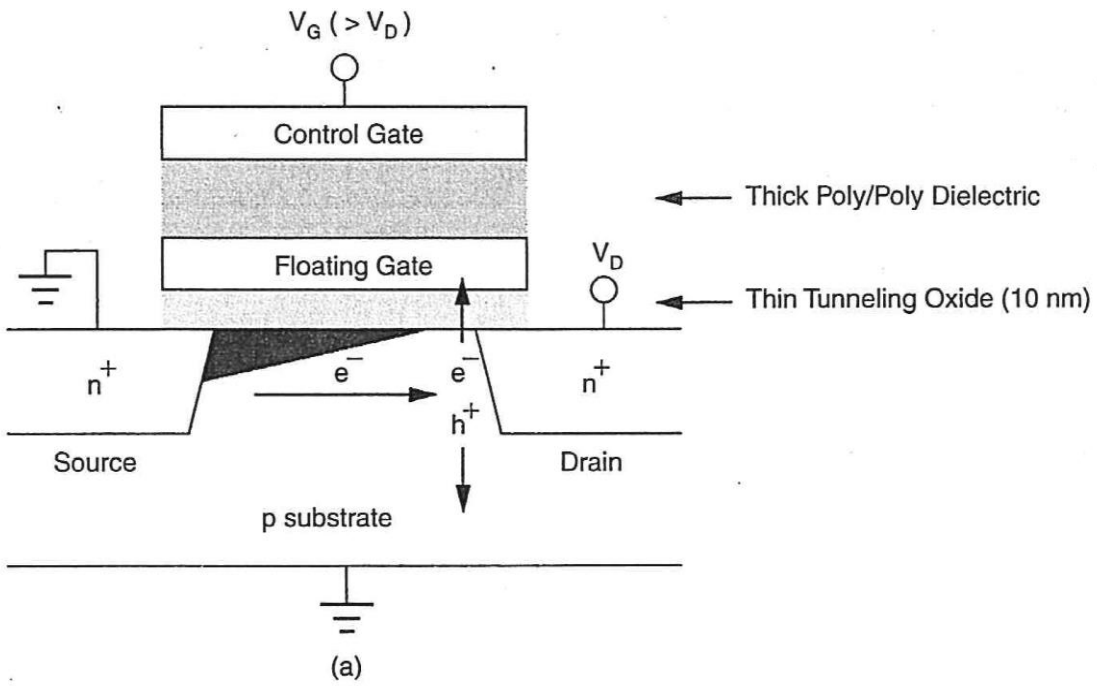


Figure 8.11 A ROM cell.

Flash Memory - (Ref. Kang and Leblebici, Ch. 10.5 by Seung-Moon Yoo)



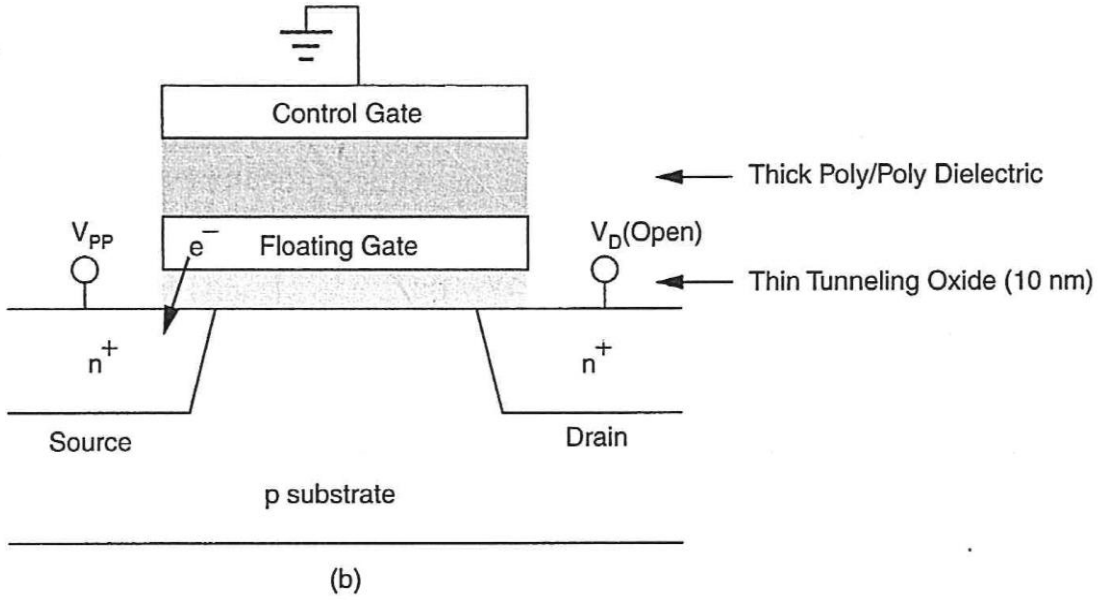


Figure 10.61 Data programming and erasing methods in the flash memory. (a) Hot-electron injection mechanism. (b) Fowler-Nordheim tunneling mechanism.

Different voltage levels are used to read and write data

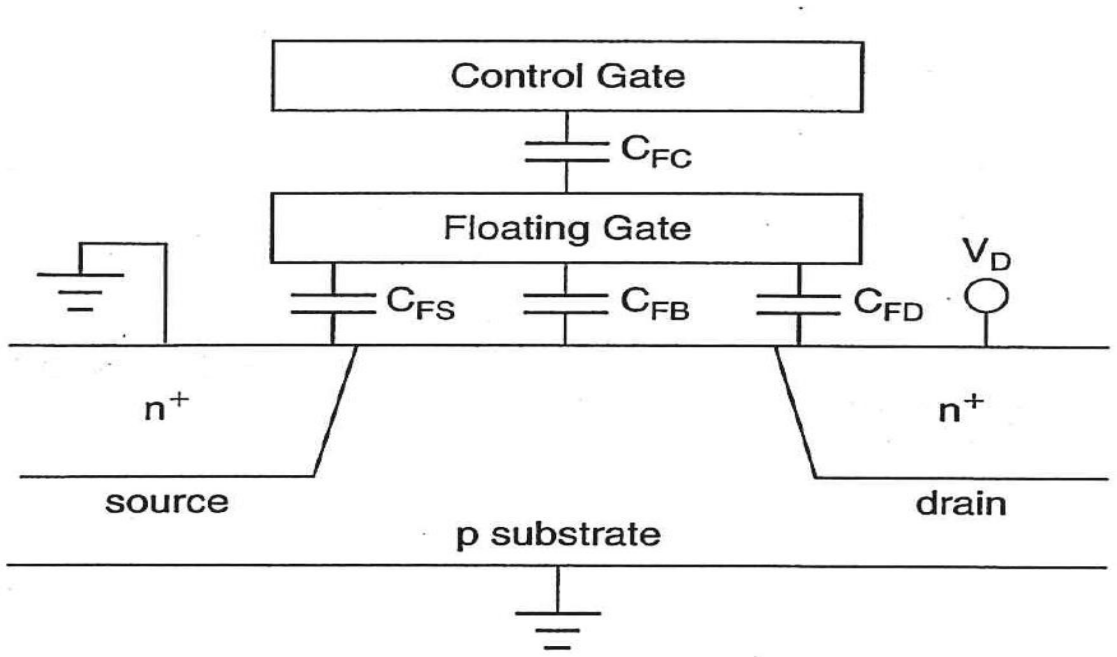


Figure 10.62 Equivalent capacitive-coupling circuit of a flash memory cell.

- When control gate voltage and drain voltage are applied, The voltage at the floating gate is a function of charge stored in the floating gate (Q_{FG}) and capacitances in the figure.

Microprocessors

Tuba Ayhan

MEF University

Direct Memory Access

CH8.4

Data transfer

- Single-word or single-byte data transfers between the processor and I/O devices

1. Read data from the I/O device.
2. Load the data into a processor register.

```
LDR R0, DATAIN
```

3. Store data into a memory location

```
STR R0, DATASTORE
```

- Read and store happens when the I/O device is ready. CPU determines that
  - by polling I/O status register or
  - by waiting for an interrupt request.
- Overhead: several program instructions must be executed involving many memory accesses for each data word transferred.
- Block transfer: Instructions are needed to increment the memory address and keep track of the word count.

Direct memory access: DMA

- DMA controller: A special control unit is provided to manage the block transfer of data directly between the main memory and I/O device.
- It transfers data without intervention by the processor.
- For each word transferred, it provides the memory address and generates all the control signals needed.
- It increments the memory address for successive words and keeps track of the number of transfers.
- It may be part of the I/O device interface, or it may be a separate unit shared by a number of I/O devices.

DMA control

- Operation must be under the control of a program executed by the processor.
- Initiate the block transfer:
  - the processor sends
    - the starting address,
    - the number of words in the block,
    - the direction of the transfer.
- End of transfer:
  - DMA controller informs the processor by raising an interrupt.

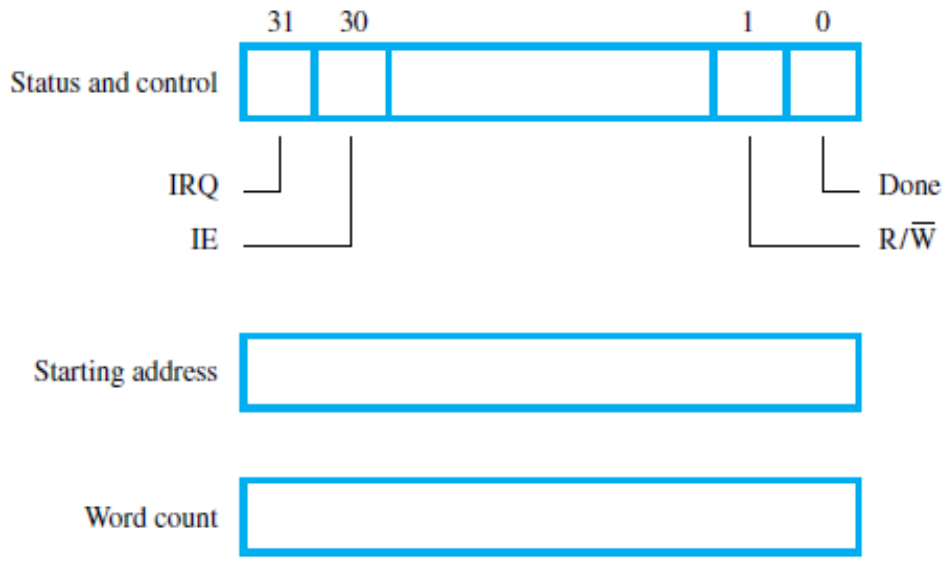


Figure 8.12 Typical registers in a DMA controller.

DMA control

- Two registers are used for storing the **starting address** and the **word count**.
- **Status and control register**: contains status and control flags:
  - The **R/W** bit determines the direction of the transfer.
    - Read operation: it transfers data from the memory to the I/O device.
    - Write operation: it transfers data from the I/O device to the memory.
  - **Done** flag indicates that the transfer is completed.
  - If **IE** flag is set, the controller raise an interrupt after it has completed the transfer, by setting the **IRQ** bit.

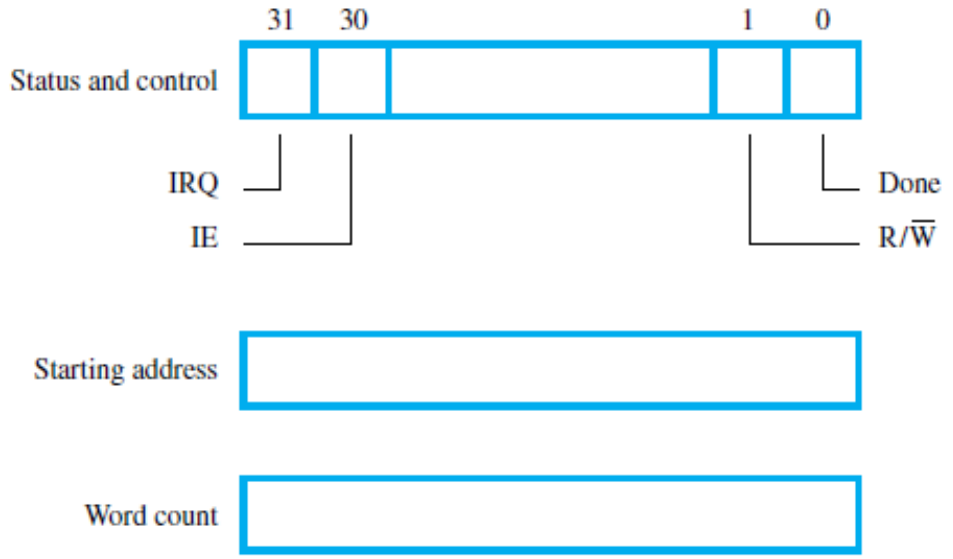


Figure 8.12 Typical registers in a DMA controller.

DMA example

- One DMA controller connects a high-speed Ethernet to the computer's I/O bus (a PCI bus).
- The disk controller, which controls two disks, also has DMA capability and provides two DMA channels. It can perform two independent DMA operations, as if each disk had its own DMA controller.
- DMA transfer from the main memory to a disk:
    1. An OS routine writes the address and word count information into the registers of the disk controller.
    2. The DMA controller proceeds independently to implement the specified operation.
    3. When the transfer is completed, it is announced by Done and/or IRQ bit in the status register of the DMA.
    4. The status register may also be used to record other information, such as whether the transfer took place correctly or errors occurred.

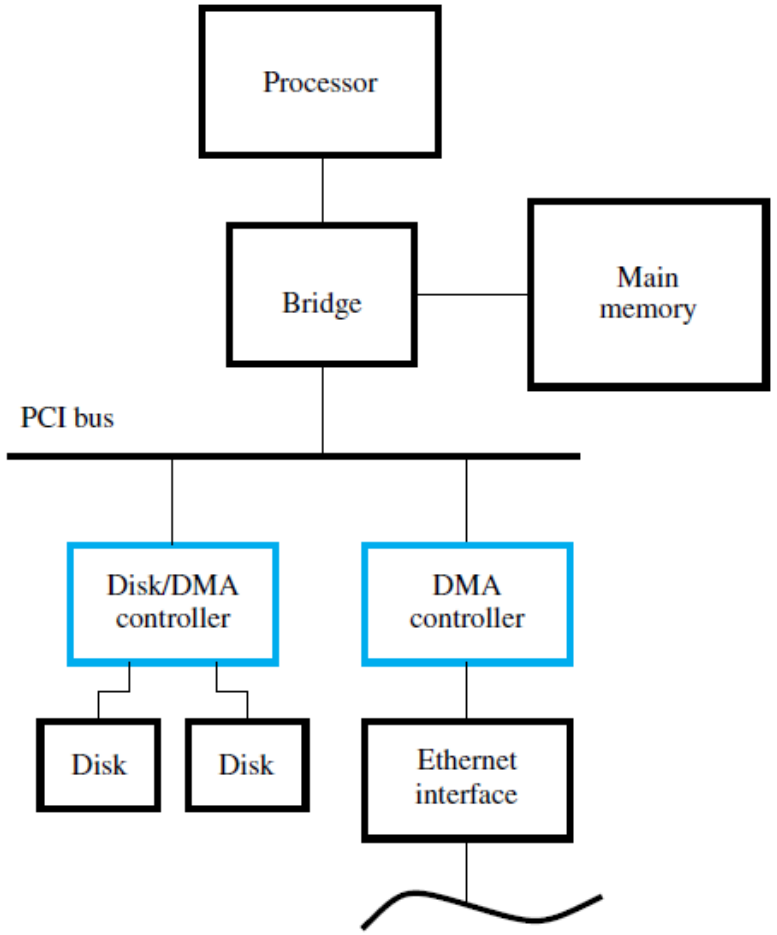


Figure 8.13 Use of DMA controllers in a computer system.

Memory Hierarchy

- The more expensive and much faster static RAM technology to be used in smaller units where speed is of the essence, such as in *cache memories*.
- The fastest access is to data held in *processor registers*.
- *Processor cache* holds copies of the instructions and data stored in a much larger external memory. Often two or more levels of cache:
  - A primary cache *level 1 L1* is always located on the processor chip. This cache is small and its access time is comparable to that of processor registers.
  - A larger, and hence somewhat slower, secondary cache is placed between the primary cache and the rest of the memory: *level 2* (L2) cache. Often, the L2 cache is also housed on the processor chip.
  - Some computers have a *level 3* (L3) cache of even larger size. An L3 cache, also implemented in SRAM technology, may or may not be on the same chip with the processor and the L1 and L2 caches.
- *Main memory*: is a large memory implemented using dynamic memory components. The main memory is much larger but significantly slower than cache memories. In a computer with a processor clock of 2 GHz or higher, the access time for the main memory can be as much as 100 times longer than the access time for the L1 cache.

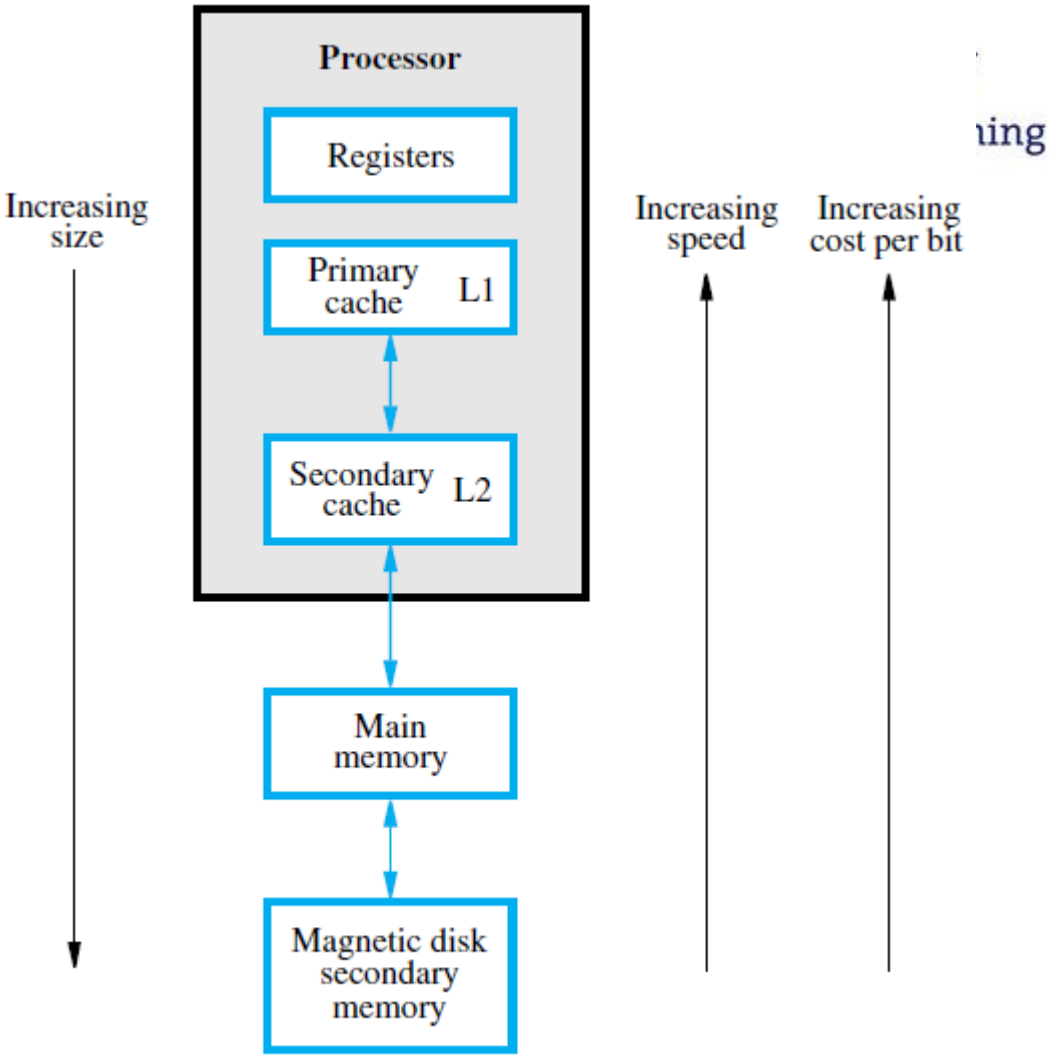


Figure 8.14 Memory hierarchy.