

07 Memory

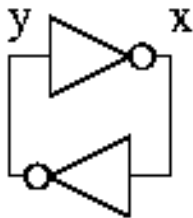
07.01 Memory devices

- General classification
- Static RAM devices
- Dynamic RAM devices
- Internal architecture
- External architecture

Classification

- First classification criterion: *volatility*
 - *Volatile* memories: Loose data when there is no power supply
 - *Non-volatile* memories: Retain data even without power supply
- Second classification criterion: *read-write* balance
 - ROM: read-only memories
 - PROM: programmable ROM
 - EPROM: erasable PROM
 - EEPROM: electrically-erasable PROM
 - FLASH: read-most non-volatile memories
 - RAM: volatile read-write random-access memories

Static RAM (SRAM) cell



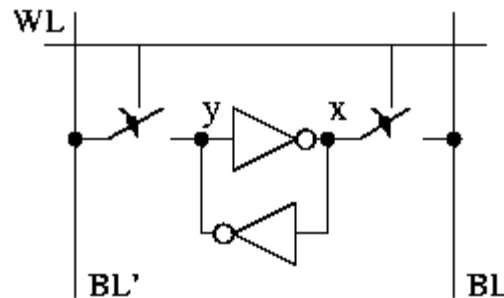
If $x=0$, the lower inverter provides $y=1$, that in its turn causes the upper inverter to sustain $x=0$

If $x=1$, the lower inverter provides $y=0$, that in its turn causes the upper inverter to sustain $x=1$

- The bistable can be used to store a bit
- The stored bit is represented by the state of the internal node x
- In order to use a bistable as a memory element we need a mechanism to
 - write a bit, i.e., change the value of x according to the bit to be stored
 - read a bit, i.e., bring the value of x outside the cell

Static RAM (SRAM)

external connections



- To make internal nodes accessible from outside the cell, two switches (i.e., transistors) are used that connect nodes x and y to two external lines, called BL (*bit line*) and BL'
- Switches are controlled by a signal called WL (*word line*)

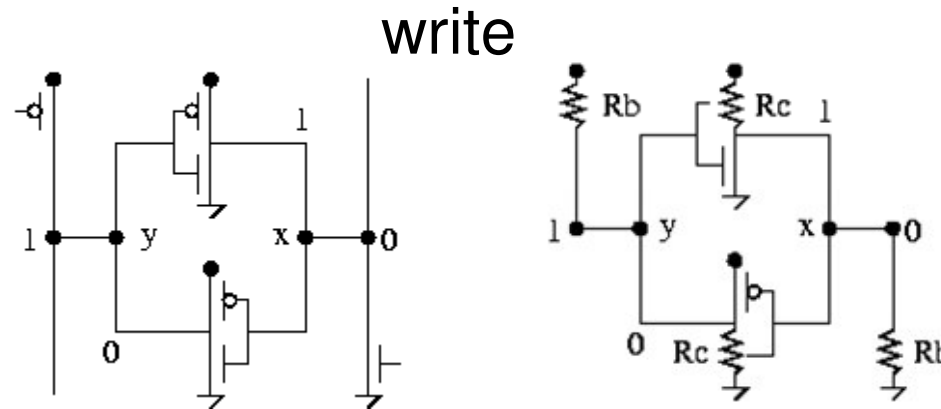
Static RAM (SRAM)

read



- To read the content of the cell, the bit lines BL and BL' are kept floating while asserting WL
- The pull-up and pull-down transistors of the internal inverters drive the bit lines bringing them to the nominal values of x and y
- The device is called static since the value of the bit is associated with the steady state of a bistable
- The pull-up and pull-down circuits drive the bit lines to 1 (0) by statically connecting them to power supply (ground)

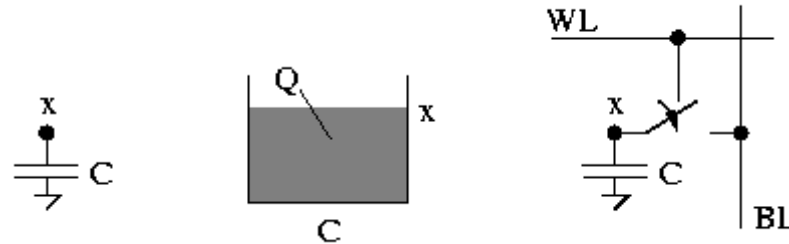
Static RAM (SRAM)



- To write a bit (say b) in the cell, the bit lines BL and BL' are driven by external drivers that provide static paths to Vdd and Ground according to the target values of b and b'
- The WL is asserted when the bit lines are externally driven
- If the value of b is not equal to the previous value of x , the write operation causes an electrical conflict between the external drivers and the internal inverters
- If R_c and R_b are the resistances of the paths provided by the cell and by the bit line to Vdd and Ground, the final value of x is determined by the path with the lower resistance
- Write operations require $R_b \ll R_c$

Dynamic RAM (DRAM)

cell



- The dynamic RAM cell is a simple capacitor, that is a passive 2-terminal component that exhibits a voltage difference (between the two terminals) proportional to the amount of charge stored

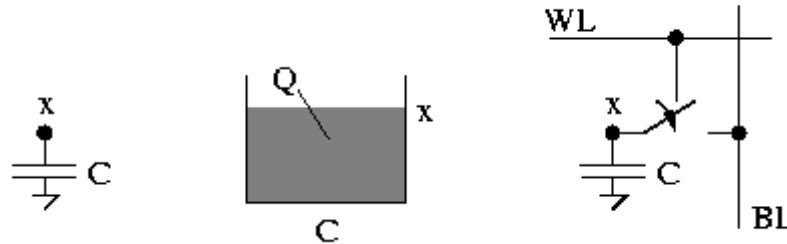
$$Q = VC$$

C is the *capacitance* of the capacitor.

- If one of the two terminals is connected to ground, V is the voltage level at the other one, that we denote by x
- A capacitor can be viewed as a cylindrical reservoir having section C . In this case, the stored charge (Q) is the volume of the fluid contained in the reservoir and x (or V) is its level.
- A transistor driven by the word line (WL) can be used in a memory device to connect the cell to the bit line (BL)

Dynamic RAM (DRAM)

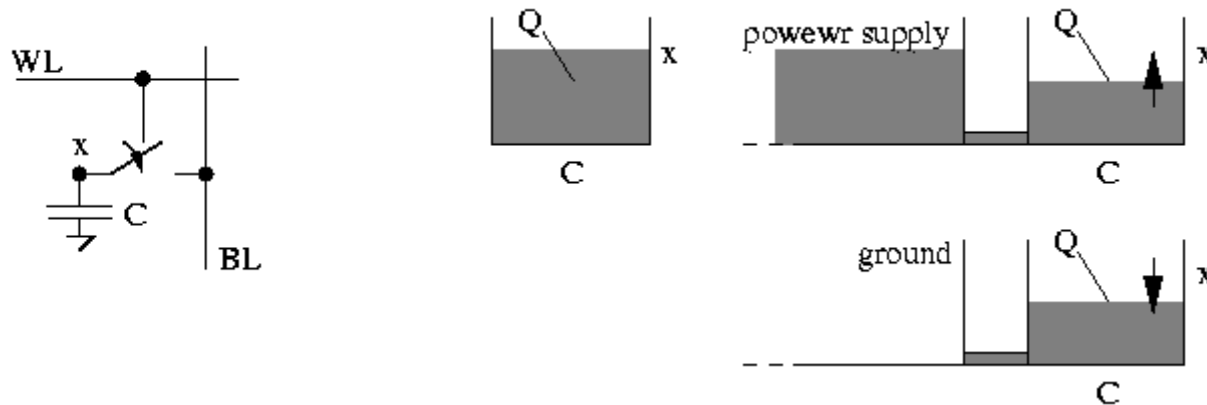
refresh



- An ideal isolated capacitor can store its charge forever.
- Actual circuits are not ideal: both capacitors and transistors are leaky devices
- DRAM cells lose their charge, thus impairing the retention of the stored data
- In order not to lose data, the content of a DRAM cell needs to be periodically *refreshed*
- Refreshing the content of a DRAM cell entails reading its value and writing it back

Dynamic RAM (DRAM)

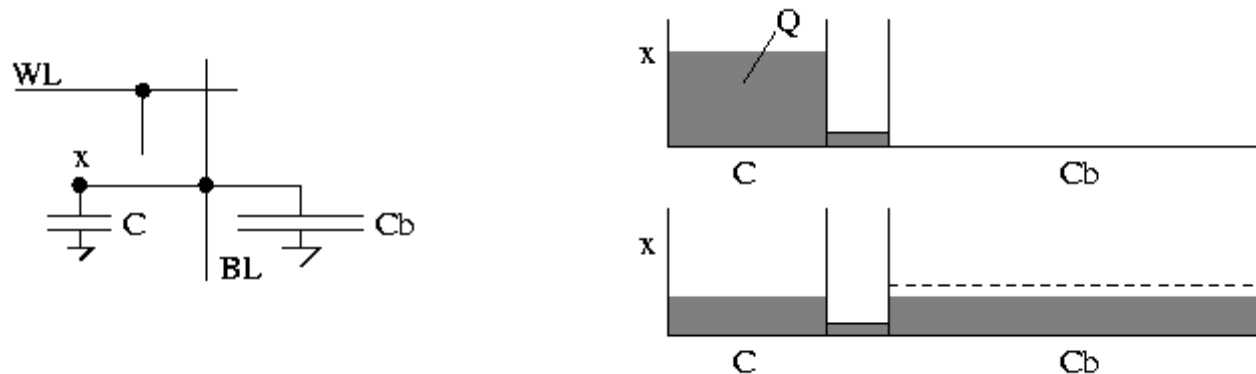
write



- To write a bit (say b) in the cell, the bit line BL is driven to the target value of b while asserting WL
- When the capacitor is connected to a driven line, it is charged or discharged in order to reach the voltage level of the driven line
- The connection of a capacitor C to power supply (ground) can be viewed as the connection of a section- C reservoir to a second reservoir with infinite section filled of fluid up to level V_{dd} (or 0).

Dynamic RAM (DRAM)

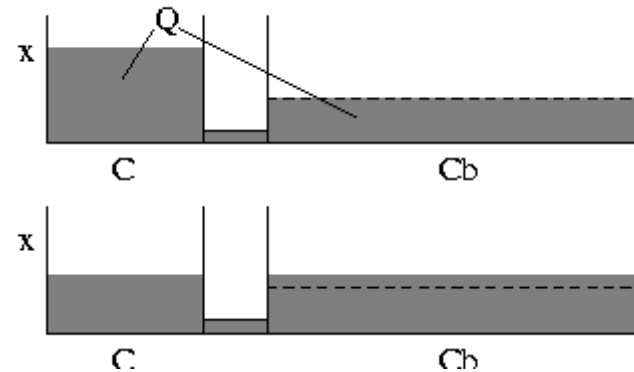
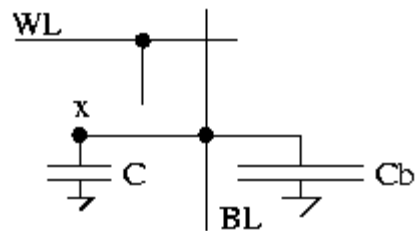
read



- To read the bit stored in the cell, the cell is connected to a floating bit line.
- Notice that the bit line has its own parasitic capacitance C_b , that is usually much larger than C
- When the connection is established, the charge Q stored in the cell (C) redistributes between C and C_b
- If C_b was originally empty, and the value stored in the cell was 1, the result of the charge redistribution process is a voltage level that is below the logic threshold $V_{dd}/2$, and is not recognized as a logic 1

Dynamic RAM (DRAM)

read

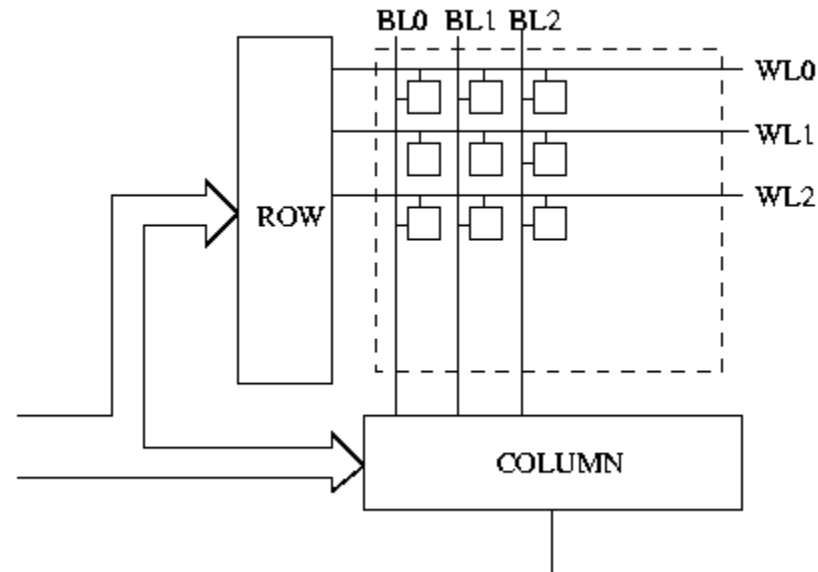


- To solve this problem, the bit line needs to be pre-charged at $V_{dd}/2$, so that the redistribution of the charge stored in the cell causes the final voltage level to be above or below $V_{dd}/2$, according to the value of the stored bit.

SRAM vs DRAM

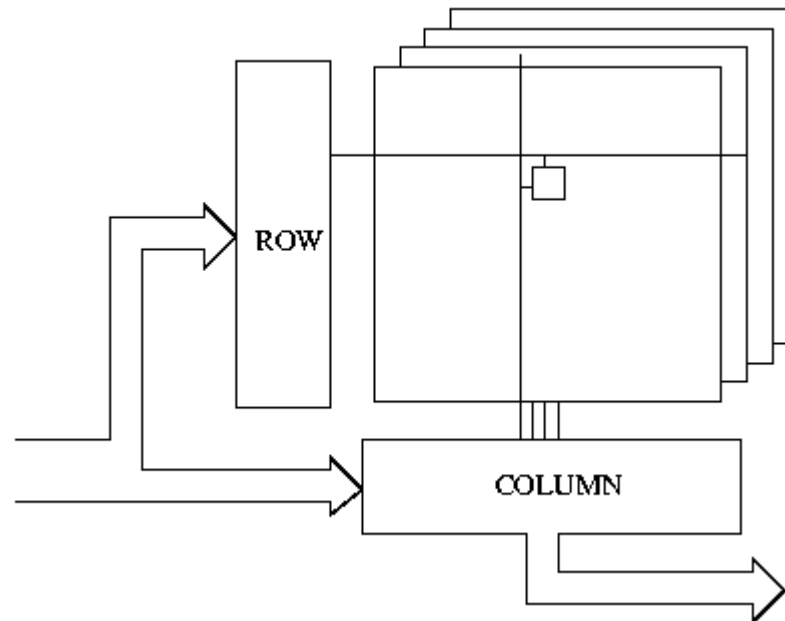
- Performance
 - SRAM devices are faster than DRAM devices since static connections to Vdd and Ground make read operations much faster than those based on charge redistributions
- Density
 - DRAM devices have higher areal density, since the elementary cell is composed of only one transistor and one capacitor, while SRAM cells required 6 transistors
 - Moreover, DRAM technologies make use of *trench* or *stacked* capacitors, that make the area of a DRAM cell close to that of a single transistor
- Cost
 - DRAM devices are much cheaper (in terms of cost per bit) than SRAM devices thanks to their higher density
- Size (capacity)
 - DRAM devices contain more bits thanks to their higher density

Internal architecture



- Let's consider a memory device composed of single-bit memory cells (parallelism 1)
- Memory cells are internally organized on a square matrix
- The square organization grant to the device a suitable shape and minimizes the complexity of the address decoders
- Addresses are split into a row address and a column address
- Row addresses are processed by a row decoder that generates mutually exclusive word lines
- Column addresses are used to select a bit line

Internal architecture



- The organization of a memory device composed of n -bit words (parallelism n) can be derived from the organization of a single-bit memory by viewing the n bits of each word as disposed on n parallel planes
- In practice, the n bits of a word are connected to n parallel bit lines that are all associated with the same column address
- Hence, each column is composed of n bit lines

Read cycle

- A read operation entails:
 - Decoding of the row address
 - Assertion of a word line
 - Activation of all the cells on the selected row
 - Driving of the bit lines
 - Decoding of the column address
 - Selection of a column
 - Driving of the output data bus
- Notice that, when reading a word, all memory cells on the same row are activated and used to drive the corresponding bit lines
- If two words that are stored in the same row are read in sequence, the row address doesn't need to be re-asserted and the first half of the read cycle can be shared. As a consequence, the second word can be read much faster than the first one.
- Most memory devices can work in *page-mode* to enable the fast reading of contiguous cells

External architecture

- Multiple memory banks can be used to realize memory systems with increased capacity and/or parallelism
- Assume that each memory device contain N words of n bits
- 2 memory devices can be used to realize a memory of $2N$ words of n bits
 - They share the first $m = \log_2 N$ bits of the address bus, while the most significant bit of the address is used to select the device
 - They are connected to the same n -bit data BUS
- 2 memory devices can be used to realize a memory of N words of $2n$ bits
 - They share the entire address
 - They are connected to separate wires of a $2n$ data bus