united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

301/1152-8

Microprocessor Laboratory
Sixth Course on Basic VLSI Design Techniques
8 November - 3 December 1999

INTRODUCTION TO VLSI ASIC DESIGN AND TECHNOLOGY
CMOS REGULAR STRUCTURES

Paulo Rodrigues S. MOREIRA
EP/MIC
CERN
1211-Geneva 23
SWITZERLAND

strada costiera, 11 - 34014 trieste italy - tel. +39 040 2240111 fax +39 040 224163 - sci_info@ictp.trieste.it - www.ictp.trieste.it

Outline

- Introduction
- CMOS devices
- CMOS technology
- CMOS logic structures
- CMOS sequential circuits
- CMOS regular structures

CMOS regular structures

- Memory classification
- Write/read cycle
- Memory architecture
- Read-only memories
- Nonvolatile read-write memories
- Read-write memories
- Sense amplifiers

Memory classification

- Memory: logic element where data can be stored to be retrieved at a later time
- Read-Only Memory (ROM)
  - The information is encoded in the circuit topology
  - The data cannot be modified: it can only be read
  - ROM's are not volatile. That is, removing the power source does not erase the information contents of the memory.

Memory classification

- Read Write Memories (RWM)
  - RWM's allow both reading and writing operations
  - RWM can be of two general types:
    - Static: the data is stored in flip-flops
    - Dynamic: the data is stored as charge in a capacitor
  - Both types of memories are volatile, that is, data is lost once the power is turned off
  - Dynamic memories require periodic "refresh" of its contents in order to compensate for the charge loss caused by leakage currents in the memory element

Memory classification

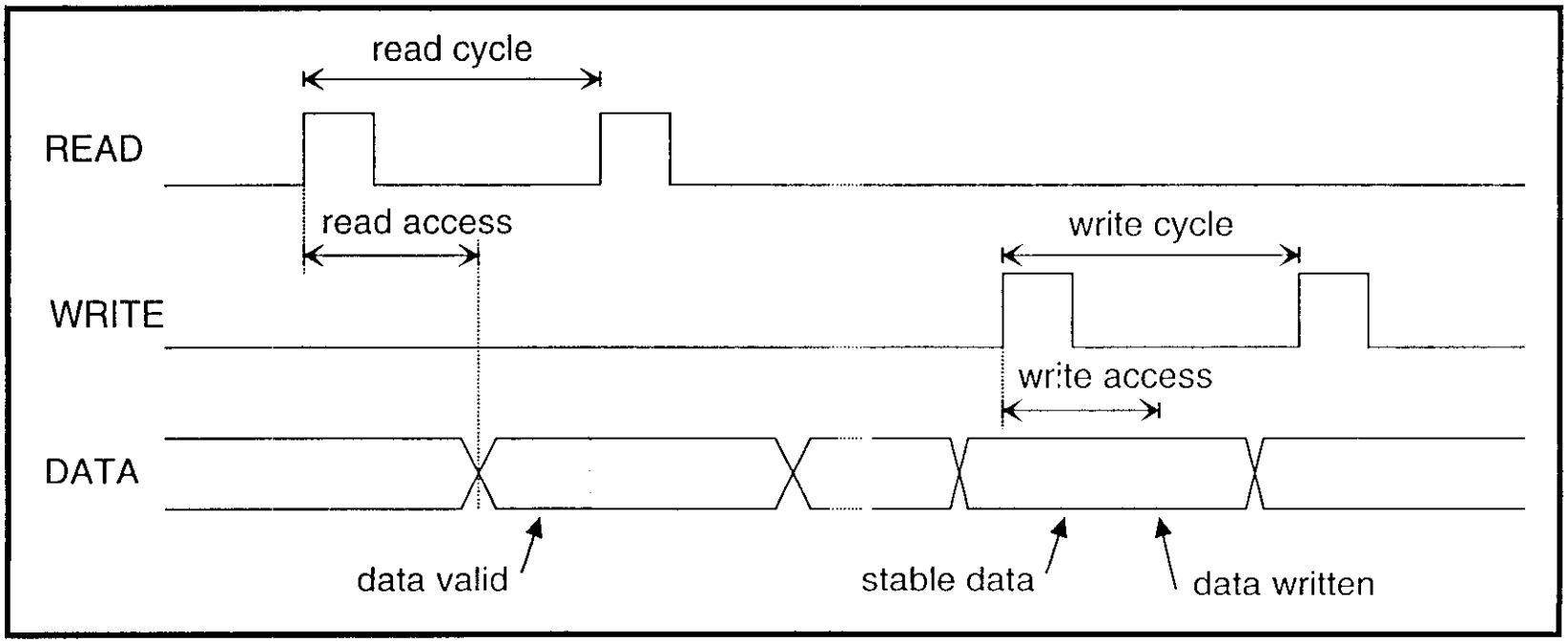
- Nonvolatile Read-Write Memories (NVRWM)
  - These are non volatile memories that allow write operations
  - However:
    - The write operation takes substantially more time than the read operation
    - For some types of NVRWM's, the write operation requires special lab equipment
  - Examples of such memories are:
    - EPROM (Erasable Programmable Read-Only memory)
    - E^2PROM (Electrically Erasable Programmable Read-Only Memory)

Memory classification

- Memories can also be classified according to the way they allow access to the stored data:
  - Random Access: memory locations can be read or written in a random order
  - First-In First-Out (FIFO): The first word to be written is the first word to be read
  - Last-In First-Out (LIFO): The last word to be written is the first word to be read (stack)
  - Shift Register: information is streamed in and out. It can work either as a FIFO or as a LIFO

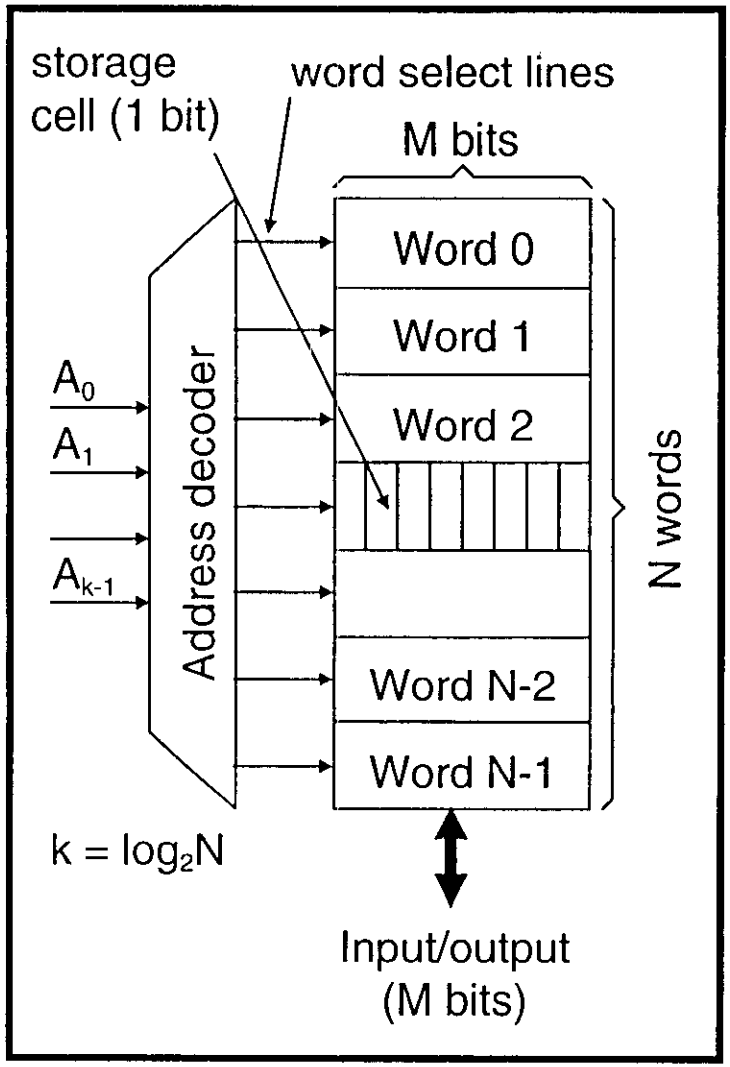
Write/read cycle

- Read-access time: delay between read request and data valid
- Write-access time: delay between write request and the actual writing
- Read or write cycle time: minimum time required between successive read or write operations



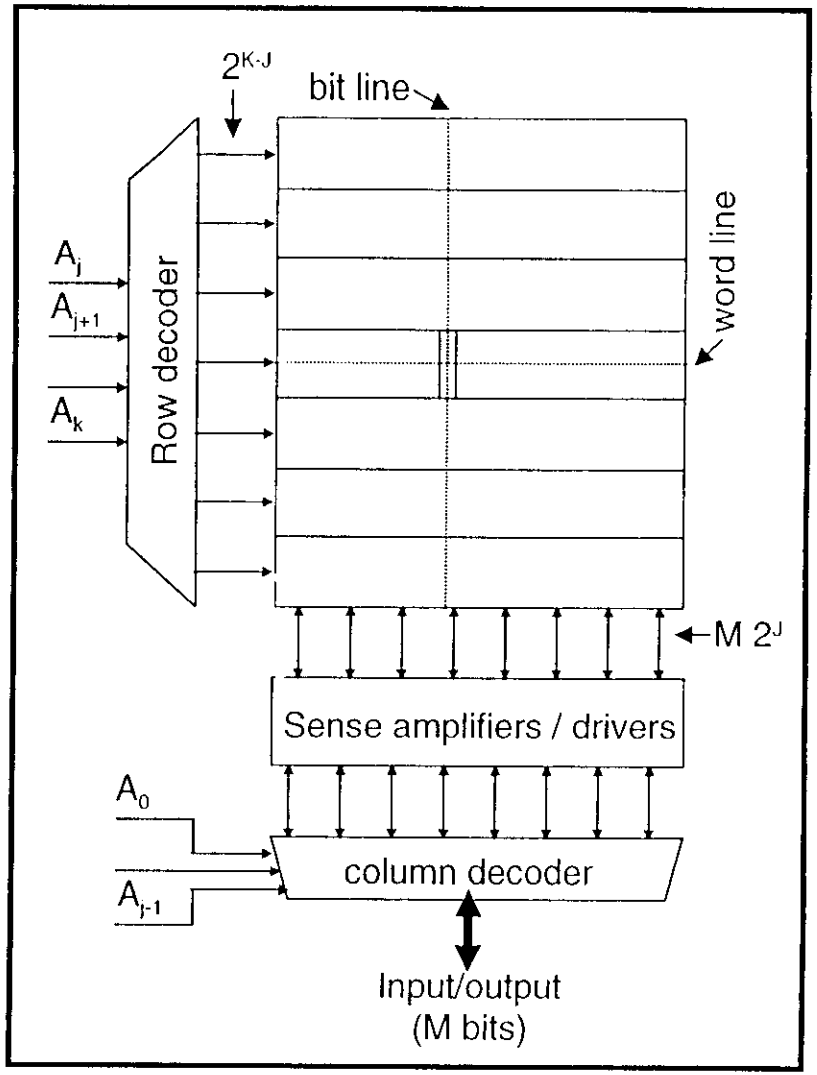
Memory architecture

- The memory is organized in N words, each of M bits wide
- One word at a time is selected for read/write using a select signal
- A decoder is used to convert a binary encoded address into a single active word select line
- This structure is not practical, it results in very big aspect ratios



Memory architecture

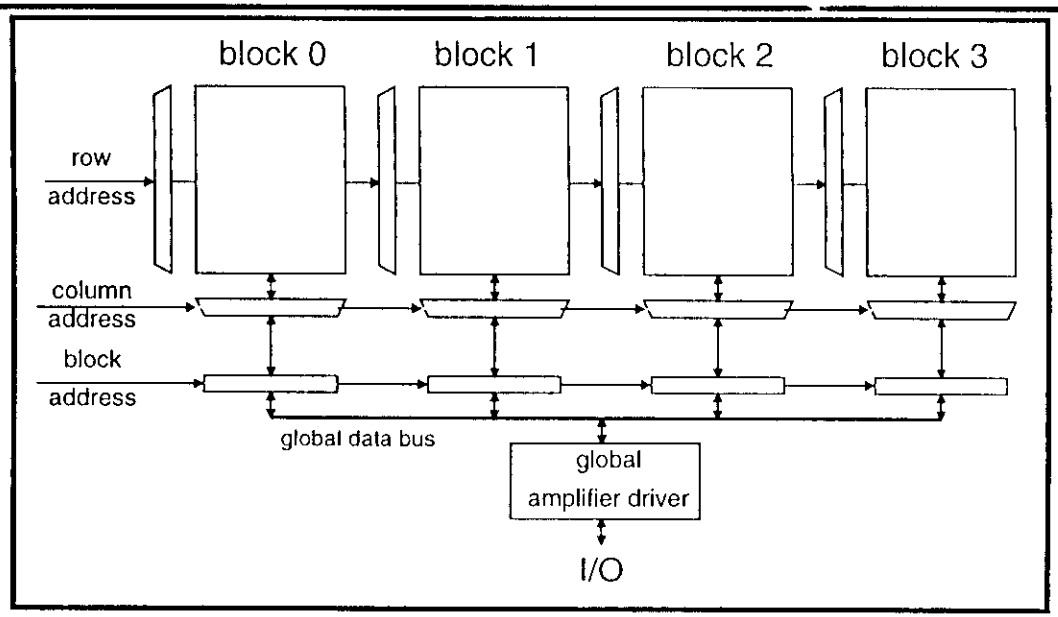
- Memories are organized to be almost square in layout:
  - Multiple words are stored in the same row and selected simultaneously
  - The correct word is then selected by the column decoder
  - The word address is split in two fields:
    - row address: enables one row for R/W
    - column address: selects a word within a row
  - Even this structure is impractical for memories bigger than 256Kbits



Memory architecture

- The silicon area of large memory cells is dominated by the size of the memory core, it is thus crucial to keep the size of the basic storage cell as small as possible
- The storage cell area is reduced by:
  - reducing the driving capability of the cell (small devices)
  - reducing the logic swing and the noise margins
- Consequently, sense amplifiers are used to restore full rail-to-rail amplitude

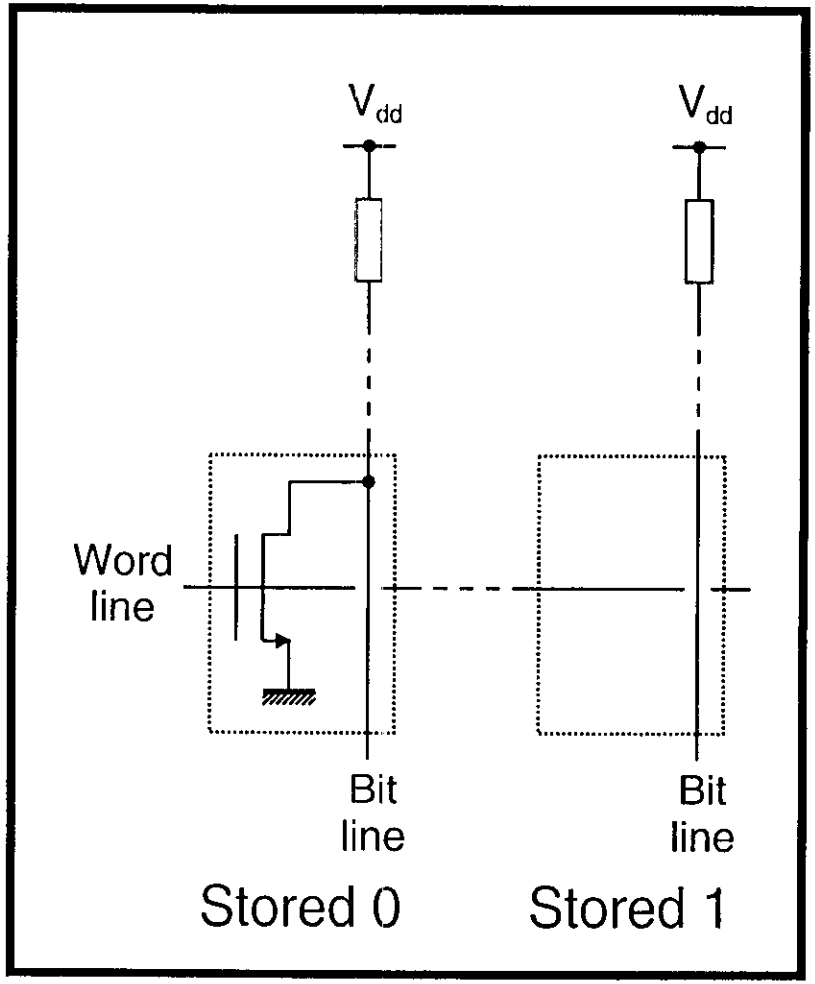
Memory architecture



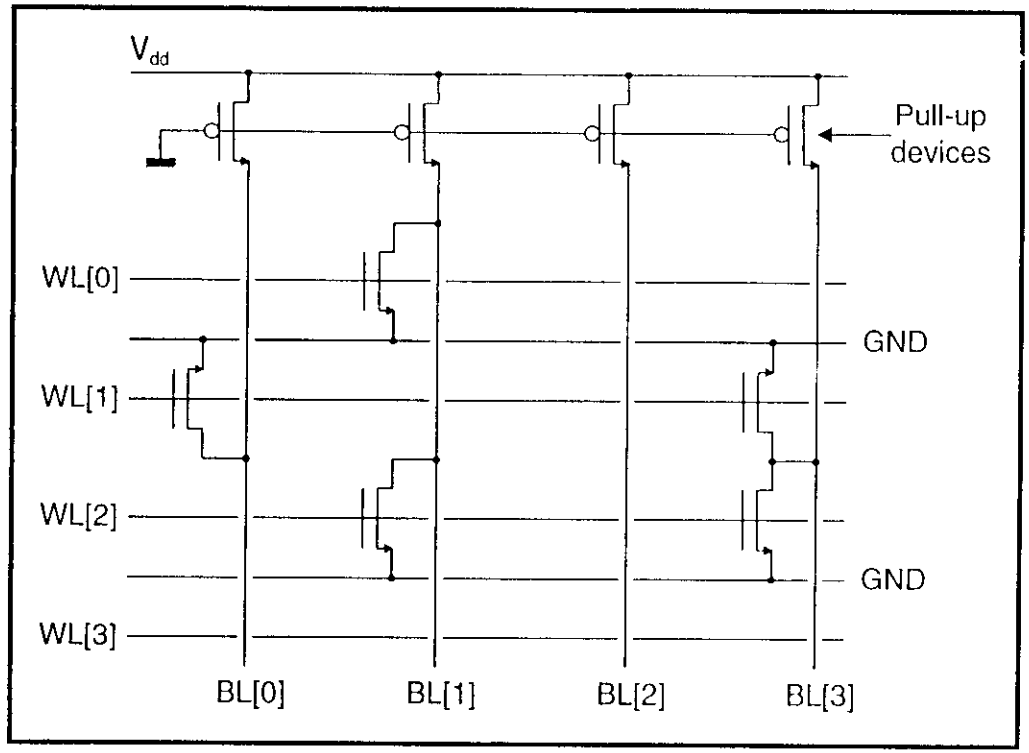
- Large memories start to suffer from speed degradation due to wire resistance and capacitive loading of the bit and word lines
- The solution is to split the memory into “small” memory blocks
- That allows to:
  - use small local word and bit lines ⇒ faster access time
  - power down sense amplifiers and disable decoders of non-active memory blocks ⇒ power saving

Read-only memories

- Because the contents is permanently fixed the cell design is simplified
- Upon activation of the word line a 0 or 1 is presented to the bit line:
  - If the NMOS is absent the word line has no influence on the bit line:
    - The word line is pulled-up by the resistor
    - A 1 is stored in the “cell
  - If the NMOS is present the word line activates the NMOS:
    - The word line is pulled-down by the NMOS
    - A 0 is stored in the cell
- The NMOS isolates the bit from the word line



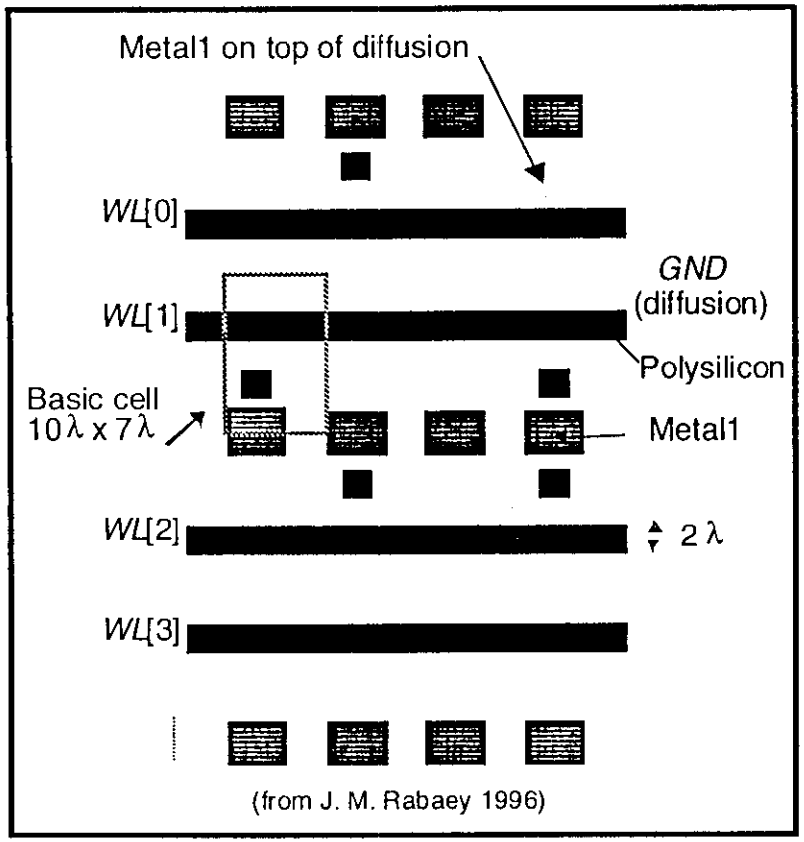
Read-only memories



- A ground contact has to be provided for every cell
  - a ground rail has to be routed through the cell
  - the area penalty can be shared between two neighbor cells:
    - the odd rows are mirrored around the horizontal axis

Read-only memories

- Use close to minimum size pull-down devices to:
  - make the cell size small
  - reduce the bit line capacitance
- R(pull-up) > R(pull-down) to:
  - ensure adequate low level
- Since for large memories the bit line capacitance can be of the order of pF's, low to high transitions will be slow
- A wider pull-up device can be used resulting in a higher V_{OL}
  - this reduces the noise margin but speeds the low-to-high transition
  - to interface with external logic, a sense amplifier is required to restore the logic levels
  - an inverter with adjusted switching threshold can be used as a sense amplifier

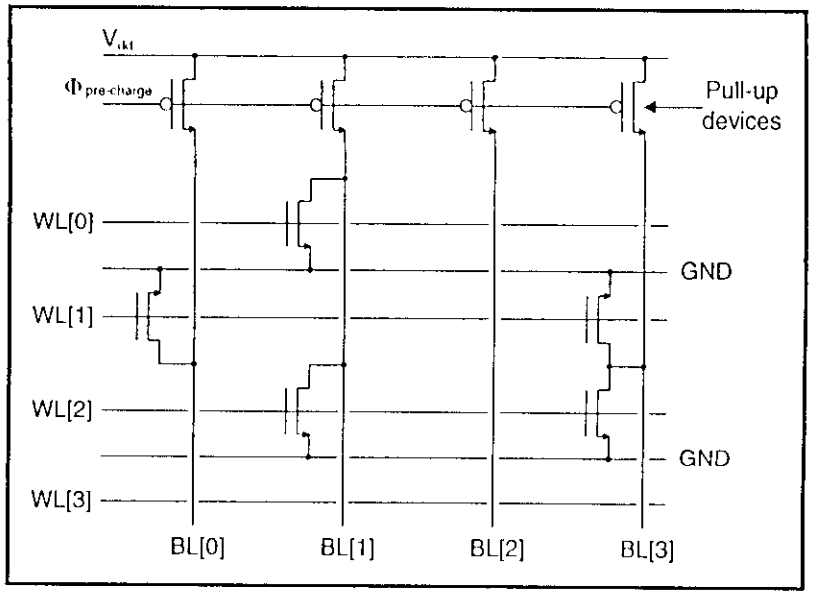


(from J. M. Rabaey 1996)

- 0 ⇒ metal-to-diffusion contact
- 1 ⇒ no metal-to-diffusion contact
- only the contact mask layer is used to program the memory array

Read-only memories

- Disadvantages:
  - V_{OL} depends on the ratio of the pull-up/pull-down devices
  - A static current path exists when the output is low causing high power dissipation in large memories
- Solution:
  - Use pre-charged logic
  - Eliminates the static dissipation
  - Pull-up devices can be made wider
  - This is the most commonly used structure



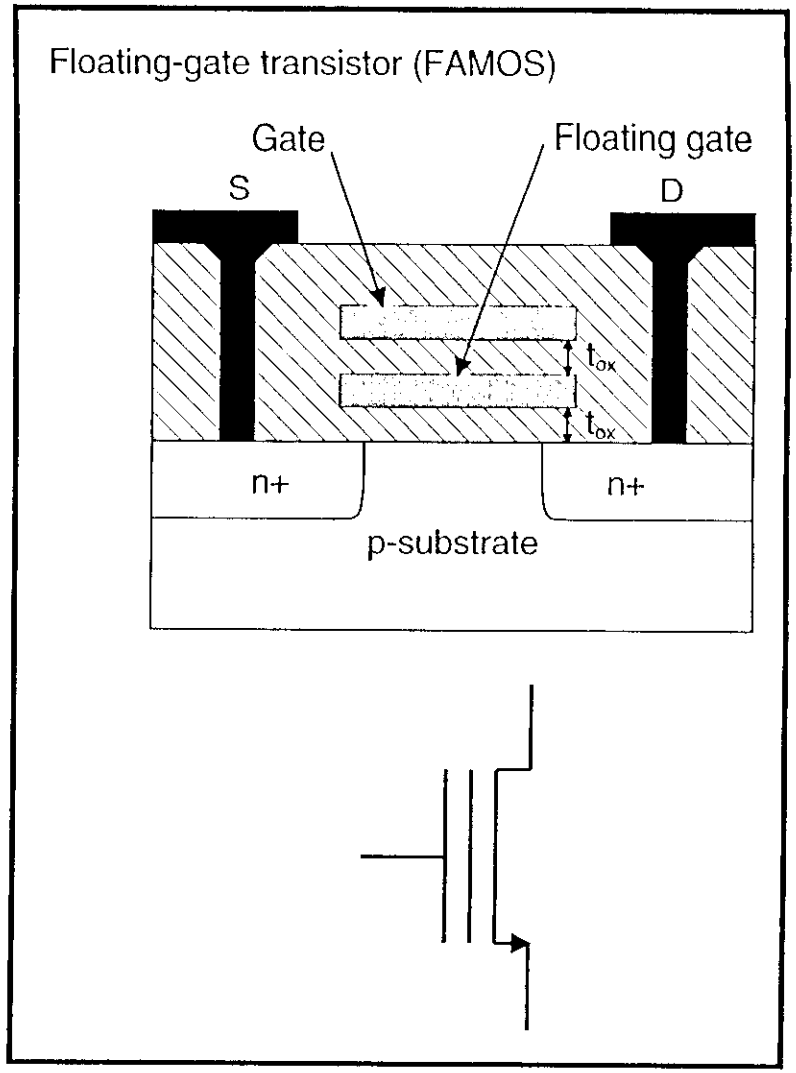
- The bit lines are first pre-charged by the pull-up devices
  - during this phase the word lines must be disabled
- Then, the word lines are activated (word evaluation)
  - during this phase the pull-up devices are off

Nonvolatile read-write memories

- The same architecture as a ROM memory
- The pull-down device is modified to allow control of the threshold voltage
- The modified threshold is retained "indefinitely":
  - The memory is nonvolatile
- To reprogram the memory the programmed values must be erased first
- The "hart" of NVRW memories is the Floating Gate Transistor (FAMOS)

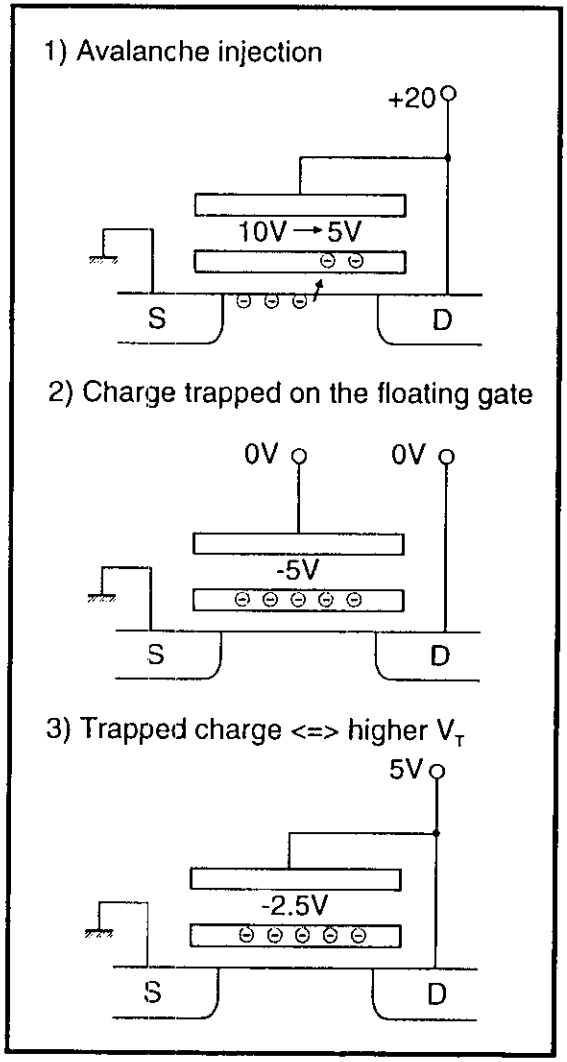
Nonvolatile read-write memories

- A floating gate is inserted between the gate and the channel
- The device acts as a normal transistor
- However, its threshold voltage is programmable
- Since the t_{ox} is doubled, the transconductance is reduced to half and the threshold voltage increased



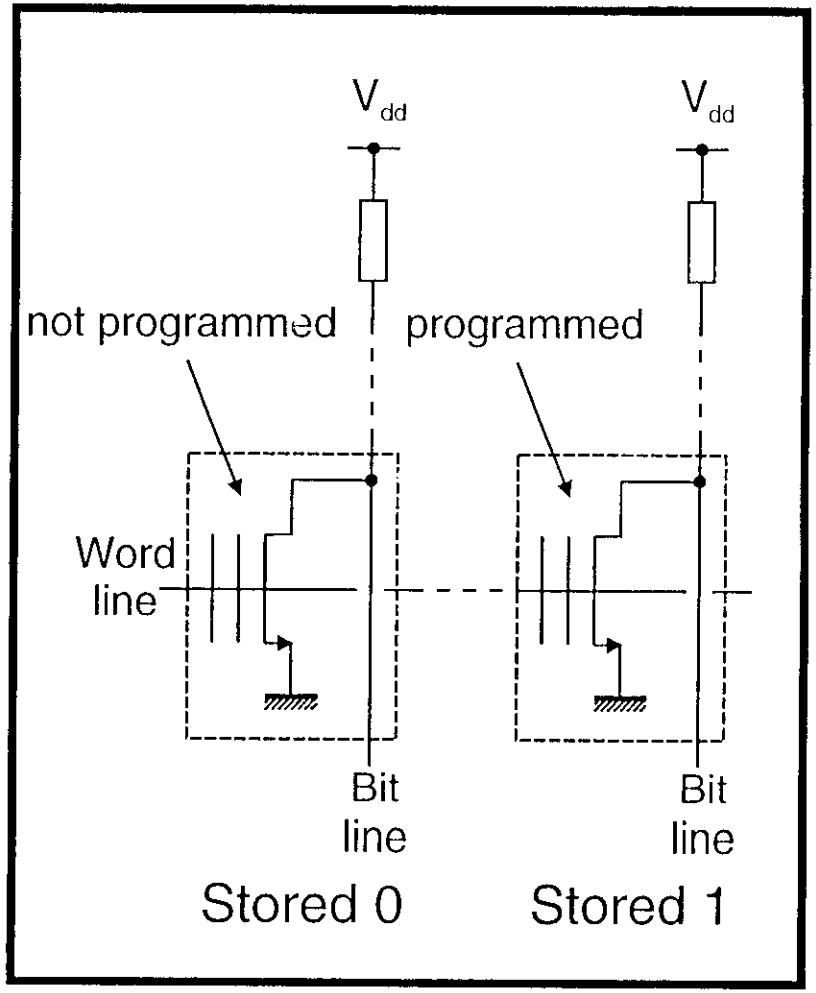
Nonvolatile read-write memories

- Programming the FAMOS:
  - A high voltage is applied between the source and the gate-drain
  - A high field is created that causes avalanche injection to occur
  - Electrons traverse the first oxide and get trapped on the floating gate (t_{ox} = 100nm)
  - Trapped electrons effectively drop the floating gate voltage
  - The process is self limiting: the building up of gate charge eventually stops avalanche injection
  - The FAMOS with a charged gate is equivalent to a higher V_T device
  - Normal circuit voltages can not turn a programmed device on



Nonvolatile read-write memories

- The non-programmed device can be turned on by the word line thus, it stores a “0”
- The word line high voltage can not turn on the programmed device thus, it stores a “1”
- Since the floating gate is surrounded by SiO_2, the charge can be stored for many years

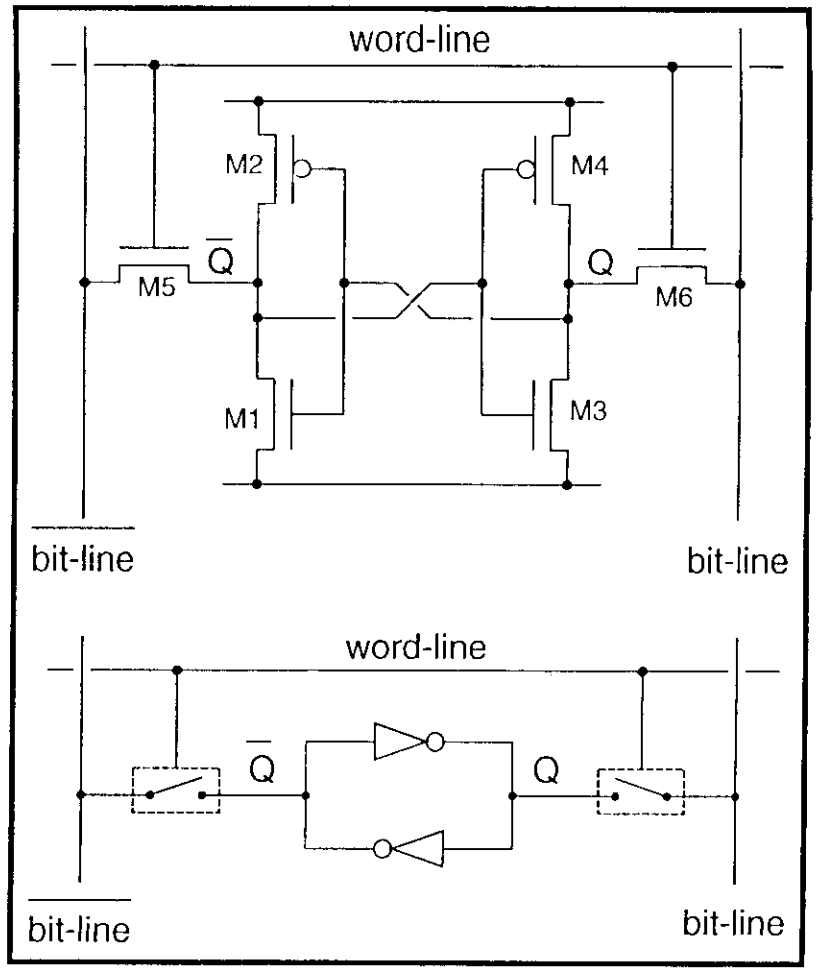


Nonvolatile read-write memories

- Erasing the memory contents (EPROM):
  - Strong UV light is used to erase the memory:
    - UV light renders the oxide slightly conductive by direct generation of electron-hole pairs in the SiO_2
  - The erasure process is slow (several minutes)
  - Programming takes 5-10μs/word
  - Number of erase/program cycles limited (<1000)
- Electrically-Erasable PROM (E^2PROM)
  - A reversible tunneling mechanism allows E^2PROM's to be both electrically programmed and erased

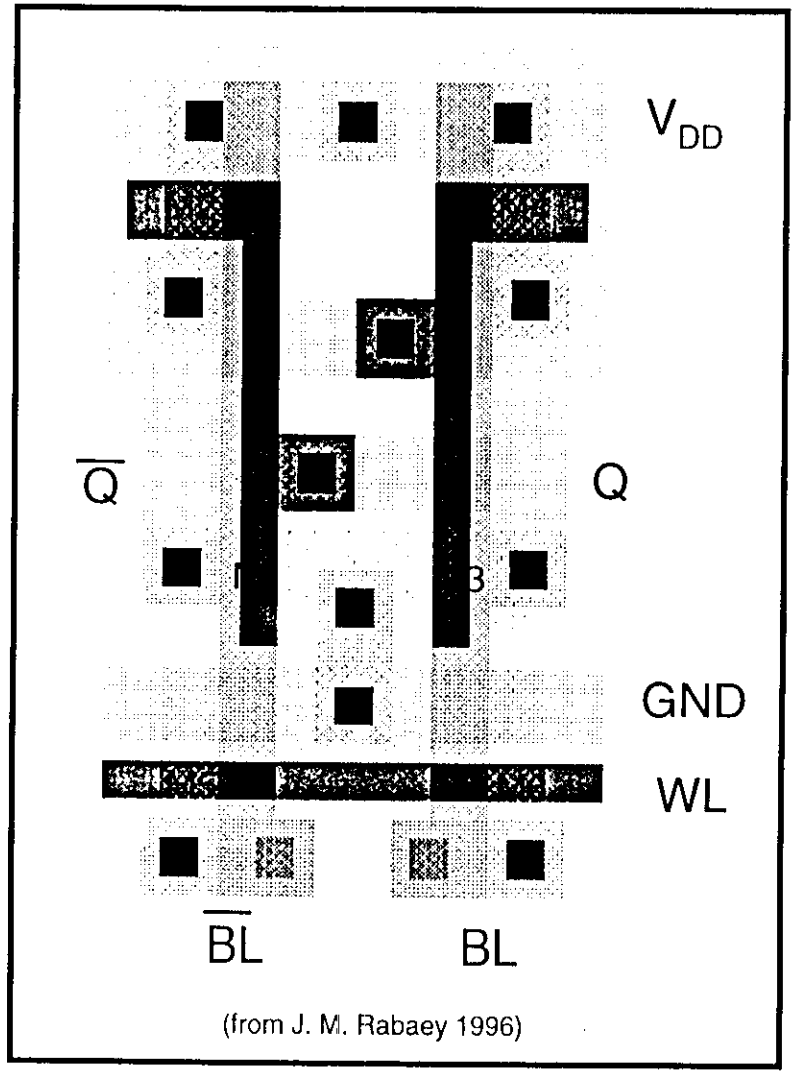
Read-write memories

- Static Read-Write Memories (SRAM):
  - data is stored by positive feedback
  - the memory is volatile
- The cell use six transistors
- Read/write access is enabled by the word-line
- Two bit lines are used to improve the noise margin during the read/write operation
- During read the bit-lines are pre-charged to $V_{dd}/2$:
  - to speedup the read operation
  - to avoid erroneous toggling of the cell



Read-write memories

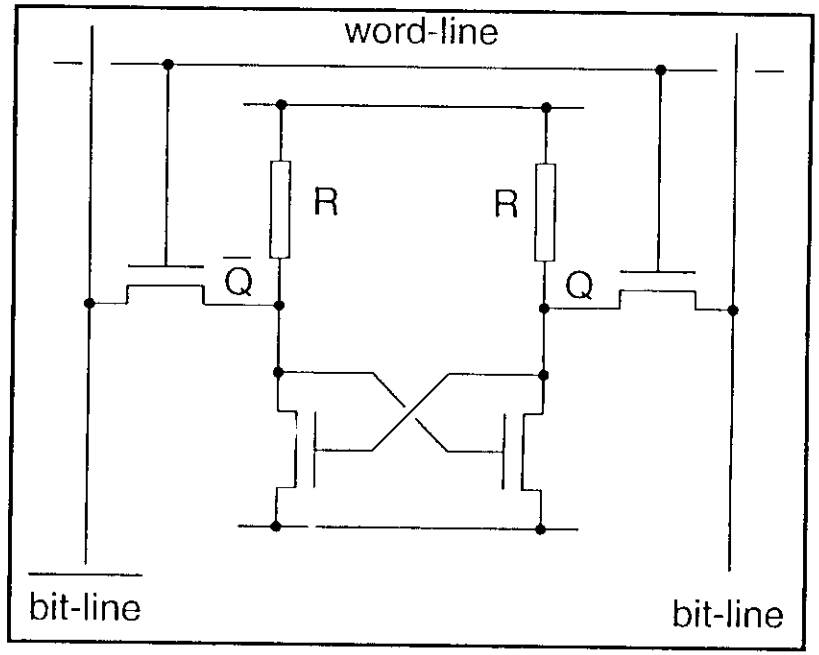
- SRAM performance:
  - The read operation is the critical one:
    - It involves discharging or charging the large bit-line capacitance through the small transistors of the cell
  - The write time is dominated by the propagation delay of the cross-coupled inverter pair
  - The six-transistor cell is not area efficient:
    - It requires routing of two power lines, two bit lines and a word line
    - Most of the area is taken by wiring and interlayer contacts



(from J. M. Rabaey 1996)

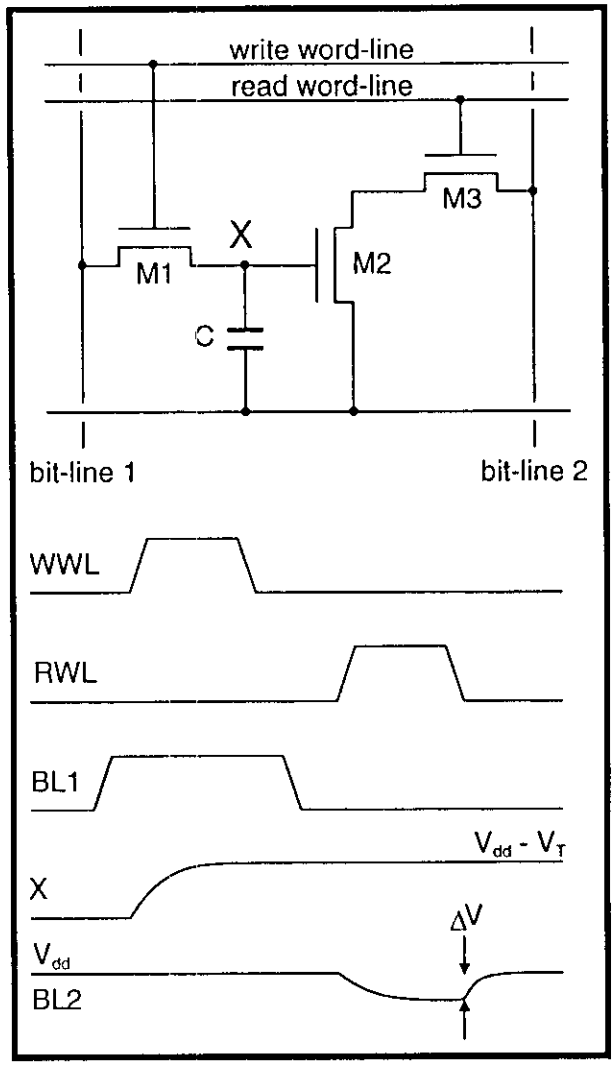
Read-write memories

- Resistive-load SRAM
  - employs resistors instead of PMOS's
  - The role of the resistors is only to maintain the state of the cell:
    - they compensate for leakage currents (10^{-15A})
    - they must be made as high as possible to minimize static power dissipation
    - undoped polysilicon $10^{12}\Omega/$
  - The bit-lines are pre-charged to V_{dd}:
    - the low-to-high transition occurs during precharge
    - the loads contribute "no" current during the transitions
  - The transistor sizes must be correctly chosen to avoid toggling the cell during read



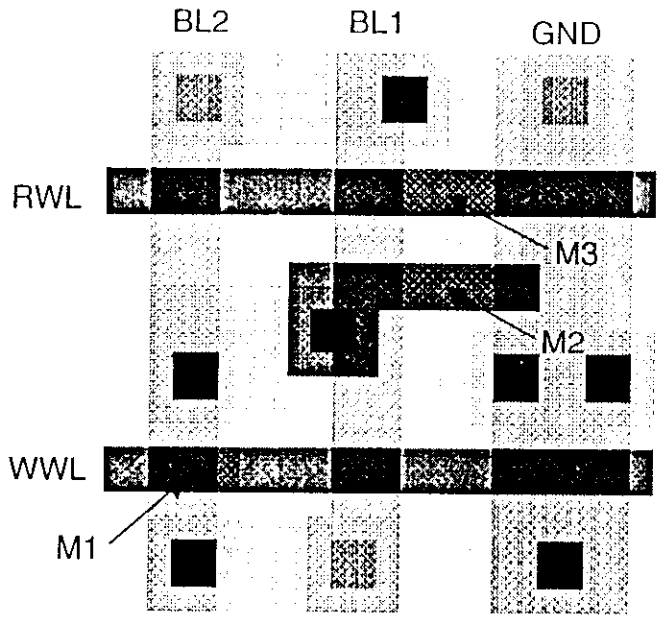
Read-write memories

- Dynamic Random-Access Memory (DRAM)
  - In a dynamic memory the data is stored as charge in a capacitor
- Tree-Transistor Cell (3T DRAM):
  - Write operation:
    - Set the data value in bit-line 1
    - Assert the write word-line
    - Once the WWL is lowered the data is stored as charge in C
  - Read operation:
    - The bit-line BL2 is pre-charged to Vdd
    - Assert the read word-line
    - if a 1 is stored in C, M2 and M3 pull the bit-line 2 low
    - if a 0 is stored C, the bit-line 2 is left unchanged



Read-write memories

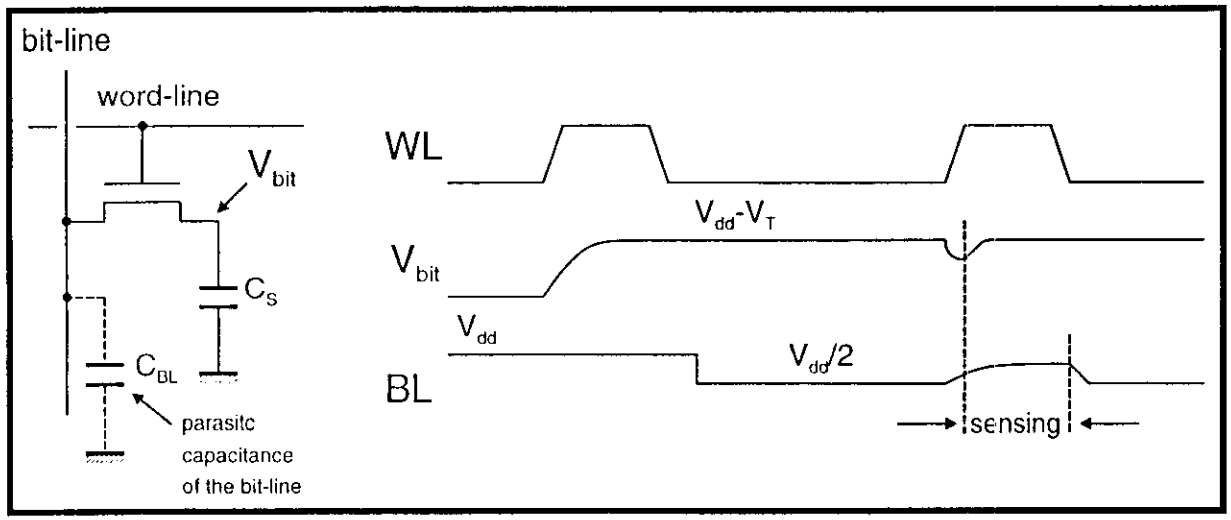
- The cell is inverting
- Due to leakage currents the cell needs to be periodically refreshed (every 1 to 4ms)
- Refresh operation:
  - read the stored data
  - put its complement in BL1
  - enable/disable the WWL
- Compared with an SRAM the area is greatly reduce:
  - SRAM $\Rightarrow$ 1092 λ^2
  - DRAM $\Rightarrow$ 576 λ^2
  - The are reduction is mainly due to the reduction of the number of devices and interlayer contacts



(from J. M. Rabaey 1996)

Read-write memories

- One-Transistor dynamic cell (1T DRAM)
  - It uses a single transistor and a capacitor
  - It is the most widely used topology in commercial DRAM's
- Write operation:
  - Data is placed on the bit-line
  - The word-line is asserted
  - Depending on the data value the capacitance is charged or discharged

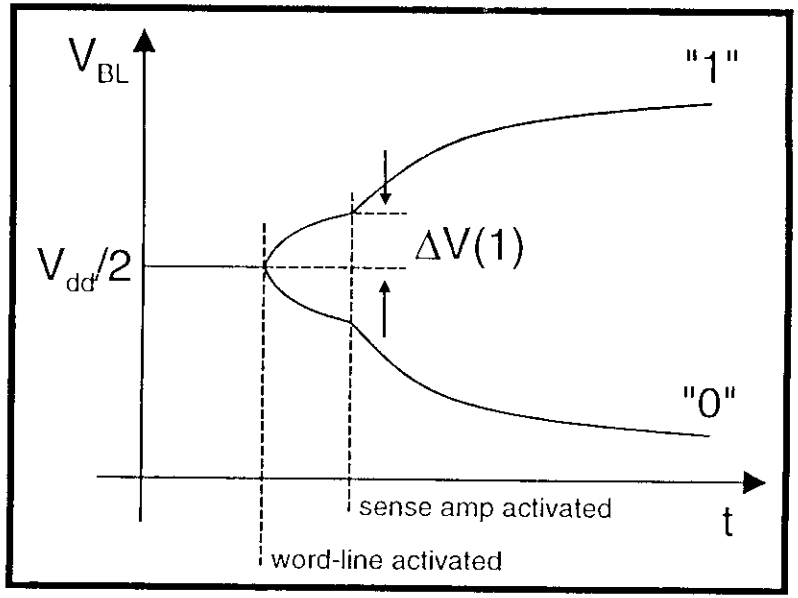


Read-write memories

- Read operation:
  - The bit-line is pre-charged to $V_{dd}/2$
  - The word-line is activated and charge redistribution takes place between C_S and the bit-line
  - This gives origin to a voltage change in the bit-line, the sign of which determines the data stored:

$$\Delta V = \left(V_{BIT} - \frac{V_{dd}}{2}\right)\frac{C_S}{C_S + C_{BL}}$$

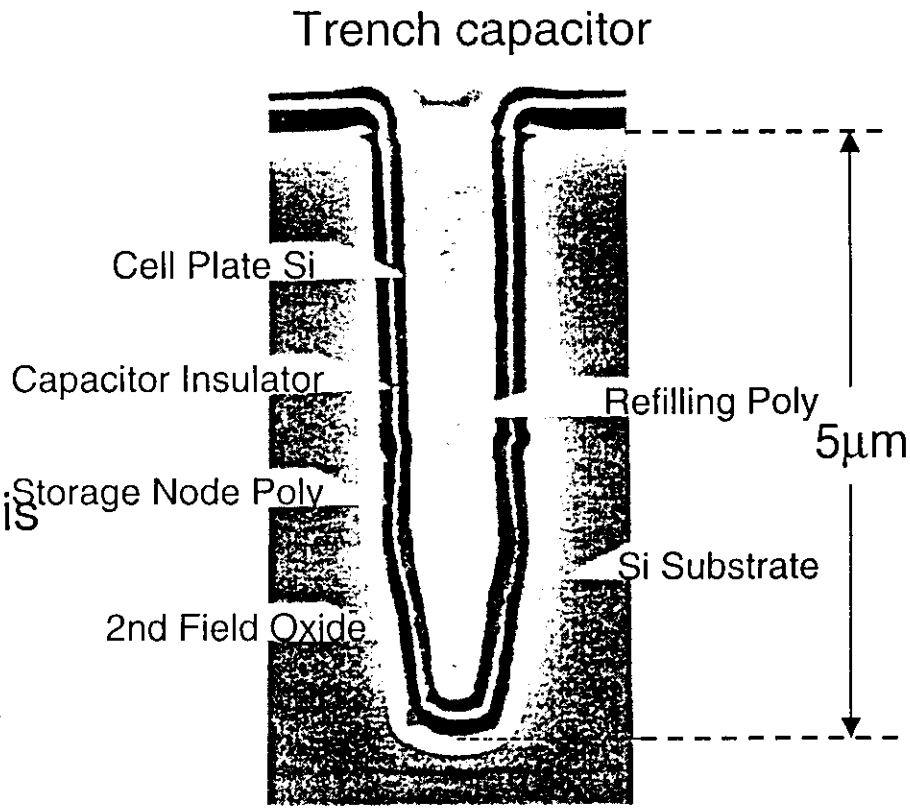
  - C_{BL} is 10 to 100 times bigger than $C_S \Rightarrow$ $\Delta V \cong 250mV$



  - The amount of charge stored in the cell is modified during the read operation
  - However, during read, the output of the sense amplifier is imposed on the bit line restoring the stored charge

Read-write memories

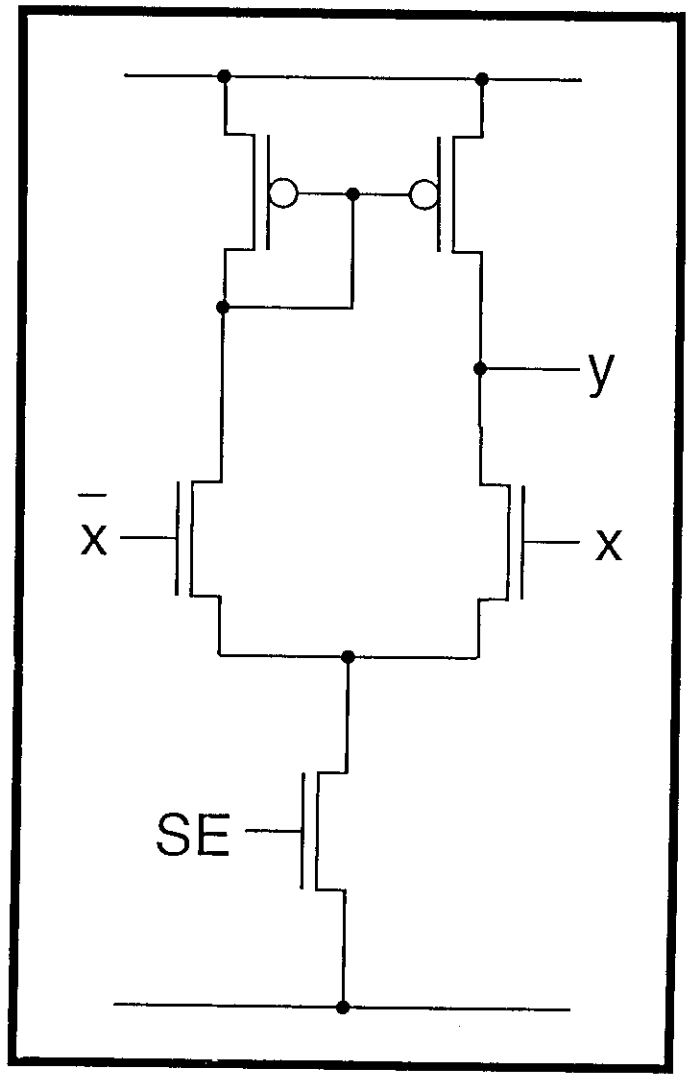
- Contrary to the previous cases a 1T cell requires a sense amplifier for correct operation
- Also, a relatively large storage capacitance is necessary for reliable operation
- A 1 is stored as V_{dd}-V_T. This reduces the available charge:
  - To avoid this problem the word-line can be bootstrapped to a value higher than Vdd



(from T. Mano et al., 1987)

Sense amplifiers

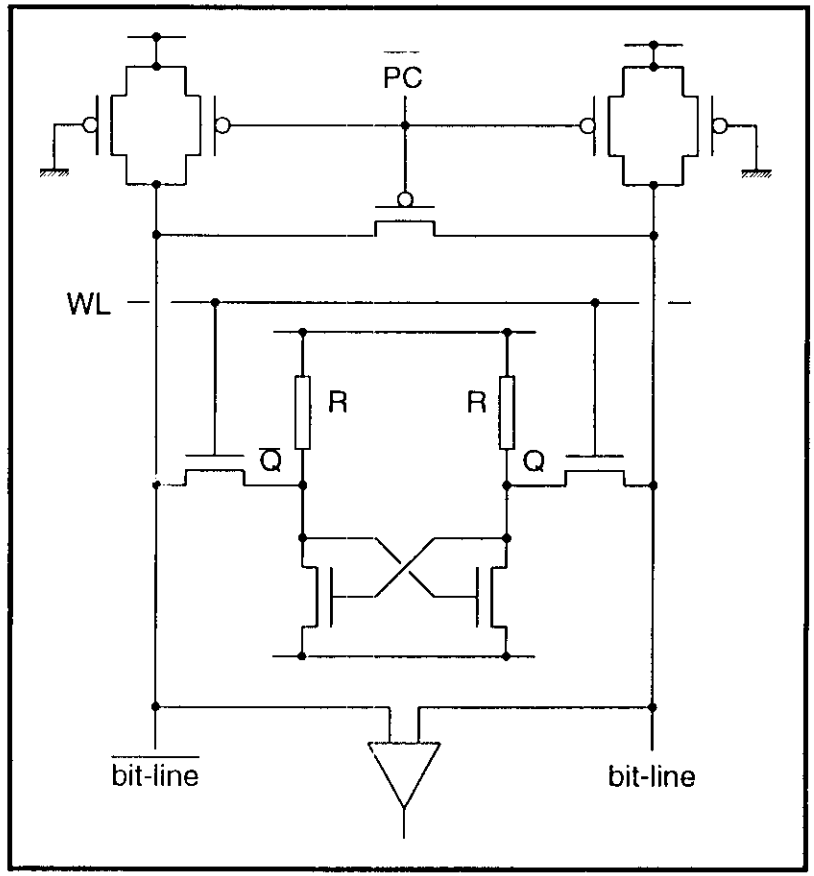
- Sense amplifiers improve the speed performance of the memory cell:
  - they compensate for the low driving capability of the cells
- Contribute to power reduction by allowing to use low signal swings on the heavily capacitive bit-lines
- They perform signal restoration in the refresh and read cycles of 1T dynamic memories
- They can be differential or single ended



Sense amplifiers

SRAM read cycle:

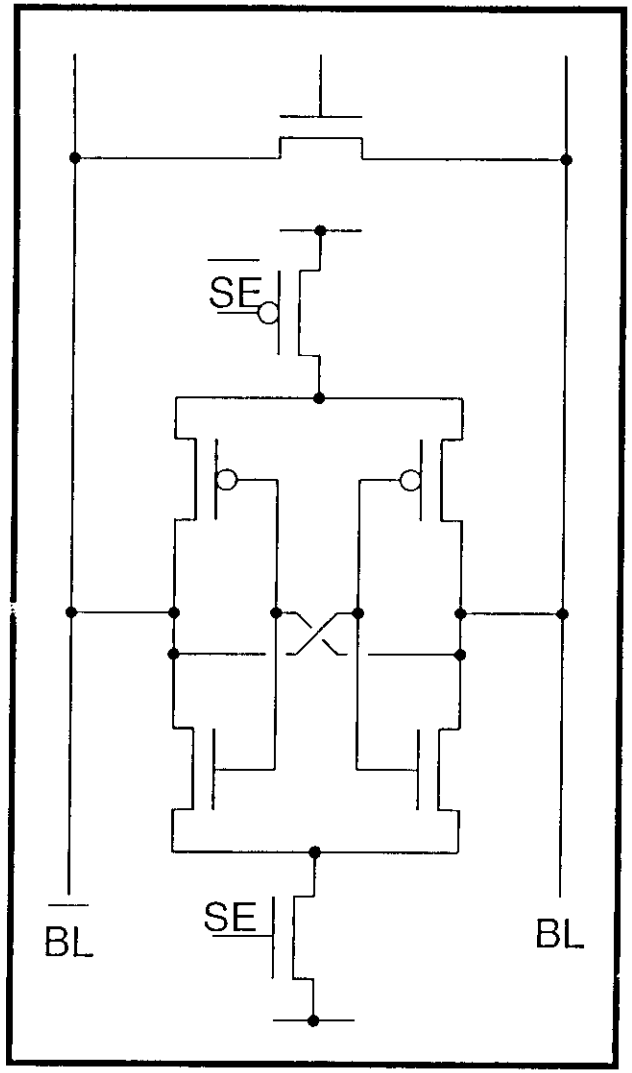
- pre-charge:
  - pre-charge the bit-lines to V_{dd} and make their voltages equal
- Reading:
  - disable the pre-charge devices
  - enable the word lines
  - once a minimum (≅0.5V) signal is built up in the bit-lines the sense amplifier is turned on
- The grounded PMOS loads limit the signal swing and facilitate the next pre-charge



Sense amplifiers

A cross-coupled inverter pair can be used as a sense amplifier

- To act as a sense amplifier:
  - The bit-lines are equalized: this initializes the flip-flop in its metastable point
  - A voltage is built over the bit lines by the selected cell
  - The sense amplifier is activated once the voltage is large enough
  - The cross-coupled pair then toggles to one of its stable operating points
  - The transition is fast due to positive feedback
- Ideal for an 1T DRAM: inputs and outputs are merged



Sense amplifiers

- The memory array is divided in two: the sense amplifier in the middle
- On each side "dummy" cells are added
- These cells serve as a reference during the reading
- EQ is asserted and both halves pre-charged to $V_{dd}/2$
- The dummy cells are also pre-charged to $V_{dd}/2$
- If a cell in one of the halves of the bit line is selected, the dummy cell on the other half is used as a reference for the sense amplifier