Budapest University of Technology and Economics
Department of Electron Devices

Technology of IT Devices

Lecture 7

Memory (Part 1)

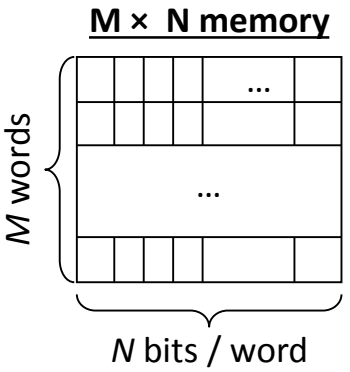
RAM memory

Memory

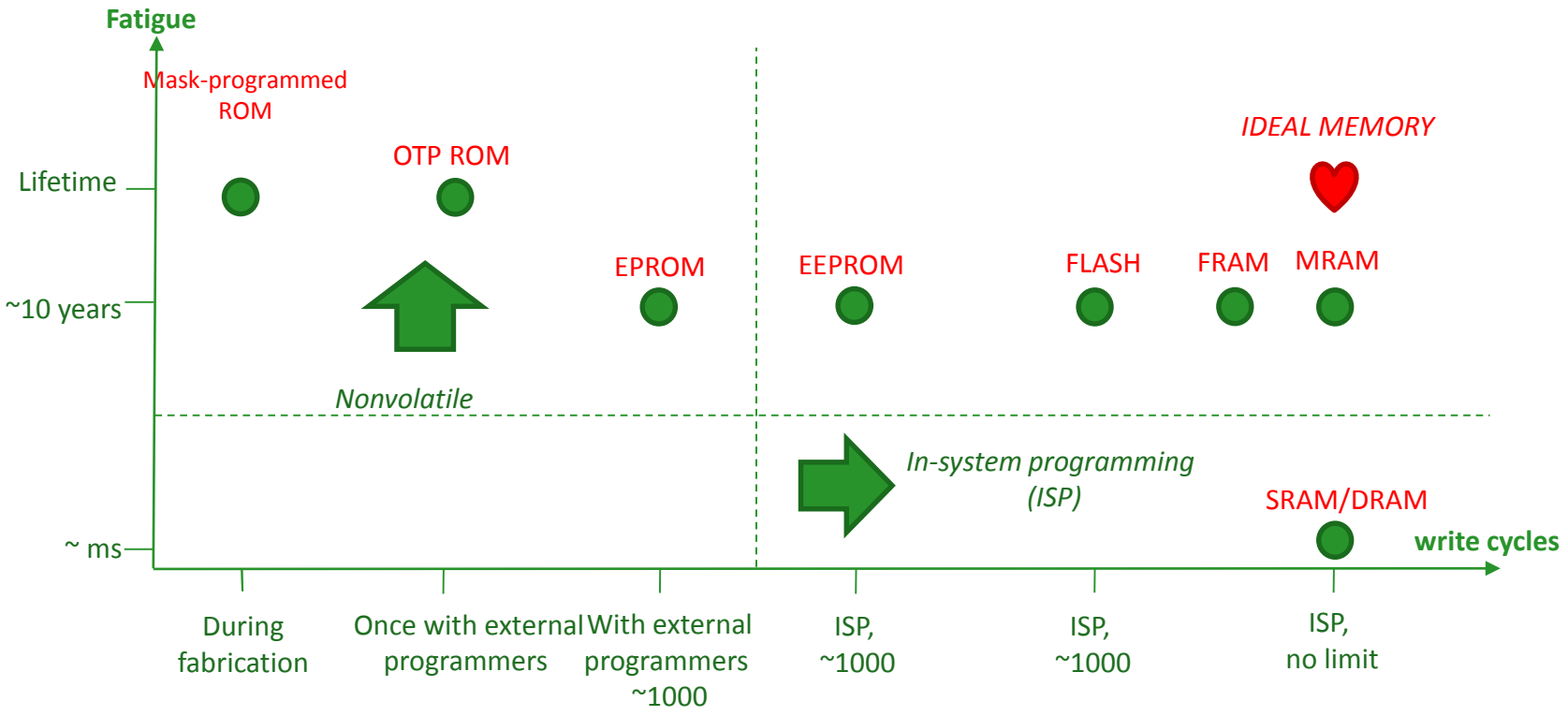
- Overview
- MOS transistor – recap and its operation in more detail
- Random Access Memory (RAM)
  - Static RAM
  - Dynamic RAM
- Content Addressable Memory (CAM cells)
- Novel memory architectures
- Examples

Semiconductor memory – basic term

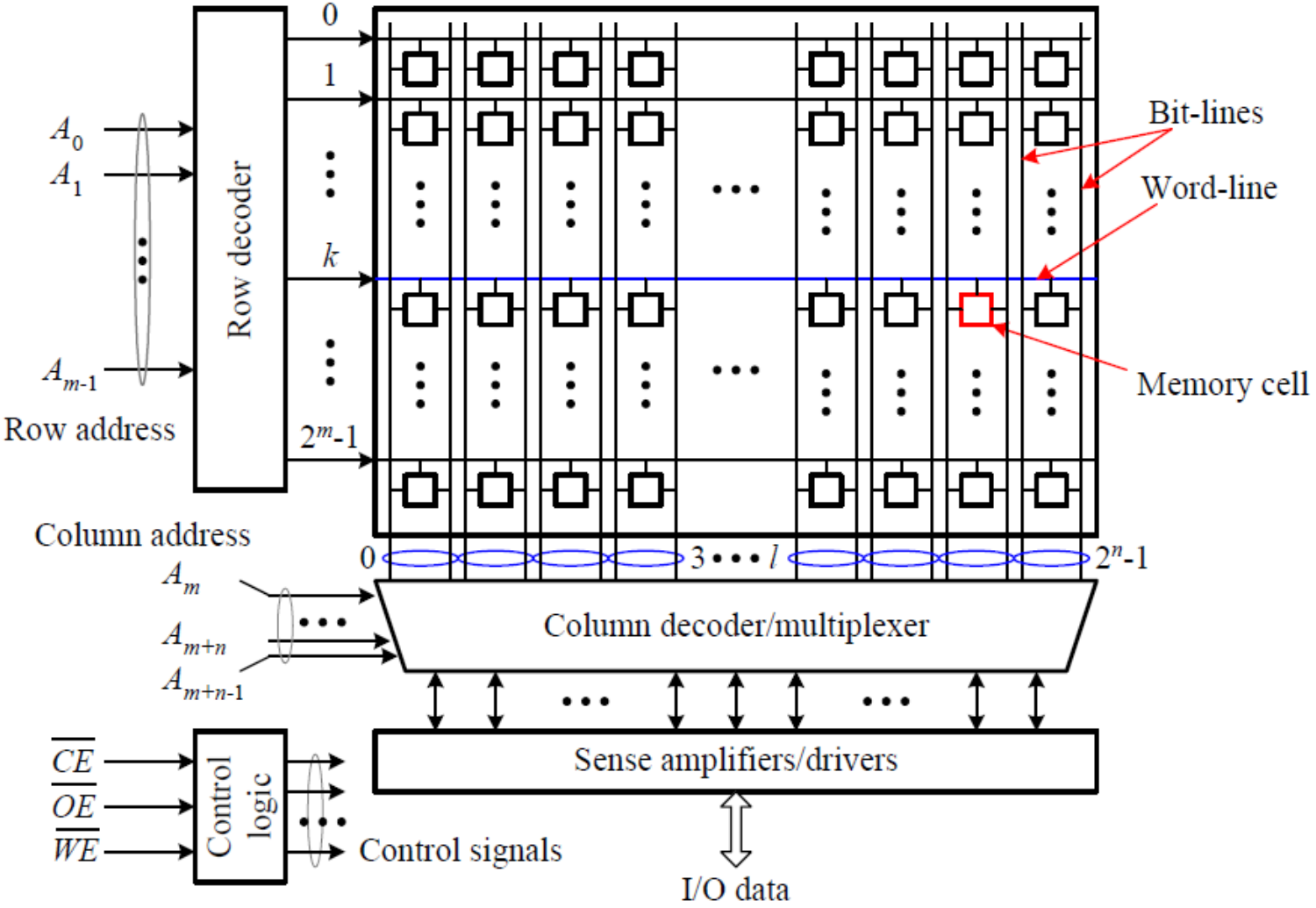
- M × N memory
  - M times N bit wide memory words. M is a power of two, N is a multiple of 8.
  - E.g.: 64k × 16 means a capacity of 1Mbit
- (Traditional) Types of Memory
  - ROM vs. RAM, read-only vs. Read/write memory
  - The names are confusing
    - EEPROMs are read/write memory, the information in NVRAM is preserved after the power supply voltage is switched off.
- Better classification:
  - Read-only vs. Read/write
  - Data storing time interval



Semiconductor Memory Classification



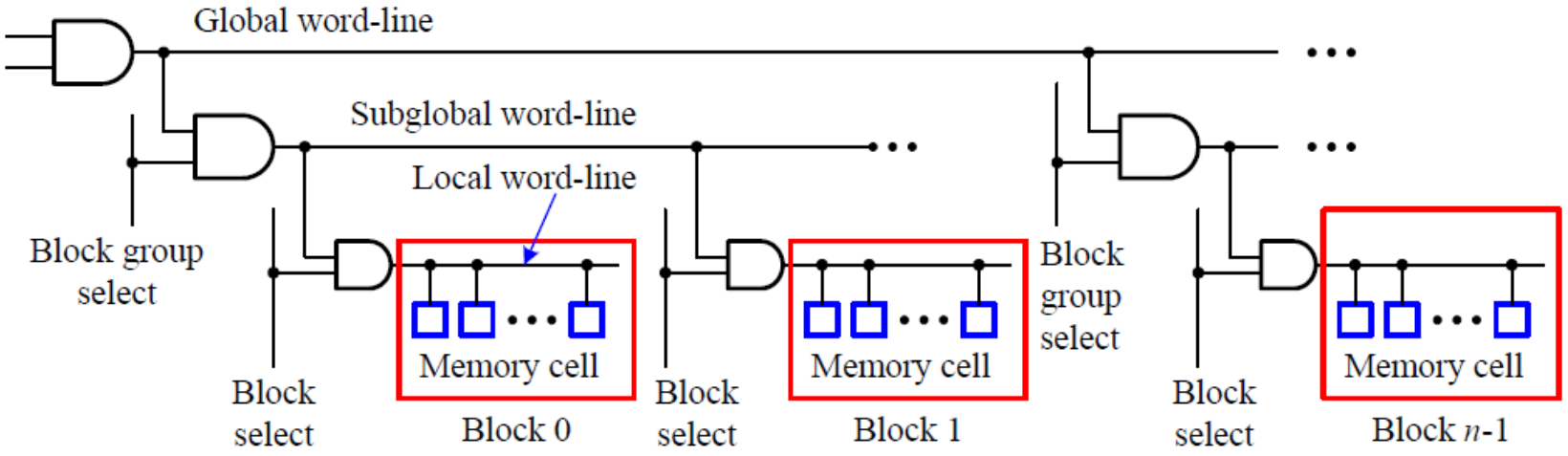
Typical Memory Architecture



Typical Memory Architecture

- Storing in a memory matrix
  - A single element of a matric stores one or more bit of information
  - We can activate a row by the **word line,** which is a part of the address generated by the row decoder.
  - The activated cells copy their content to the **bit line**
  - The other part of the address selects a single bit
- The sense amplifier is responsible for signal conditioning
  - It converts the low-level signal to a valid CMOS rail-to-rail signal
  - The transistors in the memory cells are very tiny (with the decrease of the cell size, the capacitance of the memory increases)
  - The input voltage on the sense amplifier is usually lower than 100mV, which has to be amplified

High capacity memory



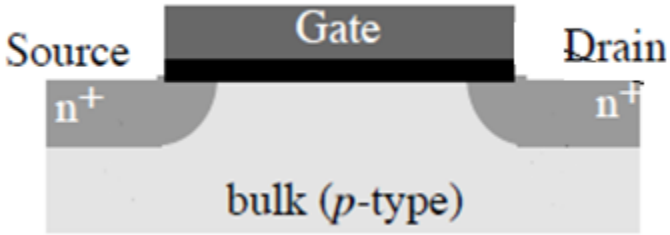
- If the number of rows is large, they are divided into partitions with hierarchical access
- The user can see **banks**

MOS transistor

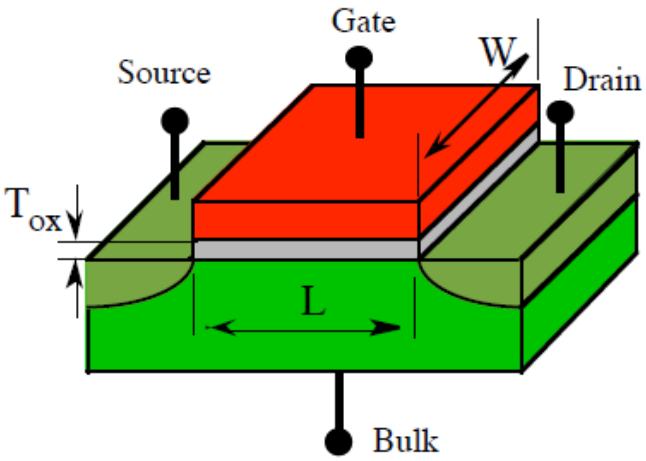
in a detailed manner

Characteristics of MOS transistors

- In digital CMOS design a MOS transistor can be substituted with a 2-state switch
- Understanding the operation of memory cells requires greater understanding
- What we know so far?
  - The current flow between the drain and source can be controlled by the gate voltage
  - If the gate-source voltage is low, there is no current
  - If the gate-source voltage is higher than a certain value (V_T – threshold voltage) the transistor conducts current
  - This is enough knowledge for understanding the digital operation
  - Now we have to know what the relationship is between the gate-source voltage and drain-source voltage

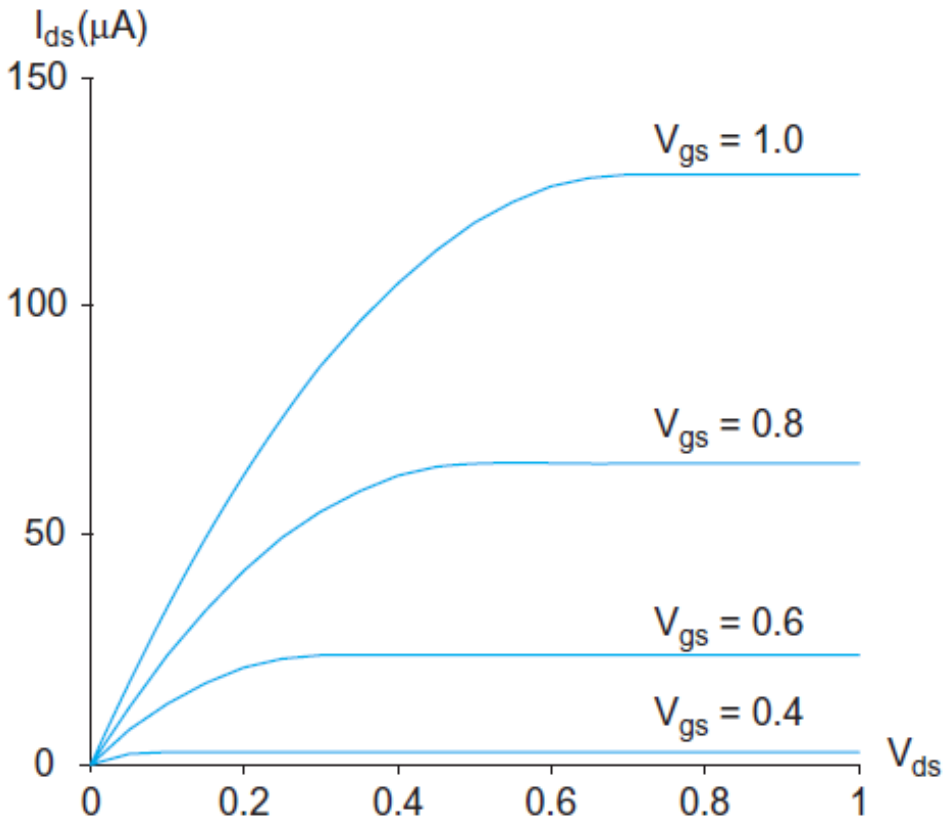


MOS transistor



- Important dimensions:
  - t_{ox}: oxide thickness (a few nm)
  - L: length of the channel (14nm – a few µm)
  - W: width of the channel (20nm – a few µm)

MOS transistor



While $V_{GS} < V_T$, the transistor has very low current (subthreshold current, leakage current)

If $V_{GS} > V_T$, the current is proportional to $(V_{GS}-V_T)^2$

- $V_{GS}-V_T$ is usually denoted by V_{on}

The characteristic equation of MOS transistors

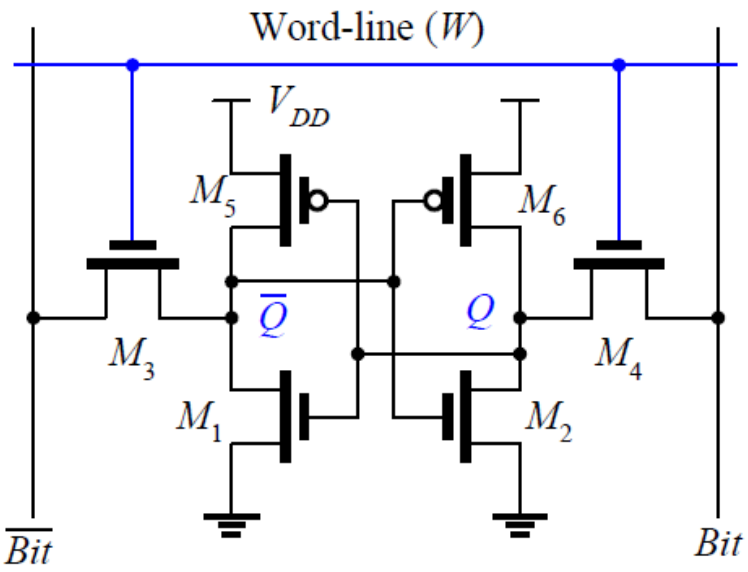
$$I_D = \frac{\epsilon_{ox}\,\mu_{eff}}{2t_{ox}}\frac{W}{L}(V_{GS}-V_T)^2 = \frac{\beta}{2}\frac{W}{L}(\mathrm{V_{GS}}-\mathrm{V_T})^2$$

- Where
  - W is the width of the gate,
  - L is the length of the gate,
  - ε_{ox}/t_{ox} is the oxide capacitance per unit area,
  - μ_{eff} is the mobility of the charge carriers in the channel,
  - V_{GS} is the gate-source voltage,
  - V_T is the threshold voltage of the device.
- Designers can only alter the value of W and L
- Wider transistors can conduct higher current
  - They have lower (channel) resistance

Static RAM (SRAM)

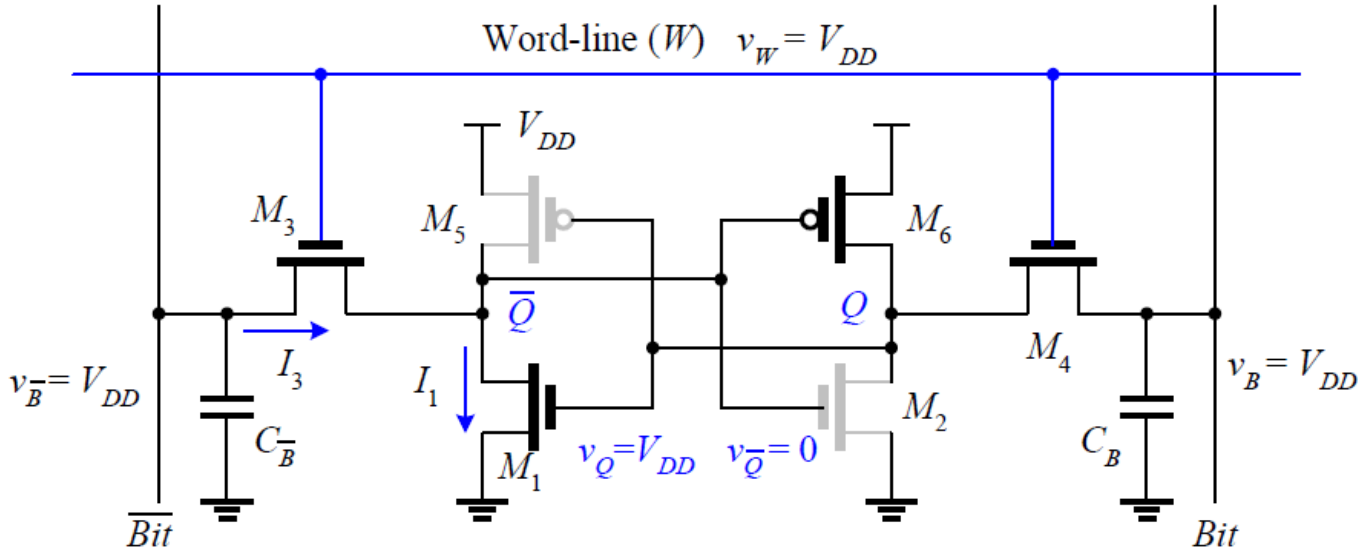
Static RAM memory

- It consists of 6 transistors
- It has two complementary bit lines
  - It is a differential logic element



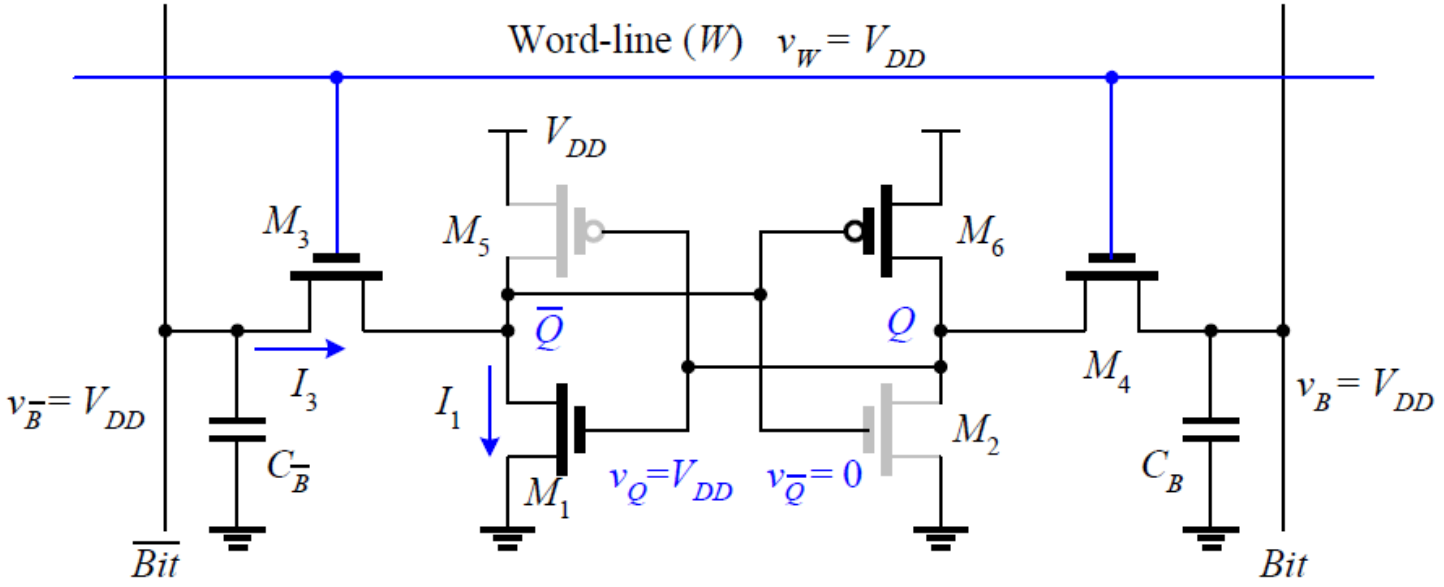
- Information is stored by two CMOS inverters in feedback.
- It is similar to SR latch, but the read/write is done through one transistor (not a complete (NAND) gate)
  - (In order to reduce the size of it)
- M3 and M4 transistors are called (access) transistors
- The operation of an SRAM cell seems simple (on the digital abstraction level), but it is more complicated

Read process of Static RAM – Step 1

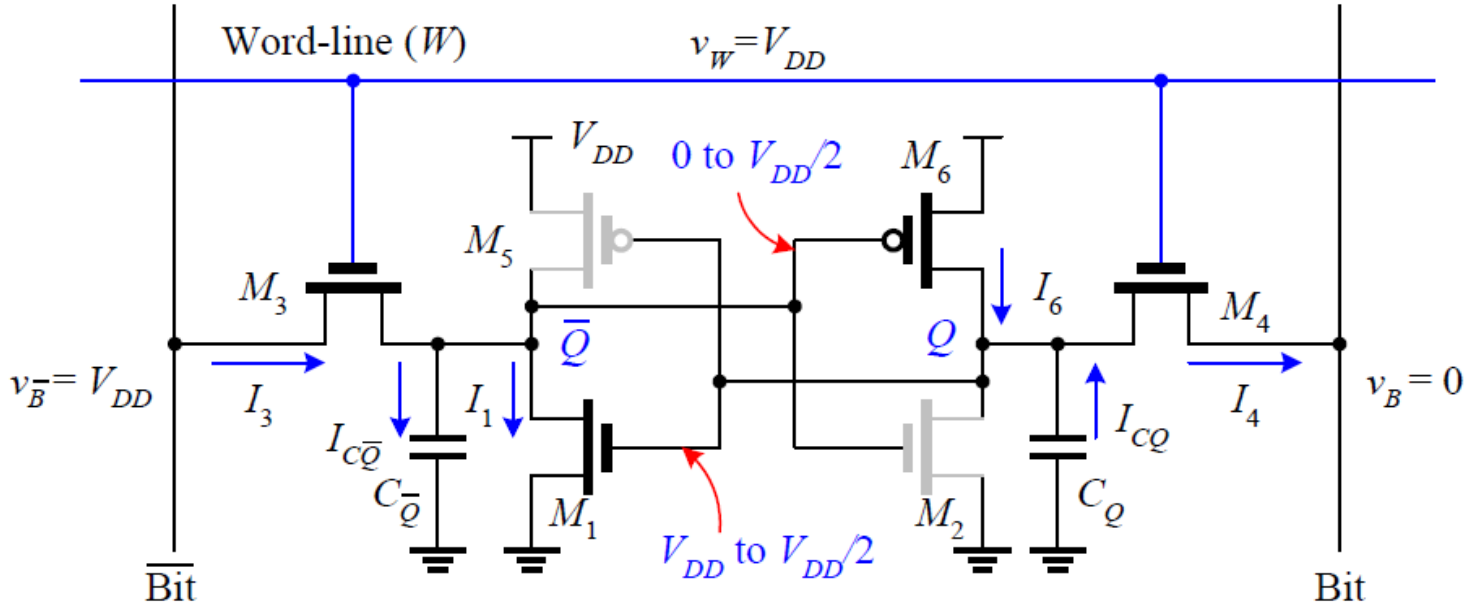


- We assume that Q is logic 1
  - The bit-lines are precharged to the power supply voltage
- The transistor M3 has a minimal channel width
- The NMOS transistors in the inverters have a two times wider channel width (they are “stronger”)
- At the beginning of the reading M3 and M1 conduct current, and the voltage level of $\overline{Q}$ will increase

Read process of Static RAM – Step 2

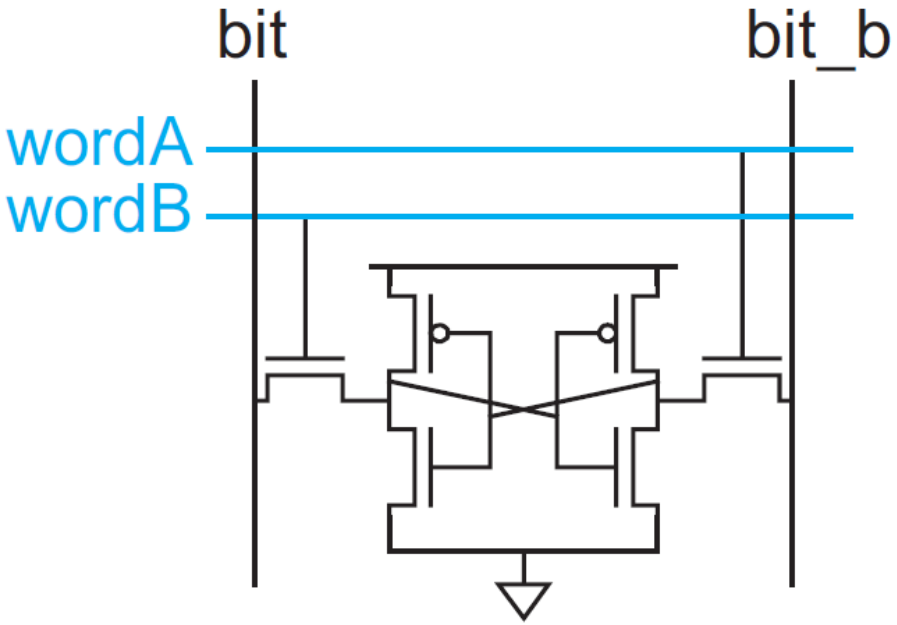


- $V_{\bar{Q}} = V_{DD} \frac{R_{M1}}{R_{M1}+R_{M3}} < \mathrm{V_{COMP}} = \frac{\mathrm{V_{DD}}}{2}$
- The change in voltage has to be lower than the threshold level of the gate, so that the content of the cell will remain unchanged
- The voltage level of the $\overline{bit}$ line decreases to the half of the original value (around 100-200mV decreases is common) through the M3 and M1 transistors
- The sense amplifier can amplify it.



We assume the information in the memory cell is logic 1, and we have to write 0

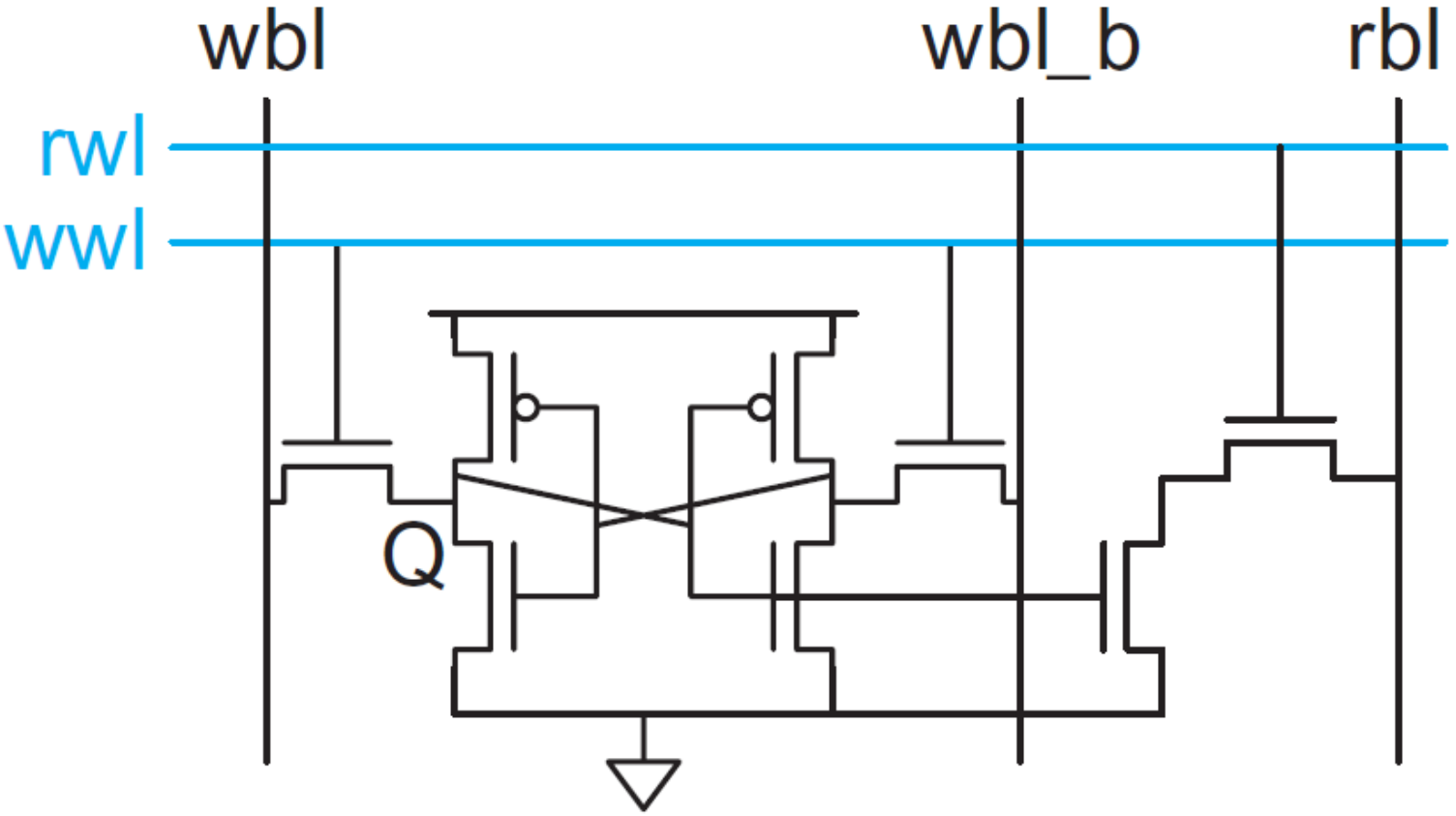
- The operation of the M3 and M1 transistors is similar to the reading case
- M4 pulls the voltage level of Q below the threshold limit
- $V_Q = V_{DD} \frac{R_{M4}}{R_{M4}+R_{M6}} < \mathrm{V_{COMP}} = \frac{\mathrm{V_{DD}}}{2}$
- The current of M1 decreases, M5 conducts, and the flipflop flips



As we saw in the previous slide, a one bit-line is enough for reading

- We can read from two cells using two bitlines
- In case of writing we need both bitlines
- Using “smart” timing we can read twice and write a bit in a cycle
- This is frequently used in CPU registers

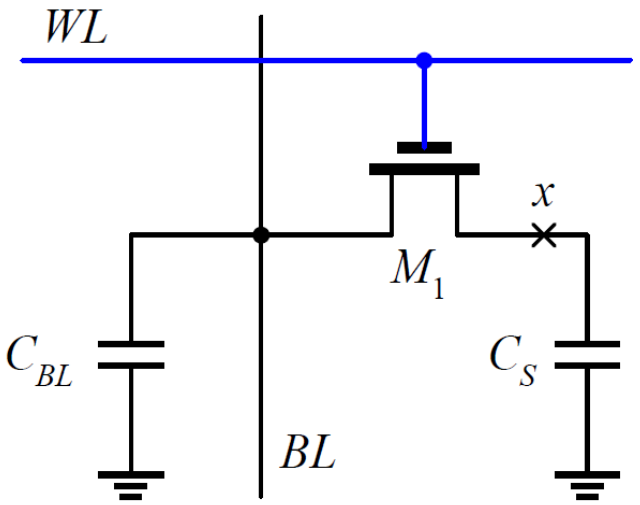
Multi-port SRAM 2.



- SRAM with dedicated read and write ports
  - It consists of 8 transistors

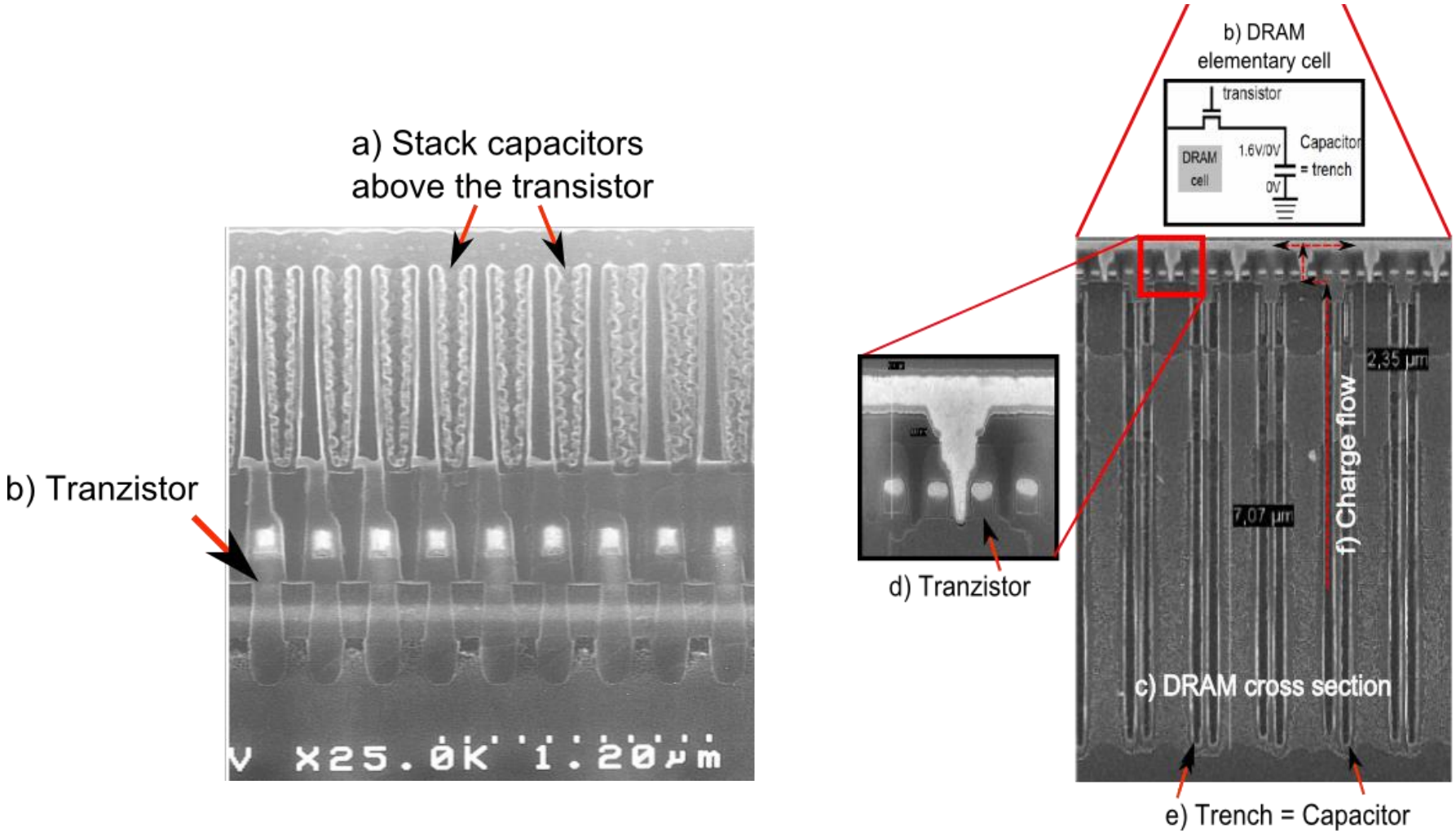
Dynamic RAM (DRAM)

Dynamic RAM cell



- The information is stored in the C_S capacitor, which is connected to the bitline by the M1 transistor.
- C_S is typically 25-40fF (there are only ~100,000 electrons in the capacitor)
- The storage capacitors have a special 3D structure
  - Trench capacitance: a trench is etched in the Si. The walls are covered with oxide and filled with poli-Si.
  - Stack capacitance: several conducting layers are separated with thin oxide layers.

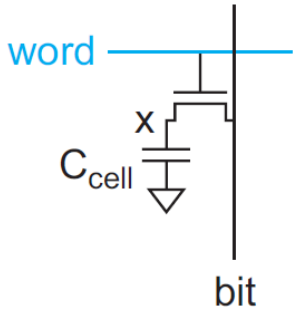
Implementations of storage capacitors (illustration)



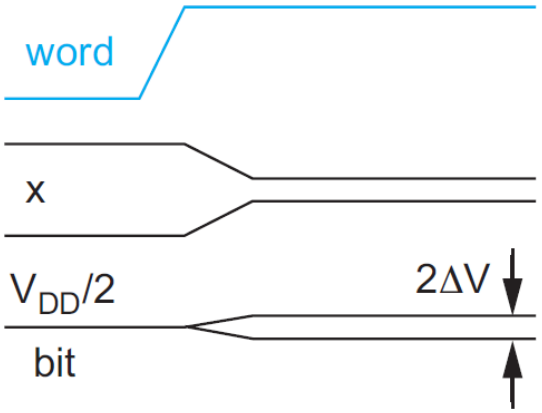
- Source: http://www.sdram-technology.info/sdram-cross-section.html

DRAM write and read

- Writing to DRAM
  - The word line is connected to VDD and the capacitor is charged to VDD or discharged to ground potential depending on the value of the bit-line
- Reading from DRAM
  - The bit-line is precharged to half of the power supply voltage
  - When the word line is activated, the capacitor is connected to the bit line
  - After charge sharing, the voltage change of the bit-line is:
  - $\Delta V = \frac{C_S}{C_{BL}+C_S} V_{DD}/2$
  - The reading process is DESTRUCTIVE!



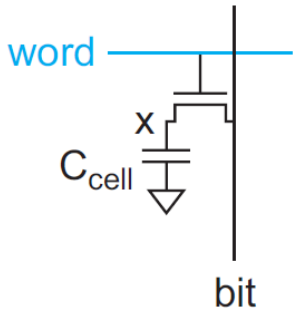
(a)



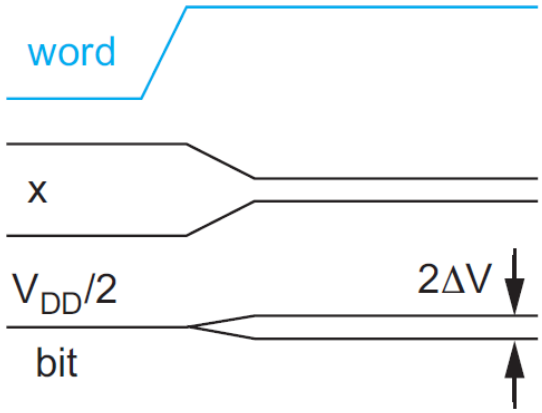
(b)

DRAM write and read

- The voltage change is ~20-50mV
- The sensing amplifier has to determine the logic value of the cell from such a tiny change. This is a very difficult task
- The charge sharing discharges/charges the capacitor to a value from which the original logic value can no longer be determined so the cell has to be rewritten after each readout.



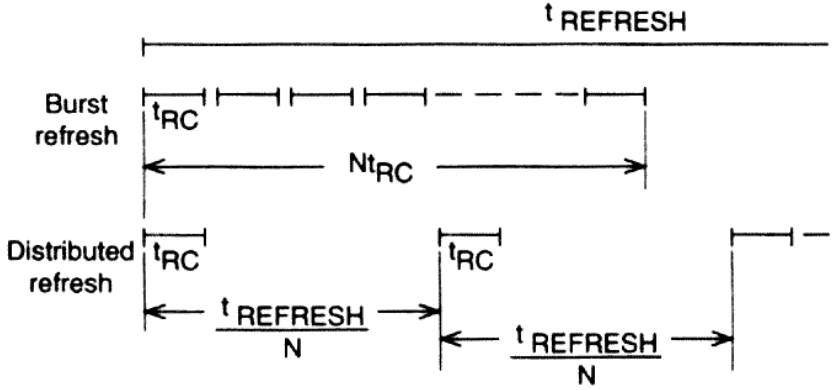
(a)



(b)

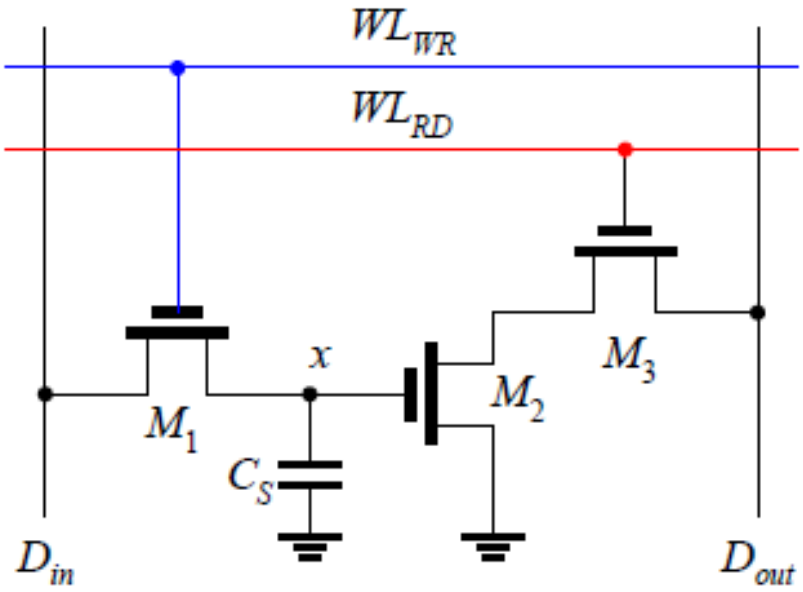
Refreshing DRAM

- The information stored in a cell fades over time due to the leakage current of the capacitor
- Thus the cells of a DRAM need to be refreshed periodically – typically every 2-4 ms.
- One line is refreshed at a time (it takes ca. 100-200 ns) (t_{RC})
  - **Burst refresh:** all the lines are refreshed in a run
  - **Distributed (hidden) refresh:** a counter stores the address of the most recently refreshed line and refreshing continues with the next line at every step.



Embedded DRAM

- The single-transistor dynamic RAM requires special technology
- In SoC devices only CMOS components are available
- Instead of a storing capacitor we can use the parasitic capacitance of the M1 and M2 transistors

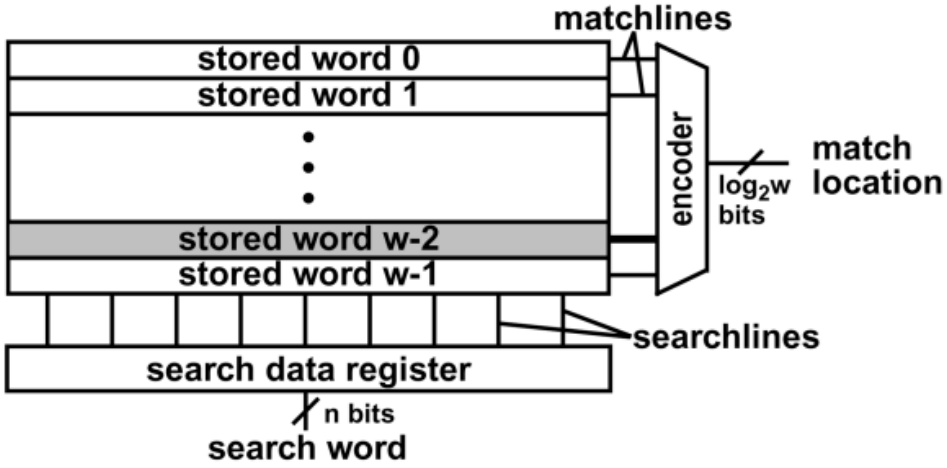


- They are used in high capacitance cache memory circuits with hidden refresh
  - (smaller than SRAM)

Budapest University of Technology and Economics
Department of Electron Devices

CAM
Content Addressable Memory

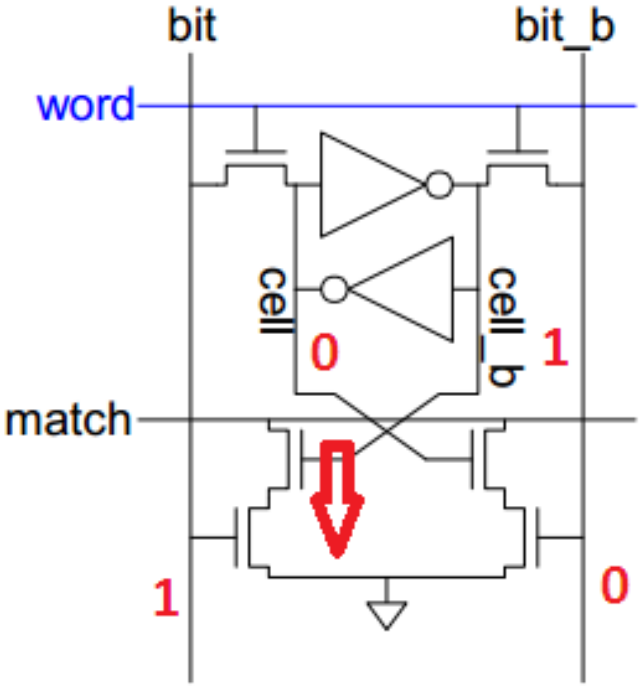
Content Addressable Memory



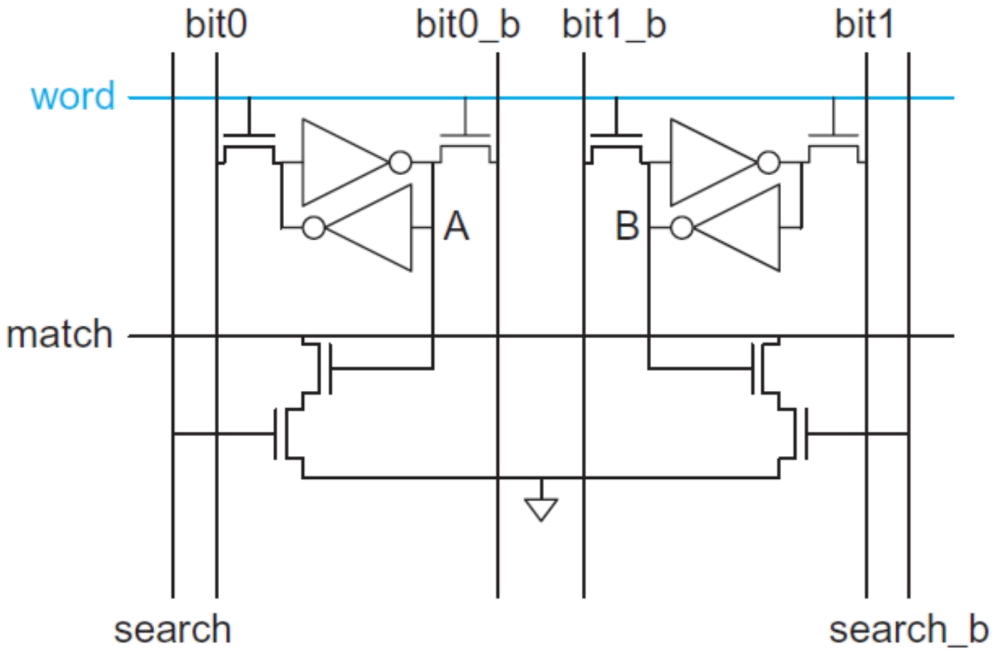
- It generates the address of the required information in a single clock cycle
  - It compares the content of the search data register to the parallel stored words
  - Only one will be activated from the matchlines, which gives the address back
- The generated address corresponds to the content which is stored in the ordinary memory (Hardware associative array)

CAM cell

- A static RAM cell and 4 additional transistors
  - 4 transistors constitute the match circuitry
  - 10 transistors per cell
- If the input bit does not coincide with that stored in the SRAM, the match line is pulled down
  - “Miss-current” exists
- When a miss occurs, the CPU retrieves the data from the main memory



Ternary CAM (TCAM)



- Realizing “don’t care”
  - The TCAM cell is doubled
  - “01” is assigned to logic 1, “10” means logic 0
  - “00” is assigned to “don’t care” state
  - “11” is not used

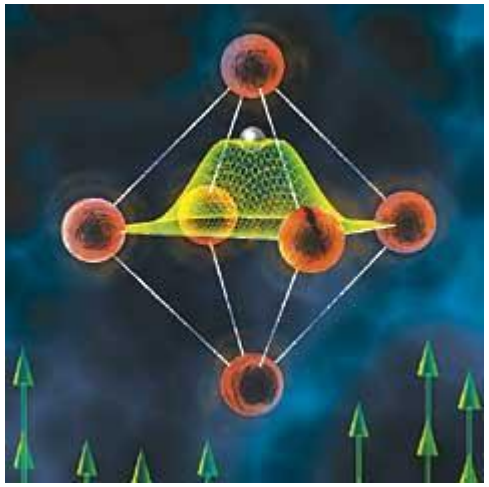
Newer memory technologies

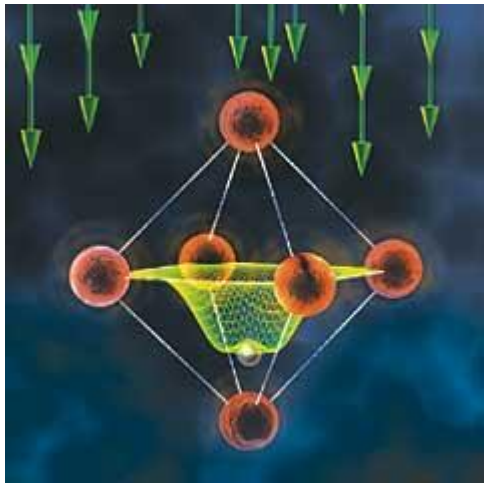
Ferroelectric RAM (FRAM)

Magnetoresistive RAM (MRAM)

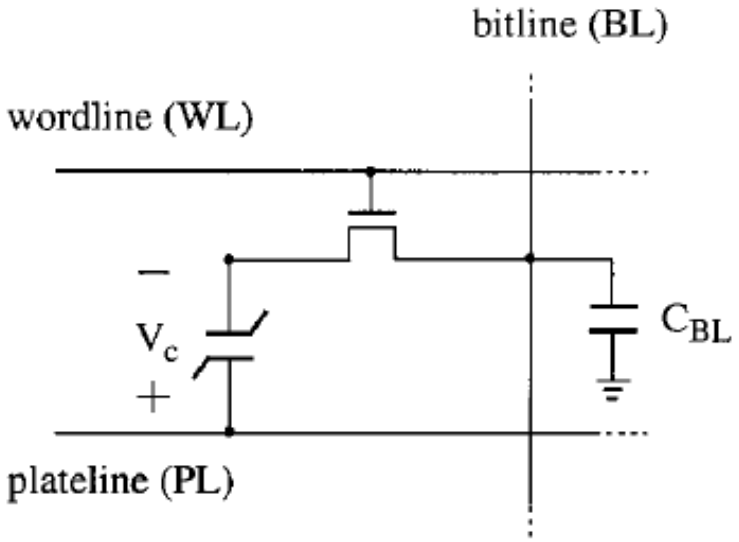
Ferroelectric RAM

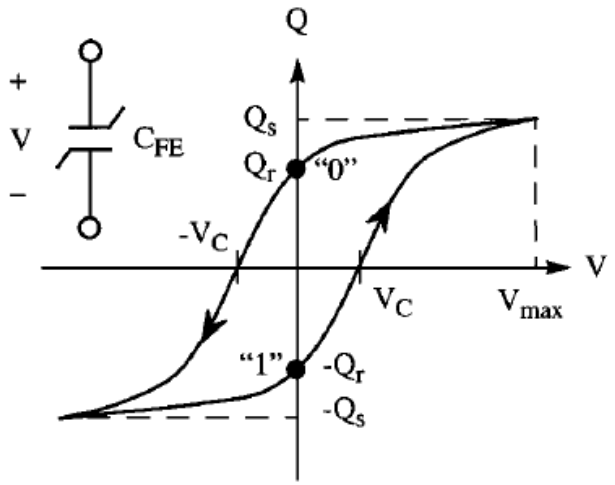
- Ferroelectric material
  - The term is used in analogy to ferromagnetism
    - in which a material exhibits a permanent magnetic moment (so not an iron-based material)
  - **Ferroelectricity** is a property of certain materials that have a spontaneous electric polarization
  - It can be reversed by the application of an external electric field
    - lead zirconate titanate (PZT). ($Pb[Zr_xTi_{1-x}]O_3$)
    - It has two stable states, because the oxygen atom can move to the center of the molecule
  - The change in polarization causes charge movement (current)
  - The idea was first published in 1952





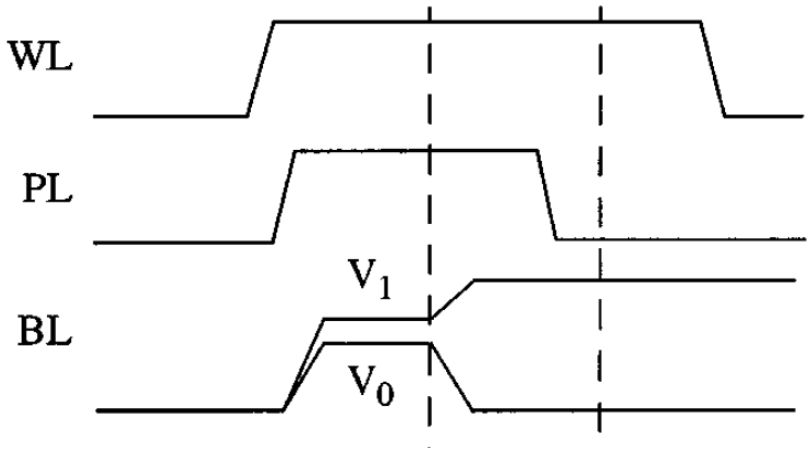
FERAM single cell





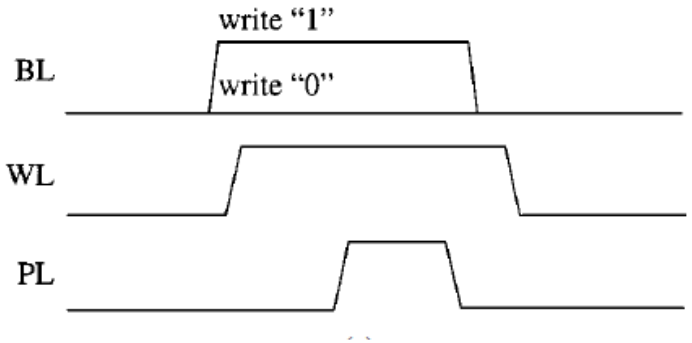
- It is similar to a DRAM cell, the storing dielectric material is the PTZ
- This is a capacitor with memory

FERAM reading



- The bitline has to be precharged
- We have to apply a pulse on the PL
- The voltage of the bitline changes (as in the case of DRAMs)
- The sense amplifier regenerates the signal
- The reading process is *destructive*

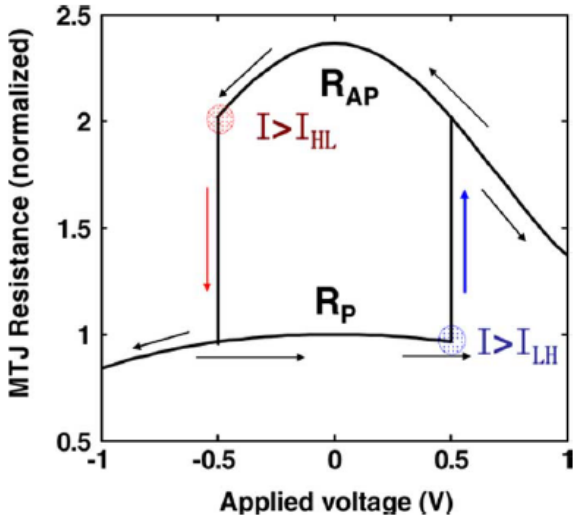
FERAM writing



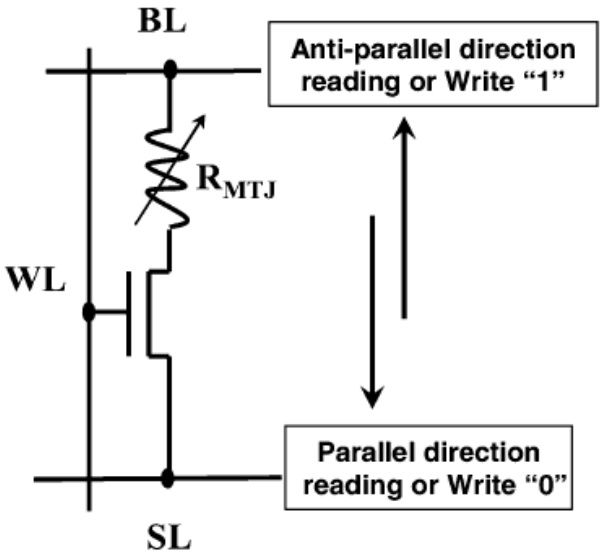
- Activate the word-line
- A pulse applied on PL writes the information of the BL into the cell

Magnetoresistive RAM

- Data in MRAM is stored by magnetic storage elements
- The elements are formed from two ferromagnetic plates
  - Each of which can hold a magnetization
  - They are separated by a thin insulating layer.
- One of the two plates is a permanent magnet set to a particular polarity
- The other plate's magnetization can be changed to match that of an external field
- The resistance of MRAM depends on the direction of the magnetization



MRAM single cell



- Reading – measuring the current of the activated cell (non-destructive)
  - BL has low voltage, SL is connected to ground
- Writing – High voltage applied between the bitline and the wordline
  - The value of the stored bit depends on the direction of the applied voltage
  - BL=0V , SL=VDD or BL=VDD, SL=0V

Comparing FRAM and MRAM

Attribute	FRAM	MRAM
Technology	1 transistor – 1 ferroelectric capacitor (1T – 1FRC)	1 transistor – 1 magnetic tunnel junction (1T – MTJ)
Read	Destructive	Non-destructive
Read Cycle Time	90ns	35ns
Read Access Time	60ns	35 ns
Write Cycle Time	90ns	35ns
Write (CE)	60ns	35ns
Fatigue	Known wear-out mechanisms 10^{14} Read / Write	No wear-out. Infinite Read / Write
Imprint	Increases with temperature and decreasing operating voltage	No Imprint
High Temperature Data Retention	Degrades above 85°C	20 years at 125°C (continuous)
Manufacturing	Complex – O_2 ambient >650°C process, H_2 sensitivity.	Standard BEOL CMOS processing
Scalability	Need 3D structures at 65nm. Imprint becomes more of an issue.	Fully Scalable

- Imprint is the preferential polarization state from previous writes to that state
- Fatigue is the increase of minimum writing voltage due to loss of polarization after extensive cycling.

RAM memory chips - Examples

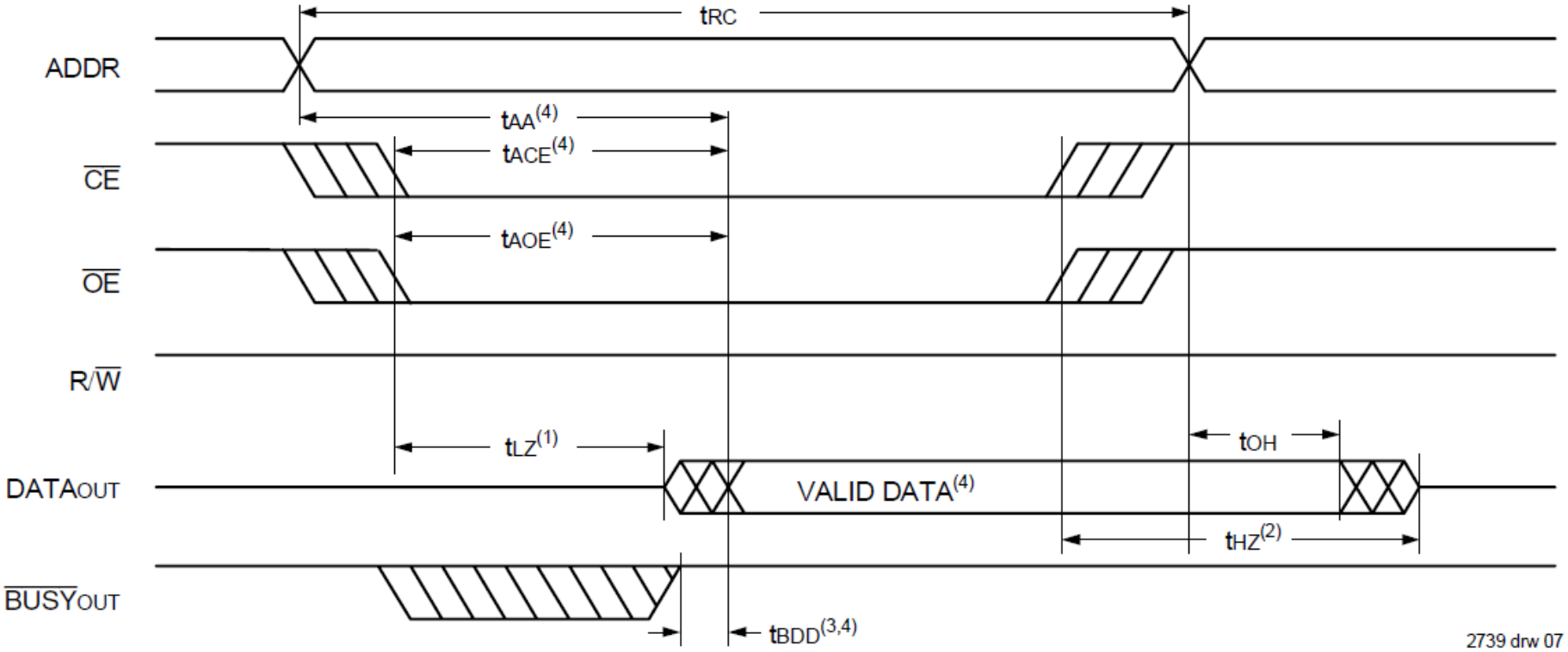
Based on google search result.

IDT 7006S/L – Dual-port static RAM

- Two ports, 16k x 8 bit SRAM
  - Left and right ports, fully symmetrical
  - Allows simultaneous reads of the same memory location
- On-chip arbitration
- HW semaphore signaling
  - Concurrent accesses of different ports handled by the hardware
- Fully asynchronous operation

Dual-Port SRAM - Example

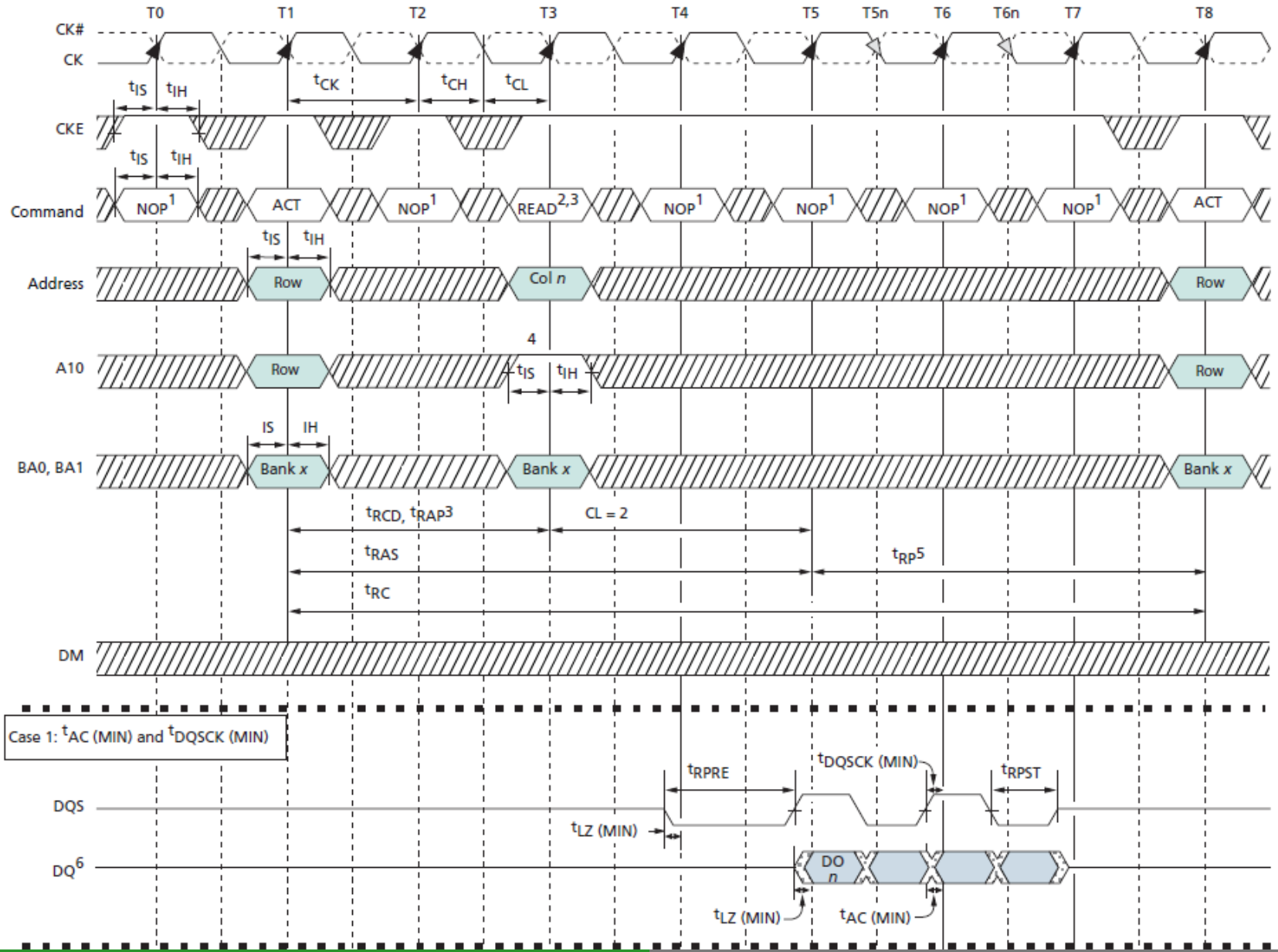
Waveform of Read Cycles(5)



Micron - MT46V64M8

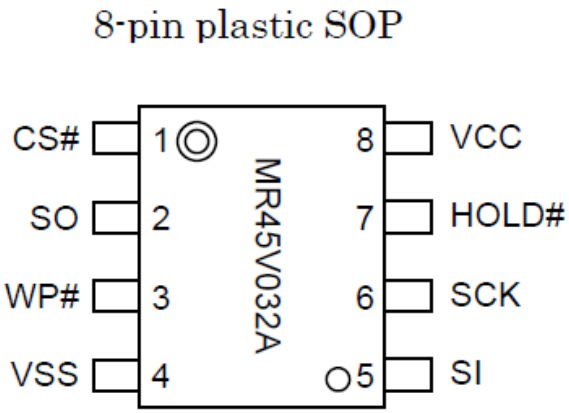
- 512MBit, 64M x 8bit.
  - Power supply: 2.5V
  - DDR 400 – Internal, pipelined double-data-rate
  - 4 banks
  - Programmable burst lengths: 2, 4 or 8
  - Auto refresh: 64ms (8192 clock cycles)
  - CL (CAS / READ latency) – 2, 2.5, 3
  - Raw address: A0-A12 (8k), Column address: A0- A9, A11 (2k)
  - Bank address: BA0, BA1
- It is very complicated to use

DDR SDRAM example - Reading

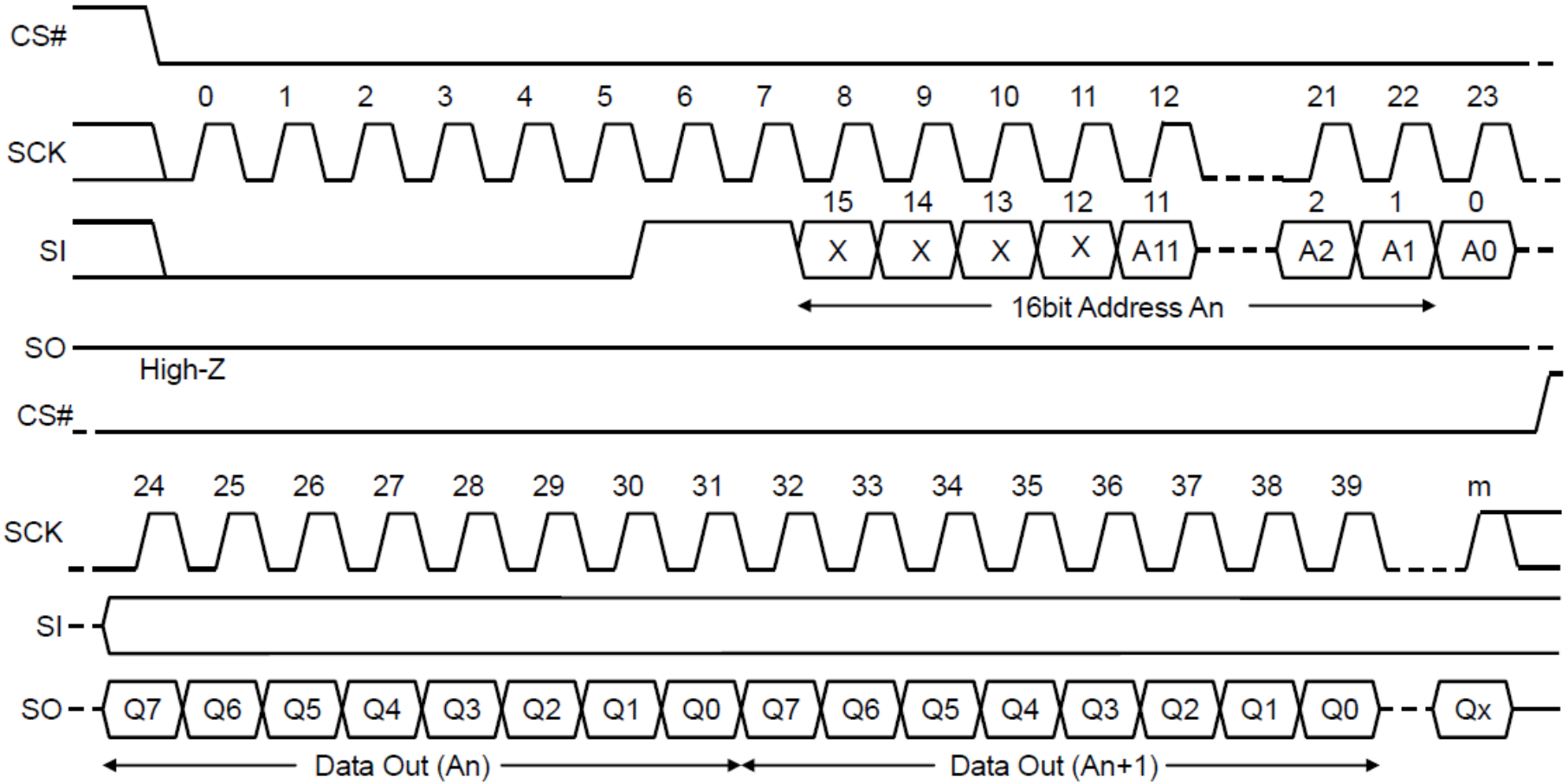


LAPIS Semi, **MR45V032A**

- 32k(4,096-Word 8-Bit) FeRAM (Ferroelectric Random Access Memory) SPI
- 4k× 8bit
- Power supply: 2.7-3.6V, (3.3V typical)
- Operating frequency: 15MHz
- Data retention: 10 years
- Read/write tolerance: 10^{12}cycles/bit
- Communication:
  - SPI (Serial Peripheral Interface)



FERAM



0x03 represents the READ command, and it is followed by the 16-bit starting address