3 Embedded Memories

3.1 The world of Memory

Semiconductor memories are vital components in modern integrated circuits. Stand-alone memories represent roughly 30% of the global integrated circuit market. Within system-on-chip, memory circuits usually represent more than 75% of the total number of transistors.

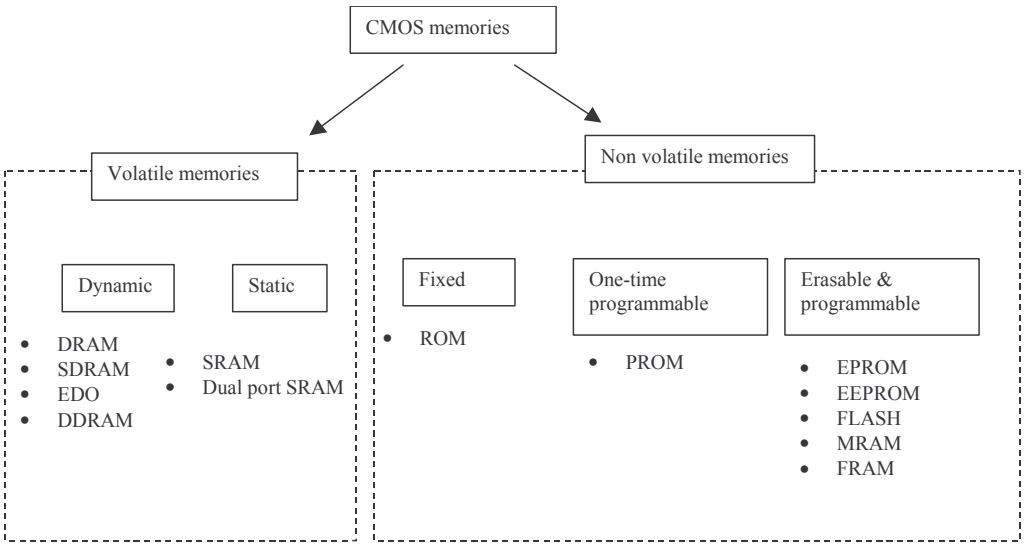


Figure 3-1 Major classes of CMOS compatible memories

Two main families of devices exist: volatile and non-volatile memories.

- In volatile circuits (Figure 3-1 left), the data is stored as long as the power is applied. The dynamic random access memory (DRAM) is the most common memory.
- Non-volatile memories are capable of storing the information even if the power is turned off (Figure 3-1 right). The read-only memory (ROM) is the simplest type of non-volatile memory. One-time programmable memories (PROM) are a second important family, but the most popular non volatile memories are erasable and programmable devices: the old electrically programmable ROM (EPROM), the more recent Electrically Erasable PROM (EEPROM, FLASH), and the new magneto resistive RAM (MRAM) and ferroelectric RAM (FRAM) memories.

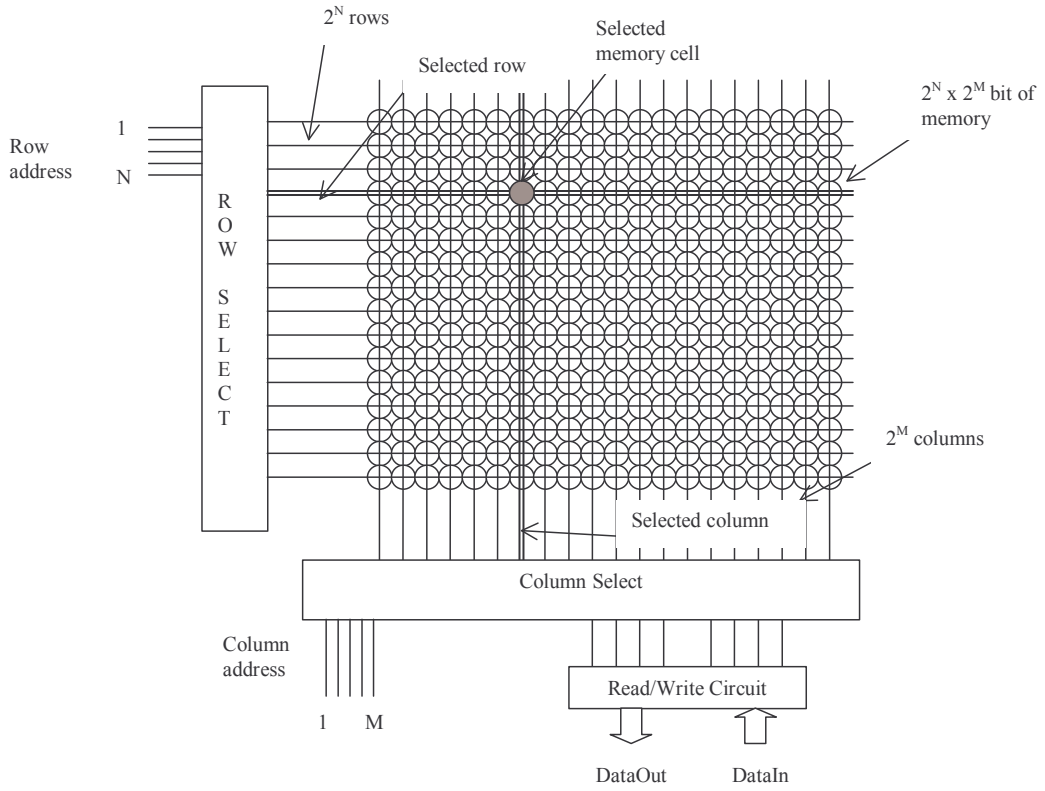


Figure 3-2 Typical memory organization

Figure 3-2 shows a typical memory organization layout. It consists of a memory array, a row decoder, a column decoder and a read/write circuit. The row decoder selects one row from 2^N, thanks to a N-bit row selection address. The column decoder selects one row from 2^M, thanks to a M-bit column selection address. The memory array is based on 2^N rows and 2^M columns of a repeated pattern, the basic memory cell. A typical value for N and M is 10, leading to 1024 rows and 1024 columns, which corresponds to 1048576 elementary memory cells (1Mega-bit).

3.2 RAM Memory

The basic cell for static memory design is based on 6 transistors, with two pass gates instead of one. The corresponding schematic diagram is given in Figure 3-3. The circuit consists again of the 2 cross-coupled inverters, but uses two pass transistors instead of one. The cell has been designed to be duplicated in X and Y in order to create a large array of cells. Usual sizes for Megabit SRAM memories are 256 column x 256 rows or higher. A modest arrangement of 4x4 RAM cells is proposed in figure 3-4. The selection lines *WL* concern all the cells of one row. The bit lines *BL* and *~BL* concern all the cells of one column.

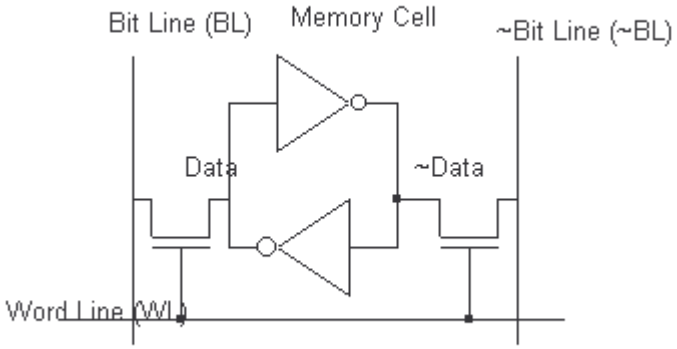


Figure 3-3: The layout of the 6 transistor static memory cell (RAM6T.SCH)

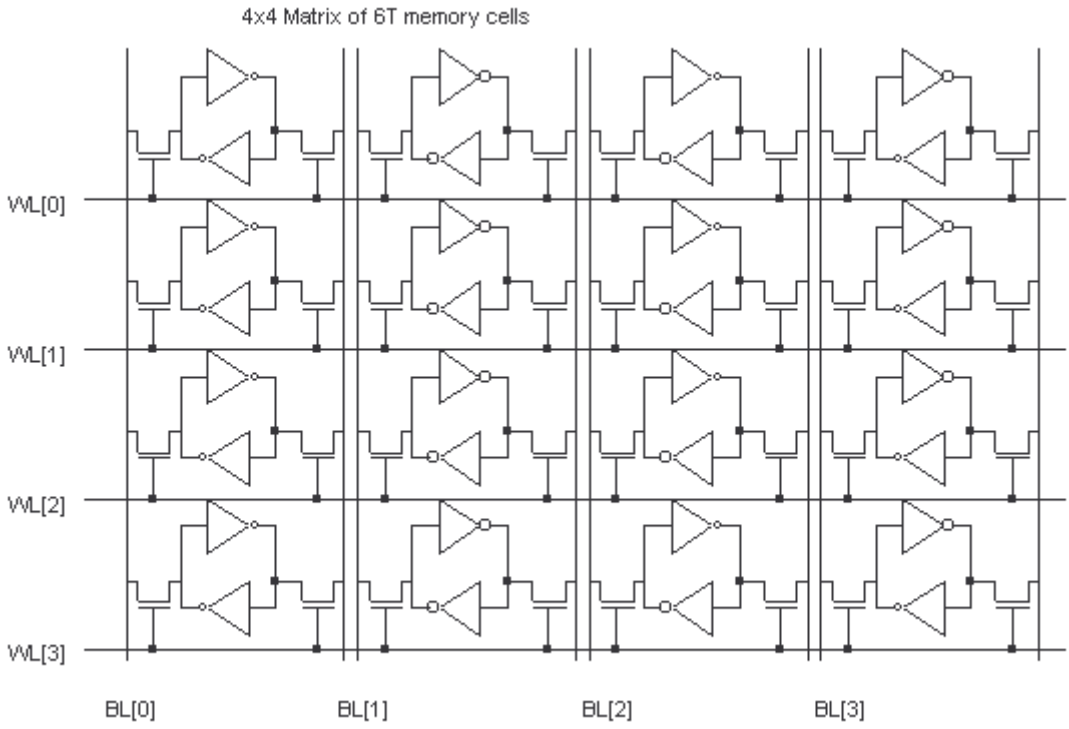


Fig. 3-4 An array of 6T memory cells, with 4 rows and 4 columns (RAM6T.SCH)

The RAM layout is given in Figure 3-5. The *BL* and *~BL* signals are made with metal2 and cross the cell from top to bottom. The supply lines are horizontal, made with metal3. This allows easy matrix-style duplication of the RAM cell.

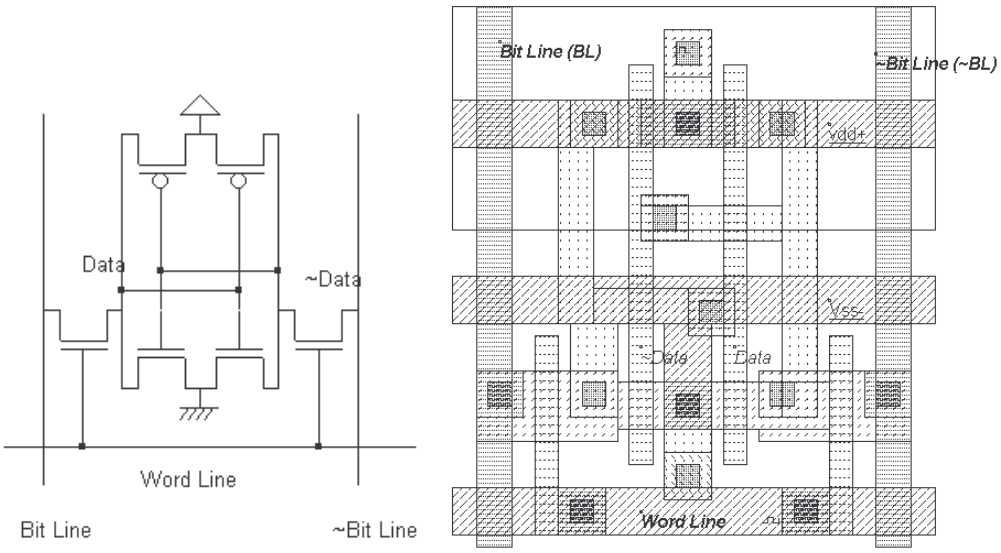


Fig. 3-5. The layout of the static RAM cell (RAM6T.MSK).

WRITE CYCLE. Values 1 or 0 must be placed on *Bit Line*, and the data inverted value on *~Bit Line*. Then the selection *Word Line* goes to 1. The two-inverter latch takes the *Bit Line* value. When the selection *Word Line* returns to 0, the RAM is in a memory state.

READ CYCLE. The selection signal *Word Line* must be asserted, but no information should be imposed on the bit lines. In that case, the stored data value propagates to *Bit Line*, and its inverted value *~Data* propagates to *~Bit Line*.

SIMULATION. The simulation parameters correspond to the read and write cycle in the RAM. The proposed simulation steps consist in writing a 0, a 1, and then reading the 1. In a second phase, we write a 1, a 0, and read the 0. The *Bit Line* and *~Bit Line* signals are controlled by pulses (Figure 3-6). The floating state is obtained by inserting the letter "x" instead of 1 or 0 in the description of the signal.

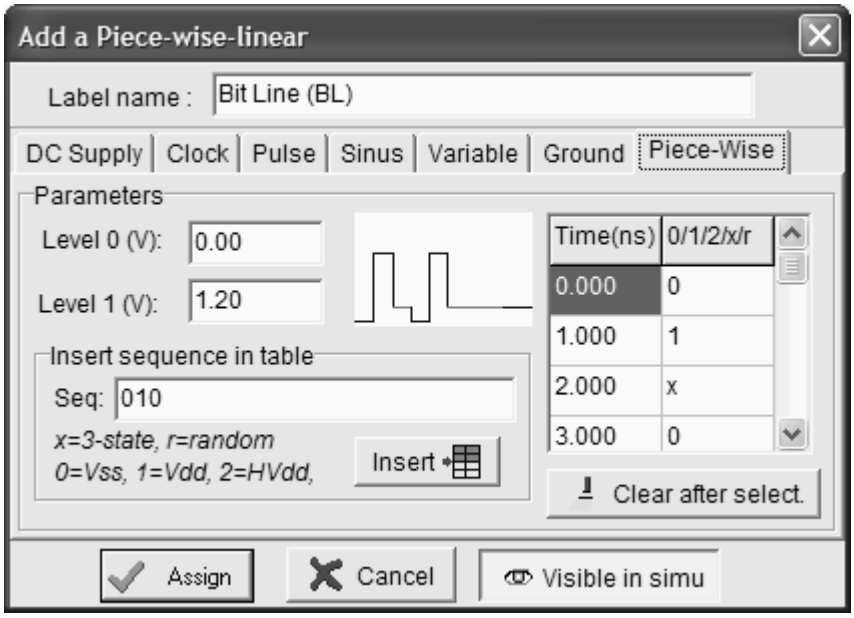


Fig. 3-6. The bit Line pulse used the "x" floating state to enable the reading of the memory cell (RamStatic6T.MSK)

The simulation of the RAM cell is proposed in figure 3-7. At time 0.0, *Data* reaches an unpredictable value of 1, after an unstable period. Meanwhile, *~Data* reaches 0. At time 0.5ns, the memory cell is selected by a 1 on *Word Line*. As the *Bit Line* information is 0, the memory cell information *Data* goes down to 0. At time 1.5ns, the memory cell is selected again. As the *Bit Line* information is now 1, the memory cell information *Data* goes to 1. During the read cycle, in which *Bit Line* and *~Bit Line* signals are floating, the memory sets these wires respectively to 1 and 0, corresponding to the stored values.

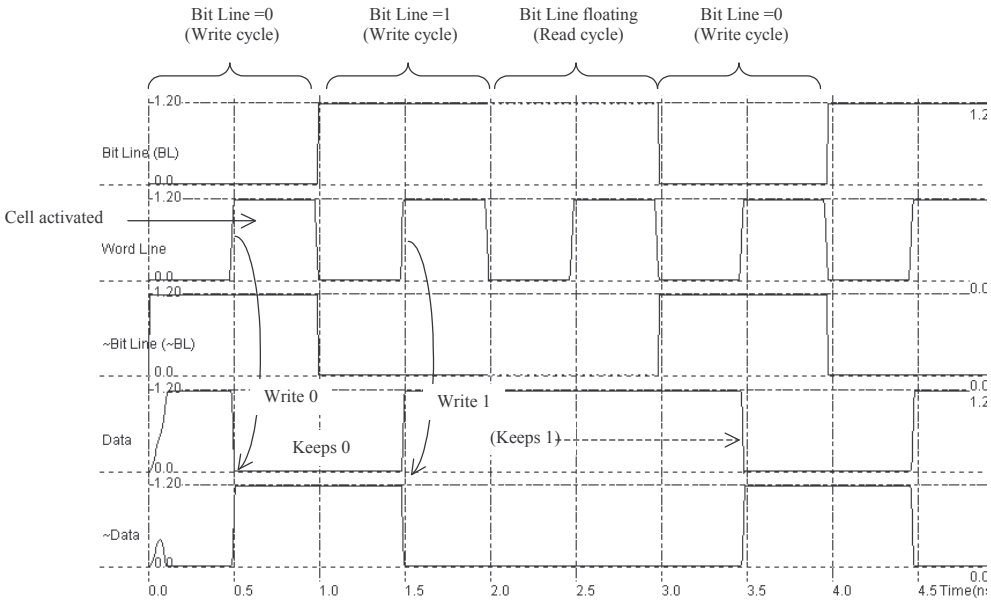


Fig. 3-7. Write cycle for the static RAM cell (RamStatic6T.MSK).

3.3 RAM Array

You can duplicate the RAM cell into a 4x4 bit array using the command **Edit → Duplicate XY**. Select the whole RAM cell and a new window appears. Enter the value « 4 » for X and « 4 » for Y into the menu. Click on « **Generate** ».

A very interesting approach to obtain a more compact memory cell consists in sharing all possible contacts: the supply contact, the ground contact and the bit line contacts. The consequence is that the effective cell size can be significantly reduced (Figure 3-8).

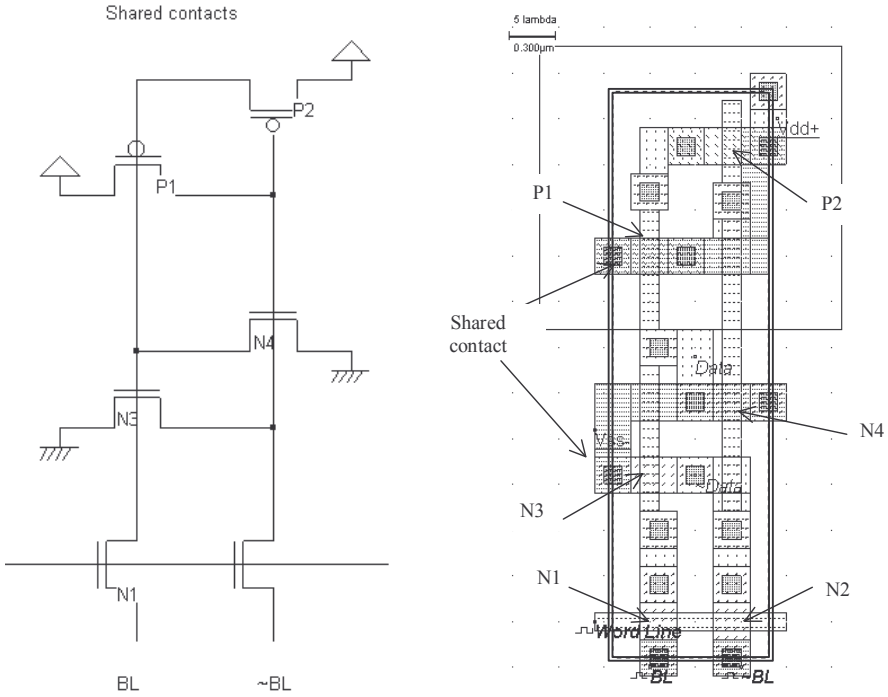


Figure 3-8. Sharing all possible contacts lead to a very compact cell design (Ram6Tcompact.MSK)

The layout is functionally identical to the previous layout. The only difference is the placement of MOS devices and contacts. We duplicate the RAM cell into a 64 bit array. The multiplication cannot be done directly by the command **Duplicate XY,** as we need to flip one cell horizontally to share lateral contacts, and flip the resulting block vertically to share vertical contacts (Figure 3-9).

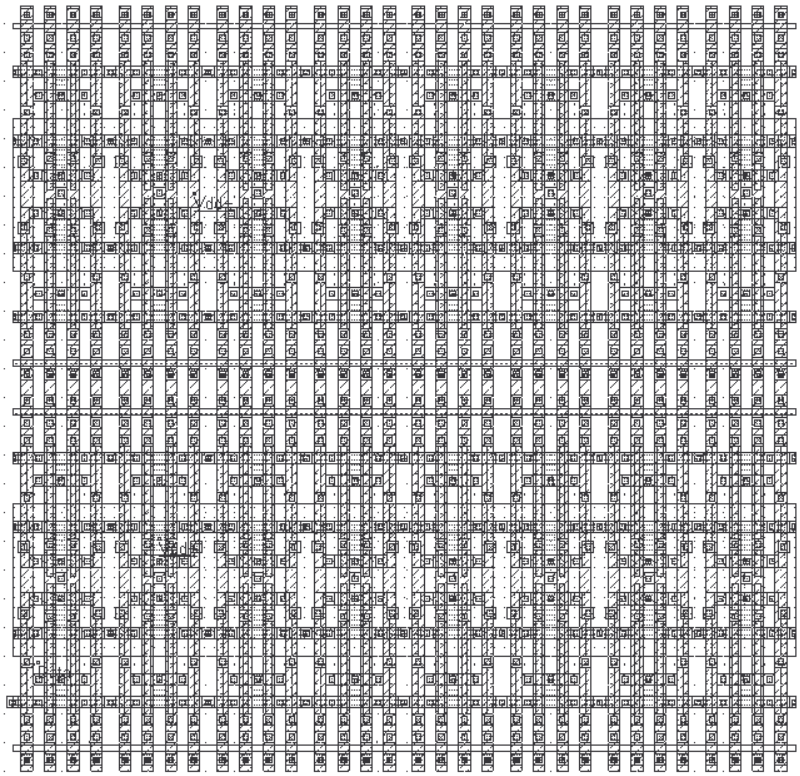


Figure 3-9. Compact 16x4 array of memory cells with shared contacts (Ram16x4Compact.MSK)

Row Selection Circuit

The row selection circuit decodes the row address and activates one single row (Figure 3-10). This row is shared by all word line signals of the row. The row selection circuit is based on a multiplexor circuit. One line is asserted while all the other lines are at zero.

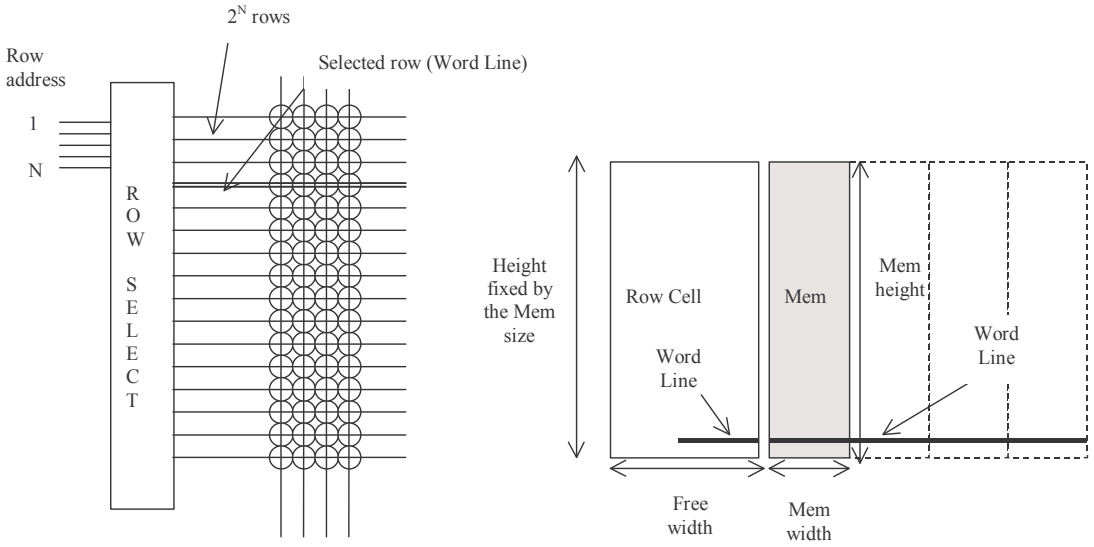


Fig. 3-10 The row selection circuit

In the row selection circuit for the 16x4 array, we simply need to decode a two-bit address. Using AND gates is one simple solution. In figure 3-11, we present the schematic diagram of 2-to-4 and 3-to-8

decoders. In the case of a very large number of address lines, the decoder is split into sub-decoders, which handle a reduced number of address lines.

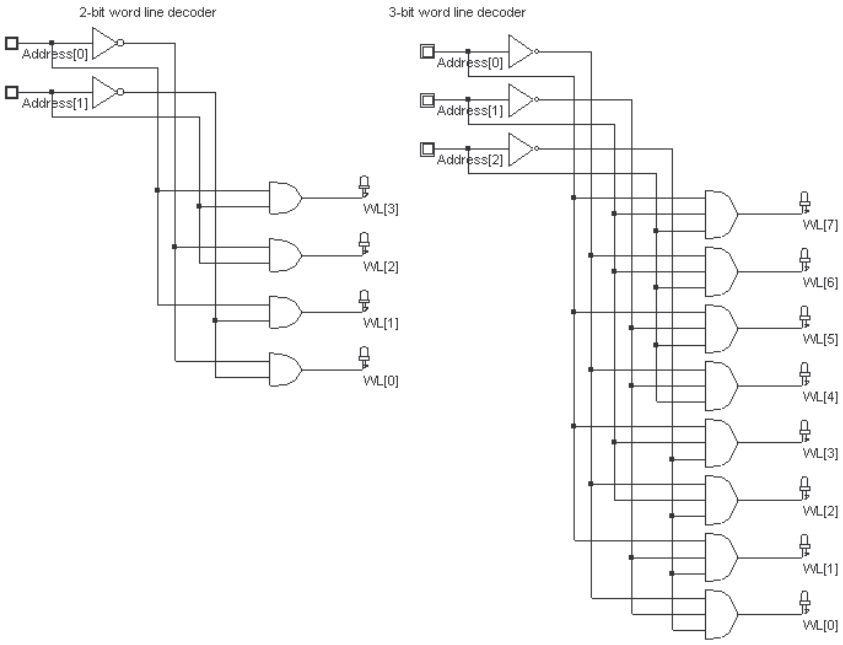


Fig. 3-11. The row selection circuit in 2 bit and 3 bit configuration (RamWordLine.SCH)

Column Selection Circuit

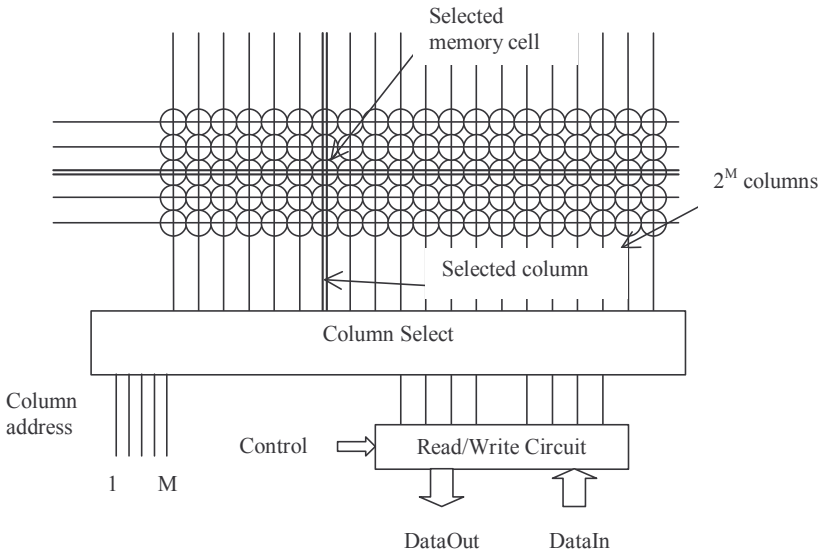


Figure 3-12. The column selection circuit principles

The column decoder selects a particular column in the memory array to read the contents of the selected memory cell (Figure 3-12) or to modify its contents. The column selector is based on the same principles

as those of the row decoder. The major modification is that the data flows both ways, that is either from the memory cell to the *DataOu*t signal (Read cycle), or from the *DataIn* signal to the cell (Write cycle). Figure 3-13 proposes an architecture based on n-channel MOS pass transistors. We consider here 4 columns of memory cells, which requires 2 address signals *Address_Col[0]* and *Address_Col[1]*. The n-channel MOS device is used as a switch controlled by the column selection. When the nMOS is on and *Write* is asserted, (Figure 3-13) the *DataIn* is amplified by the buffer, flows from the bottom to the top and reaches the memory through *BL* and ***~BL***. If *Write* is off, the 3-state inverter is in high impedance, which allows the information to be read on *DataOut*.

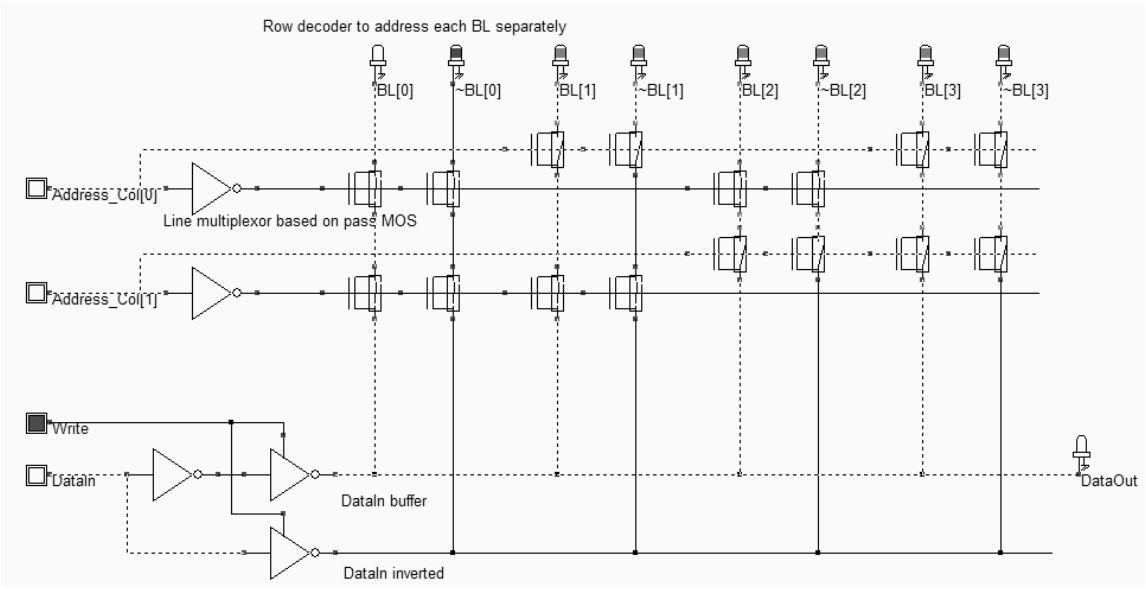


Figure 3-13. Row selection and Read/Write circuit (RamColumn.SCH)

A Complete 64 bit SRAM

The 64 bit SRAM memory interface is shown in figure 3-14. The 64 bits of memory are organized in words of 4 bits, meaning that *DataIn* and *DataOut* have a 4 bit width. Each data *D0..D15* occupies 4 contiguous memory cells in the array. Four address lines are necessary to decode one address among 16. The memory structure shown in figure 10-44 requires two address lines *A0* and *A1* for the word lines *WL[0]..WL[3]* and two address lines *A2* and *A3* for the bit line selection. The final layout of the 64 bit static RAM is proposed in Figure 3-15.

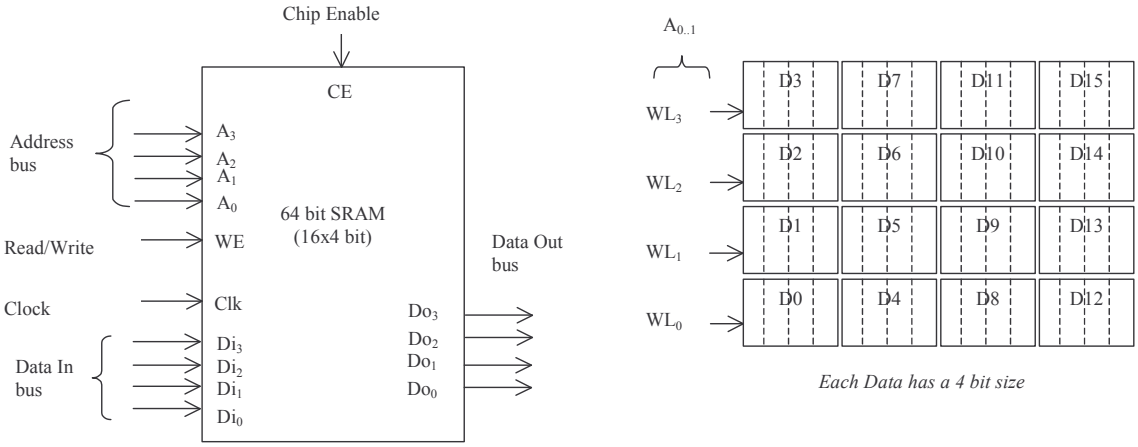


Figure 3-14. The architecture of the 64 bit RAM (RAM64.MSK)

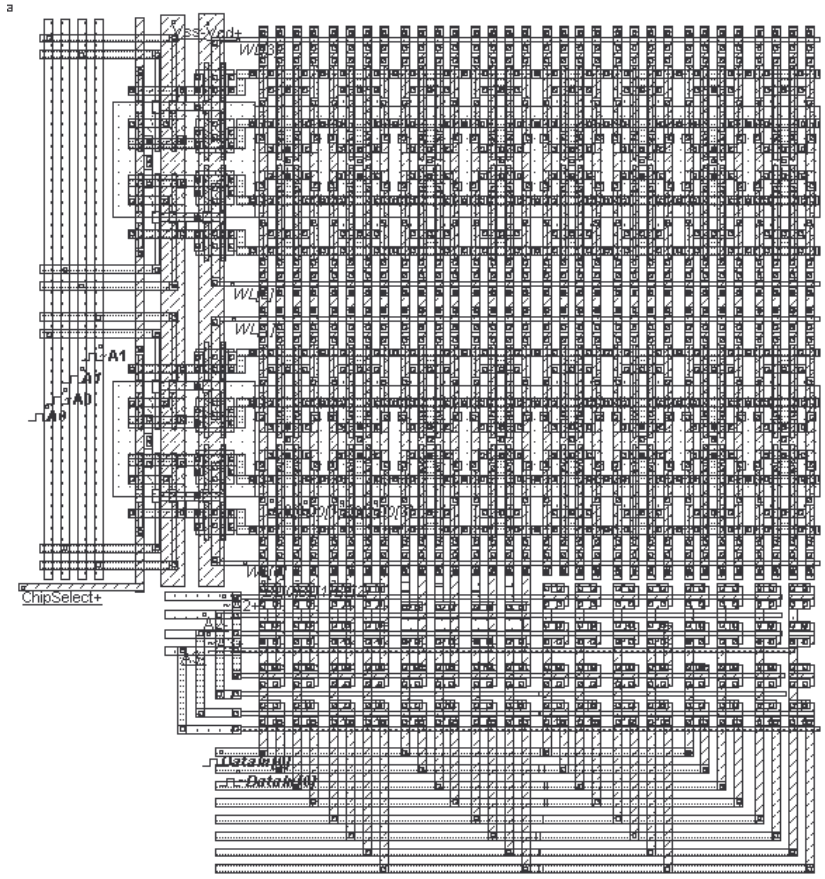


Figure 3-15. The complete RAM layout (RAM64.MSK)

3.4 Dynamic RAM Memory

The dynamic RAM memory has only one transistor, in order to improve the memory matrix density by almost one order of magnitude. The storage element is no longer the stable inverter loop, as for the static

RAM, but only a capacitor *Cs*, also called the storage capacitor. The DRAM cell architecture is shown in figure 3-16.

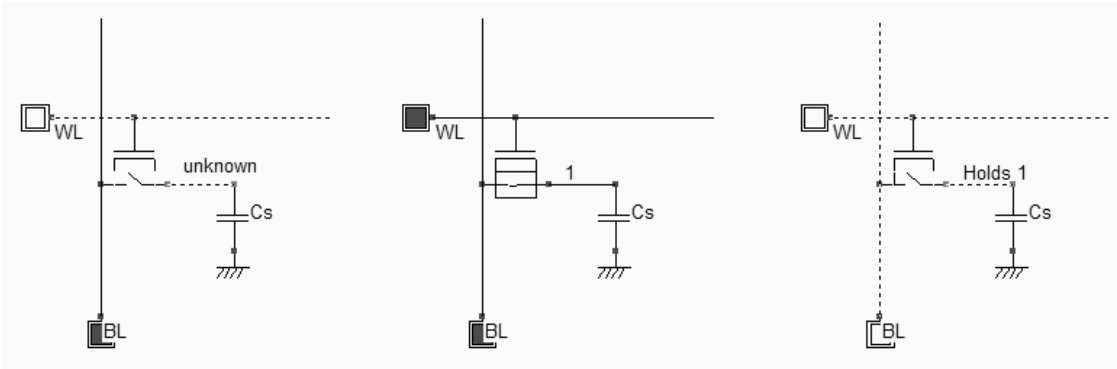


Figure 3-16: Simulation of the Write cycle for the 1 transistor dynamic RAM cell (RAM1T.SCH)

The write and hold operation for a "1" is shown in figure 3-17. The data is set on the bit line, the word line is then activated and *Cs* is charged. As the pass transistor is n-type, the analog value reaches *VDD-Vt*. When *WL* is inactive, the storage capacitor *Cs* holds the "1".

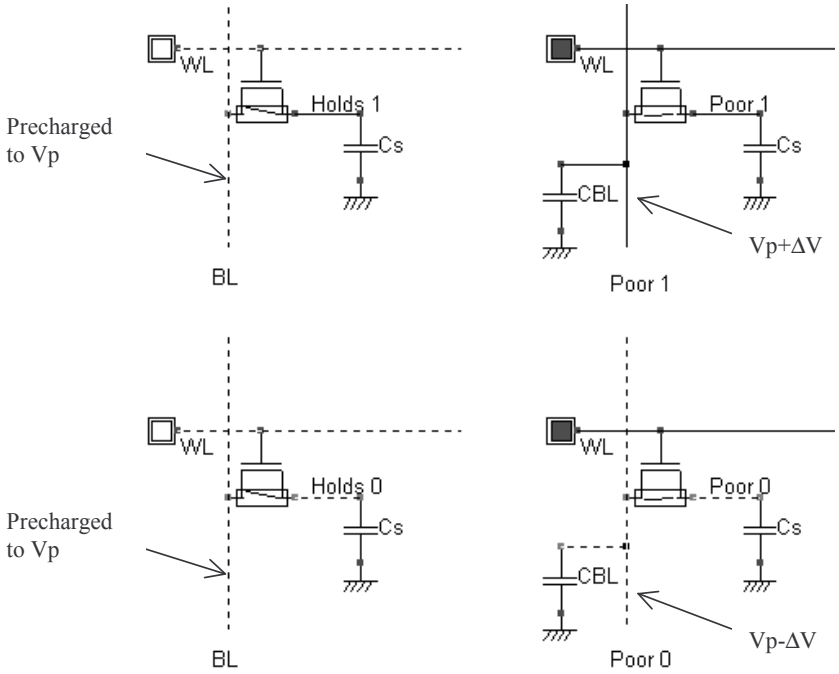


Figure 3-17: Simulation of the Read cycle for the 1 transistor dynamic RAM cell (RAM1T.SCH)

The reading cycle is destructive for the stored information. Suppose that *Cs* holds a 1. The bit line is precharged to a voltage *Vp* (Usually around VDD/2). When the word line is active, a communication is established between the bit line, loaded by capacitor *CBL*, and the memory, loaded by capacitor *CS*. The charges are shared between these nodes, and the result is a small increase of the voltage *Vp* by *ΔV*, thanks to the injection of some charges from the memory.

Commercial dynamic RAM memories use storage capacitors with a value between 10fF and 50fF. This is done by creating a specific capacitor for the storage node appearing in figure 3-18 left thanks to the following technological advances: the use of specific metal layers to create the lower plate and external walls of the RAM capacitor, an enlarged height between the substrate surface and metal1, and the use of high permittivity dielectric oxide. The silicon dioxide SiO2 has a relative permittivity ε_r of 3.9. Other oxides, compatible with the CMOS process have a higher permittivity (Higher "K") : Si_3N_4 with ε_r equal to 7, and Ta_2O_5 with ε_r equal to 23.

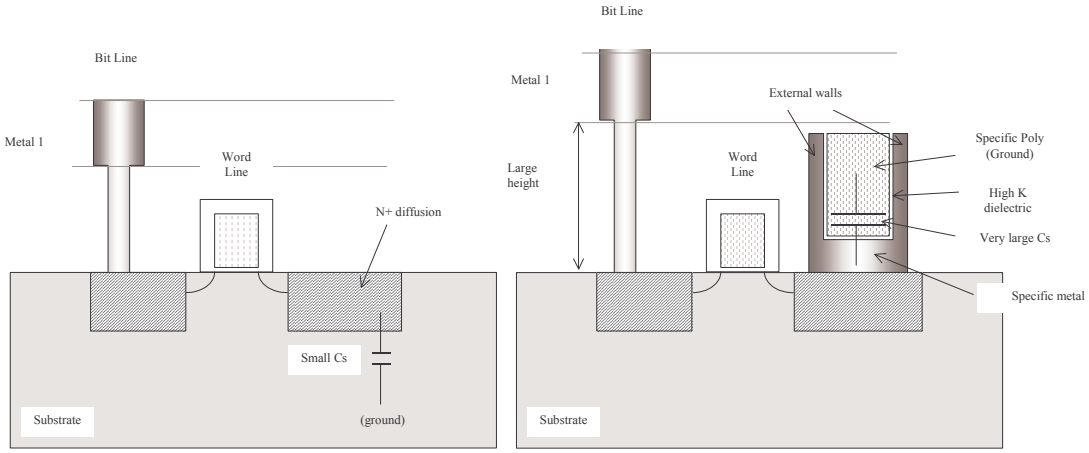


Figure 3-18: Increasing the storage capacitance (Left: junction capacitor, right, embedded capacitor)

The cross-section of the DRAM capacitor is given in figure 3-19. The bit line is routed in metal2, and is connected to the cell through a metal1 and diffusion contact. The word line is the polysilicon gate. On the right side, the storage capacitor is a sandwich of conductor material connected to the diffusion, a thin oxide (SiO2 in this case) and a second conductor that fills the capacitor and is connected to ground by a contact to the first level of metal. The capacitance is around 20fF in this design. Higher capacitance values may be obtained using larger option layer areas, at the price of a lower cell density.

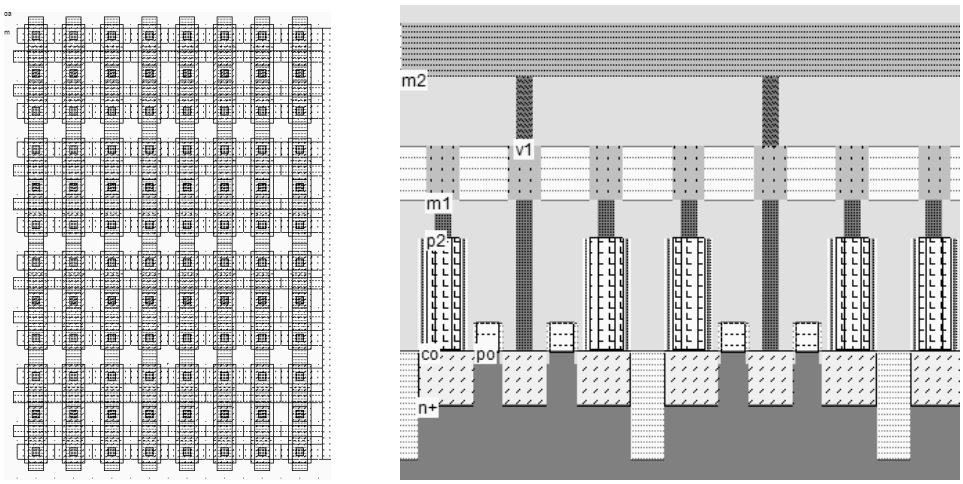


Figure 3-19: The stacked capacitor cell compared to the diffusion capacitor cell (DramEdram.MSK)

3.5 EEPROM

The basic element of an EEPROM (Electrically Erasable PROM) memory is the floating-gate transistor. The concept was introduced several years ago for the EPROM (Erasable PROM). It is based on the possibility of trapping electrons in an isolated polysilicon layer placed between the channel and the controlled gate. The charges have a direct impact on the threshold voltage of a double-gate device. When there is no charge in the floating gate (Figure 3-20, upper part), the threshold voltage is low, meaning that a significant current may flow between the source and the drain, if a high voltage is applied on the gate. However, the channel is small as compared to a regular MOS, and the Ion current is 3 to 5 times lower, for the same channel size.

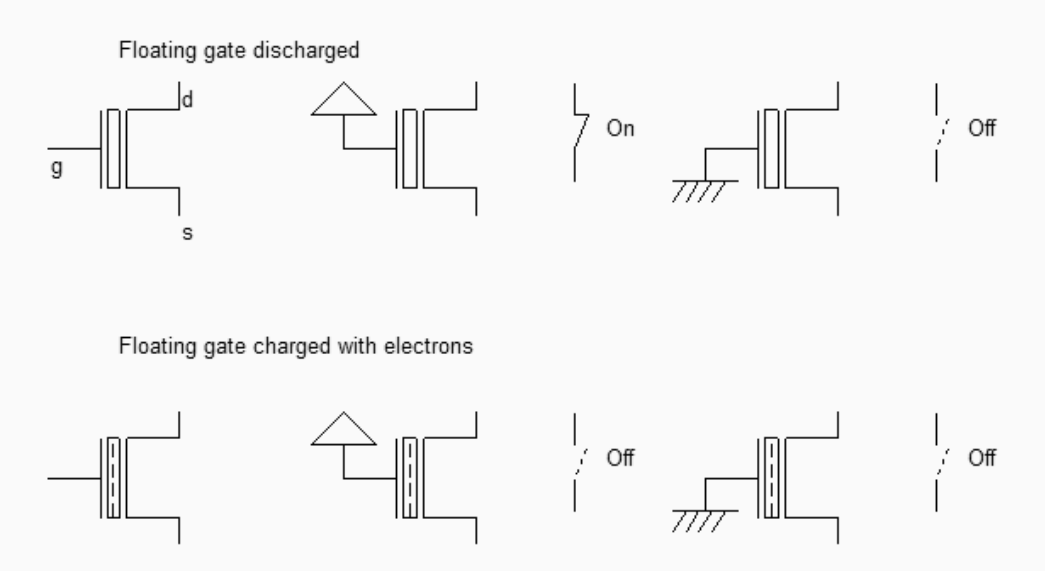


Fig. 3-20: The two states of the double gate MOS (EepromExplain.SCH)

When charges are trapped in the floating polysilicon layer (Figure 3-20, lower part), the threshold voltage is high, almost no current flows through the device, independently of the gate value. As a matter of fact, the electrons trapped in the floating gate prevent the creation of the channel by repealing channel electrons. Data retention is a key feature of EEPROM, as it must be guaranteed for a wide range of temperatures and operating conditions. Optimum electrical properties of the ultra thin gate oxide and inter-gate oxide are critical for data retention. The typical data retention of an EEPROM is 10 years.

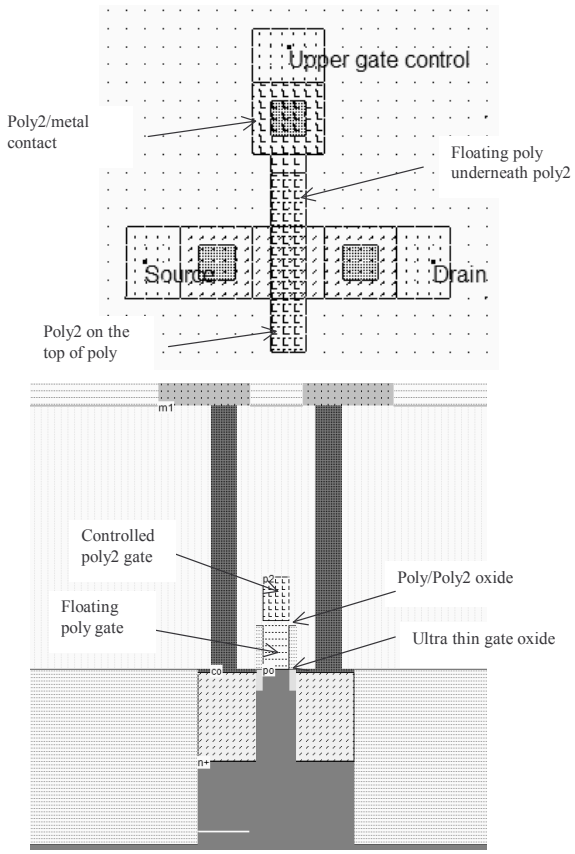
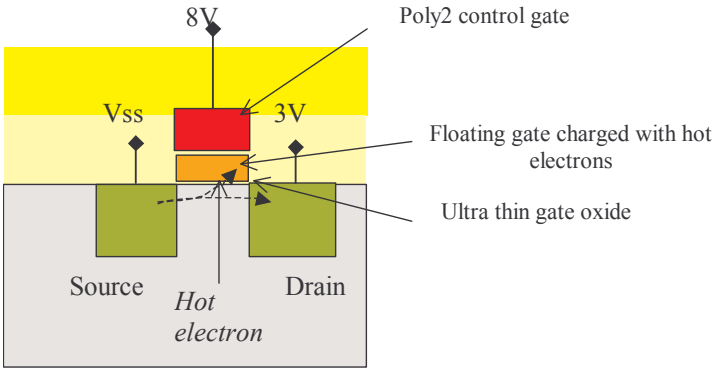


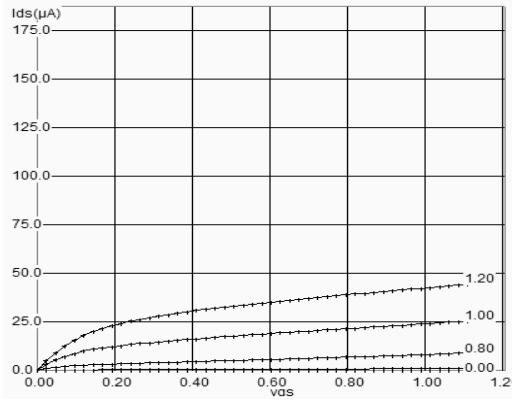
Fig. 3-21: The double gate MOS generated by MICROWIND3 (Eeprom.MSK)

The double gate MOS layout is shown in figure 3-21. The structure is very similar to the n-channel MOS device, except for the supplementary *poly2* layer on top of the polysilicon. The lower polysilicon is unconnected, resulting in a floating node. Only the *poly2* upper gate is connected to a metal layer through a *poly2/metal* contact situated at the top. The cross-section of figure 3-21 reveals the stacked *poly/poly2* structure, with a thin oxide in between.

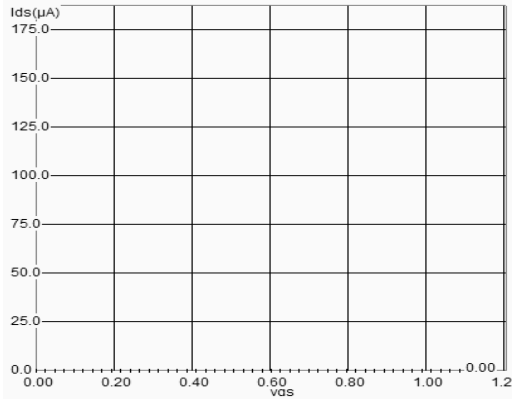
Double-gate MOS Charge

The programming of a double-poly transistor involves the transfer of electrons from the source to the floating gate through the thin oxide (Figure 3-22). Notice the high drain voltage (3V) which is necessary to transfer enough temperature to some electrons to become "hot" electrons, and the very high gate control to attract some of these hot electrons to the floating poly through the ultra thin gate oxide. The very high voltage varies from 7V to 12V, depending on the technology. Notice the "++" symbols attached to the upper gate and drain regions which indicate that a voltage higher than the nominal supply is used.





(a) Initial state: floating gate discharged



(b) Final state: floating gate charged

Fig. 3-22: Double-gate MOS characteristics without (a) and with charges (EepromCharge.MSK)

At initialization (Figure 3-22-a) no charge exists in the floating gate, resulting in a possibility of current when the poly2 gate voltage is high. However, the device is much less efficient than the standard n-channel MOS due to an indirect control of the channel. The maximum current is small but significant. The programming operation is performed using a very high gate voltage on poly2, here 8V. The mechanism for electron transfer from the grounded source to the floating polysilicon gate, called tunneling, is a slow process. In MICROWIND3, around 1000ns are required. With a sufficiently positive voltage on the poly2 gate, the voltage difference between poly and source is high enough to enable electrons to pass through the thin oxide. The electrons trapped on the floating gate increase the threshold voltage of the device, thus rapidly decreasing the channel current. When the gate is completely charged, no more current appears in the *Id/Vd* characteristics (Figure 3-22-b).

Double-gate MOS Discharge

The floating gate may be discharged by ultra violet light exposure or by electrical erasure. The U.V. technique is a heritage of the EPROM which requires a specific package with a window to expose the memory bank to the specific light. The process is very slow (Around 20mn). After the U.V exposure, the threshold voltage of the double gate MOS returns to its low value, which enables the current to flow again. In MICROWIND3, the command **Simulate → U.V exposure to discharge floating gates** simulates

the exposure of all double gate MOS to an ultra violet light source. Alternatively, the charge can be accessed individually using the command **Simulate→ Mos characteristics**. Changing the *Charge* cursor position modifies dynamically the MOS characteristics.

For the electrical erase operation, the poly2 gate is grounded and a high voltage (Around 8V) is applied to the source. Electrons are pulled off the floating gate thanks to the high electrical field between the source and the floating gate. This charge transfer is called Fowler-Nordheim electron tunneling (Figure 3-23).

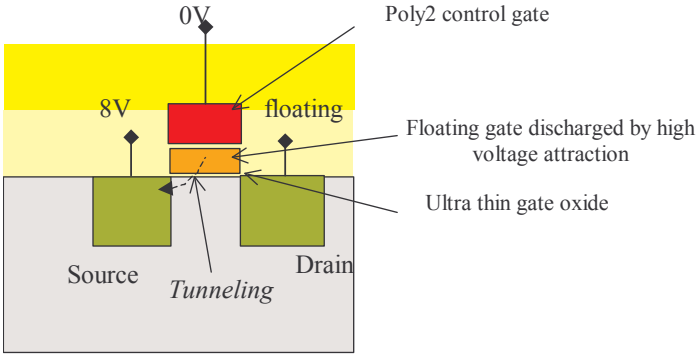


Fig. 3-23. Discharging the double gate MOS device (EepromDischarge.MSK)

The basic structure for reading the EEPROM information is described in the schematic diagram of figure 3-24. After a precharge to VDD, and once *WL* is asserted, the bit line may either drop to VSS if the floating gate is empty of charges, or keep in a high voltage if the gate is charged, which disables the path between *BL* and the ground through the EEPROM device. In the case of figure 3-24 left, the floating gate has no charge, so *BL* is tied to ground after the precharge, meaning that *DataOut* is 1.

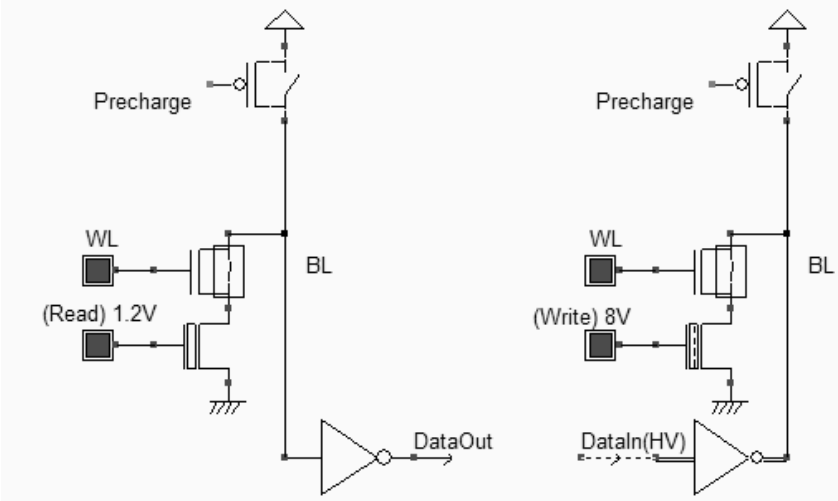


Fig. 3-24 Reading and writing in the EEPROM (Eeprom.MSK)

The write operation consists in applying a very high voltage on the gate (8V), and to inject a high or low state on *BL*. A zero on *DataIn* is equivalent to a high voltage on *BL,* which provokes the hot electron effect and charges the floating gate. In contrast, a one on *DataIn* keeps *BL* low, and no current flows on the EEPROM channel. In that case, the floating gate remains discharged.

3.6 Flash Memories

Flash memories are a variation of EEPROM memories. Flash arrays can be programmed electrically bit-by-bit but can only be erased by blocks. Flash memories are based on a single double poly MOS device, without any selection transistor (Figure 3-25). The immediate consequence is a more simple design, which leads to a more compact memory array and more dense structures. Flash memories are commonly used in micro-controllers for the storage of application code, which gives the advantage of non volatile memories and the possibility of reconfiguring and updating the code many times.

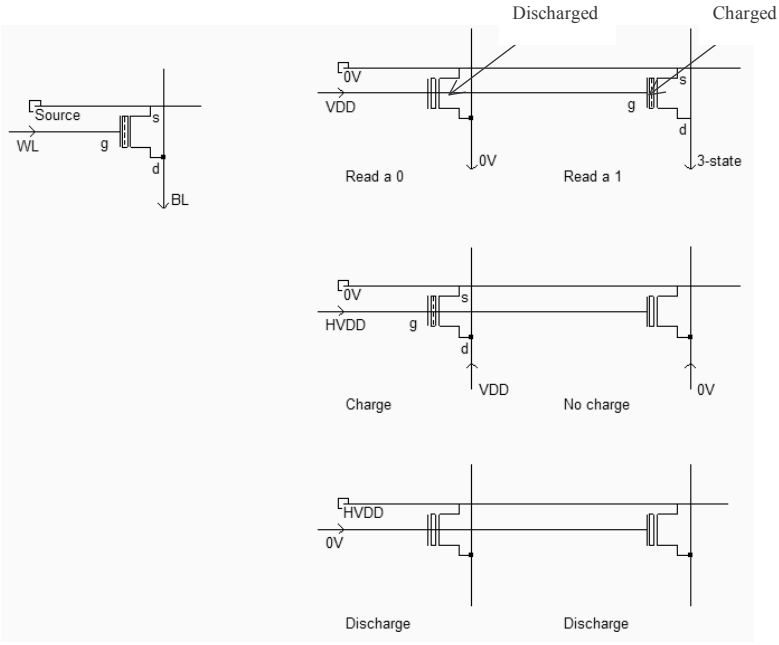


Fig. 3-25 : The flash memory point and the principles for charge/discharge (FlashMemory.SCH)

The Flash memory point usually has a "T-shape", due to an increased size of the source for optimum tunneling effect [1]. The horizontal polysilicon2 is the bit line, the vertical metal2 is the word line which links all drain regions together. The horizontal metal line links all sources together. It is a common practice to violate usual design rules, in order to achieve a more compact layout. In the case of figure 3-26, the poly extension is reduced from 3 lambda to 2 lambda.

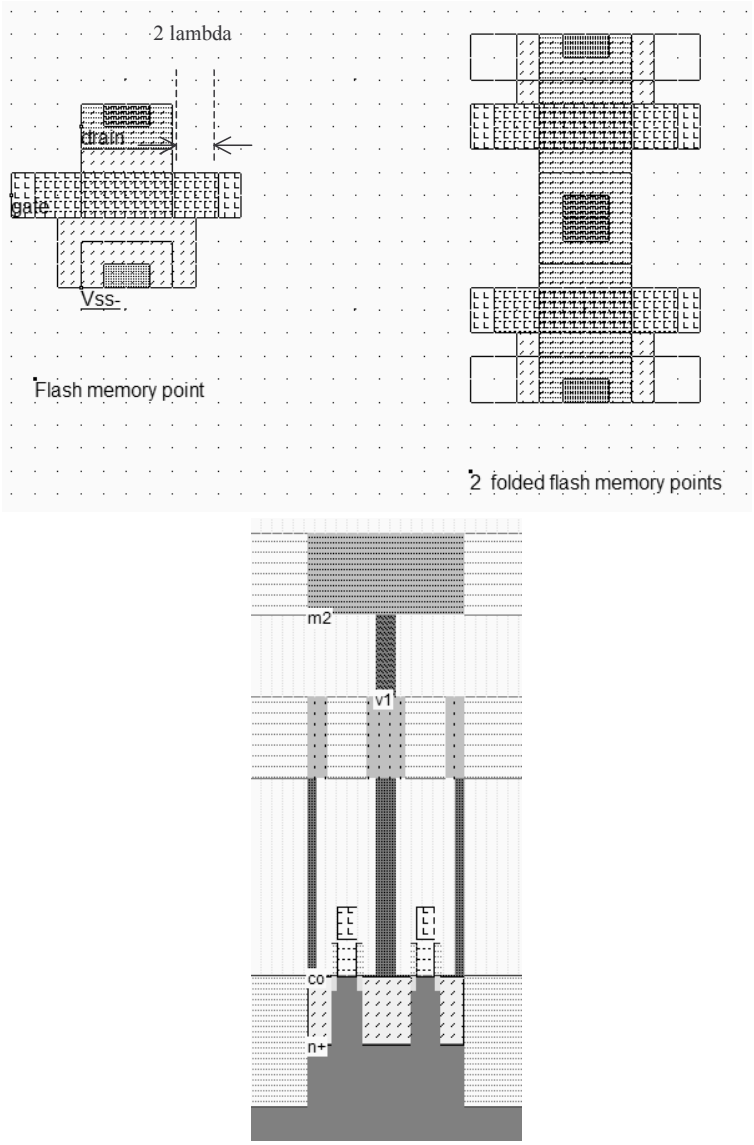


Fig. 3-26 The flash memory point and the associated cross-section (Flash8x8.MSK)

3.7 Ferroelectric RAM memories

Ferroelectric RAM memories are the most advanced of the Flash memory challengers [2]. The FRAM is exactly like the DRAM except that the FRAM memory point is based on a two-state ferroelectric insulator, while the DRAM relies on a silicon dioxide capacitor. Mega-bit FRAM are already available as stand alone products. However, FRAM embedded memories have been made compatible since the 90nm CMOS technology. The MICROWIND3 software should first be configured in 90nm to access the FRAM properties using the command **File → Select Foundry**.

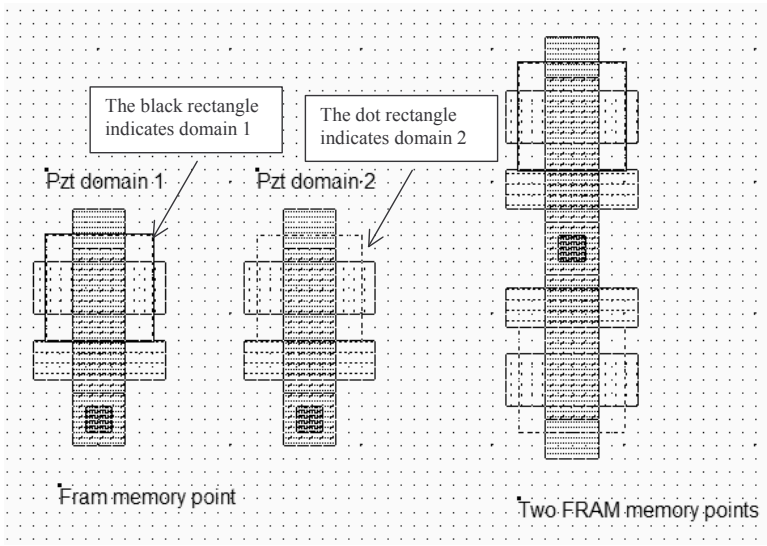


Fig. 3-27. The two domains of the FRAM memory (FramCell.MSK)

The 2D cross-section (Figure 3-28) shows the ferroelectric crystalline material made from a compound of lead, zirconium and titanium (PZT). The chemical formulation of PZT is $PbZr_{1-x}Ti_xO3$. Adjusting the proportion of zirconium and titanium changes the electrical properties of the material.

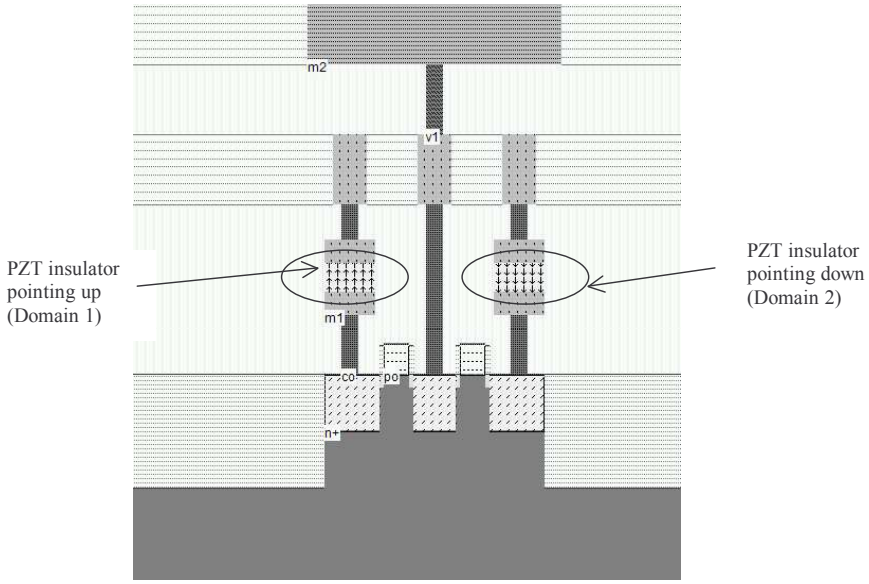


Fig. 3-28. The two domains of the FRAM memory (FramCell.MSK)

The PbTiO3 molecular structure is given in figure 10-90. It is equivalent to a cube, where each of the eight corners is an atom of lead (Pb). In the center of the cube is a titanium atom, which is a class IVb element, with oxygen atoms at its ends, shared with neighbors. The two stable states of the molecule are shown in the figure 3-29. The titanium atom may be moved inside the cell by applying an electrical field. The remarkable properties of this insulator material are: the stable state of the titanium atom even without any electrical field, the low electrical field required to move the atom, and its very high dielectric constant (Around 100).

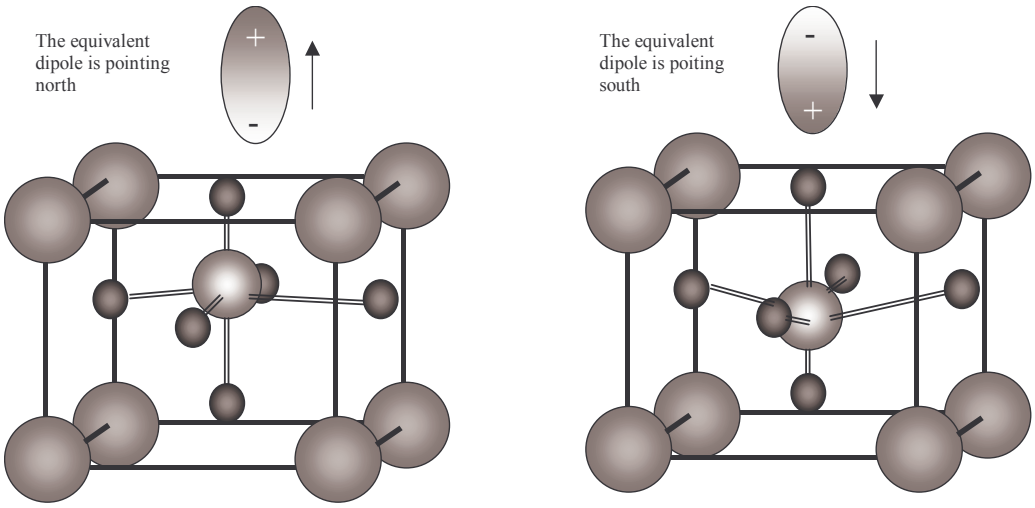


Fig. 3-30. The two domains of the structure which change the orientation of the equivalent dipole

The PZT capacitor behavior is usually represented by an hysteresis curve shown in figure 3-30. In the X Axis, the electrical field applied to the electrodes is displayed. The Y axis represents the dipole orientation for each molecule. It can be seen that if a minimum field is applied on the capacitor, the polarization changes. An inverted electrical field is required to change the state of the material.

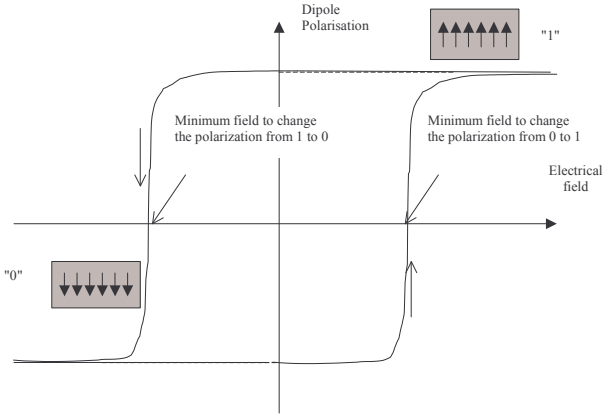


Fig. 3-20 : The hysteresis curve of the PZT insulator

Consequently, the write cycle simply consists, for a 1, in applying a large positive step which orients the dipoles north, and for a zero in applying a negative voltage step, which orients the dipoles south (Figure 3-31).

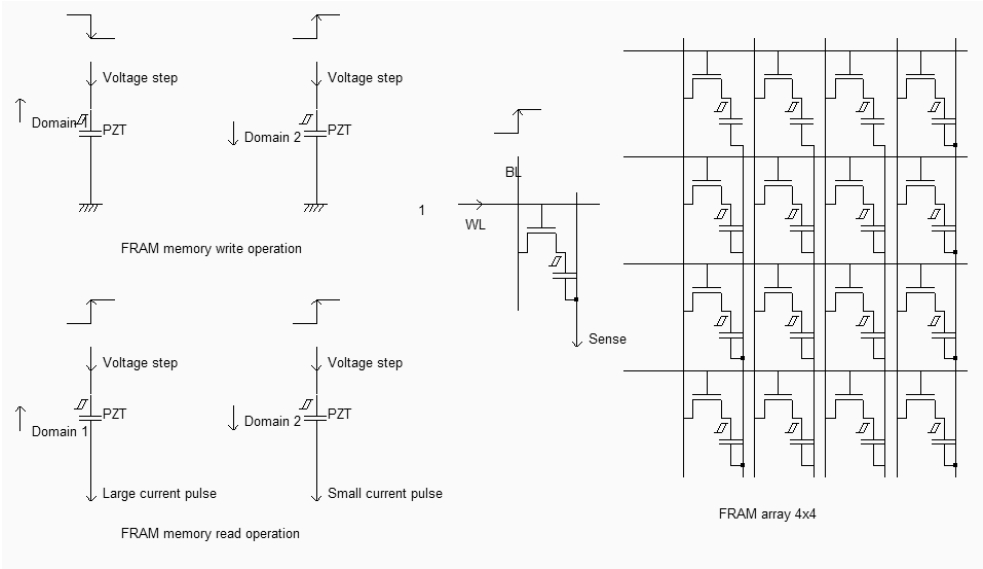


Fig. 3-31 The FRAM circuit principles and architecture (Fram4x4.SCH)

To read the domain information, an electrical field is applied to the PZT capacitor, through a voltage pulse. If the electric field is oriented in the opposite direction of the elementary dipole and is strong enough, the inner atom orientation is changed, which creates a significant current which is amplified and considered as a 1. If the electric field is oriented in the same direction as the elementary dipole, only a small current pulse is observed which is considered as a 0. Reading the logical information is equivalent to observing the current peak and deciding whether the current peak is small (0), or large (1). Notice that the read operation destroys the data stored in the PZT material, as for the DRAM. Just after the memory information is read, the logic information must be written back to the memory cell.

3.8 Memory Interface

All inputs and outputs of the RAM are synchronized to the rise edge of the clock, and more than one word can be read or written in sequence. The typical chronograms of a synchronous RAM are shown in figure 3-32. The active edge of the clock is usually the rise edge. One read cycle includes 3 active clock edges in the example shown in figure 3-32. The row address selection is active at the first rise edge, followed by the column address selection. The data is valid at the third fall edge of the system clock.

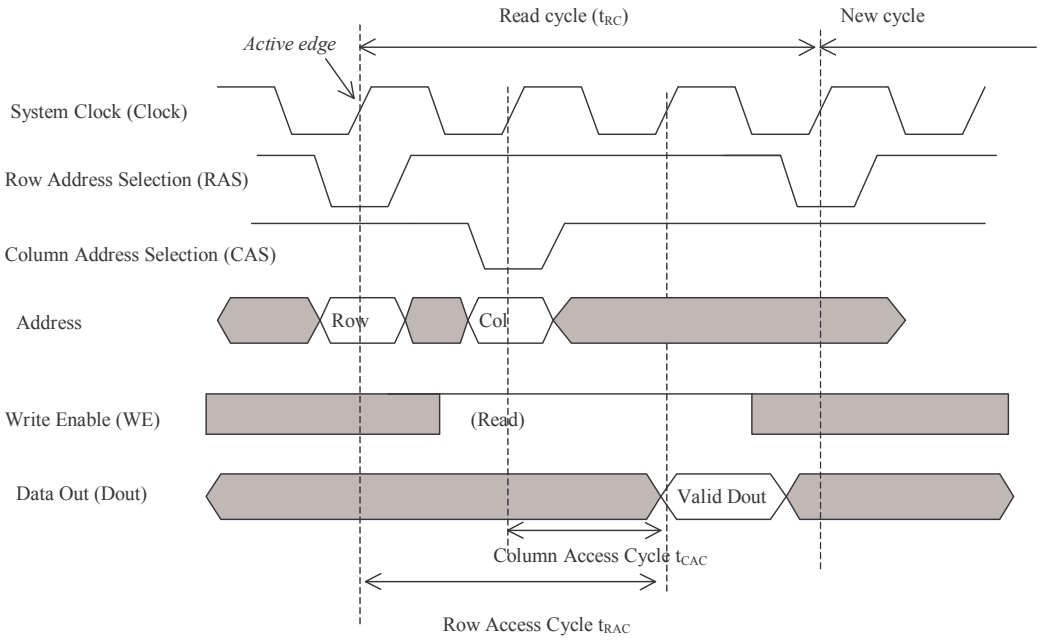


Figure 3-32: Synchronous RAM timing diagram

Double data Rate memories involve both the rise and fall edge of the clock [1]. Furthermore, a series of data from adjacent memories may be sent on the data bus. Two contiguous data are sent, one on the rise edge of the clock, the other on the fall edge of the clock. This technique is called "burst-of-two". An example of double data rate and burst-of-two data in/out is proposed in figure 3-33. Notice that *Data In* and *Data Out* work almost independently.

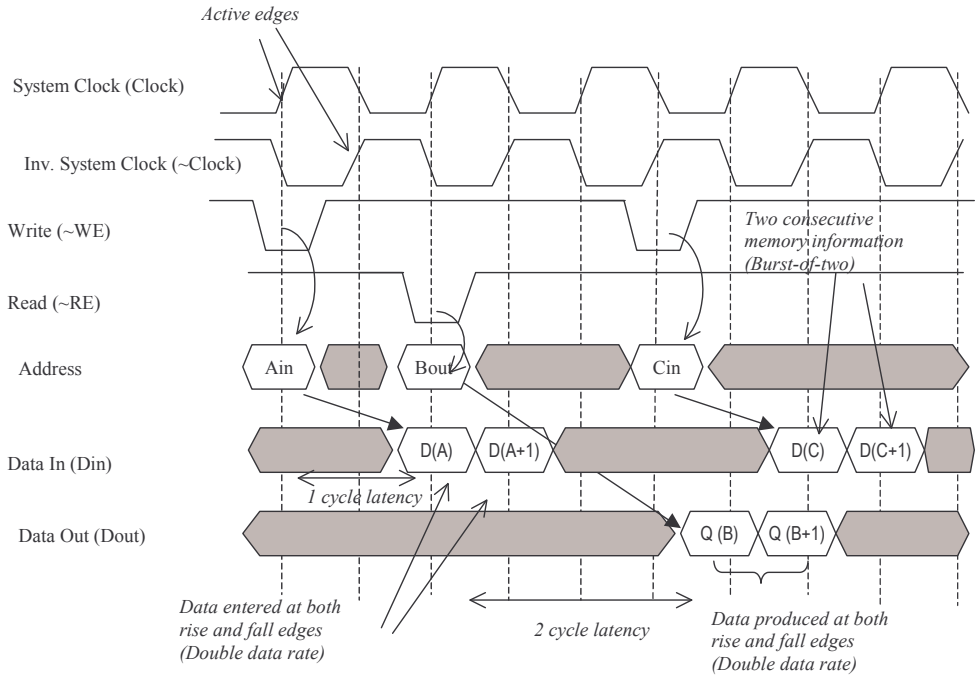


Figure 3-33: Double data rate diagram

3.9 EXERCISES

- Compare the leakage current on a DRAM cell for the following technologies : 0.35µm, 0.12µm and 90nm.
- Given a 4x4 EEPROM memory array, create the chronograms to write the words 0001, 0010, 0100 and 1000, and then to read these values.
- Modify the ROM array to write the word "Welcome".

References

[1] A. K. Sharma “Semiconductor Memories, Technology, Testing and Reliability”, IEEE Press, ISBN 0-7803-1000-4, 1997

[2] L. Geppert, “The New Indelible Memories,” IEEE Spectrum, v. 40, no. 3, Mar. 2003, pp. 49-54.