
The Dynamic Storage Cell

The main subassemblies of a dynamic random-access memory, or DRAM, are the same as those of a conventional random-access memory. However, the attention of researchers and designers focused particularly on those that made it specific, namely the storage cell and the sense amplifier. This chapter details the operating principle of its cell as well as the different structures that have been invented.

1.1. Reminder

The storage area of a random-access memory is a grid of storage cells (cross-point array architecture) with a unit capacity of one bit. A cell is accessed exclusively by coincidence addressing (see section VI-2.1). The address makes it possible to exclusively activate the row and the column at whose intersection this cell is located.

In read mode, the delivered electrical signal is interpreted by the sense amplifier (SA). In write mode, the information is sent into the cell.

1.2. Storage principle

Volume 2 (Darche 2026) studied the storage principle of SRAM, which is based on the self-maintenance of information through the use of a latch, or flip-flop (FF). In a DRAM, binary information is stored in a capacitor C_s (s for storage), or storage node (SN), in the form of an electrical charge (stored-charge memory) Q_s . This principle has been described in Holt (1952). The capacitance of the storage capacitor is here in the order of a few tens of femtoFarads (fF or 10^{-15} F). The number of electrons is in the order of 3×10^5 and 3×10^6 . Figure 1.1(a) presents the basic

electrical schematic of a memory cell. One of the connections of the capacitor C_s is connected to the substrate tied to ground, to the supply voltage V_{DD} , or to another potential, generally $V_{DD}/2$ (see below for the reason, after equation [1.3]), depending on the fabrication technology used. The connected line is called the plate line (CPL). The other connection, named SNC for SN contact, is connected to the cell select transistor (select transistor) T_s . This transistor, used as a controlled switch (see Figure 1.1(b)) driven by the wordline (WL), connects the cell to the data column, or bitline (BL). The contact point between the BL and the transistor connection is called the BL contact, or BLC. The inventor of this principle, Dr. Dennard, was employed by the company IBM (Dennard 1968).

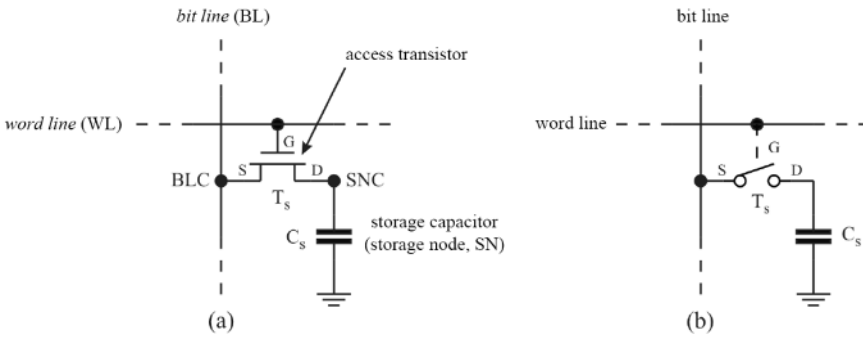


Figure 1.1. A memory cell with a transistor and a capacitor and its equivalent electrical diagram

The arbitrary logical equivalence is as follows (positive logic):

$$V_S = 0 \text{ V} \Leftrightarrow \text{“logic 0” stored} \quad [1.1]$$

$$V_S = V_{DD} - V_{GS(th)} \approx V_{DD} \Leftrightarrow \text{“logic 1” stored} \quad [1.2]$$

where $V_{GS(th)}$ is the conduction threshold voltage between the gate and the source of the select transistor.

The charge Q_{C_S} stored in the capacitor C_S is equal to:

$$Q_{C_S} = C_S \times V_{C_S} \quad [1.3]$$

In read mode, in the ideal case, it would then suffice to interpret the voltage of the data column once the cell is connected to it through the select transistor T_s ,

which is conducting (see Figure 1.2). Unfortunately, this column also exhibits a parasitic capacitance C_{BL} (connections of this parasitic component shown as dashed lines in the figure). It is previously precharged to $V_{DD}/2$ in order to limit the voltage excursion of the column and thus reduce the read time (Lu and Chao 1984).

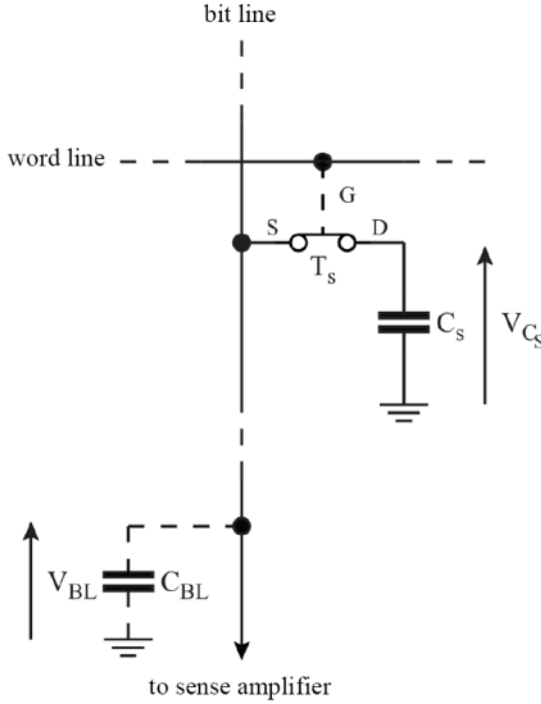


Figure 1.2. Reading a cell

When the cell is selected, the two capacitors are in parallel. Their electrical charges are added. The total charge Q_{tot} is equal to:

$$Q_{tot} = (C_S \times V_{CS}) + \left(C_{BL} \times \frac{V_{DD}}{2}\right) = (C_S + C_{BL}) \times \left(\frac{V_{DD}}{2} + \Delta V\right) \quad [1.4]$$

The transistor T_s is considered ideal. There is therefore no series voltage drop. After the transient period during selection, the voltage $V_{BL} = V_{sense}$ is equal to V_{CS} . We could deduce that:

$$\Delta V = \left(\frac{C_S}{C_S + C_{BL}}\right) \times \left(V_{CS} - \frac{V_{DD}}{2}\right) \quad [1.5]$$

The charge transfer ratio T is defined as being equal to:

$$T = \frac{C_S}{C_S + C_{BL}} \quad [1.6]$$

This ratio generally varies between 5 and 25% depending on the stored information. For the 16 Kib generation, the ratio C_{BL}/C_S varied from 3 to 6. For the 64 Kib generation, this ratio increased to between 8 and 10. A simplification could be made by considering $C_{BL} \gg C_S$ in the expression of T [1.6]. We could then write:

$$T = \frac{C_S}{C_S + C_{BL}} \approx \frac{C_S}{C_{BL}} = T' \quad [1.7]$$

The difference in the data column voltage ΔV detected by the sense amplifier (bitline sense amplifier (BSA), or simply SA) is then equal to:

$$\Delta V = T \times \left(V_S - \frac{V_{DD}}{2} \right) \approx T' \times \left(V_S - \frac{V_{DD}}{2} \right) \quad [1.8]$$

Two values, $\Delta V(1)$ and $\Delta V(0)$, have to be calculated as a function of the state of the stored bit:

$$\Delta V(1) = \frac{C_S}{C_{BL}} \times \left(\frac{V_{DD}}{2} - V_{GS(th)} \right) \quad [1.9]$$

$$\Delta V(0) = \frac{C_S}{C_{BL}} \times \left(-\frac{V_{DD}}{2} \right) \quad [1.10]$$

The same derivation with the cell capacitor whose one terminal is connected to V_{DD} can be found in Cand et al. (1986).

Compared with a precharge of the column to V_{DD} in NMOS technology (negative (channel) metal oxide semiconductor, full V_{DD} bitline sensing scheme) or to ground (Gnd) in PMOS technology (positive (channel) MOS, grounded bitline sensing scheme), the choice of $V_{DD}/2$ (half- V_{DD} bitline sensing scheme) makes it possible to limit the voltage excursion (voltage swing) and thus the coupling noise between lines. Current spikes (di/dt) are also reduced, resulting in limited ground-line noise (bouncing noise) due to the inductance of the connections. The total consumption of the memory array is also reduced. The dielectric of the capacitor is less stressed because of a lower electric field. This value is, for example, chosen in 4 Kib NMOS memories (Stein et al. 1972a, 1972b; Foss and Harland 1975). The transition to a single supply voltage with the 64 Kib version resulted in the precharge voltage being returned to V_{DD} (Ahlquist et al. 1976a, 1976b). With CMOS technology (complementary MOS (Shimohigashi et al. 1980; Masuda et al. 1981; Shimohigashi et al. 1982a, 1982b)) and starting from a capacity of 1 Mib, it returned to half the supply voltage (Lu and Chao 1984). Figure 1.3 shows the

evolution of the column voltage as a function of time. The exponential-like shape of the signal is characteristic of the (dis)charging of a capacitor. For the 16 Kib generation, if the average storage voltage is in the order of one volt, the detected differential voltage is then in the order of 200 mV. For the 64 Kib generation, for an average storage voltage on the order of one volt, the detected differential voltage dropped to 80 mV. The sense amplifier then amplifies this differential voltage.

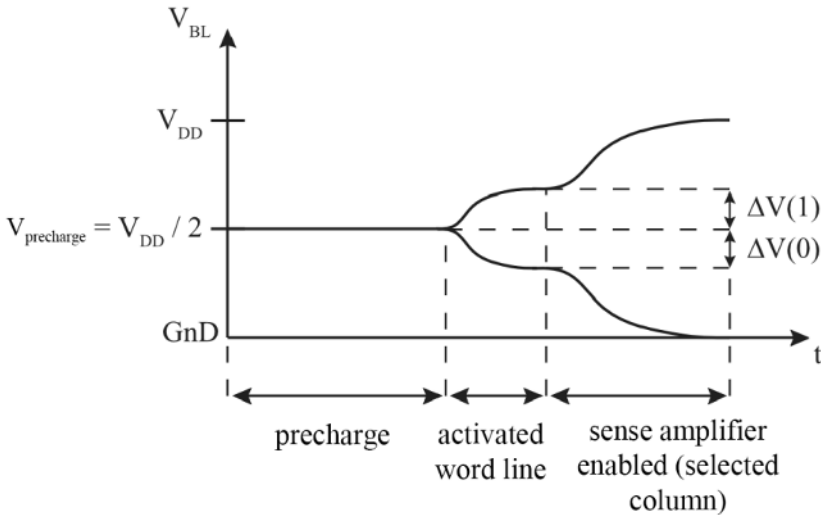


Figure 1.3. Evolution of column voltage as a function of time

The major drawback of this structure is that a read operation destroys the stored information (DRO, destructive read-out), because the capacitor discharges when a “logic 1” is stored. It is therefore necessary to rewrite it (write-back operation) from the information that was read by the sense amplifier. This operation is called regeneration (see the end of section 2.3). The cycle time is thus lengthened.

Moreover, since the capacitor and the select transistor are not perfect, there exists a leakage current in the order of a few nA that discharges it. It is therefore also mandatory to periodically refresh the electrical information by a read–rewrite operation (order of magnitude of the period: a few milliseconds to several tens of milliseconds).

The two operations, regeneration and refreshing of the information (see section 3.2.2), obviously have to be carried out as transparently as possible for the accessor, but they have a non-negligible time cost.

It should be noted that the parasitic capacitor may not be in a transistor. Hetherington et al. (1992) described a structure based on a thyristor with a storage time of 16 s at room temperature.

1.3. Cell structures

The cell of the first MOS-technology DRAM presented at the IEEE International Solid-State Circuits Conference (ISSCC) comprised three transistors (Regitz and Karp 1970a, 1970b), as shown in Figure 1.4. The storage capacitor C_S was originally the parasitic capacitance $C_{\text{Gate_Substrate}}$ that exists between the gate (G) and the substrate of a MOS transistor (transistor T_2 in the figure). To avoid two select lines, the gates of transistors T_1 and T_3 could be connected, as was the case for the memory developed by Regitz and Karp (1970a, 1970b). At the end of the column was the sense amplifier responsible for reading and amplifying the data signal (see Figure 1.4). T_2 isolated and amplified the information. Palfi (1971) also described a similar three-transistor cell.

As shown in Figure 1.5, wordline selection differs depending on whether access is performed in read or write mode. The designers, to avoid dissipating energy in static biasing circuits, used the parasitic capacitance of the column C_R (a value 100 times greater than that of a cell) as an energy reservoir. Before performing an access, the memory is precharged to a reference voltage V_{ref} by means of transistors T_4 and T_5 . Recall that this operation consists of connecting all columns to V_{DD} (in our example), to $V_{\text{DD}}/2$, or to another voltage reference, with all wordlines disconnected (i.e. in the so-called “high-impedance” state). This approach is called “dynamic” (see sections V1-2.2.1 and V1-2.2.2). During a row access, only one row is activated. At rest, there is therefore no current consumption. Consumption occurs only during precharge, when a current exists at column switching because of the parasitic capacitance of the precharged line. The operation of the memory is therefore faster. In write mode, the “wordline write” signal selects transistor T_1 , which passes the binary information from the write bitline into transistor T_2 , the latter playing the role of the storage cell. To read, the signal Φ switches transistors T_4 and T_5 , which precharge the bitlines by means of their parasitic capacitance C_{BL} . At the end of this operation, the cell is selected by the “wordline read” signal, which switches transistor T_3 of all the cells of a wordline. If the cell contains a “1” (storage capacitor C_S charged), the voltage V_{CS} makes T_2 conduct, which then discharges C_{BL} , the parasitic capacitance of the read bitline, through T_3 . The state of the column is the logical complement of that of the cell. It is then interpreted by the SA at the end of the column.

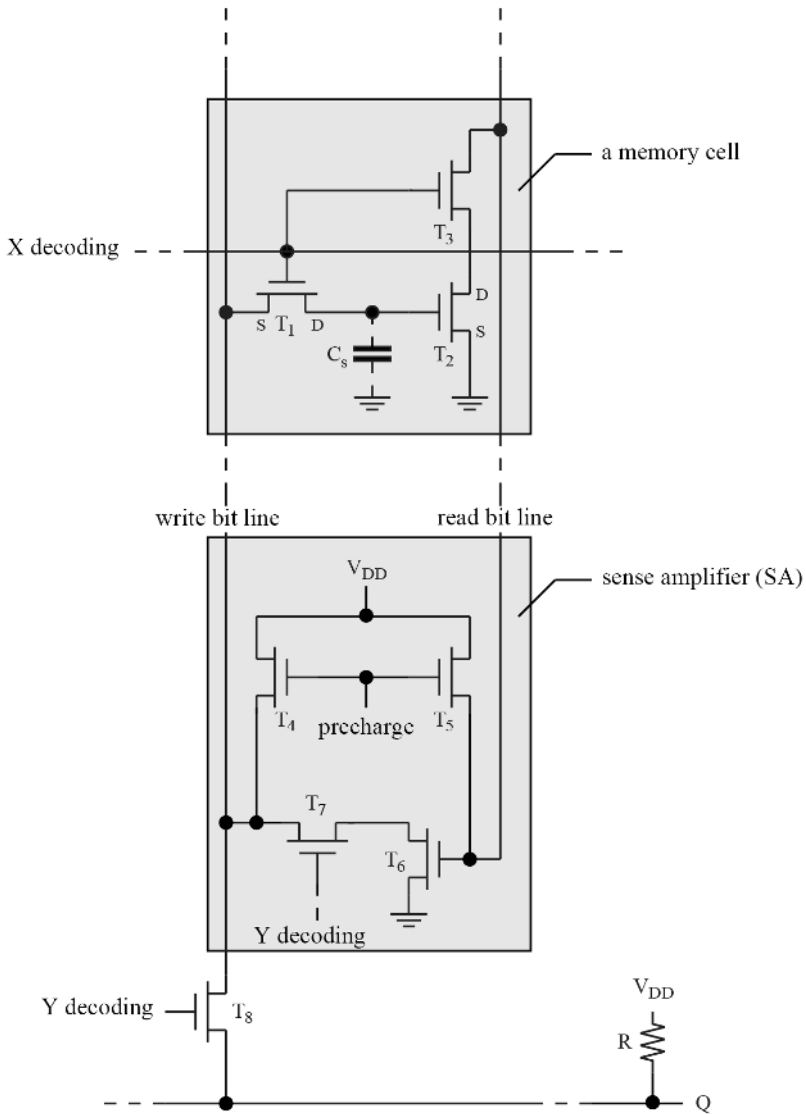


Figure 1.4. Electrical diagram of the memory cell of the first DRAM (according to Regitz and Karp (1970a, 1970b))

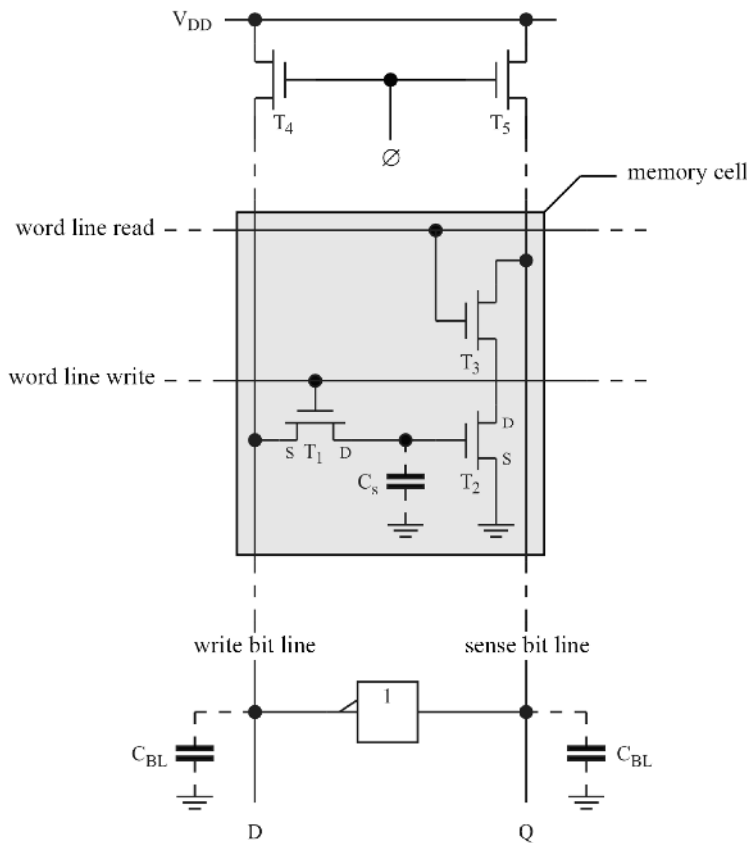


Figure 1.5. *Electrical diagram of the 3T memory cell of memory 1103 from Intel (P channel)*

Itoh (2001) classified 3T versions according to the number of lines required to control a cell, with the ground line shared by an adjacent cell counting as one half. He identified three categories, namely 4.5 (see Figure 1.6(a), industrial version), 3.5 with its versions (Martino and Croxon 1972; see Figure 1.6(b); Regitz and Karp 1970a; see Figure 1.7(a)), and finally 2.5 lines (Karp et al. 1972; see Figure 1.7(b)).

Figure 1.7 presents the sequence of cell categories in decreasing order of the number of lines.

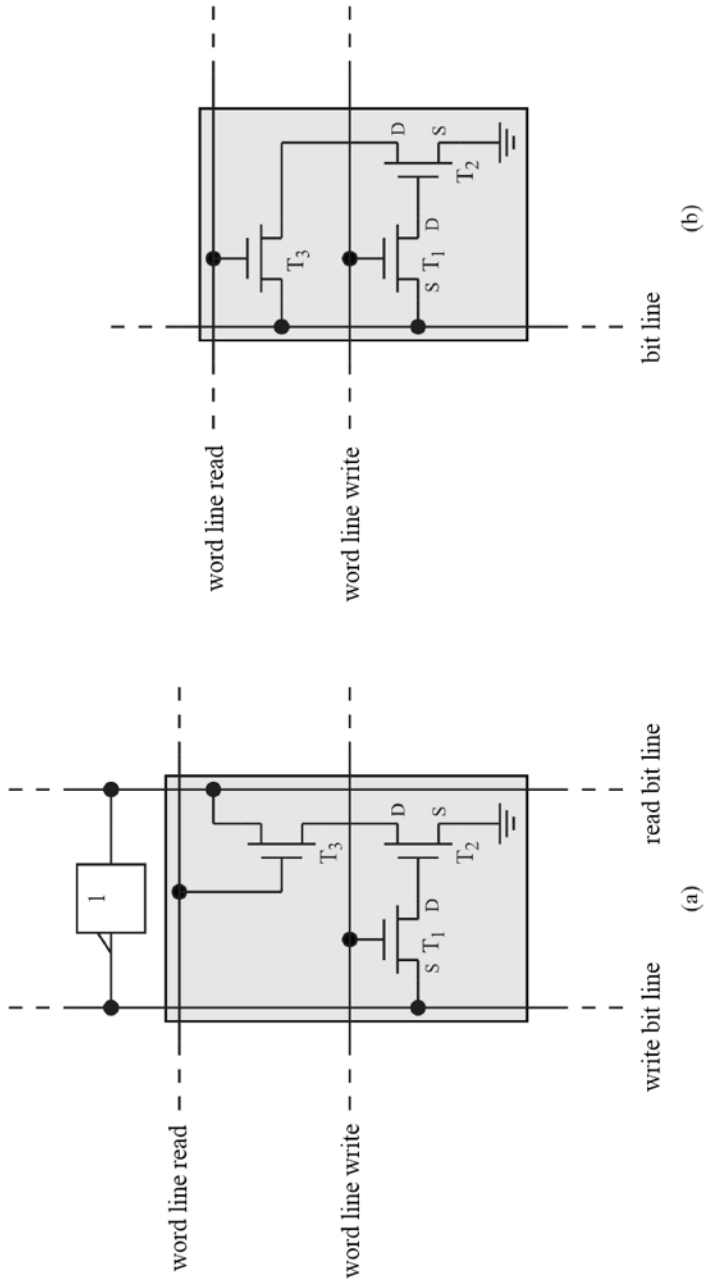


Figure 1.6. 3T cells with a) 4.5 lines and b) 3.5 lines (after Itoh (2001))

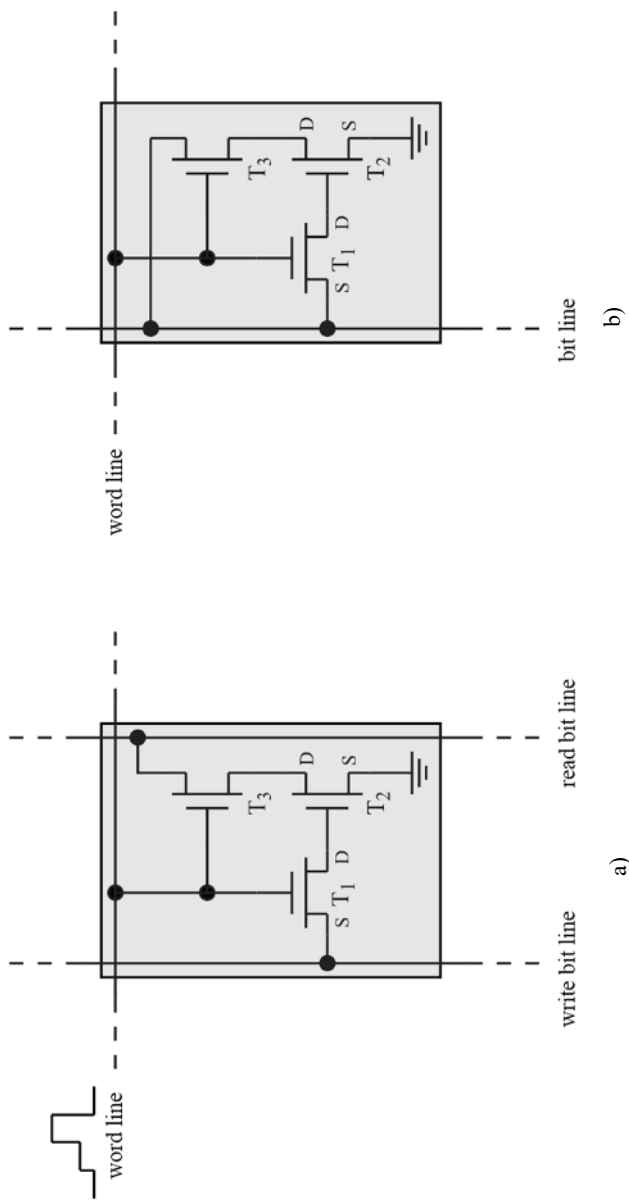


Figure 1.7. 3T cells with a) 3.5 lines and b) 2.5 lines (after Itoh (2001))

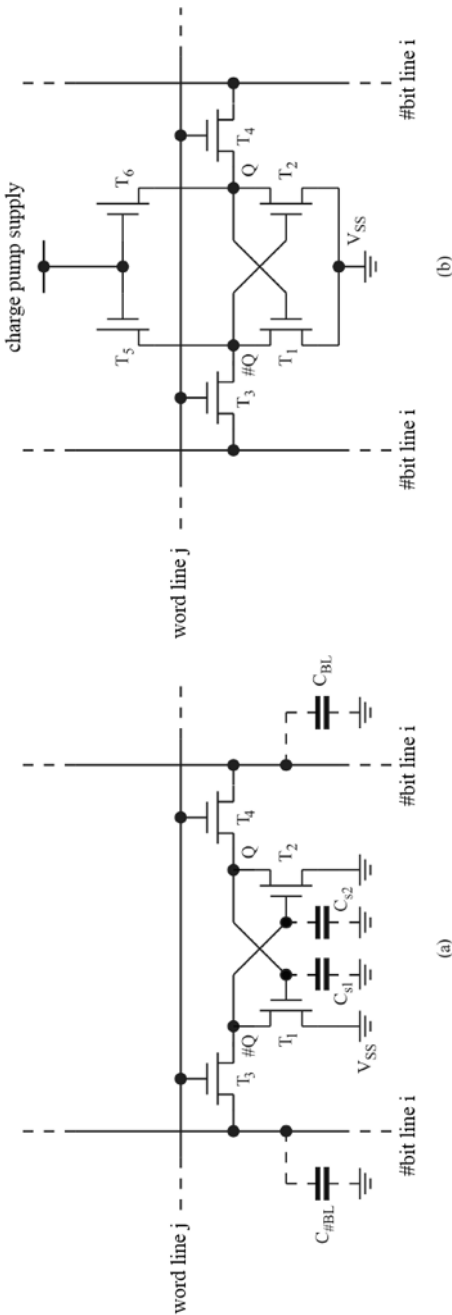


Figure 1.8. Electrical diagrams of a conventional 4T pump memory cell (after Burke and Michon (1972))

A simpler way to classify cells is to count the transistors and the rows and columns forming buses. For example, the 1103 memory has a 3T cell and four buses (see Figure 1.6(a)), whereas the one next to it is a 3T-3-bus cell. It is even possible, in counting bus lines, to distinguish data lines and bitlines and thus, in total, three parameters for the cell.

Another configuration with two storage capacitors could be considered, as shown in Figure 1.8(a). This is that of the AMS6002 memory. The connections of the parasitic capacitors C_S and C_{S2} (not shown) of transistors T_1 and T_2 , as well as those of columns C_{BL} and $C_{\#BL}$, are shown as dashed lines. Obviously, because of leakage currents, it is necessary to refresh the information. For this purpose, the normal and complemented bitlines are set to the high state. This configuration is interesting because it is possible to refresh the entire memory array rather than one row at a time (see section 3.2.2). To avoid having to bias both columns, one proposal is charge pumping (Brugler and Jespers 1969), which compensates for leakage currents (see Figure 1.8(b)).

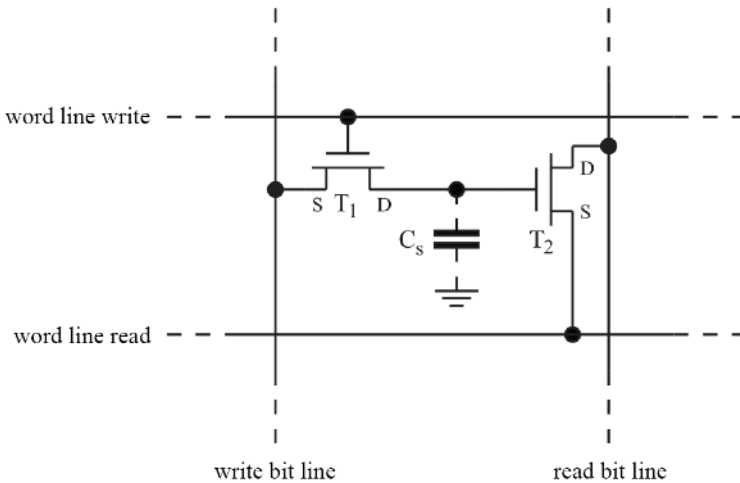


Figure 1.9. Dual transistor (2T) memory cell with separate read and write rows and columns

Figure 1.9 shows a two-transistor cell. Transistor T_1 is used as a switch for writing. Transistor T_2 is used as the memory cell (connections of the parasitic capacitance $C_S = C_{\text{Gate_Substrate}}$ shown as dashed lines) as well as a switch for reading. This second function made it possible to eliminate one transistor, which reduced the size of the cell. The input resistance of the transistor is in the order of $\text{G}\Omega$ ($10^9 \Omega$).

The RC time constant (product of resistance R and capacitance C , in seconds) is therefore in the order of a millisecond or several tens of milliseconds. This means that this parasitic capacitance, once charged and in the absence of a voltage, would discharge through the input resistance with an RC time constant.

There exist cells of the 2T–2C type (twin cell) that store one bit of information in complementary form in two cells. This makes it possible, among other things, to increase the noise margin at the cost of a reduction in cell density.

For more information, see, for example, Matsumiya et al. (2002) and Tsukikawa and Ito (2005).

We could also cite the 3T–1D cell proposed by Luk et al. (2006), which enables non-destructive readout (see Figure 1.10). A MOSFET in a gated-diode (GD) configuration is both the storage node (SN) and an amplifier (Luk and Dennard 2005). Transistors T_1 and T_2 control writing and reading, respectively. The two bitlines could be connected if only a single port is required.

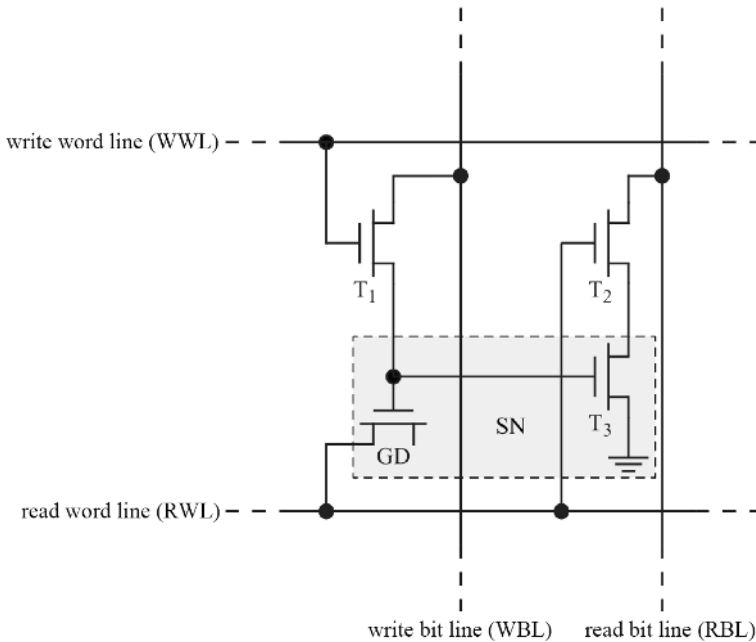


Figure 1.10. 3T–1D memory cell

1.4. Storage capacitor

Originally, the capacitor was a parasitic element of a transistor. Starting with the 4 Kib generation (Boonstra et al. 1973), and as described in the original patent of Dennard (1968), it was implemented as a component in its own right, with a transistor connecting it to the bitline. With the disappearance of this “memory transistor”, density increased. However, the absence of the amplification transistor (T_2 in Figure 1.9) meant that the read operation then became destructive. Several capacitor technologies were developed in order to reduce or eliminate the surface area occupied by this component. Each of them was developed for a given range of capacitance (see Figure 5.3). But despite scaling down, the capacitance of C_S had to remain constant in order to maintain an acceptable noise margin. We may cite the stacked capacitor, with a capacitance of 15–20 fF, and the trench capacitor, with a capacitance of 30–40 fF. These structures are described in section 5.5.

1.5. Leakage currents

Physically, leakage currents exist in transistors. They do not pose a problem in a static cell because the information is self-maintained. By contrast, they can discharge the storage capacitor. In the MOSFET, we must distinguish subthreshold (switching-threshold) leakage currents and gate leakage currents. The most critical situation is when C_S is charged. These currents increase with temperature and with supply and control voltages. They also depend on the technology node (lithographic feature size; see section 1.5 of Darche (2021)).

The subthreshold leakage current subthreshold conduction or subthreshold leakage, is the current that flows between the drain and source when the MOSFET operates in the subthreshold or weak-inversion region. In our case, this concerns T_1 , and the current $I_{DS(sub)}$ flows in the write line. The gate leakage current $I_{DS(sub)}$ is the current that passes through the gate insulator gate-oxide leakage. Here, this again concerns T_1 and also T_2 .

These leakage currents make it necessary to periodically regenerate the charge of C_S , whether it is parasitic or a component in its own right. By selecting a row, all cells are refreshed in parallel. Each column has a refresh amplifier that amplifies, with inversion, the signal of the read bitline in order to reapply it to the write bitline. Total memory refresh consists of performing this operation for each row (see section 3.2.2).

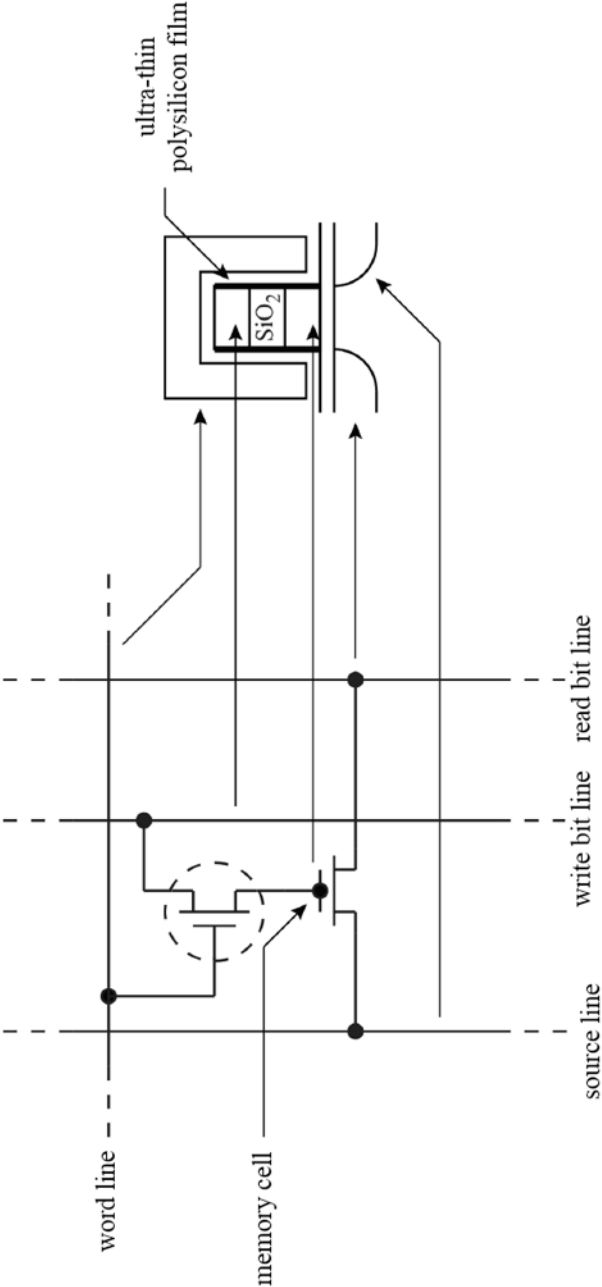


Figure 1.11. Equivalent electrical diagram and schematic section of a SESO-type memory cell

Osabe et al. (2000) proposed improving leakage current by means of the SESO (single-electron shut-off) transistor. The latter exhibits a leakage current of 10^{-19} A, i.e. less than one electron per 100 ms. It is a field-effect transistor, or FET, with a drain–source connection made of polysilicon deposited in an ultra-thin layer (thickness of 2 nm). Figure 1.11 presents the electrical schematic of an SESO cell, the SESO transistor being marked by the dashed circle. A prototype memory cell (normalized size of $4\text{--}6\text{ F}^2$) produced by the company Hitachi exhibits a retention time of 3,000 s before refresh for a total node capacitance of 6 fF. It foreshadowed future cells that would use a single electron to store one bit (see Volume 6).

1.6. Conclusion

The storage principle of a DRAM is the storage of electrical charges in a capacitor. Since the latter is not perfect, a discharge current exists that destroys the information. To stem this electron leakage, it is necessary to refresh, i.e. to regenerate the information by periodically rewriting it. Much research effort has focused on improving this capacitor in terms of capacitance and leakage current. Thanks to this storage principle, the price per bit is lower than for SRAM because its cell is smaller, 1T-1C versus 6T or 4T-2R. It is, however, slower than its counterpart because of its storage principle.

Chapter 2 presents the internal organization of the DRAM.