Computing with SRAMs

Static RAM (SRAM) has a shorter data access latency compared to DRAM.

SRAM is a matured memory technology.

Data in SRAM is also volatile so SRAM is a type of temporary data storage.

But periodic refreshing is not needed.

SRAMs are used for the on-chip cache and register files of most modern CPUs.

SRAM has larger and more complex bit cells compared to DRAM.

SRAM cells are arranged in an array for efficient data access.

In an SRAM array, an wordline connects to all cells in a row via access transistors.

Likewise, *bitline* and $\overline{bitline}$ spans one column and connects to all cells in the column.

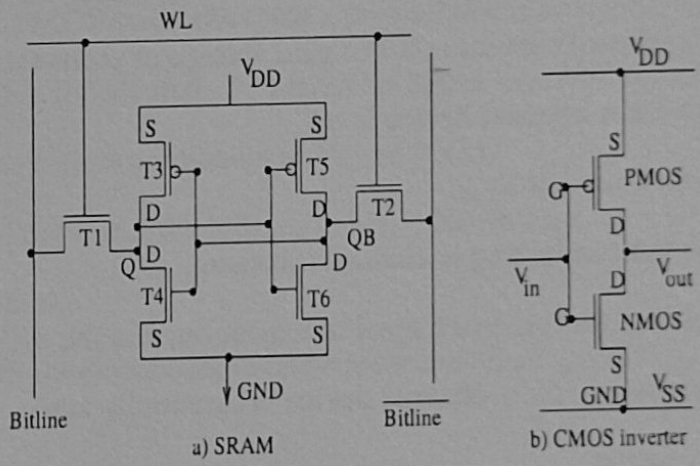


Figure 25: 6-T SRAM cell

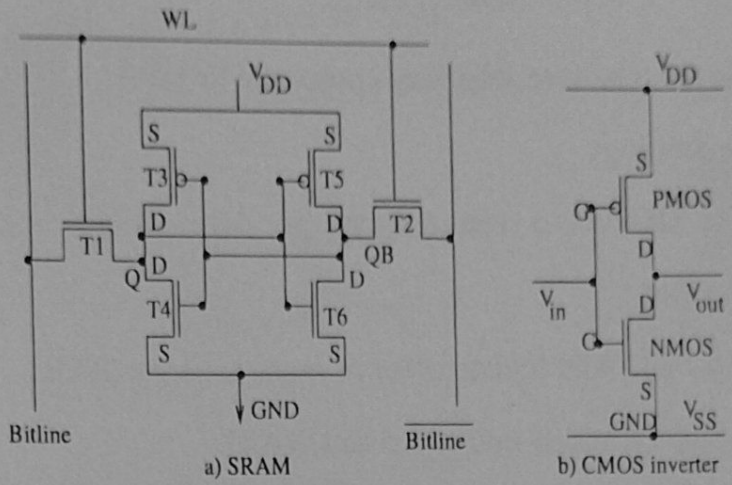
Some peripheral logic is placed around each SRAM array, assisting data access from the array.

Row address decoder: to activate the correct wordline as per row address.

The bitline peripherals include sense-amplifiers, bitline prechargers, and write drivers.

Typical SRAM has a 6-transistor cell structure as shown in Figure 25(a).

Cell is made of a pair of cross-coupled inverters (Figure 25(b)) and two access transistors.



In Figure 25(b), the drains (D) of two transistors are connected together and form output.

The input terminal is the common connection to transistor gates.

p-channel device has a negative threshold gate voltage V_T, but for n-channel V_T is positive.

If $V_{in} = 0$, gate voltage for the n-channel device is 0, but gate voltage of p-channel device is $-V_{DD}$.

Thus p-channel device is on, but the n-channel device is off.

Therefore, full voltage V_{DD} is measured at V_{out}.

That is, V_{DD} appears across the nonconducting n-channel transistor.

Alternatively, a positive value of V_{in} turns the$n$-channel transistor on, and the p-channel off.

The output voltage measured across the 'on'n-channel device is essentially zero.

Thus, the circuit operates as an inverter.

The beauty of this circuit is that one of the devices is turned off for either condition.

Since the devices are connected in series, no drain current flows

except for a small charging current during the switching process from one state to the other.

In SRAM, for each column the *bitline* and $\overline{bitline}$ pair is connected to a precharge circuit.

Function of it is to pull-up the bit lines to V_{DD} level and perfectly equalized them.

A typical precharge circuit is shown in Figure 26(a).

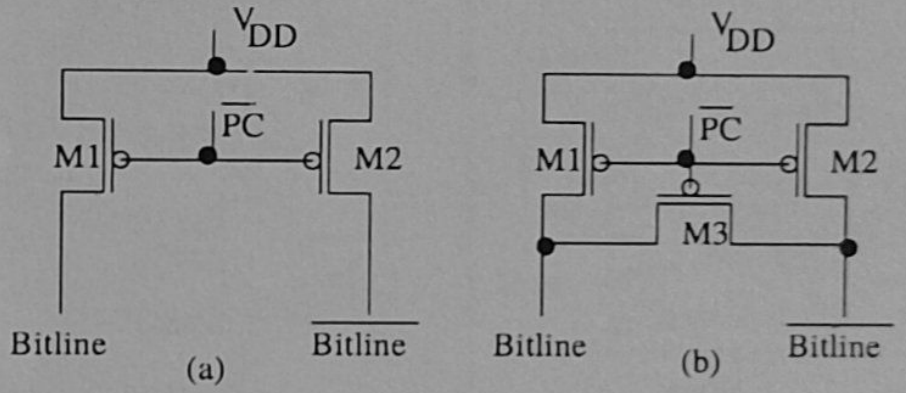


Figure 26: Precharge circuit

It is composed of a pair of PMOS transistors and a precharge circuit enable signal (PC).

When both the transistors are in ON state, that is, PC is active low, bitlines are connected to V_{DD}.

Recent 3-transistor precharge circuit is shown in Figure 26(b).

In Figure 26(b), M1 and M2 connect the bitlines to V_{DD} for pull-up.

M3 equalizes both the bitlines.

In precharge circuit, PMOS transistors are used because they have good V_{DD} passing capacity.

The Cross-coupled inverters in Figure 25(a), forms a latch and holds the binary information.

If output of one inverter is high (corresponding to the nMOSFET being OFF, and pMOSFET ON).

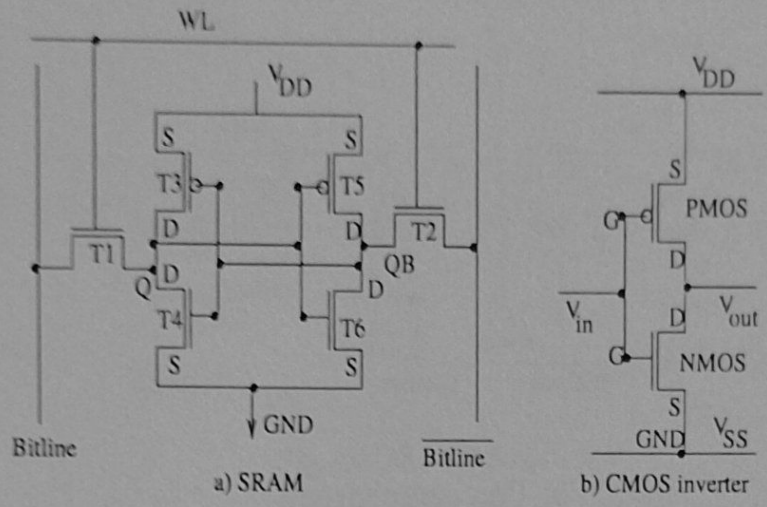


Figure 27: 6-T SRAM cell

That high voltage is fed to input of the other inverter, and output of the other inverter will be low.

This is one logic state (say '1') of the SRAM.

Conversely, the other stable state of latch can be considered to be the other logic state (say '0').

Cell is accessed through two access transistors whose gates are controlled by the wordline.

Access transistors also provide isolation from the other neighbouring circuits during hold state.

Read

Let assume that the internal data storage at nodes Q and QB are 0 and 1 respectively.

To read the cell content

The *bitline* and $\overline{bitline}$ are both precharged to the same voltage (say, V_{DD}).

In some SRAM designs bitlines are precharged to intermediate level of 0 and V_{DD}.

Now, the wordline (WL) is set to high -that is, access transistors are turned ON.

When WL rises from 0 to 1, the precharged *bitline* is discharged through the T1 and T4.

A voltage differential now develops between *bitline* and $\overline{bitline}$

The sense amplifier converts the differential signal to a logic-level output (here it is 0).

The voltage differential is amplified until the voltage separation is V_{DD}.

Write

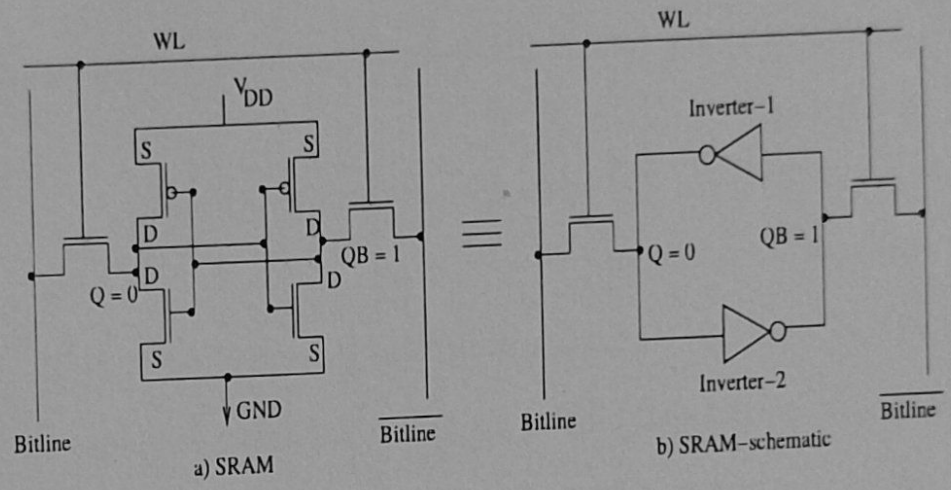


Figure 28: SRAM write

Let assume that internal nodes Q and QB are at 0 and 1 (Figure 28).

Following sequence of steps are for write 1 to the cell

Initially, wordline (WL) = 0.

Precharge the bitlines to V_{DD}.

After prechrage, both the bitlines are disconnected from V_{DD}.

Wordline is activated to high (data enters the bitcell during this step).

Place data value (here it is 1) on *bitline* and the complementary data value (that is, 0) on $\overline{bitline}$.

$\overline{bitline}$ connected to storage node QB via access transistor, is driven to ground by write driver.

While bitline is remained held at V_{DD} to pull node Q to high via access transistor.

As Q and QB flip their states de-assert the wordline (WL) back to 0.

A quantitative measure of stability of SRAM cell is the noise margin

Which is the maximum level of electrical noise that does not cause data corruption in cells.

Digital computing

The logical operations in SRAM perform element-wise logical operations.

These are AND, OR, NOR on the data stored on multiple wordlines in an SRAM array.

In-SRAM computing can be implemented by

activating the wordlines simultaneously and sensing on the pair of bitlines.

At a compute cycle, two wordlines are activated.

Figure 29 provides a summary of the implementation of in-SRAM logical AND/NOR operations.

If one bit-cell stores the data 0, the precharged bitline will be pulled down to below V_{ref}.

If all the bit-cells store the data 1, the bitline will remain at the precharged high value.

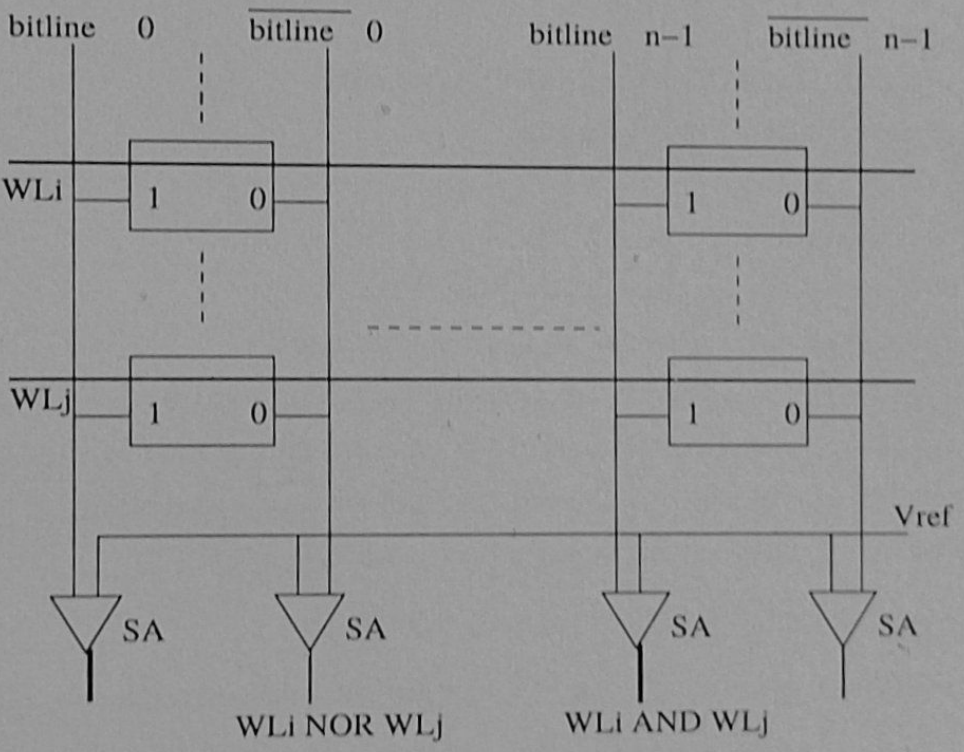


Figure 29: In-SRAM logical AND and NOR operations

By sensing voltage on bitline with a single-ended sense-amp

the logical AND can be computed on the data stored in the bit-cells.

Similarly, logical NOR can be computed by sensing the $\overline{bitline}$, as only when all the data are 0

The complement data are all 1 and the $\overline{bitline}$ will remain high.

The extra hardware required includes:

Additional row decoder to activate two wordlines

Single-ended sense-amps (as opposed to the differential sense-amps in a normal SRAM array).