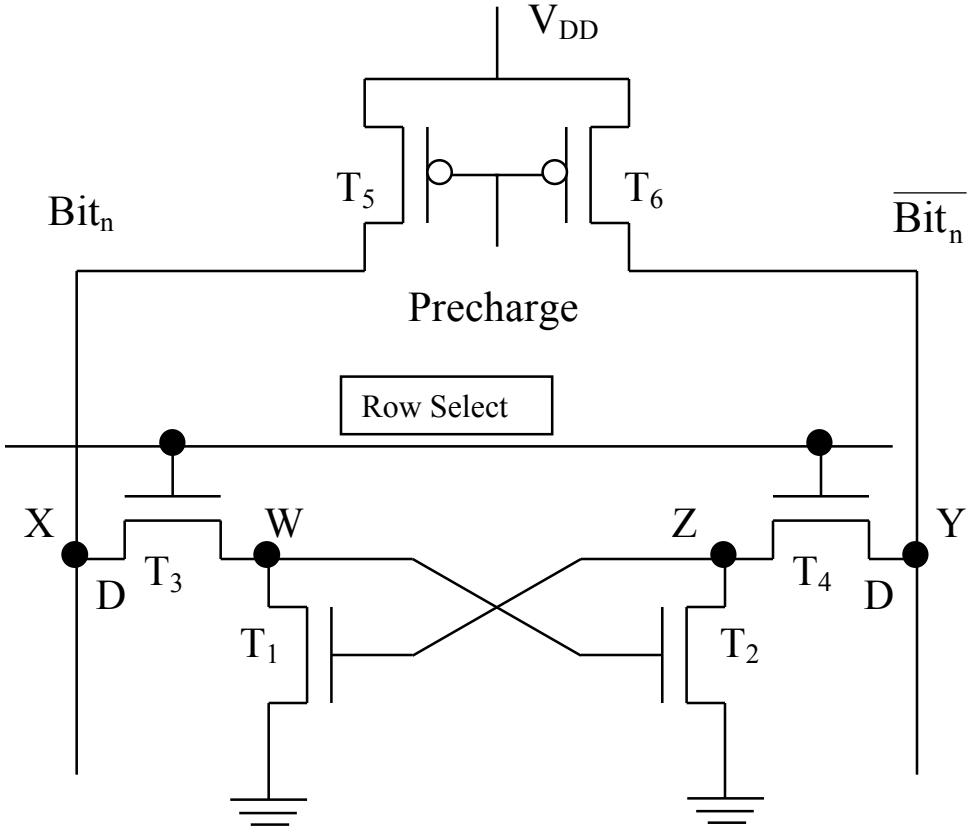
Memory Systems

The emphasis of this section will be on the basic understanding of dynamic memory cells. Only DRAM with 1 T will be discussed in more details. The 4 transistor dynamic memory cell is as shown with precharge transistors as it helps to understand SRAM cell.



Before any read write operation is performed, both the bit_n (Bit_n) and $\overline{bit_n}$ ($\overline{Bit_n}$) lines are changed to V_{DD}. The transistors T_5 and T_6 are doing this precharging. The write operation is now performed by turning column select line on (This part of the circuit is not shown). One of the bit_n and $\overline{bit_n}$ data is now pulled low. Still there is no storage done. To affect the storage, let us assume that $bit_n = 0$. Now we turn on row select line.

The way T_1 and T_2 are connected, they will be forced in complementary states. Hence both T_1 and T_2 cannot be on at the same time. In this case, ϕ is written at W and 1 is written at Z. Now when row select line goes low, both T_3 and T_4 are turned off and information written is held by the gate capacitances C_{g1} and C_{g2}. These capacitances are discharging through the leakage currents of many reversed bias junctions and should be high enough to hold their values until at least the next refresh is reached.

To read stored data, both bit lines are again precharged to 1. Since Z was holding 1, T_1 is on. When row select signal goes high, T_3 is also turned on and Bit_n line is discharged to ϕ through T_1 and T_3. $\overline{Bit_n}$ line is still held at 1. Charge sharing issues are important for all RAM cells, but they will be discussed at the end.

6 T Static RAM Cell

The cell schematic is as shown. The sense amplifier which aids discharging of bit lines is as shown. The p-channel transistors present in the SRAM cell help to hold the written information indefinitely by supplying the lost charge. As the lost charge is small, they can have any reasonable size. By adding a sense amplifier in each column of bits, the memory cell size can be reduced as sense amplifier is used to discharge the bit lines appropriately. During precharge, the sense signal is low and the sense amplifier plays no role. The precharging can be done very fast by selecting large p channel devices. For discharging a bit line, sense signal is turned high so that the

line is discharges quickly (bit_n in our case). Also, low bit_n turns on T12 heavily and helps sustain the $\overline{bit_n}$ line high as high line has to source the current to pull up the storage node. The static memory is organized as shown for a complete memory system.

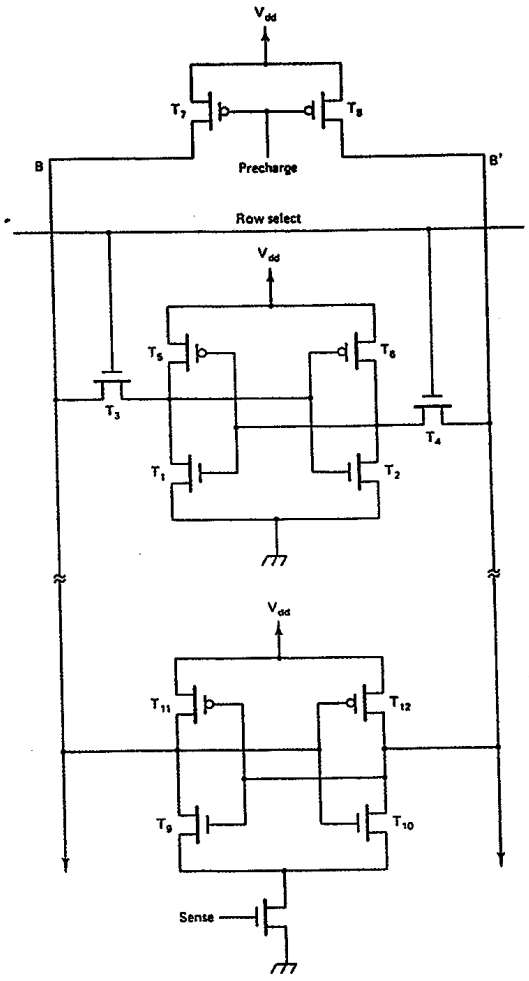


Figure 8 2 Six-transistor RAM memory cell and a sense amplifier

Memory Organization

The organization is of 2^k coordinate 1 bit SRAM where address of each bit is encoded in total of k bits. The address is decoded by the row and column decoder to select the proper bit. Initially both the bit lines are precharged high. The W / L for precharge transistors T_7, T_8, T_{13} and T_{14} are made very large for fast precharging. The read operation is performed by setting R / $\overline{W}$ input. This enables the row decoder selecting a particular row. The column decoder will enable a particular column. Depending on the information stored at the location defined by row and column, one of the bit / $\overline{bit}$ lines will be pulled to zero. The sense amp in each column helps to source the current for the high bit line and sink the current for low bit line. The information from $\overline{bit}$ line is transferred to the output via transmission gates T_{17} and buffer S.

A write is performed by selecting $\overline{CE} = 0$ and R / $\overline{W} = 0$. This enables the transmission gates T_{15} and T_{16} but disables T_{17}. The desired input is put on the input pin and buffered

to give input and $\overline{\text{input}}$ to the circuit I/O bus. The precharge on one of the I/O lines discharges through the transmission gates. The column decoder then selects a particular column whose one of the bit /$\overline{\text{bit}}$ lines is discharged to zero through I/O / $\overline{\text{I/O}}$ line and one of the transmission gates T_{15} or T_{16}. The row decoder then goes high for a particular row and the information is written at the location of the intersection of row and column decoder.

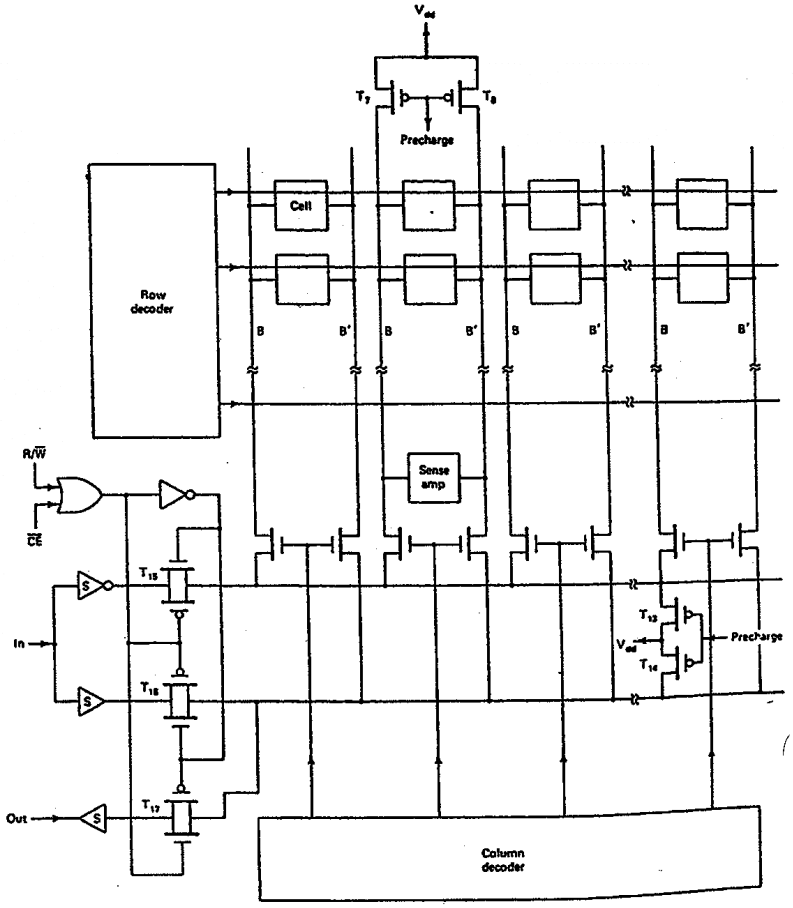
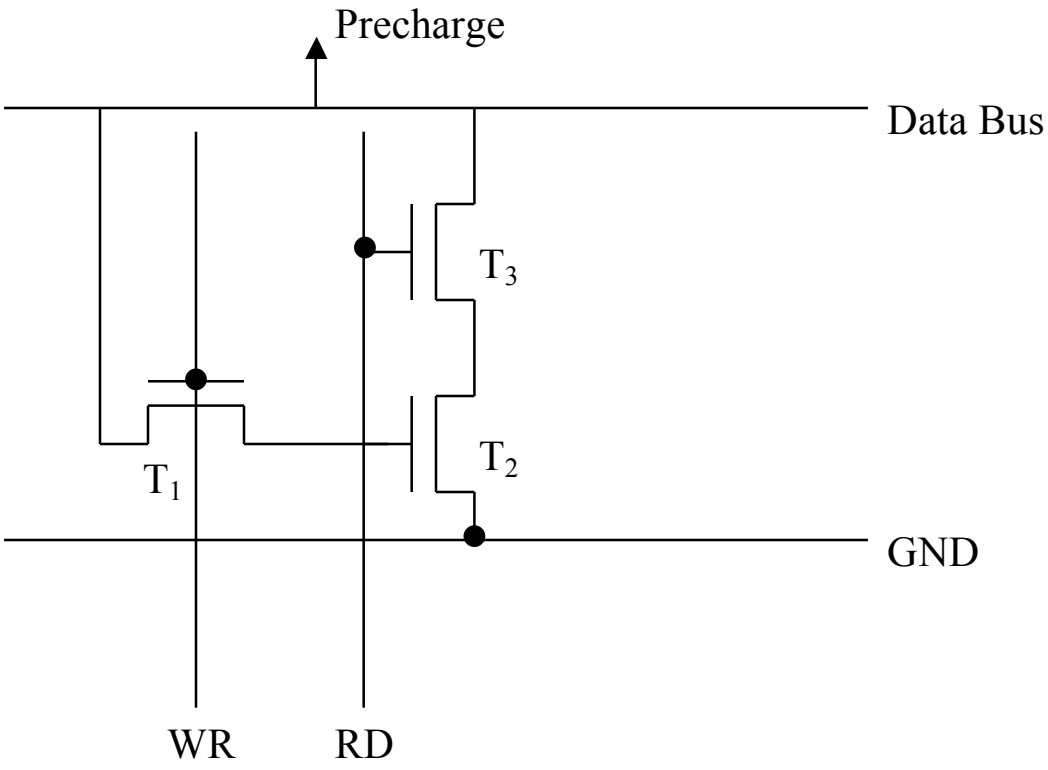


Figure 8.3 Organization of a static RAM

3 T DRAM Cell

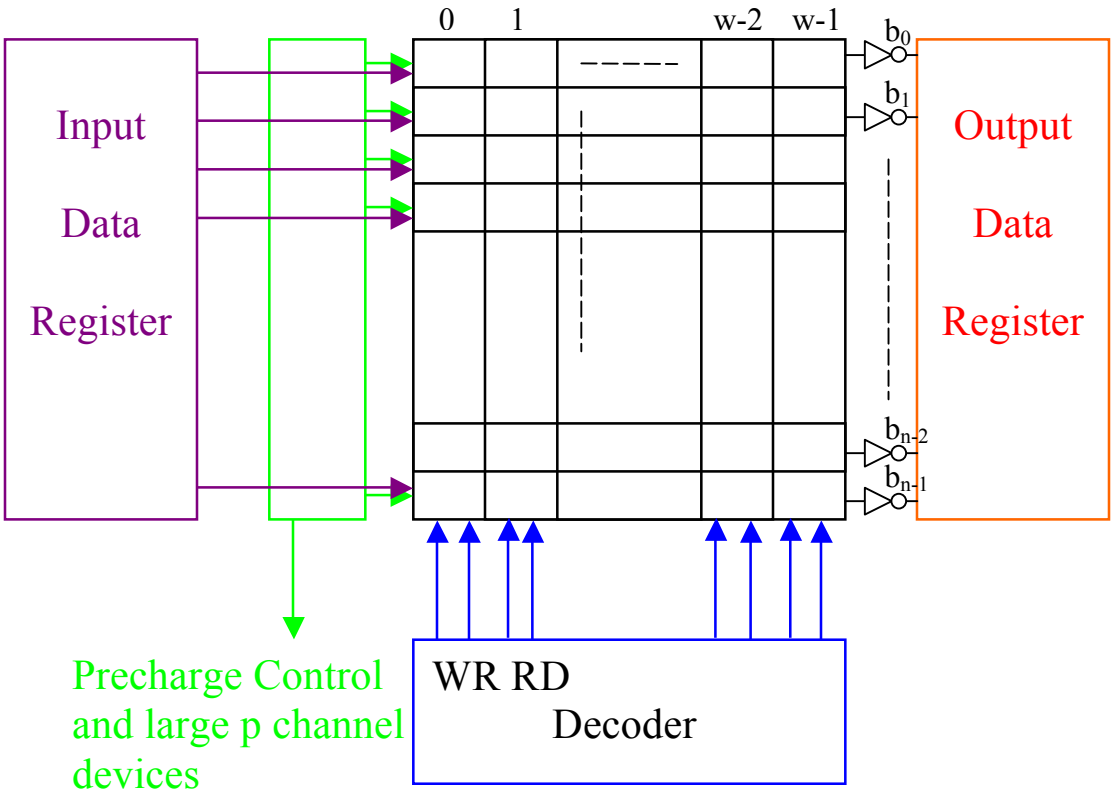
The 3 T DRAM cell is as shown.



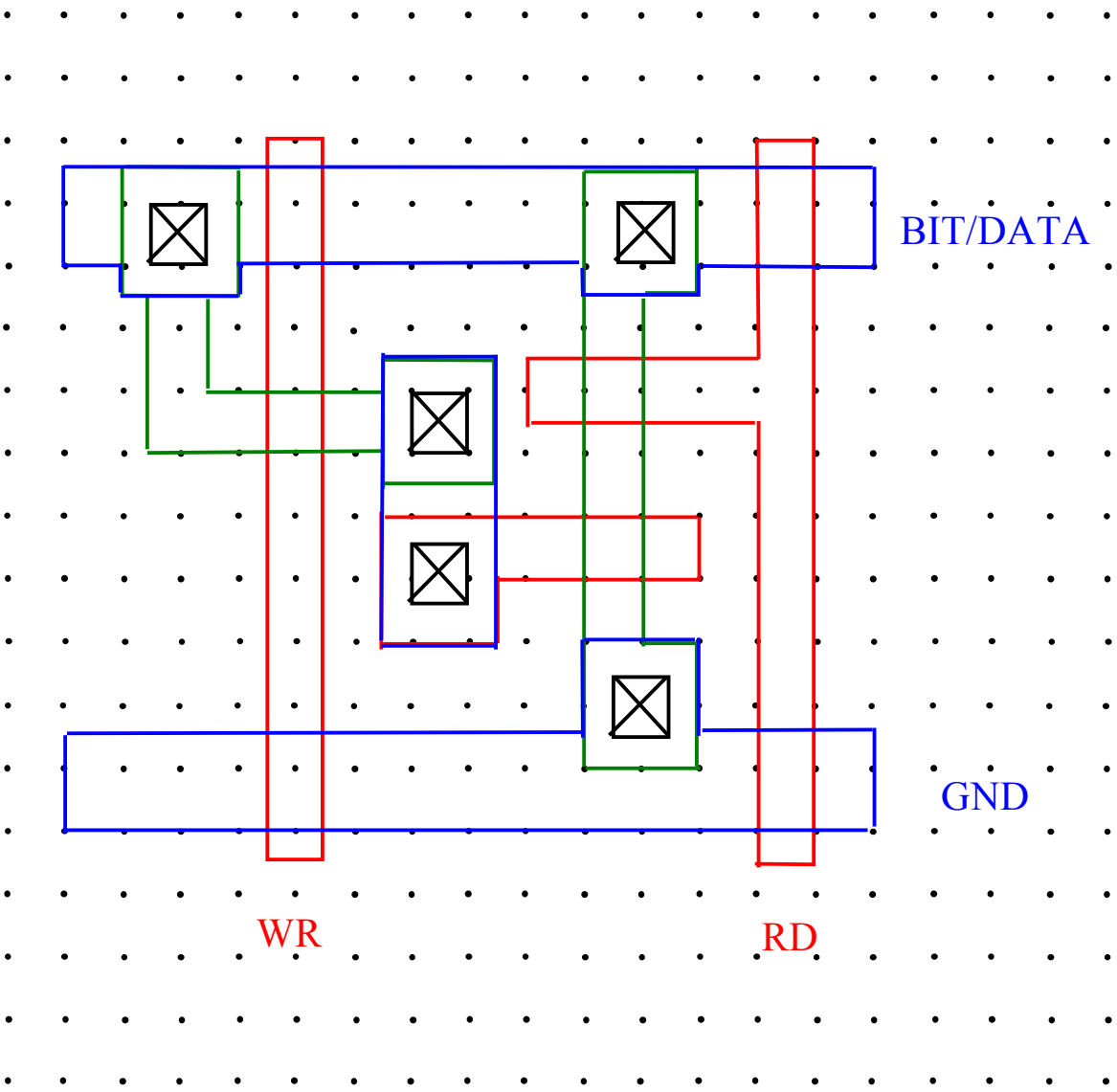
(i) To perform write operation, WR is set high and RD is set low. Also the bit line is precharged high. The line will be pulled low by the I/O bus (not shown) depending on the data to be written. Since T_1 is on, the data on the data bus is transmitted to the gate of T_2.

(ii) When both WR and RD are low, T_1 and T_3 are both off. Thus, T_2 will hold the information as a charge on its gate capacitance.

(iii) To perform a read operation, the bus again is precharged high. Then WR is set to ϕ and RD to 1. If stored data was 1, both T2 and T3 will turn on to pull data bus low. If data were ϕ, T2 will be off and bit line or data bus will stay high. Thus the bit information is actually inverted in this type of DRAM cell. The information needed from outside this cell is the WR, RD signals, precharge and data. The layout strategy is to run WR, RD along poly throughout the design, GND, bit line in metal and have one data block, one precharge block, one set of inverters as the data is inverted and the address decoders. An n bit, w Word memory is organized as shown.



The basic cell layout is as shown and it is very compact.



3T Memory Cell Layout

The cell is dynamic in the sense that the gate of T_2 may leak charge to loose information when it is in state 1. The typical refresh times are in milliseconds. When refreshing all the bits are read and written back to restore charge

which makes refresh somewhat clumsy. This cell arrangement is known as 3T, 2 Address, 1 bit line inverting cell arrangement. A two bit line arrangement is also possible which can be configured to give a non-inverting cell arrangement which can get rid of the output inverters. The transistor arrangement is as shown.

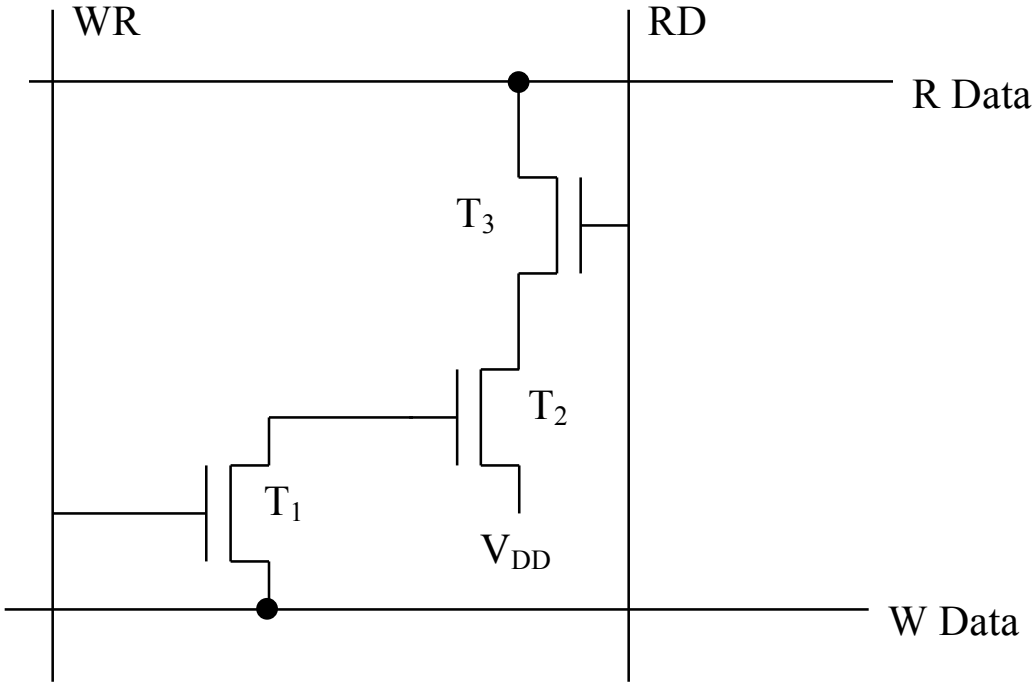


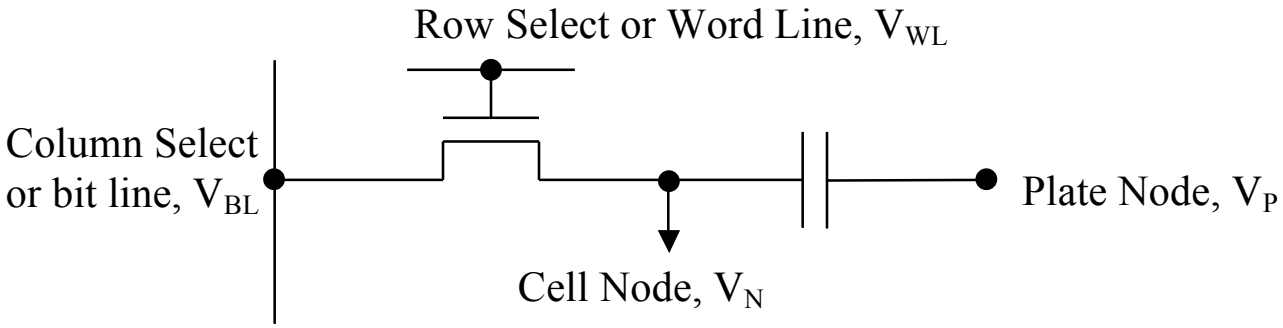
Figure : Non inverting 3T Cell

To write data, the data is put on the W data bus and WR is turned high. Then T_1 transmission gate transfers data to the gate of T_2. To read data, R data line is set to ϕ. Then RD

line is taken high. If the data stored is ϕ, T_2 and T_3 will be off and R data stays ϕ. Otherwise, T_2 and T_3 will turn on to put V_{DD} on R data bus.

1T DRAM Cell

This is the popular memory cell used in today high density DRAMs as it is the most compact cell.



Write Operation: The main objective here is to force cell node voltage V_N to applied bit line voltage V_{BL}. The data to be written is placed on the bit line first which is driven to desired level. Then word line is turned on. The access transistor is fully turned on during write i.e. $V_{WL} > (V_{BL,\,max} + V_{TN} + \Delta V_{TN,\,BG})$ where ΔV_{TN} is the increase in the threshold voltage due to back gate or body bias as source-substrate are not shorted. This way, there is

no loss in transmitted voltage which normally occurs in the case of a standard one transistor transmission gate.

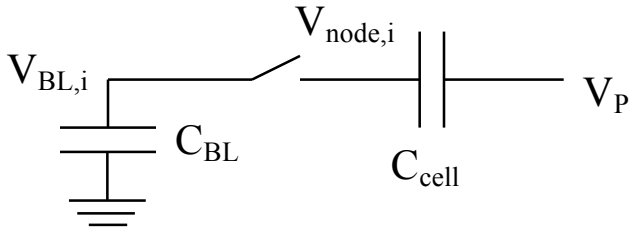
Storage

The data is stored at cell node when Row select is set back to zero. The cell is dynamic as V_N gradually leaks off due to the drain junction leakage, capacitor leakage, transistor sub threshold leakage when V_{WL} is low, and due to ionizing radiation. The cell normally holds written data for hold time. This time is in the range of 16-128 ms. Typical refresh overhead is about 0.6% and is done row by row.

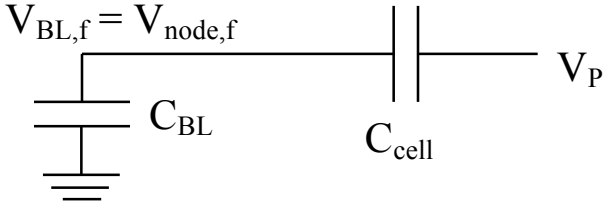
Read Operation

This is quite a tricky operation for this cell due to charge sharing problem as there is no power voltage connection. To read, the bit line is floated at precharge voltage $V_{BL,i}$. Normally, $V_{BL,i} = \frac{V_{DD}}{2}$ for large DRAM chips. The transistor is then turned on with high V_{WL}. With high V_{WL},

the transistor is treated as a short circuit switch. Let us say $V_{node,i}$ is the voltage at the storage node representing data. Before the transistor is turned on, the equivalent circuit is as shown where C_{BL} is the bit line capacitance.



The circuit after the transistor is on is



Hence the voltages basically equalize. Since transistor is treated as a zero resistance, charge is conserved.

Initial charge = $V_{BL,i}\, C_{BL} + V_{node,i}\, C_{cell}$ if $V_P = 0$.

Final charge = $V_{BL,f}\, C_{BL} + V_{BL,f}\, C_{cell}$

Charge conservation gives

$$V_{BL,i}\, C_{BL} + V_{node,i}\, C_{cell} = V_{BL,f}\, C_{BL} + V_{BL,f}\, C_{cell}$$

$$\therefore V_{BL,f} = \frac{V_{BL,i}\, C_{BL} + V_{node,i}\, C_{cell}}{C_{BL} + C_{cell}}$$

$$\therefore V_{BL,f} - V_{BL,i} = \frac{V_{BL,i}\, C_{BL} + V_{node,i}\, C_{cell} - V_{BL,i}\, C_{BL} - V_{BL,i}\, C_{cell}}{C_{BL} + C_{cell}}$$

$$\therefore \Delta V_{BL} = \frac{C_{cell}}{C_{BL} + C_{cell}} \left(V_{node,i} - V_{BL,i}\right)$$

Hence if percharge is at $\frac{V_{DD}}{2}$ i.e. $V_{BL,i} = 2.5V$ with V_{DD}= 5V

and $V_{node,i}$ = 5V = Logic 1, then

$$\therefore \Delta V_{BL} = \frac{C_{cell}}{C_{BL} + C_{cell}} (5 - 2.5) = \frac{2.5\, C_{cell}}{C_{BL} + C_{cell}}$$

Normally, to achieve a compact cell, C_{cell} is small and due to stacking of bits in a column C_{BL} is large. Hence, normally, $C_{BL} \cong 6 - 12 * C_{cell}$.

With $C_{BL} = 12C_{cell}$

$$\therefore \Delta V_{BL} = \frac{2.5}{13} = 0.192 \text{ V}$$

Typical ΔV_{BL} is in the range of 100 – 200 mV.

Hence, since V_{node} and V_{BL} are equalized after a read and change in V_{BL} is only 100 – 200 mV, the reading is a

destructive process as $V_{node,f}$ is no longer maintaining its previously written value of logic high or low.

Hence there is a need of additional circuitry that will restore logic. Also a sense amplifier like that in SRAM organization is needed to pull-up or pull-down the bit line after sensing a change in its voltage. Bit line is not driven to an appropriate logic value at all by the stored data.

Row/Column Decoders

For efficient implementation, there are 2^n row lines and 2^n column lines. We will consider 8 bit by 8 bit arrangement. An 8 rows and 8 columns memory structure will implement this. As the location of the column and row of a bit is between 1 and 8, respective row and column locations can be determined by a 3 bit address. If address of a bit is m bits, it has to be decoded by using mostly first $\frac{m}{2}$ bits for row and next $\frac{m}{2}$ bits for column. In effect, considering a

simple 6 bit address system, 3 bit address have to be encoded into row or column numbers 1 through 8. The figure shows tree decoder that implements the above function. Column decoder basically uses row decoder and a transmission gate to select a particular column. For example, address 010 turns on T_2 to place V_{DD} at N1 followed by T_3 to place V_{DD} at N6 and finally T_{13} to decode address to 3 i.e. setting signal for 3rd word line or row to V_{DD}. All other outputs are set to zero.

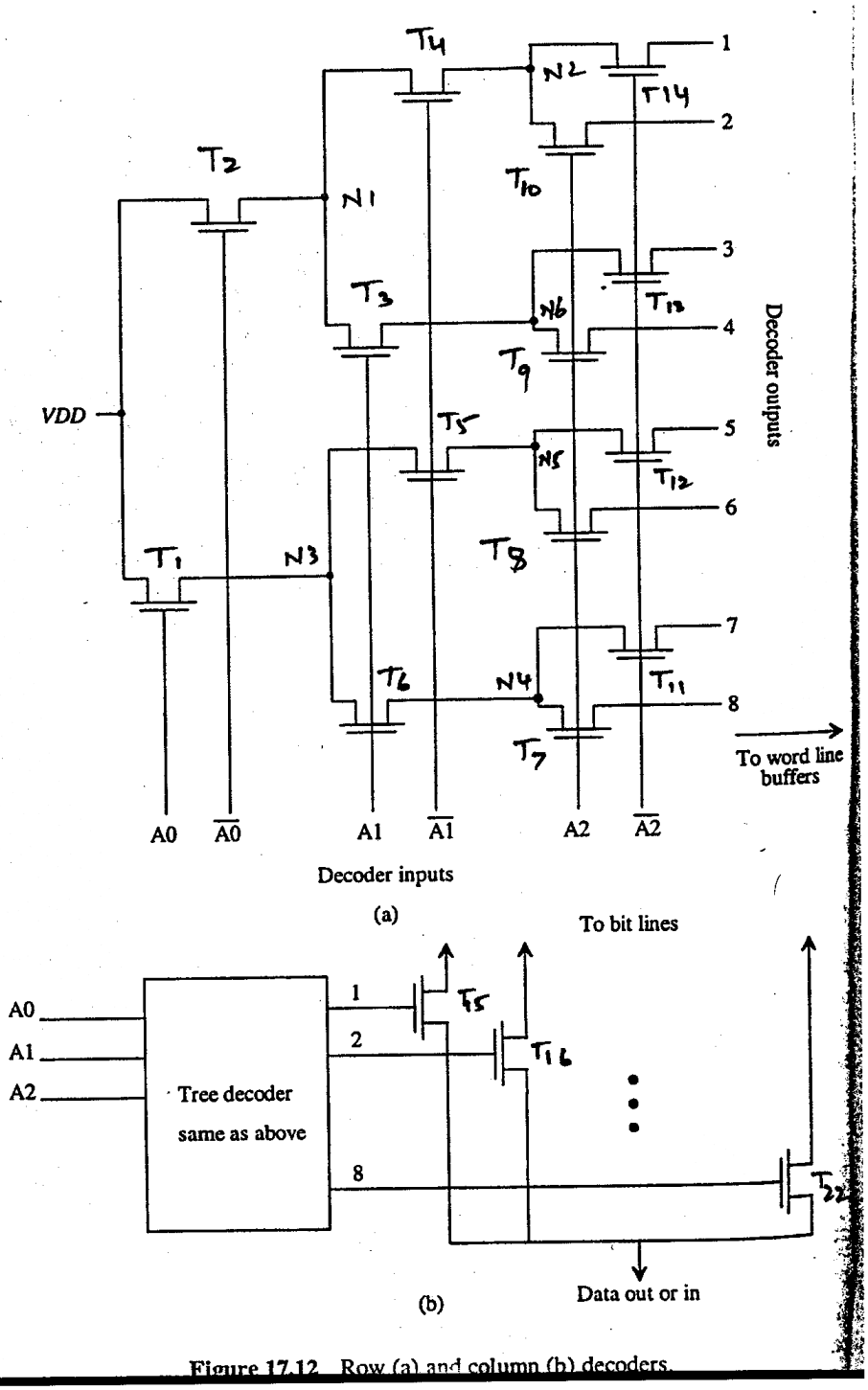
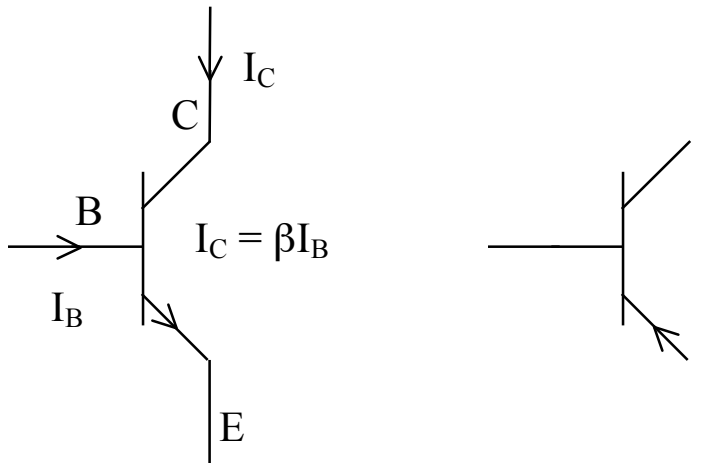


Figure 17.12 Row (a) and column (b) decoders.

BiCMOS Design

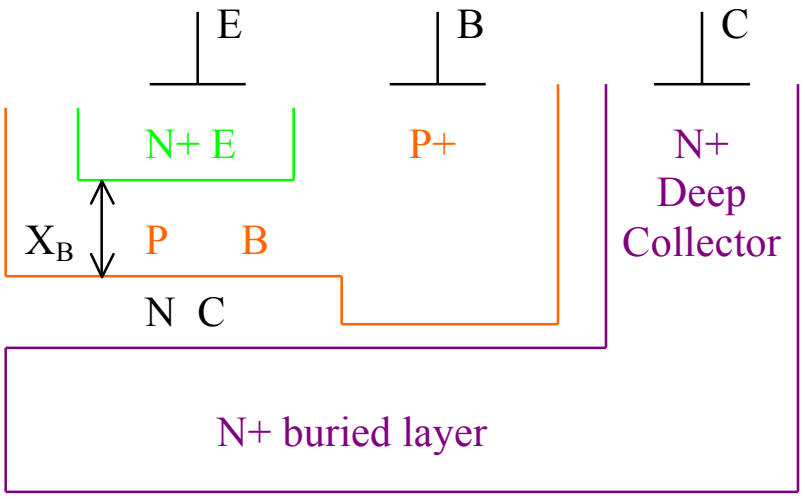
Bipolar Device Operation



NPN PNP

Bipolar Transistor

In typical BJT operation, BE junction is forward biased and BC junction is reversed biased and β is generally high ($50 \leq \beta \leq 500$). The NPN or PNP indicate the doping sequence of the device. The cross-section and layout of the basic transistor is now discussed.



The basic considerations between the layers are:

(i) The base width should be low (X_B) for high β.

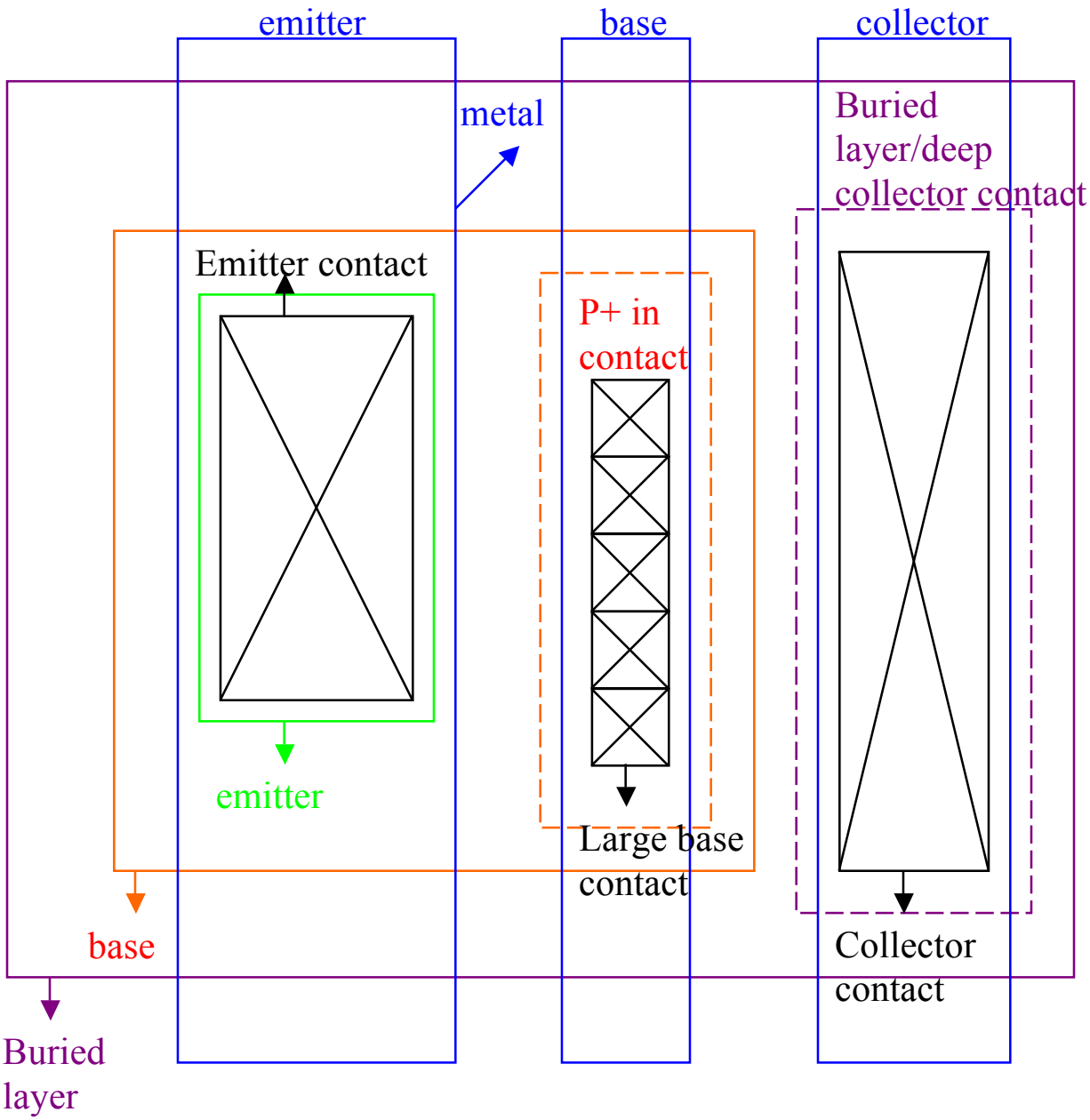
(ii) Emitter should be heavily doped (~ 10^{19}), the base medium doped (~ 10^{17}) and the collector low doped (~ 10^{15}) for high currents.

(iii) The base/collector junction should be able to sustain the power voltages without punching emitter-collector.

(iv) Collector resistance should be low which gives rise to the buried layer under the low doped collector (~ 10^{19}) and the deep N+ collector contact which connects metal to the buried layer. There are several other considerations which are not discussed.

As for a single NPN transistor layout, the buried layer is basic design mask outside which there is isolation. Inside this buried layer, there is base implant/diffusion after leaving space for collector contact.

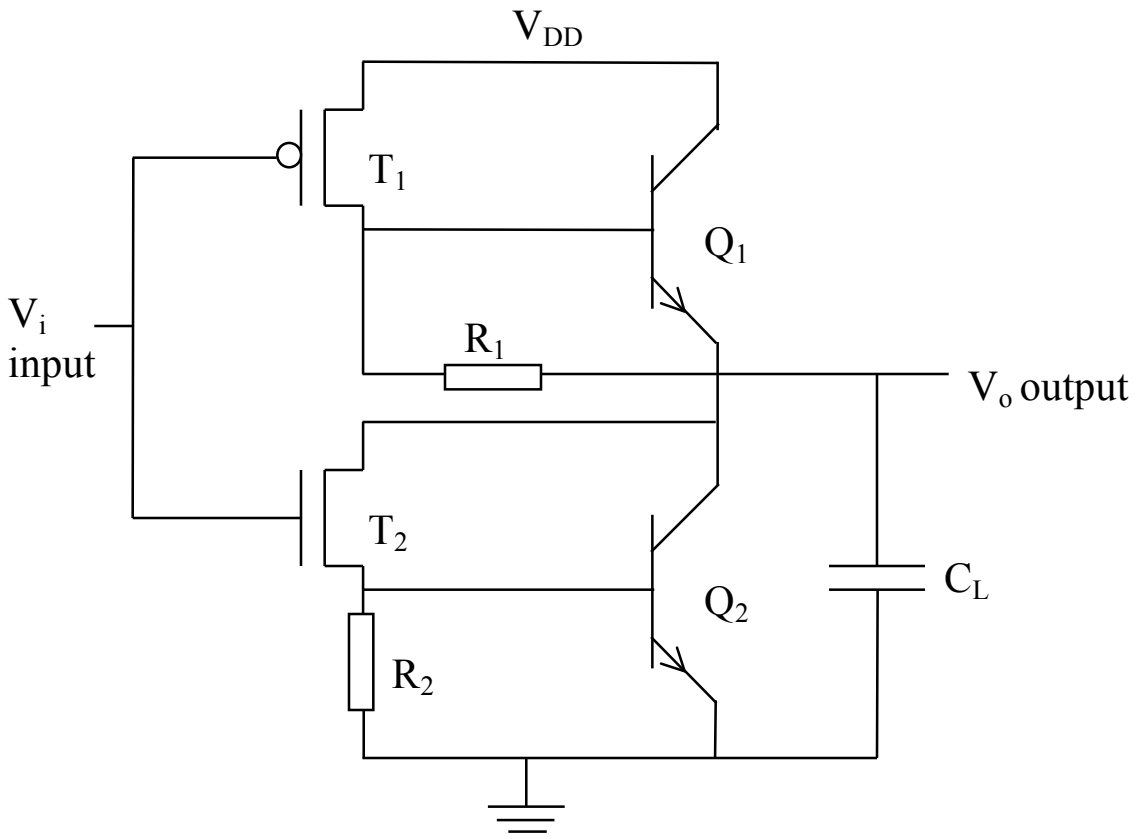
Inside base diffusion, there is emitter diffusion. Then the contacts go in their respective features with metals covering them. Hence the basic bipolar transistor layout is as shown.



BiCMOS Digital Design

BiCMOS technology tries to combine advantages of Bipolar and CMOS technology. BiCMOS design retains the CMOS capabilities of the low quiescent power and the bipolar advantage of the high output drive current capability. The main disadvantages are that the process complexity and the area (for low load situations) is increased. One should, however, note that the drive capability per unit area is much better for BiCMOS than CMOS.

BiCMOS circuits use CMOS circuit to implement the logic and a bipolar circuit to drive the load. As before, the first example we look at is the BiCMOS inverter.



BiCMOS Inverter Circuit

In this circuit, since there is no DC path to supply base currents, the DC power consumption is very low. Let us now see how this circuit operates as an inverter.

Consider the case when Vi switches from 0 to V_{DD}. As soon as $V_i = V_{DD}$, T_2 turns-on and starts to discharge C_L with a high drain current comparable to CMOS inverter current. This current flows through R_2 and generates

sufficient voltage across Q_2 base-to -emitter junction to forward bias the junction. Now, Q_2 turns on and discharges C_L at a much faster rate. The base current for Q_2 is also provided by T_2. Note that is always voltage across Q_2 base-to -emitter junction < V_O. Eventually, V_O falls to such a low value that voltage across R_2 is not sufficient to forward bias Q_2 base-emitter junction. Then the discharge process continues through T_2 and R_2 relatively slowly till V_O is pulled down to zero. The operation for low input is similar.

R_1 and R_2 also serve as bleeding resistors. For example consider the situation when $V_{BQ1} = V_O = V_{dd}$. Let V_i now increase to V_{dd}. T_2 turns on and discharges V_O to ground through Q_2 / T_2 – R_2 as explained earlier. If R_1 was not there (it is removed to make an open circuit in its place), V_{BQ1} will stay at V_{dd} and will tend to turn on Q_1 which may take away some of the discharge current or it will be in very unstable state with a large forward bias but no current. However, R_1 discharges base node of Q_1 (bleeds the base charge) to zero and overcomes this problem.

In the resistor-less implementation, the resistors are replaced by bleeding devices as shown in the resistor-less example.

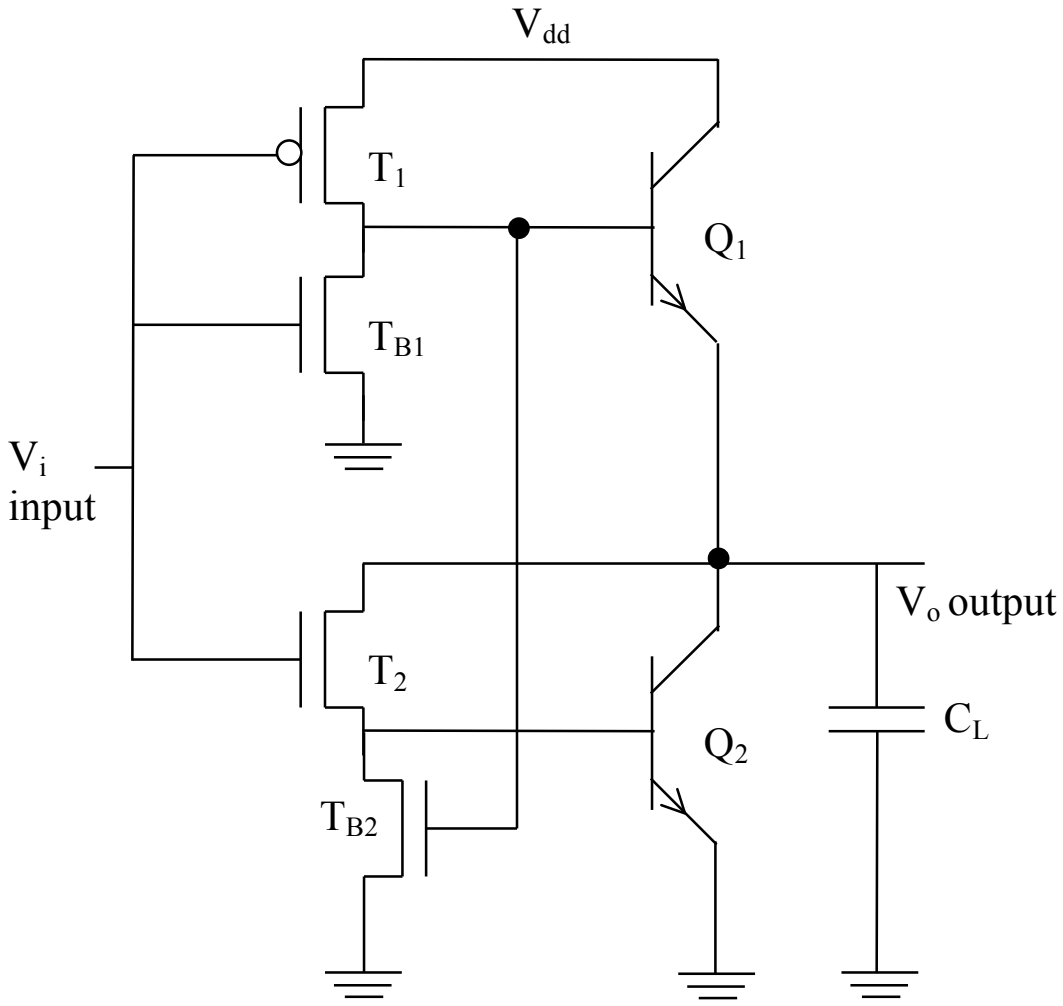
Due to the presence of 2 MOSFETs, 2 resistors, 2 BJTs in the circuit and full 0 to V_{DD} output swing, this circuit is known as 2 MOSFET, 2 R, 2 BJT rail to rail BiCMOS inverter circuit.

Resistor-less implementations of the inverter are possible. However, they suffer from a serious drawback that full 0 to V_{DD} swing is not available. One such implementation is shown.

For more implementations, please refer to the following.
"BiCMOS Technology and Applications"
edited by Antonio R. Alvarez, Kluwer Academic Publishers, 1989)

Most of the BiCMOS treatment is taken from this book.

Resistor-less BiCMOS Inverter

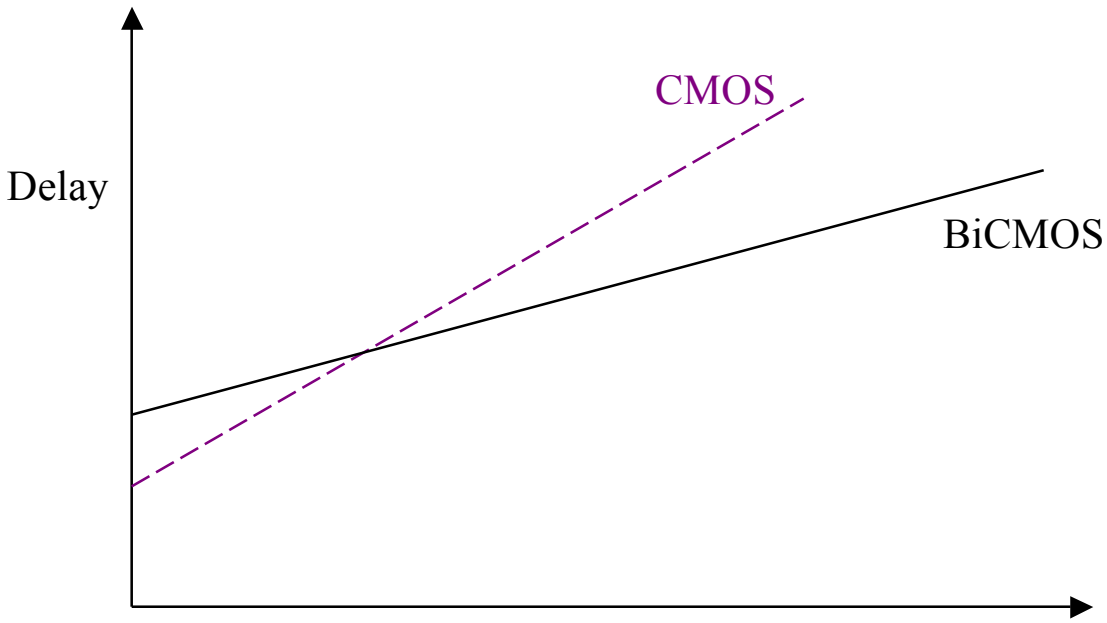


Here when V_{in} goes high from zero, the base of Q_1 is at V_{dd} and V_O is high. T_{B1} turns on and bleeds the base of Q_1 to ground. T_2 turns on and provides base current to Q_2 and discharges C_L to a low value.

Since V_{BQ2} can never exceed V_O, Q_2 eventually turns off when V_O is 0.2 – 0.3 V range. At this point the discharge process stops. In the other case of $V_{in} = 0$, T_1 provides the base current to Q_1 and charges C_L to (V_{dd} – 0.3V) to (V_{dd} – 0.2V) range. T_{B2} bleeds the base of Q_2 to ground.

Hence such an inverter does not have rail to rail swing. An advantage of this structure is that the bleeding devices do not take current from T_1 or T_2 and hence base current is higher. However apart from voltage swing problems, these circuits have larger area for complex gates.

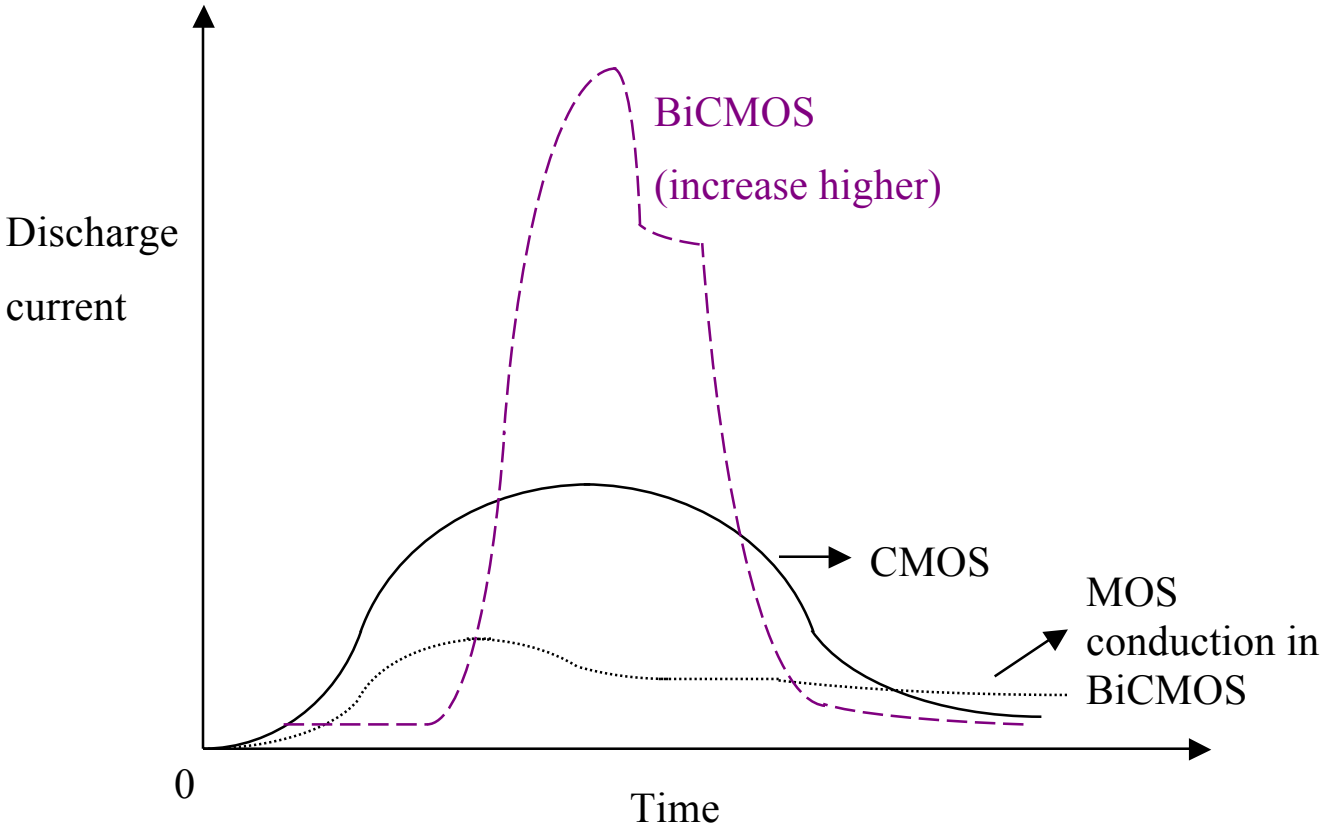
BiCMOS Inverter Delay Characteristics



This type of delay characteristics results from the fact that the bipolar device collector current (which is the capacitance discharge current) is exponentially dependent on V_{be} i.e.

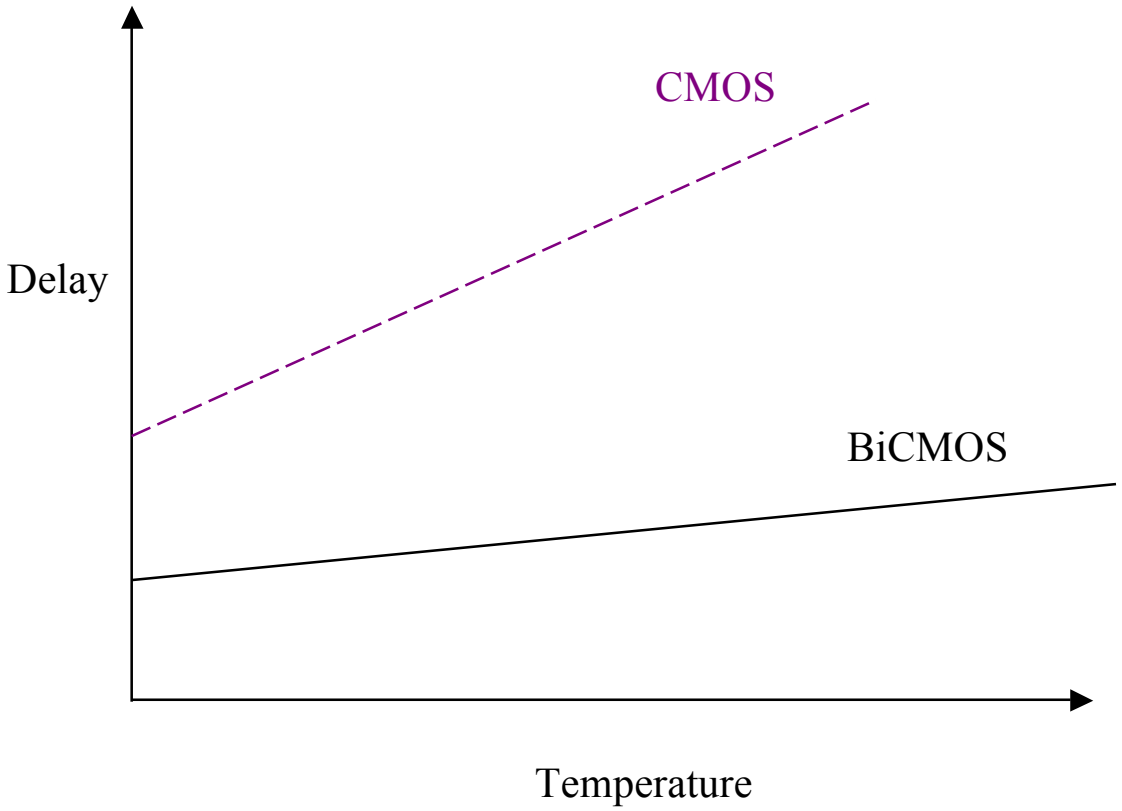
$$I_C = I_S \left[e^{\frac{V_{be}}{V_T}} - 1 \right]$$

hence for a small V_{be}, the discharge current is large provided $V_{be} \sim 0.6 - 0.7$ V. After the base-emitter voltage falls below 0.5V, the discharge is entirely by MOS devices. Before this, discharge is controlled by the high bipolar device current. The MOS discharge current at the most had a quadratic dependence on the voltages and hence is not very fast. A complete delay analysis and sizing analysis for BiCMOS circuits is possible which we will not address. Only a plot which shows discharge current for CMOS and BiCMOS is shown for a ramp input for a comparison.

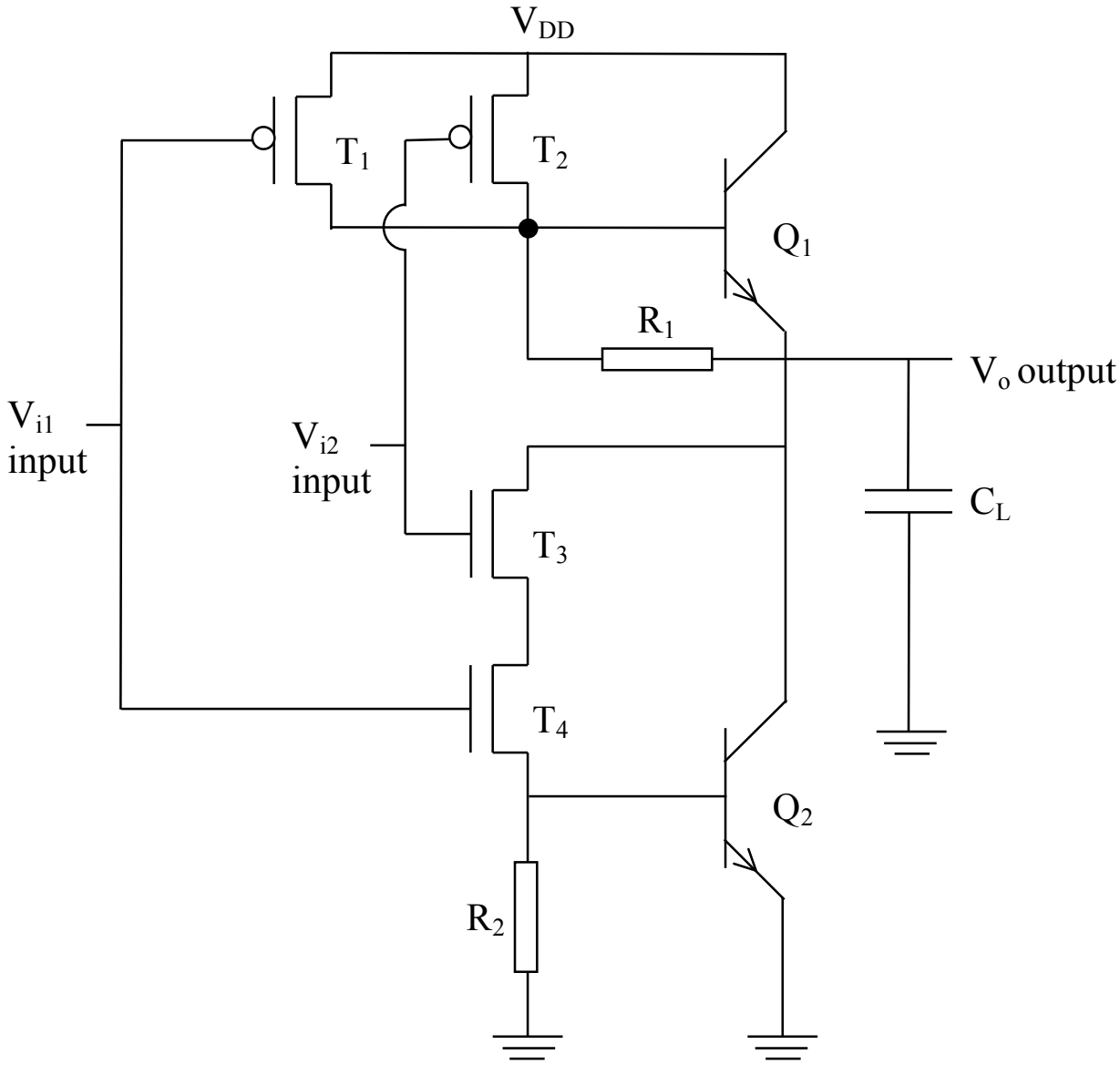


It is clearly seen that the bipolar device can sustain much higher discharge current during most of the discharge process. The difference is typically 3 times.

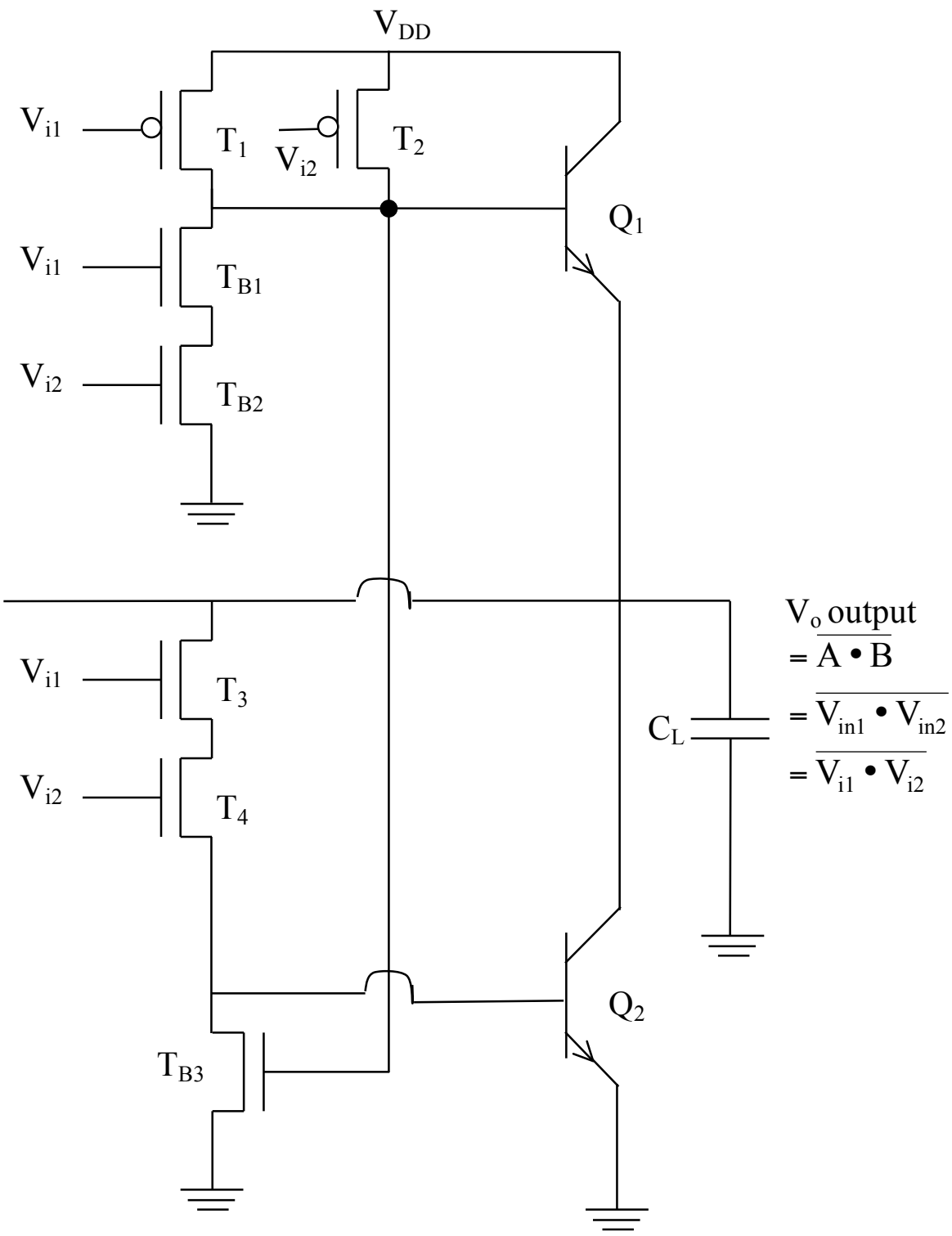
An advantage of BiCMOS is that the timing is less sensitive to the reduction in power voltage and the timing is relatively less sensitive to temperature as shown below. Clearly, bipolar and MOS dependencies compensate to have this behavior.



The implementation of other combinational logic in BiCMOS is straightforward knowing the fact that the logic function is carried out by the CMOS devices. The circuit on the next page shows 4 MOSFET, 2 resistors, 2 BJT 2 input NAND implementation with full swing. In the circuit, the logic part is same as 2 input CMOS NAND. The following page shows the resistor-less circuit which does not have full $0 - V_{DD}$ swing at the output. Any other logic can be implemented in the same fashion.



BiCMOS 2 Input NAND



Resistor-less 2 input BiCMOS 2 Input NAND

Implementation of DRAM and SRAM which are much faster is possible. We will not discuss them. This basically concludes the course and the basic introduction to BiCMOS.