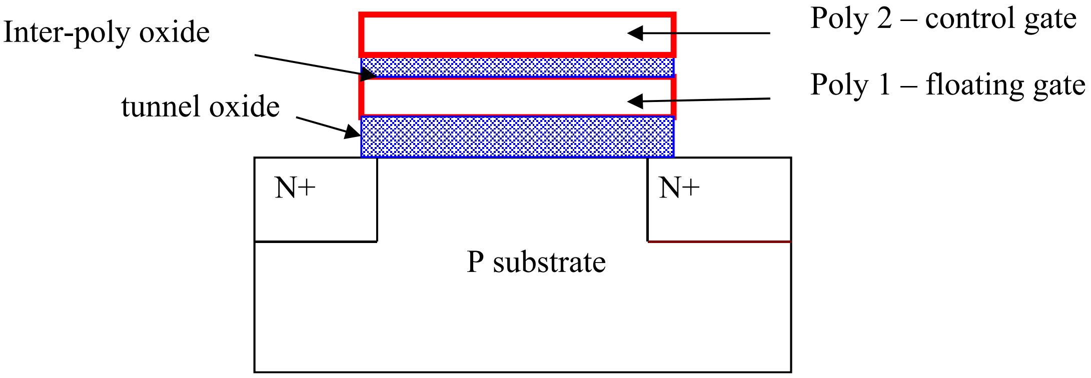
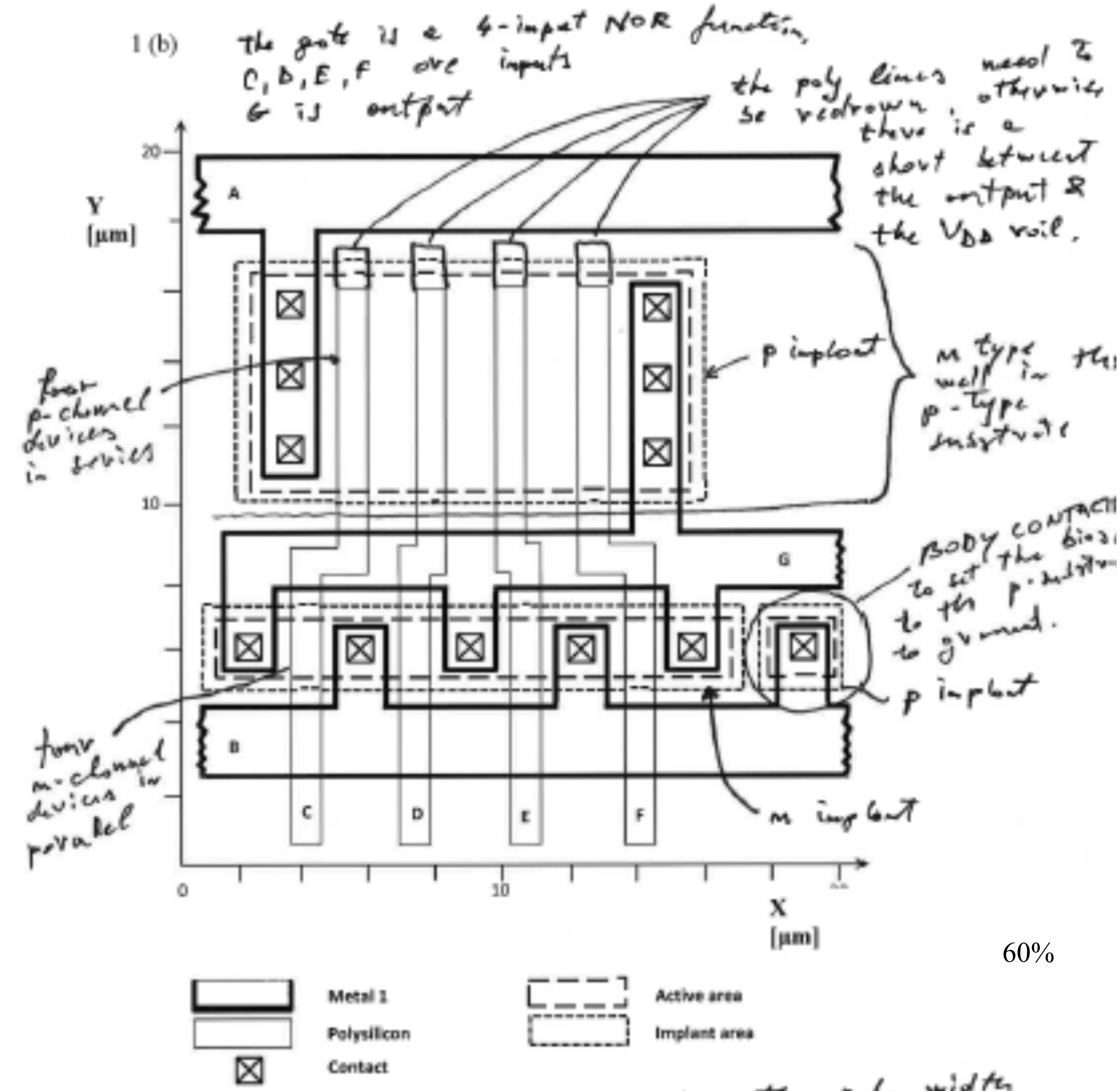
1. (a) The bulk resistivity of copper is about 1.67 Ωcm. This is almost half of that of Aluminium. Thus, the sheet resistance of a layer of Copper would be half of that of Al with the same thickness. In addition Copper has superior resistance to electromigration and thus can allow higher current densities than Al. This means that layers of metal interconnect. which in practical designs occupy a considerable area, can be made narrower, while still being more conductive. Copper however cannot be patterned and etched using a standard process. The *Damascene* process is the solution for advanced interconnections. The process eliminates the need for metal etch and dielectric gap fill that becomes very challenging as dimensions continue to shrink. The *Damascene process* is based on defining a trench pattern; etching through the dielectric material, depositing copper on the surface through electroplating, and planarising the surface through CMP (Chemical Mechanical Polishing).

Good quality capacitors can be obtained by using an additional layer of polysilicon. The polysilicon layers need not be separated by a thick layer of dielectric (like subsequent layers of metals) and therefore a very thin layer of oxide can be used between the two poly layers to make large capacitors occupying a very small area. These are useful both in analogue circuits where RC filters or switched capacitors are used (as in RF) or in dynamic memory cells.

Reprogrammable memories (EEPROM) can also be made using two polysilicon layers separated by a thin inter-poly oxide. [20%]



1 (b)



The gate is a 4-input NOR function, C, D, E, F are inputs G is output

the poly lines need to be redrawn, otherwise there is a short between the output & the $V_{DD}$ rail.

four p-channel devices in series

p implant

n type well in p-type substrate

BODY CONTACT to set the bias to the p-substrate to ground.

p implant

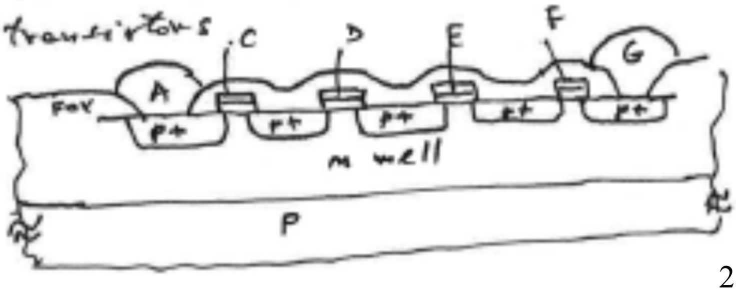
four n-channel devices in parallel

n implant

60%

- the linewidth is approximately the poly width in the n-channel & p-channel transistors ≃ 1μm
- the p-channel transistors have lower mobility hence they need to have longer width (4X) to have equivalent transconductance with the n-channel transistors

Cross-section



20%

2. a) Advantages:

- excellent electrical isolation between devices or blocks of devices (less leakage, no latch-up)
- less area consumed (no buried layers)
- (alternatively one can mention the high junction temperature)

Disadvantages:

- self-heating (the buried oxide layer acts as a thermal barrier)
- expensive (the SOI wafer can be 5-10 times more expensive than a standard bulk silicon wafer)

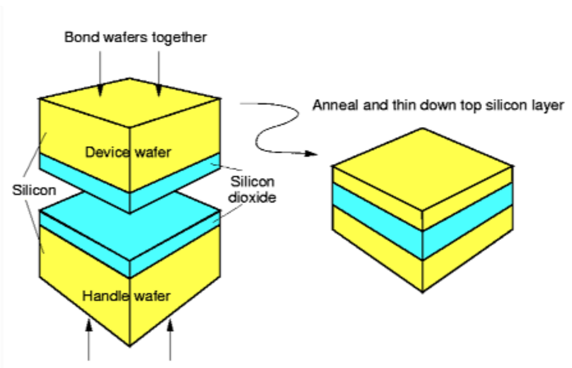
[20%]

*(b) Unibond - SmartCut*

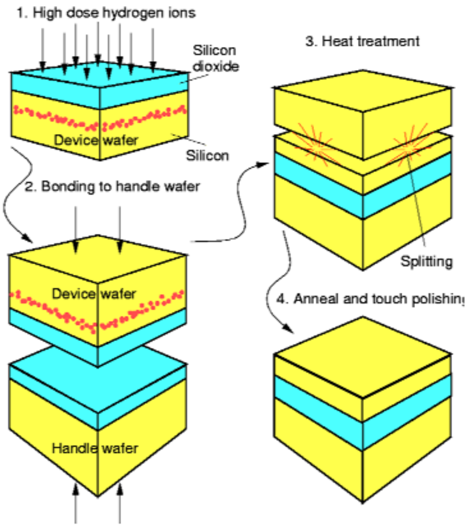
The device wafer, which has a layer of silicon dioxide on top of it, is implanted with a high dose of hydrogen ions (between $3.5 \times 10^{16}$ and $1 \times 10^{17}$ $\mathrm{cm}^{-2}$), after which it is bonded to the handle wafer. A heat treatment at 600 °C divides the wafers along the line of the implanted hydrogen, leaving behind a thin and uniform silicon-on-insulator layer on the handle wafer which requires only a final high-temperature anneal and touch polish to yield the finished wafer.

Wafer bonding is much cheaper than SmartCut. Smart cut however is much better in defining accurately the thickness of the SOI layer

**Wafer bonding**



**Unibond-SmartCut SOI technology**



20%

2 (c) (i) **Advantages:**

- The presence of a silicon window (gap thorough the buried oxide) allows better heat dissipation to the substrate. This allows cooler operation and therefore higher performance (note that the channel mobility decreases with the increase in temperature).
- The presence of the p-well/n-substrate could enhance the breakdown of the p-channel transistor.

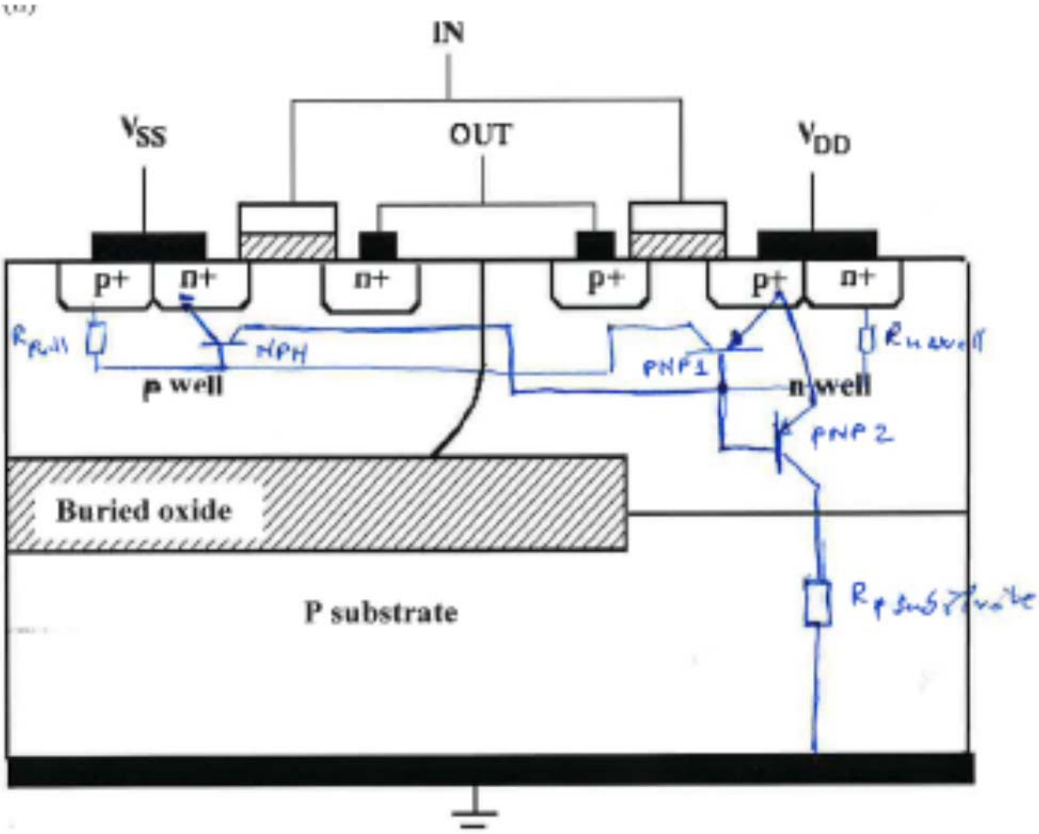
**Disadvantages:**

- The isolation is not as effective as in full SOI, as leakage through the substrate will occur

There is an additional pnp parasitic transistor which could result in latch-up or increased injection of carriers (holes) into the substrate.

[10%]

(ii)



- NPN transistor 1
  - Emitter: n+ source of the n-channel MOSFET
  - Base: p well
  - Collector: n well
- PNP transistor 1
  - Emitter: p+ source of the p-channel MOSFET
  - Base: n well
  - Collector: p well
- PNP transistor 2 (the presence of this additional transistor could result in more severe latch-up issues)
  - Emitter: p+ source of the p-channel MOSFET
  - Base: n well

[30%]

o Collector: p substrate

(iii) The ‘latch-up’ parasitic structure is composed of three bipolar transistors (NPN & PNP1 and PNP2) in a thyristor configuration. PNP1 and PNP2 have a common base and common emitter, with their base connected to the collector of the NPN transistor. The PNP transistors are more likely to be turned-on (leading eventually to latch-up) because the $R_{n\text{-well}} > R_{pwell}$.

The latch-up condition can be written as :

$I_{trigger} = V_{pnp\text{-on}} / (\alpha_{NPN} \cdot R_{n\text{-well}}) = 0.7\ V / (\alpha_{NPN} \cdot R_{n\text{-well}})$

[20%]

3. (a) **Clock Skew**. In many VLSI systems operations are synchronised to a Master Clock. This might be generated by on-chip circuitry or introduced from outside via an input pad. The clock is distributed to all parts of the circuit by means of interconnect, which may be as long as 1-2 chip diameters
The interconnect introduces R-C delay (and the L element also introduces distortion). Hence different destinations on the chip receive clock signals delayed by different amounts. These different delays relative to the master clock are called CLOCK SKEW.
Clock skew can therefore arise from the following:

- different lengths and types of interconnect between the master generator and locations where the clock is used
- passage through different numbers / configurations of control gates
- the need for extra inverters to form $\overline{\varphi}$ , e.g. for transmission gates

In design of sequential circuits, designers need to specify a minimum HOLD time to guarantee proper latching of data to alleviate the effects of clock skew

Clock skew is reduced by:

- keeping interconnect paths short and direct
- avoid the use of high resistivity conductors (e.g. polySi) for all but the shortest interconnect runs
- split clock lines into short segments separated by buffers
- user of pipe-lining – an enabled T-gate may be placed in series with a signal to compensate for delays in other paths
- use of silicide, copper, to minimise inserted resistance
- use of organoSi glass or SoI to minimise capacitance to substrate and hence the delay. [30%]

**Numerical Part**

(b) The given formula, $T = \dfrac{rcl^2}{2}$, is in terms of *R*/unit length, *C*/unit length

Alternatively, $T = \dfrac{1}{2}RC$ , where *R* = total resistance, *C* = total capacitance

Total resistance 10 mm $= \dfrac{10\times10^3}{1}\times 40 \;=\; 4\times10^5\,\Omega$

Total capacitance 10 mm $= 2\times10^{-10}\times10^{-2} \;=\; 2\,\text{pF}$

Hence for 10 mm trace, $T = \dfrac{1}{2}\times4\times10^5\times2\times10^{-12} =$ <u>400 ns</u>

(i) Using a double-width interconnect is expected to double the conductance, and to increase the capacitance. The capacitance will not quite double, since there are both area and peripheral components, of which the peripheral component scarcely changes with the doubled width. This is more noticeable with smaller geometries. Hence the delay falls slightly, since $T \propto RC$.

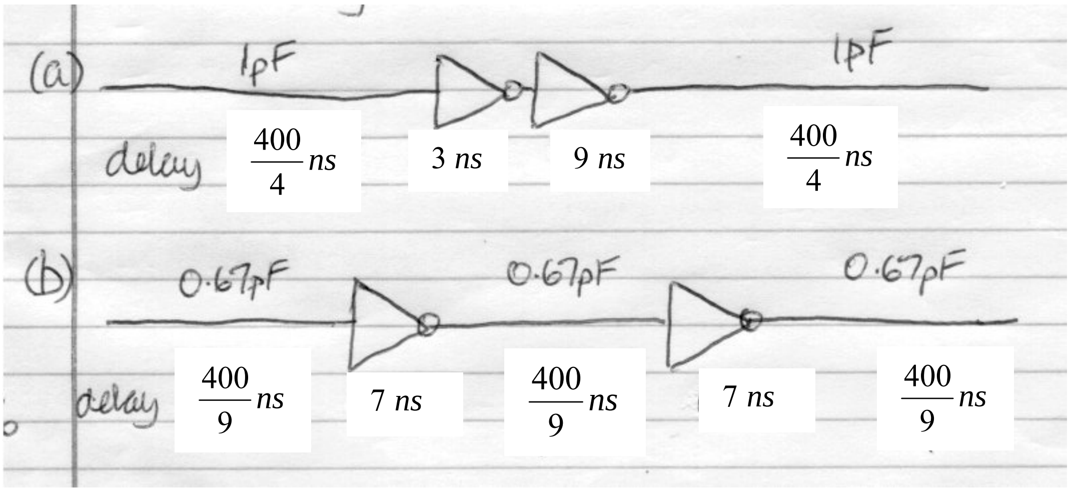
(ii) Using the silicide reduces the resistance by a factor 40/4 = 10, but leaves the capacitance effectively unchanged since there is no change in geometry. Hence *T* is reduced by a factor 10, to 40 ns. [40%]

(c) The use of buffers to separate the clock line into shorter segments *may* reduce the overall delay, but this depends on the relative delays arising from the increased number of shorter segments and from the buffers driving the capacitance of the line.

A number of configurations are worth considering. They must have an even number of gates, since a non-inverted clock is mandated.

In (a) two inverters are inserted at the centre of the line. The first drives effectively zero line capacitance so its delay is 3 ns. The second drives 1pF so its delay is 9 ns. Each 5 mm segment has 100 ns delay. The total is 100 + 3 + 9 + 100 = 212 ns – too much.

In (b) the two inverters are placed so they break the line into 3 equal segments.



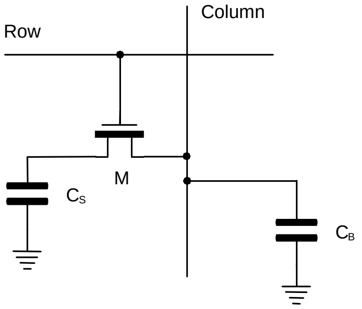
Each inverter is driving a capacitance of 0.67 pF, and its delay will be (3 + 0.67 × 6) = 7 ns. Each 3.3 mm segment will have delay 400/9 ns or 44.4 ns.

Hence the total delay will be: 44.4 + 7 + 44.4 + 7 + 44.4 ns = 147.3 ns.

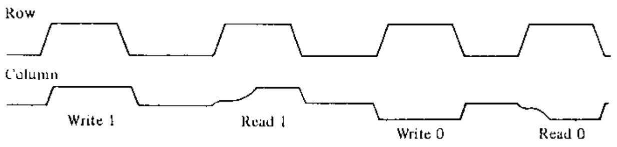
This comfortably meets the requirement stated, i.e. a factor of two reduction, and is the simplest arrangement to do so. [35%]

4. (a) A DRAM cell may be achieved using an n-channel MOS transistor in association with parallel capacitor elements. This facilitates an extremely small cell and leads to high memory densities. $C_S$ is a parasitic element typically O(20fF). Much effort has gone towards fabricating capacitors with the highest possible C and minimum area, e.g. trench capacitor.

Reading and writing are accomplished by applying logic high to the gate of M via the row/address line in order to select the cell. The cell must periodically be refreshed (O(10ms)) because of charge leakage from $C_S$. Data can be written into the cell by forcing logic 0 or logic 1 on the column/bit-line while the cell is selected. $C_S$ charges to this value, which is retained when the cell is deselected. When reading the cell is selected by applying logic high to the row line, making it conduct. The column line is connected to a sensitive comparator. Since $C_S$ is very small and $C_B$ may be significant (O(1 pF)), a charge sharing analysis shows that the potential change seen on the column line may be 1 mV or less. Design of suitable sensing comparator in a noisy environment is a great challenge. Normally a regenerative amplifier is used and the column line is precharged to the mean of the logic levels. [40%]



DRAM Cell



Representative timing diagram

(b) Owing to charge sharing the potential ΔV appearing on the bit line at the sense amp input is << 3V. Assume $C_B$ and the sense amp are precharged to $V_{DD}/2$ or 1.5 V and $C_S$ is charged to logic 0 or logic 1, 0 V or 3 V. First find ΔV in terms of capacitances using conservation of charge and assuming $C_S$ is at 3 V.

$$\Delta V = \frac{C_S \times 3 + C_B \times 1.5}{C_S + C_B} - 1.5.$$ Now subst ΔVmin = 10 mV and $C_S$ = 20

$$10^{-2} = \frac{20 \times 3 + 1.5\,C_B}{20 + C_B} - 1.5$$ (all C in fF)

Hence $60 + 1.5\,C_B = 1.51 \times (20 + C_B)$ and $60 = 30.2 + (1.51 - 1.50)C_B$

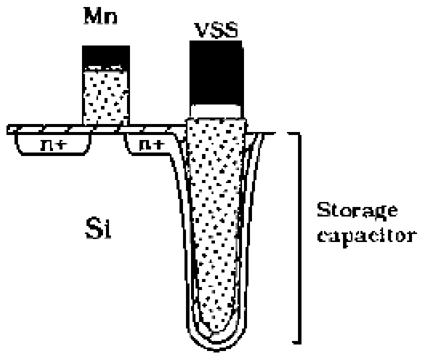
$$\frac{29.8}{0.01} = C_B \rightarrow C_B = 2980 \text{ fF}$$

This sets the max allowable length of the bit line in terms of its capacitance. The element of bit line passing over each cell contributes 2 × 0.5 fF due to the 2 μm length of bit line itself and 5 fF due to the MOSFET drain-substrate capacitance. If there are N cells in the direction of the bit-line, total C = $N \times (5 + 2 \times 0.5) = 6.0 \times N$ fF. Hence Nmax = 2980/6 ~ 500.

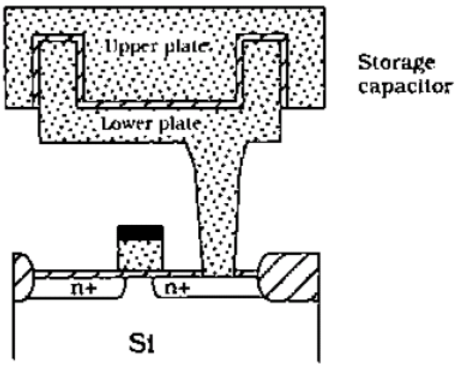
Since the array is said to be square its max size is $500^2$ ~ 250.0 kbits.

We have considered only the bit-line capacitance here and not accounted for the capacitance at the input to the amplifier. Memory designers use a number of additional circuit 'tricks' in order to facilitate much larger memories. [40%]

(c) Trench and stacked capacitors



Trench capacitor



Stacked capacitor

Current technologies call for 1Gb+ on a DRAM. For large capacity memories, less than 30% of the chip area is concerned with addressing, refresh, buffers etc, and it is the size of the one transistor cell that determines overall chip size. To achieve high density, we must minimise the area of the one-transistor cell.

The calculation in (c) shows how critical are parasitic capacitances and hence layout in the development of a successful high density DRAM. The charge sharing analysis clearly shows that the relative magnitudes of $C_S$ and $C_B$ determines the size of array that can be achieved. However, it is difficult to see how $C_S$ can be increased without increasing the size of the cell, and hence inflating $C_B$ (an undesired side-effect). Two methods now popular to circumvent this difficulty involve use of: (i) trench capacitor or (ii) stacked capacitor. Each approach seeks to maximise $C_S$ without increasing the cell area.

In (i), an enhanced storage capacitor is obtained by use of a large junction capacitance fabricated in the form of a conical trench which may extend up to about 8 μm into the substrate (in comparison with about 0.5 μm horizontally). The trench capacitor abuts the MOSFET source electrode, and in the example shown above lies beneath the $V_{SS}$ rail. In advanced trench structures, the junction may be constrained to follow a folded (tree-like) profile to maximise the capacitance obtained.

In (ii) the capacitor is fabricated above the MOS transistor by use of an overlaid polysilicon layer; poly 1 forms the lower plate and provides a via, connecting the plate to the MOSFET source. The upper plate is formed of a second polysilicon layer, with thin oxide forming the dielectric between the plates. In this stacked structure, the capacitor may occupy virtually the full area of the 1-bit cell without requiring an increase in dimensions. The capacitor may be folded vertically to further enhance the capacitance. [20%]

5.(a) Input pad structures are primarily required to protect MOSFET inputs from:

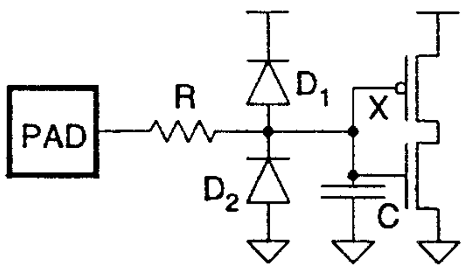
- over and under-voltages
- consequential latchup conditions
- electrostatic discharge

In addition they may contain inverting circuitry, or Schmitt trigger circuitry if the input signals to be fed to the circuit are not known to be proper CMOS level signals.

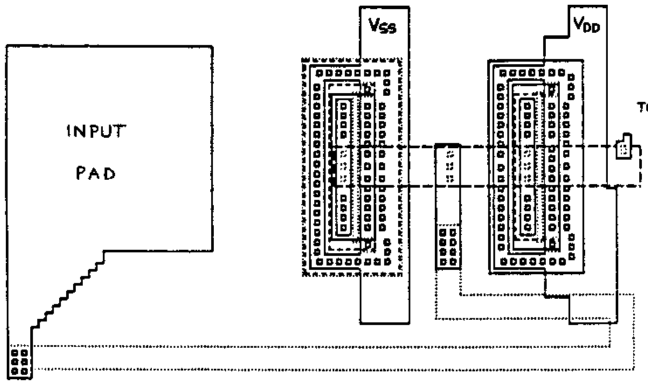
The pads are the squares of metal, generally 60-150 μm square, that are connected to the pins of the package with bonding wires. The word *pad* is often used to also include the circuitry that is used to interface the CMOS logic within the IC (typically composed of near minimum-geometry transistors) to the outside world.

Gate oxide thicknesses in modern processes are O(20 nm) thick with breakdown voltages of 5 V or so. Input resistances may exceed $10^{12}$ ohms. Since the gate electrode typically has capacitance of a few fF, only a very small packet of charge is required to generate voltages far in excess of $V_{breakdown}$.

The human being is often modelled (for evaluation of 'electrostatic risk') as a capacitance of ~100 pF charged to ~1.5 kV in series with a resistance of a few kΩ. The energy available is sufficient to vaporise a considerable volume of Silicon. Protection can be achieved with the circuit below:



The diodes become forward-biased as Vpad exceeds $V_{DD}$ + 0.7, $V_{SS}$ – 0.7 sinking excess current to the rails.
R may be implemented as a strip of polySi 20-100 sq long.



The diodes are provided with guard rings and generous well/substrate taps to minimise carrier injection.

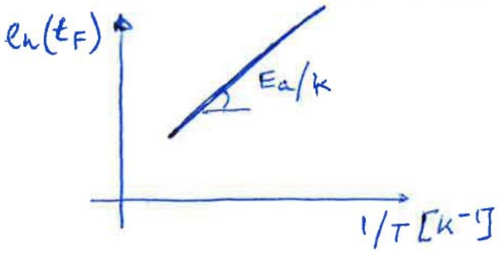
C is a parasitic capacitance due to the diodes and the parallel-plate capacitance arising from the interconnect used, etc.

The presence of the diodes reduces the input resistance of the circuit to ~$10^{10}$ ohms. The resistor and the input capacitance of the first stage of the circuit will present an RC time-constant. If this time constant is unacceptable, the value of the resistor can be reduced, but this will reduce the voltage capability of the protection circuit. Protection circuits should have a capability of at least 2kV; 8 kV capability is possible with careful design. Selection of values and hence dimensioning are necessarily a compromise. Excessive R and C will give good protection but will delay legitimate digital edges and cause slower rise/fall.

Often Punch-thru devices are used in place of the diodes (very short MOSFETs with closely spaced S & D, and no gate, which avalanche at a few volts).

(b) (i) The activation energy can be easily extracted by plotting the $\ln(t_F)$ function of 1/$T$. The slope of the linear curve is equal to $E_a / k$:

$$\ln(t_F) = \ln(c_t) + E_a / kT$$

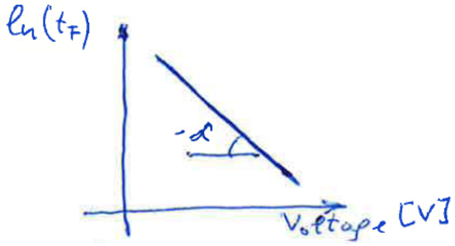


where:

$t_F$ = time to failure  
$c_t$ = constant,  
$E_a$ = activation energy,  
$k$ = Boltzmann's constant (8.6e-5 eV/K)  
$T$ = absolute temperature

The stress factor (e.g. voltage coefficient) can be easily extracted by plotting the $\ln(t_F)$ function of the stress (Voltage) at constant temperature. The slope of the linear curve is equal to $-\alpha$ :

$$\ln(t_F) = \ln(const) - \alpha S$$



where:

$t_F$ = time to failure  
$const$ = constant,  
$S$ = stress (Voltage)  
$\alpha$ = stress coefficient (Stress Voltage coefficient) [20%]

(ii) The acceleration factor can be calculated as $AF = t_{F_{use}} / t_{F_{test}}$ :

AF = $\exp[(E_a / k)(1/T_{use} - 1/T_{test}) + \alpha(V_{test} - V_{use})]$

= $\exp[(0.8/8.6\times10^{-5})(1/(273+40)) - (1/(273+125)) + 0.03(400-200)]$

= $561\times403$

= 226083

Average Failure rate in accelerated test (125 °C, 600V) = $1\times10^{-3}$ /h

Average Failure rate in the field of use (40 °C, 400V) = $1\times10^{-3} / 226083$

= $1.33\times10^{-8}$

≅ 4.4 FIT [30%]