4B7 VLSI Design, Technology & CAD 2003

Answers

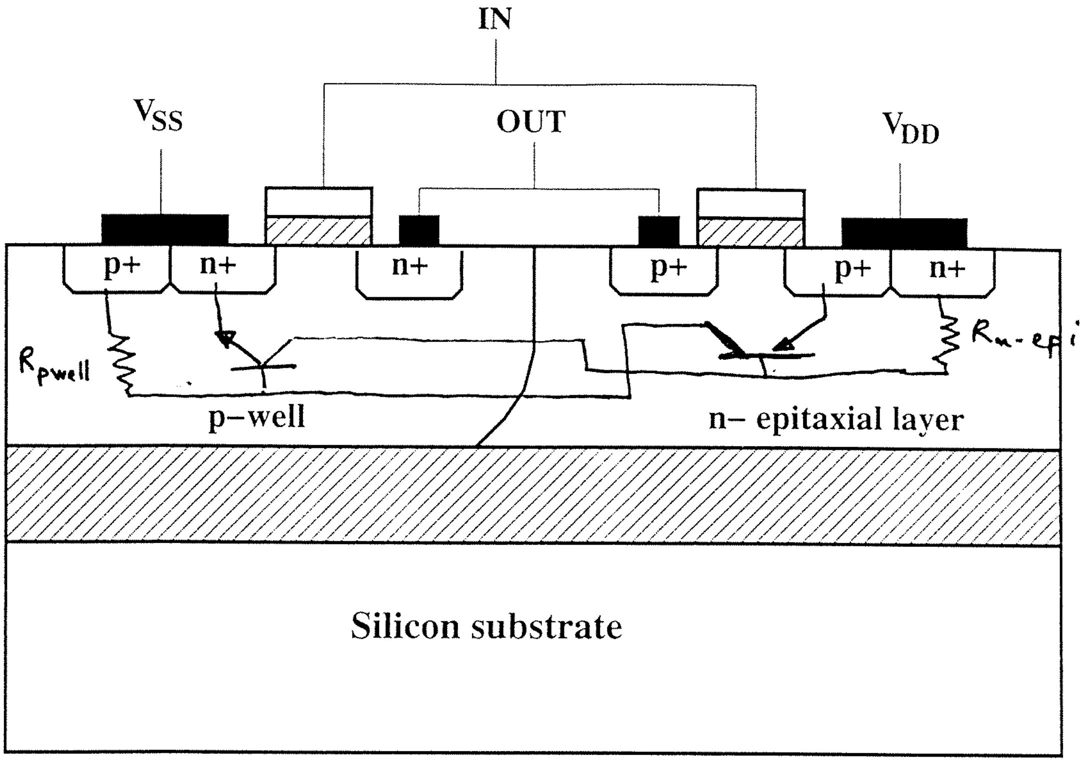
1. (a) The extra process steps required in SOI CMOS technology are [redacted] related to the fabrication of the SOI substrates. The fabrication of the SOI substrates can be done through one of the three well known methods (1) SIMOX, (2) Wafer bonding (3) Unibond. Following the SOI substrate fabrication, the process flow is similar to that used in standard bulk CMOS technologies.

The SOI technology offers (i) an increased level of isolation since the active devices are isolated from the substrate through the buried insulating layer, (ii) a considerably reduced (or suppressed) latch-up effect. (iii) The presence of the buried oxide also minimises the leakage currents and reduces cross-talk between adjacent devices. (iv) The SOI CMOS can operate at higher junction temperatures than the standard SOI without experiencing latch-up or parasitic bipolar effects. (v) Finally, owing to improved lateral trench isolation and vertical buried oxide isolation, there is no need for buried layers and lateral isolation rings, thus minimising the area consumption.

The main drawbacks of the SOI technology are (i) the high cost of the SOI substrates (approximately 5 times that of standard silicon wafers) and (ii) self-heating (the buried oxide acts as a thermal barrier, thus trapping the heat at the surface of the device) [30%]

(b) The twin-tub technology is based on a highly conductive substrate onto which an epitaxial layer is grown. This is followed by the n-well and p-well formation. The values of the parasitic p-well and n-well resistors which are placed between the base and emitter of the parasitic NPN and PNP transistor respectively are considerably reduced, thus providing a more effective short to the base-emitter junction. The technology also provides an effective sink (conductive substarte) for collecting the hole current of the parasitic pnp transistor. [20%]

©



NPN parasitic transistor
emitter : n+ source of the n-channel MOS
base : p-well
collector : n-epi layer

PNP parasitic transistor
emitter : p+ source of the p-channel MOS
base : n-epi [30%]
collector : p well

- The 'latch-up' parasitic structure is comprised of two bipolar transistors (NPN & PNP) displaced laterally in a thyristor configuration. The PNP is more likely to be turned on first because $R_{n-epi} > R_{pwell}$ and because α_{NPN} is generally greater than α_{pnp}.
- the latch-up condition can be written as: $I_{trigger} = \frac{V_{pnp-on}}{\alpha_{NPN} \cdot R_{n-epi}} \simeq \frac{0.7V}{\alpha_{NPN} \cdot R_{n-epi}}$ [20%]

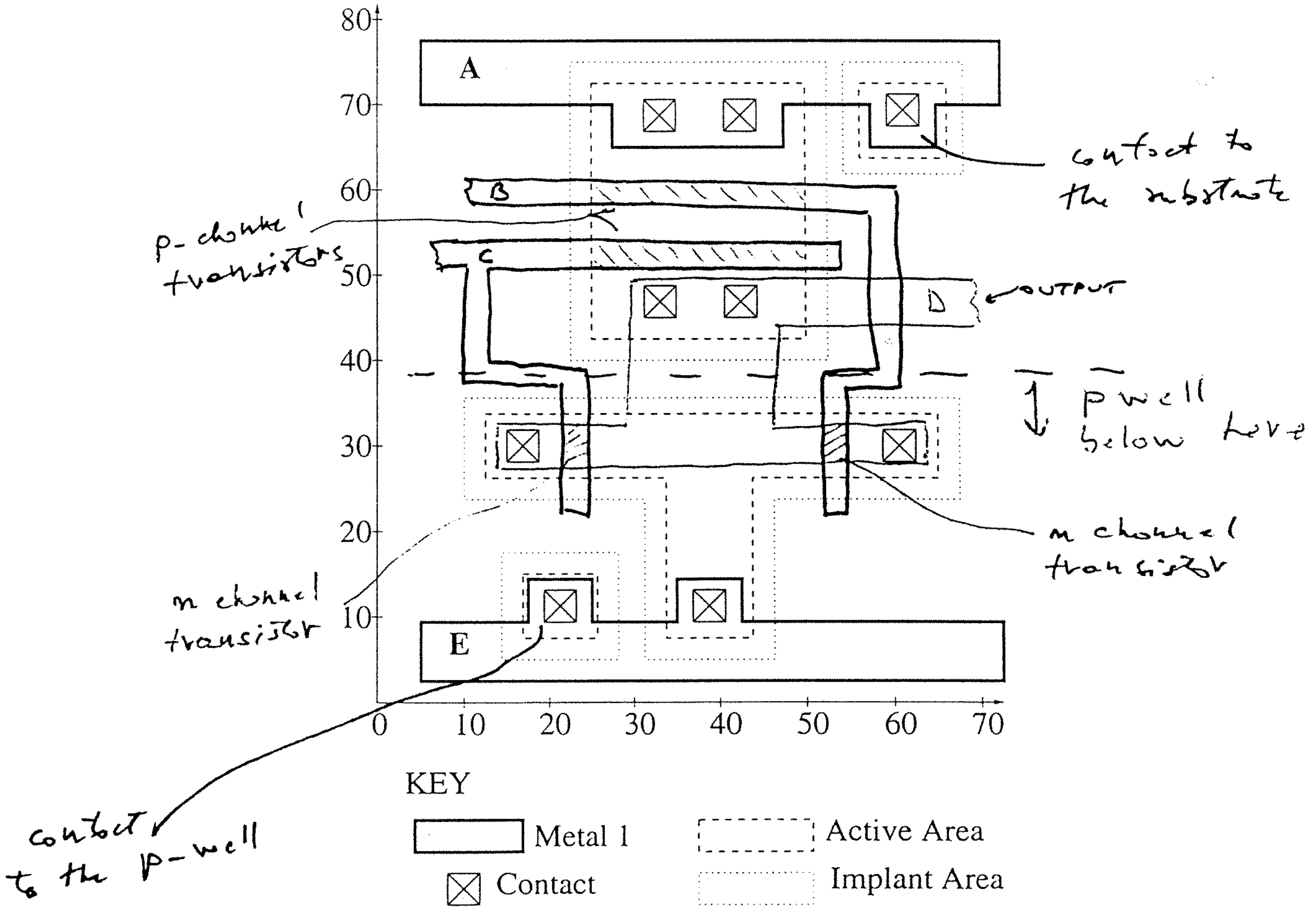
4B7 VLSI Design, Technology & CAD 2003

2. (a) Photolithography is the process of imprinting a geometric pattern from a mask onto a thin layer (~µm) of material called a *photoresist* (a radiation-sensitive material). First, a resist is usually either spin-coated or sprayed onto the silicon wafer and then a mask placed above it. Second, in optical lithography, UV radiation is used to change the solubility of the photoresist in a known solvent. Positive photoresists become more soluble on the exposure to the UV light whereas negative photoresists become less soluble due to a polymerisation process.

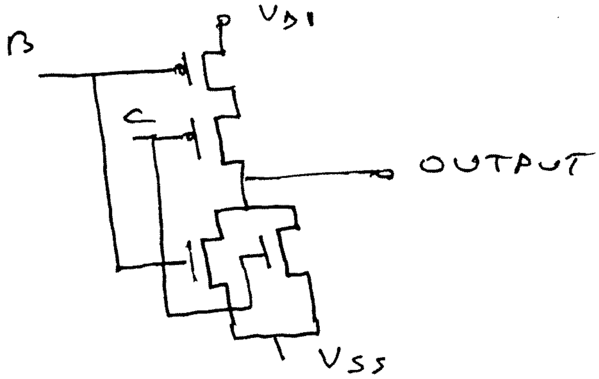
After the uncured photoresist has been dissolved away by washing it in an organic solvent, the exposed SiO_2 layer is then etched away by an HF solution and th remaining polymerised resist is burnt off. The photolithographic sequence is repeated for each masking step. Subsequent steps are aligned to previous steps through an alignment scheme.

The photolithographic resolution is dictated by the minimum poly gate feature. Today this is limited to 0.13 µm, but in the next two years it is expected to go to 0.09 um. The e-beam lithography offers greater precision and finer lines but it is still too expensive for commercial applications. [30%]

(b)



- This is a two input NOR gate



[25%]

see diagram on prev page

- The upper transistors are p-channel (connected in series to VDD) and the lower transistors are n-channel (connected in parallel with the source connected to Vss).

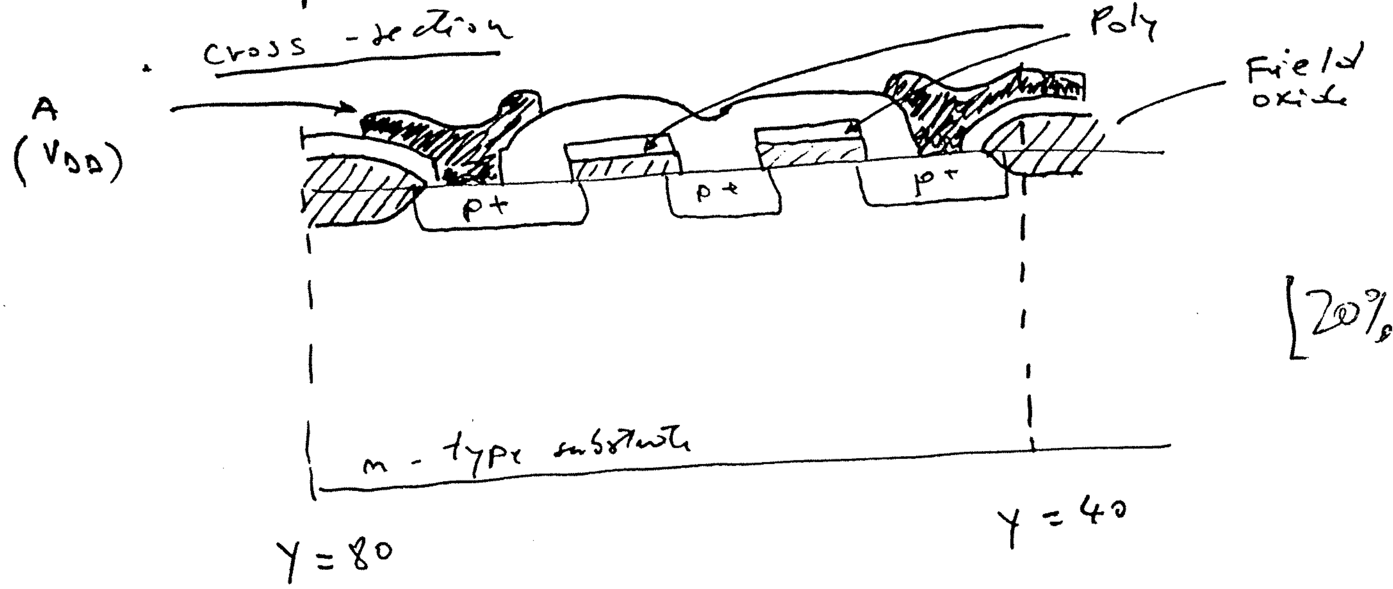
[15%]

- The upper transistors have lower mobility (hole mobility is lower than electron mobility). To improve the transconductance and obtain a more balanced gate, the width of the upper transistors is longer than that of the lower (n-channel) transistors.

- If the substrate is n-type then we must have a p-well where the n-channel transistors are placed.

10% + diagram

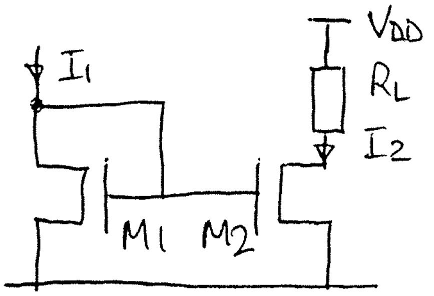
- cross-section



[20%]

2003 4B7 Qn 3

Current mirror circuit



Both transistors are arranged to operate in their saturation region. Note that M1 is by definition, since $V_{DS} = V_{GS}$ and $V_{DS} > V_{GS} - V_T$

M_2 may be held in saturation by attention to choice of the load R_2. Assume both M1 & M2 are in saturation. Then:-

$$I_{DS} = \frac{1}{2} \frac{\mu \epsilon}{t_{ox}} \frac{W}{L} (V_{GS} - V_T)^2$$

For M1 $I_1 = K \frac{W_1}{L_1} (V_{GS} - V_T)^2$

For M2 $I_2 = K \frac{W_2}{L_2} (V_{GS} - V_T)^2$

If we may assume M2 and M1 have same process transconductance K, threshold V_T; and noting both have same V_{GS}, we see

$$\frac{I_2}{I_1} = \frac{W_2}{L_2} \cdot \frac{L_1}{W_1} \quad \text{or} \quad I_2 = I_1 \frac{W_2}{L_2} \times \frac{L_1}{W_1} \qquad (1)$$

This is a current mirror. $I_2 = I_1$ if the transistors have identical dimensions. If I_2 is to be different from I_1, say, $6 I_1$ then

$$\frac{W_2}{L_2} \cdot \frac{L_1}{W_1} = 6$$

[30%]

Typically L is held fixed and W varied. Then $W_2 = 6 W_1$

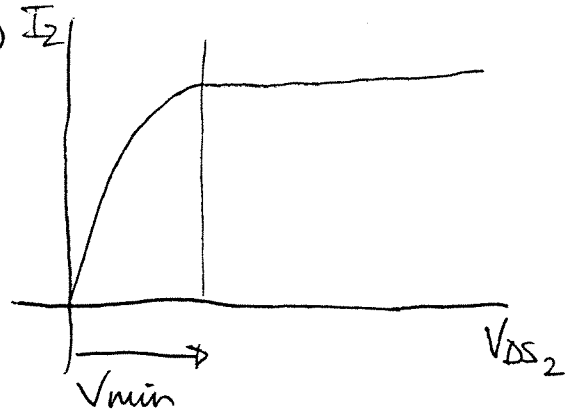
(b) Fabrication tolerances tend to result in systematic line width variations (bias too large, or too small, 2003 4B7 Q3

generally independent of the design dimension. If features are of the same size, these may be expected to TRACK; however, if $W_1 \neq W_2$ they may not. Hence to minimise this effect & ensure I_2 is as close as possible to the design value, it is preferred to lay out M2 as a set of paralleled transistors (6 in the sample given) of the same dimensions as M1. V_T may vary across the chip - keep devices close together [20%]

(c) From ① $\frac{W_2}{L_2} \times \frac{L_1}{W_1} = 4$. If L remains fixed, $W_2 = 4 W_1$

The choice of dimensions for the transistors needs to take account of power dissipation or energy density. If $V_{DS} \sim 5V$, then the dissipation in M2 is $5 \times 400 = 2mW$. For larger dissipations larger values of L and W achieving the same aspect ratio may be needed [30%]

(d)



Typical output characteristic. It can be seen that there is a minimum V_{DS_2} consistent with correct operation. The gradient of the characteristic in saturation is determined by the channel length modulation characteristic of the devices, λ, typically $0.05\ V^{-1}$, and the output resistance $\sim 1/\lambda I_D = \frac{1}{g_{ds}}$. Maximum output resistance is thus achieved with low currents. The output resistance may be raised by inserting additional resistance r in the source of M2. It may be shown [20%] that the new output resistance is:-

$$r_{out} = g_{m2} r \times \frac{1}{g_{ds}}$$

r is chosen such that $g_{m2} r > 1$

But note that this will increase the drop out voltage Vmin

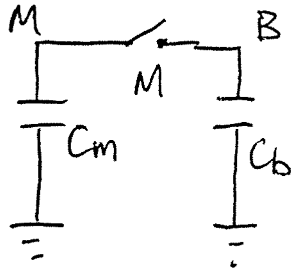
2003 4B7 Qn 4

(a) Clock skew – in sequential systems operation may need to by synchronised to a master clock. In a VLSI circuit different parts of the chip may receive the clock signal by different lengths of interconnect, so that the timing of the critical edges may differ. Clock skew may therefore arise from other sources, including:-

- differential delay along interconnect
- passage through different numbers of controlling gates or buffers
- need for extra inverters to form $\bar{\phi}$ from ϕ

The effects are as for a mistimed discrete circuit. Clock pulses may arrive too late to latch data before it decays to an unknown state. Several approaches may be used to minimise clock skew, including pipe-lining, use of buffers to split clock lines into smaller segments, and avoiding use of polysilicon for clock lines. Use of SOI also reduces delay through reducing parasitic capacitance [10%]

(b) A dynamic memory cell and its associated data line can be modelled as a pair of capacitors with an interconnecting switch (M). We must consider what is the resultant potential on the bus, V_s, immediately after the switch closes.



$Q_m = C_m V_m$ $\quad Q_b = C_b V_b$

Total charge $Q_{tot} = Q_b + Q_m$

Total capacitance $C_{tot} = C_b + C_m$ [30%]

Immediately after M conducts, the resultant voltage is

$$V_s = \frac{C_m V_m + C_b V_b}{C_m + C_b}$$

if $V_m = V_{DD}$ and V_b is initially V_{SS} then

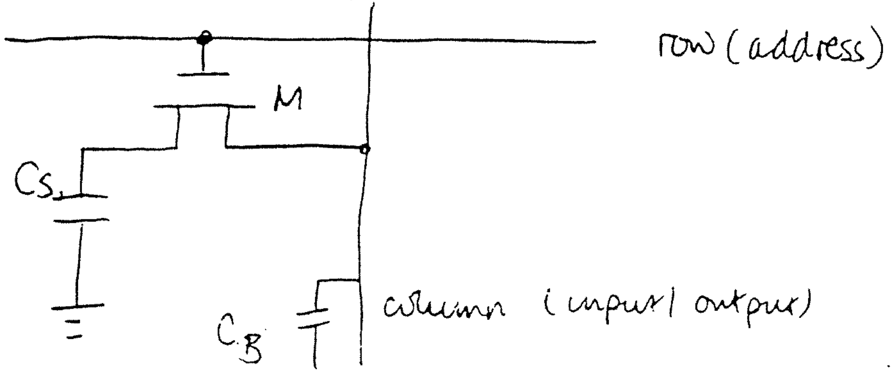
$$V_s = \frac{C_m}{C_m + C_b} V_{DD}$$

and for the potential at C_m to be reliably transferred, $C_m >> C_s$

This is the opposite of what is actually the case. A much

Q4 4B7 VLSI Design, Technology & CAD

Qn (b) A DRAM cell may be achieved using an n-channel MOS transistor in association with parasitic capacitor elements. This facilitates an extremely small cell and leads to high memory densities



C_S is a parasitic element, typical 50fF. Much effort has gone towards fabrica capacitors with the highest possible C and minimum area, eg trench cap.

Reading & writing are accomplished by applying logic high to the gate of M via the row/address line to select the cell. The cell must be refreshed periodically (O(10ms)) because of charge leakage from C_S

Data can be written into the cell by forcing logic 1 or 0 on the column/bit-line while the cell is selected. C_S charges to this value, which is retained when the cell is deselected.

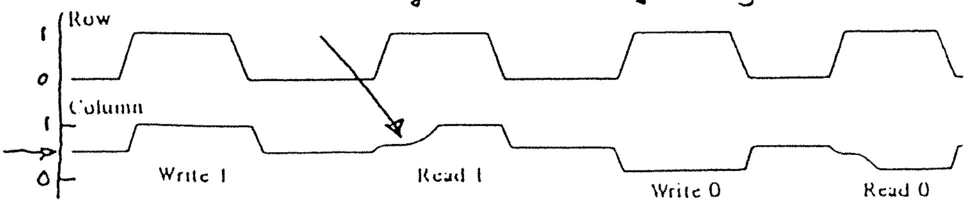
When reading the cell is selected by applying logic high to the row line, making M conduct. The column line is connected to a sensitive comparator.

Since C_S is very small and C_B may be significant (O(1pF)), a charge sharing analysis shows that the potential change observed on the column line may be of order mV. Design of suitable sensing comparator in a noisy environment is a challenge.

Normally a regenerative amplifier is used, ~~and~~ the column line is precharged to the mean of the logic levels eg 2.5v

Small change due to charge on C_S

[30%]



precharged to $V_{DD}/2$

4B7 Q4

(d) Assume the following stage is designed to switch at $V_{SW} = V_{DD}/2$ and that the capacitor is charged to $V_{DD} = 5v$ as stated. Hence, if C loses half its charge, having been set to logic 1, it will (incorrectly) indicate logic 0. Leakage is at a fixed rate of 0.1 nA, assumed independent of potential.

Hence time taken to discharge to 2.5v is

$$\frac{C \times (5-2.5)}{I_{leak}} = \frac{60 \times 10^{-15} \times 2.5}{0.1 \times 10^{-9}}$$

$$= 1.5 \times 10^{-3} = 1.5 ms$$

The cell must be refreshed more often than this.

If the bus line is at 2.5v, and the memory capacitor is charged to 5v, then using charge sharing:-

$$V_{sense} = \frac{(0.06 \times 5 + 1.5 \times 2.5) \times 10^{-12}}{(1.5 + 0.06) \times 10^{-12}}$$

$$= \frac{0.3 + 3.75}{1.56} = 2.59_6 \text{ V}$$

[30%]

Hence the change in potential observed is 96mV

4B7 VLSI Design, Technology & CAD

5(a) The resistance of a uniform rectangular slab of conducting material is written

$$R = \frac{\rho}{t}\frac{l}{w} \quad (1)$$

where ρ is the resistivity of the material, t its thickness, l & w are the conductor length & width

This may be rewritten

$$R = R_s.\left(\frac{l}{w}\right) \quad (2)$$

where $R_s = \rho/t$ and incorporates material as well as the thickness.

Rs may thus be viewed by the circuit designer as a process constant, since neither ρ nor t may be controlled by the circuit designer, whereas l & w may.

The units of Rs are Ω/square, being the resistance of a square of the material (of arbitrary side)

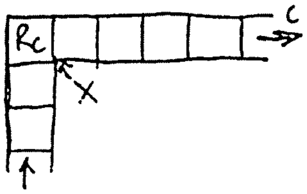
Thus to obtain the resistance of a conductor of rectangular form (2) may be used. For a conductor formed from a series of abutted rectangles an expression like:

$$R = R_s \sum_i \frac{l_i}{w_i}$$

may be used.

[30%]

Where corners appear the pattern of equipotentials in the conductor is distorted. A finite element analysis shows that the inserted resistance is very sensitive to the curvature at X, which may not be well defined for many cases.



However, a satisfactory approximation is obtained by taking the resistance of a corner square Rc as 0.66 Rs.

A similar approach can be used to evaluate the effective resistance of MOSFET channels formed into serpentines or other folded structures

4B7 VLSI Design Technology & CAD

5(b) An MOS transistor consists electrically of charge stored in the dielectric layers, in the surface/surface interfaces, and in the substrate (or well) itself. Switching an* MOST from OFF to ON consists of applying a gate potential to neutralise these charges and allow the underlying semiconductor to undergo an inversion due to the E-field from the gate.

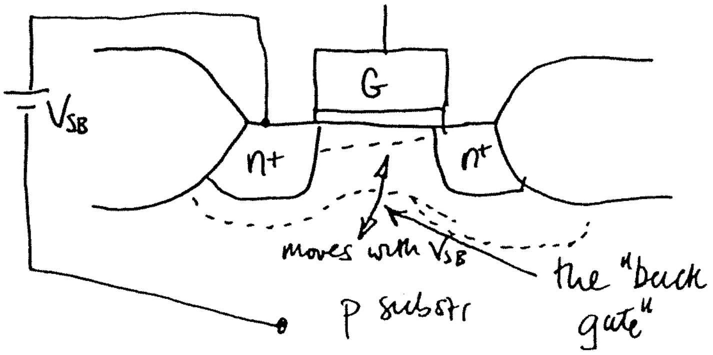
NB enhancement mode - The threshold voltage Vt can be written:-

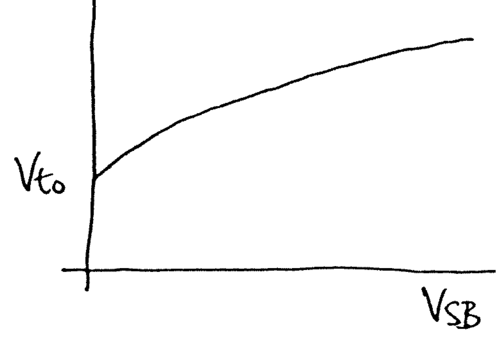
$$V_t = \phi_g + \frac{Q_B - Q_{ss}}{C_o} + 2\phi_{fN}$$

ϕ_g is the W.F. between gate & Si (very small)
ϕ_{fN} is the Fermi potential between inverted surface & bulk Si
C_o is the capacitance per unit gate area
Q_{ss} is the charge density at the Si:SiO2 i/f in the channel
Q_B is the charge in the depletion region beneath the gate oxide.

With the exception of Q_B, these are dependent only on physical parameters & process params. However Q_B depends on ϕ_{fN} and the potential between the transistor source and the substrate, V_{SB}. This is the so called BODY EFFECT

Increasing V_{SB} causes the channel charge to be depleted; the perceived effect is that V_t is raised, for a single transistor.





Change in V_t is given by

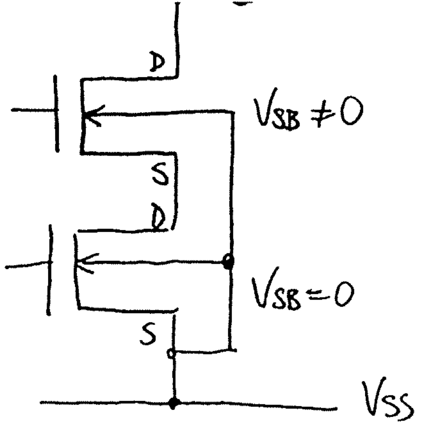
$$V_t = V_{to} + \gamma V_{SB}^{\frac{1}{2}}$$

for nMOS devices, where V_{to} is the threshold voltage for $V_{SB} = 0$

γ is typically 0.5 → 1.5, being process dependent.

Where transistors are connected in series, as in 2/3/4... input logic gates, hand computation of the transfer func is difficult. The SPICE simulator can model this effect accurately

4B7 VLSI Design, Technology & CAD



The upper transistor has a higher V_T than the lower, owing to body effect.

This means that for multi input gates, the switching level ($V_{DD}/2$) for an inverter) is raised (NAND gates) or lowered (NOR gates)

This has the subsidiary effect of eroding the noise margins in a corresponding way.

The switching level (and noise margins) will change according to which input, or combinations of inputs, change in a transition.

As far as transient response is concerned, transistors exposed to significant body effect will have a lower apparent conductance in the ON-state (assuming a fixed V_G) owing to the elevated V_T. As a result, multi input gates will exhibit slower rise/fall times when parasitic capacitances are charged / discharged thru series transistors subject to B.E.

[40%]

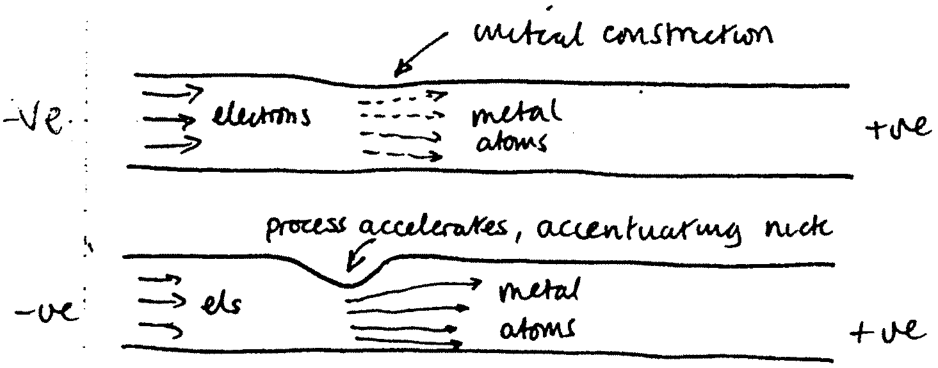
The designer can compensate for this affect by selecting devices of greater W/L in proportion to the higher resistance in the on-state of affected devices. Such compensation would probably have to be done in the light of detailed analogue simulation.

VLSI Design, Technology & CAD

5(c) Electromigration can result in voids appearing in metal lines carrying current, with consequent risk of device failure.

As current flows through a metal line, the electrons constantly bombard the metal atoms. Under severe conditions the momentum of the electrons is sufficient to push the metal ions aside and cause them to drift towards the positive terminal, resulting in the development of local voids at the negative end.

As more atoms are pushed away, the void becomes larger, increasing the current density in the remainder of the cross section, and hence increasing the momentum of the eletrons; as a result, the process is accelerated there. Eventually, the conductor will fail at that point as the process gets more and more rapid compared with the remainder of the conductor.



The designer can minimise the risk of electromigration-induced failure by keeping current densities low, which demands that all power rails be adequately broad. Maximum J is typically 10^9 Am^{-2} for aluminium, translating to a typic 'rule of thumb' of 1mA current maximum per μm width.

Other factors may affect the rate of electromigration:-

- grain size of metal
- temperature
- duty cycle (AC or DC current flow)
- metal type

[30%]

as well as current density