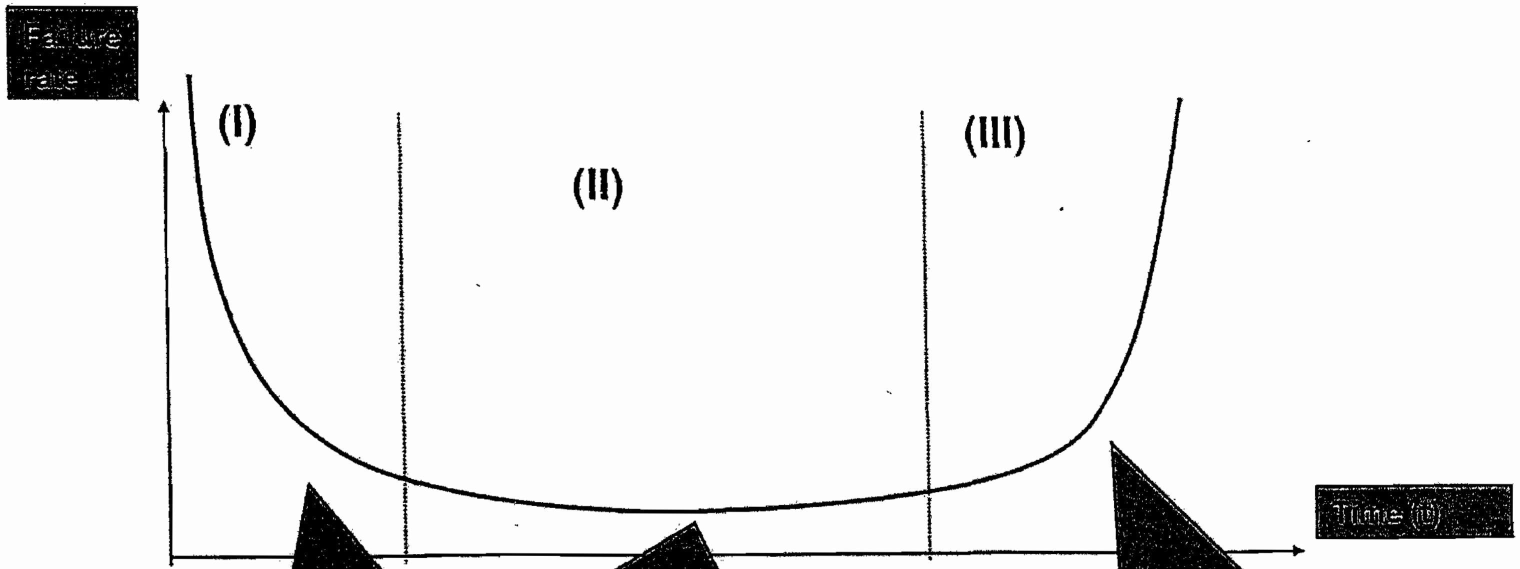
2008 IIB 4B7 VLSI DESIGN, TECHNOLOGY AND CAD

DR D M HOLBURN

Q1 (a)

VLSI Reliability Physics and Failure Mechanisms

- Bath-tub curve : Typical failure rate curve of VLSI products



Period I
1. Early failure period (infant mortality)
2. Due to extrinsic failure (defect in design or manufacturing)
3. weibull distribution

Period II
1. Useful life period(random fail)
2. Failure rate is constant
3. Exponential distribution
4. Reliability test must assure that region (III) is far away from this period

Period III
1. Wear-out failure period
2. Wear-out due to intrinsic failure such as fatigue or aging
3. Lognormal distribution
4. Most of the root cause is GOI

(b)

		Failure mechanisms	**Failure signature**
(i)	Oxide film:		
		Mobile ions	decreased breakdown voltage
		Pinholes	
		Interface states	short-circuit
		Time dependent dielectric breakdown	increased leakage currents
		hot carrier injection	gain reduction or Vth drift
(ii)	Metallization		
		Scratch or void damage	
		Mechanical damage	open circuit
		Non-ohmic contact	short circuit
		Step coverage	increased resistance
		Weak adhesion	
		Improper thickness	
		Corrosion	
		Electromigration	
		Stress migration	
(iii)	Passivation		
		Pinhole or crack	decreased breakdown voltage
		Thickness variation	short circuit
		Contamination	increased leakage current
		Surface inversion	gain reduction or Vth drift
			Noise deterioration
(iv)	Die or wire bonding		
		Die detachment	open circuit
		Die crack	short circuit
		Wire bonding deviation	increased leakage
		Off-centre bonding	
		Damage under pad	
		Disconnection	
		Loose wire	
		Contact between wires	

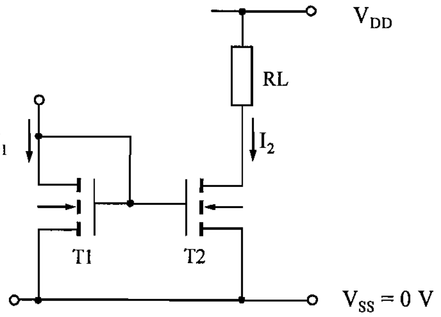
(c) Notes on three failure modes:

Electromigration: metal tracks are multigranular with oxide films at some places between grains, and currents have very high densities. The field lines can be distorted by the oxide layers and gradients of fields can produce forces on the metal so that the metal may move between grains. This happens at the highest field gradients which occur with small grains and near defects, forming voids and hillocks. The initial flow of material exacerbates the problem left behind, and the metal motion avalanches towards fusing.

Time dependent dielectric breakdown: good quality thermal oxide films have a dielectric breakdown strength of 10MV/cm and more. Over time, and under conditions of repeated use, the films can breakdown at lower electric field intensities. Hot electrons can form traps and under electrical stress can modify the properties of the oxide, so that eventually the films will breakdown under normal operating conditions.

Hot carrier Injection: under high fields, some electrons and holes can gain substantial energy before scattering. The can have enough energy to overcome the barrier between the Si and the thin gate oxide film. In addition to forming traps as above, their electrical presence in traps can affect the threshold voltage so that a space charge or inversion layer forms even when the device has no gate bias. The threshold voltage shifts the transconductance degrades, and the overall field distribution in the device and circuit can be altered leading to breakdown.

Q2



Current mirror circuit
Both transistors are arranged to operate in the saturation mode. Note that T1 is by definition in that mode since $V_{DS} = V_{GS}$ and $V_{DS} > V_{GS} - V_T$. T2 must be held in saturation by suitable choice of R_L.

(a) Assume T1 and T2 are in saturation mode. Then

$$I_{DS} = \frac{1}{2}\frac{\varepsilon_0\varepsilon_r}{t_{ox}}\mu\frac{W}{L}(V_{GS}-V_T)^2$$

$$\text{For T1, I1} = k\frac{W_1}{L_1}(V_{GS}-V_T)^2 \quad \text{where } k = \frac{1}{2}\frac{\varepsilon_0\varepsilon_r\mu}{t_{ox}}$$

$$\text{For T2, I2} = k\frac{W_2}{L_2}(V_{GS}-V_T)^2$$

Note that T2 has the same k, VGS and VT as T1

$$\text{Hence } \frac{I_2}{I_1} = \frac{W_2}{L_2}\cdot\frac{L_1}{W_1} \text{ and } I_2 = I_1\cdot\frac{W_2}{L_2}\cdot\frac{L_1}{W_1}$$

(b) Fabrication tolerances tend to result in systematic linewidth variations, bias too large or too small, largely independent of the design dimension. If features are of the same size, these TRACK. However, if W2 ≠ W1, they may not track. Hence to minimise this undesired effect and ensure that I2 is as close as possible to 6 I1, design the circuit such that T2 consists of 6 identical paralleled transistors of the same dimensions as T1. This overcomes systematic variations, but leaves unaffected any statistical variations in W, L, which have a direct effect on the ratio I2/I1

(c) This is a current mirror, and I2 = I1 if the transistors are of the same size. If I2 is required to be different from I1, say, 4I1, then

$$\frac{W_2}{L_2}\cdot\frac{L_1}{W_1} = 4$$

It is normal to keep L constant and vary W. If so, W2 = 4 W1

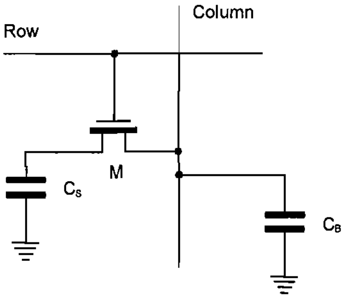
(d) To extend the circuit, note that the n-channel device cannot in this configuration drive a load connected to VSS, but a p-channel device can. Hence, use a further transistor T3, identical to T1, to generate a current sink equal to I1. Use this to drive a further 10:1 mirror implemented in p-channel devices.

Very few knew how to extend the cct

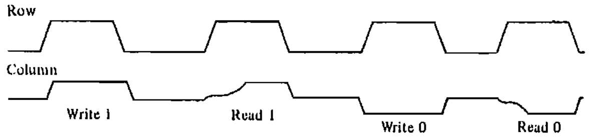
W/L ratios may of course be scaled as needed to provide greater current capability. Note that T2 and T5 should be implemented as multiple paralleled devices if the precise ratios I2/I3 and I3/I1 must be maintained.

Origin of CS and CB not known by many

Q3 (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. CS 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. CB is the capacitance due to the drain/source of M and the bus interconnecting all cells: may be significant o(1pF).
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 CS. Data can be written into the cell by forcing logic 0 or logic 1 on the column/bit-line while the cell is selected. CS 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 CS is very small and CB may be significant, a charge sharing analysis shows that the potential change observed 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.



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 CB and the sense amp are precharged to VDD/2 or 1.5 V and CS 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 CS 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 CS = 30

$$10^{-2} = \frac{30 \times 3 + 1.5\, C_B}{30 + C_B} - 1.5$$ (all C in fF)

Hence $90 + 1.5\, C_B = 1.51 \times (30 + C_B)$ and: $90 = 44.7 + (1.51 - 1.50)C_B$

$$\frac{45.3}{0.01} = C_B \to C_B = 4530 \text{ fF}$$

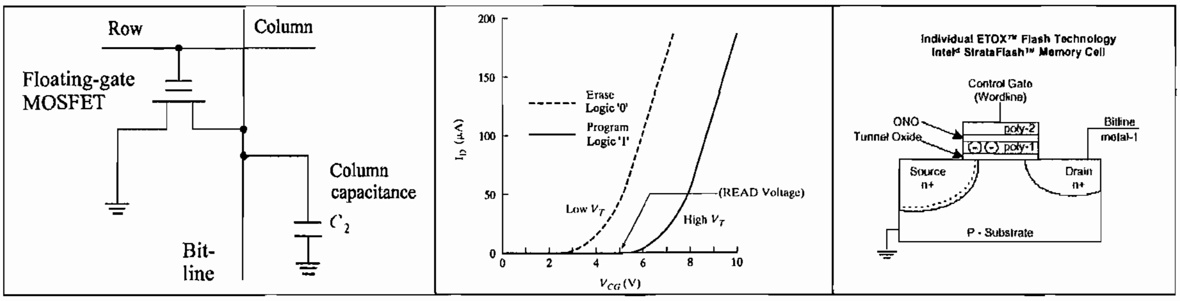
** Many overlooked this ...*

.. and C at input to sense amp

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 5fF due to the MOSFET* and 5 ×0.6 fF due to the 5 μm length of bit line. If there are N cells in the direction of the bit-line, total C = $N(5 + 5 \times 0.6) = 8 \times N$ fF. Hence Nmax = 4530/8 = 566.

Since the array is said to be square its max size is 566^2 ~ 320.3 kbits. We've only considered bit-line capacitance here & not accounted for any C at input to amplifier.

(c) **'Flash' memory** - an important type of non-volatile memory, yet has density and speed of operation associated with the DRAM. It has a very simple structure and compact layout - see diagram. The cell closely resembles the one-transistor DRAM cell, except that there is no storage capacitance, and the MOSFET used has an additional *floating gate* between the control gate electrode and the channel. The dielectric separating the floating gate from the control gate is typically a 'sandwich' comprising oxide-nitride-oxide (ONO). The floating gate is electrically isolated, but is capacitatively coupled both to the control gate and to the underlying silicon.



Write operation - consists of placing carefully measured amounts of charge on the floating gate so as to 'program' the MOSFET to have two different values of V_T.

- If the floating gate contains a large electronic charge, the MOSFET has a higher value of V_T (measured at the control gate) and can be considered to be 'programmed' to the **logic '1'** state.
- If the charge is removed from the floating gate, the MOSFET has a lower value of V_T and the cell can be considered to be 'erased' to the **logic '0'** state.

Transferring charge in – a high electric field is applied to the drain (bit-line) and to the control gate (row) so that the MOSFET is in saturation. The carriers in the pinch-off region are then highly energetic (hot). If the kinetic energy of the electrons is sufficiently high, a few can become sufficiently hot to be scattered into the floating gate. Once in the floating gate, electrons become trapped in a potential well, and can remain indefinitely without being discharged.

Erase operation - involves removing charge from the floating gate. This is achieved through use of *Fowler-Nordheim* tunneling between the floating gate and source electrode. The control gate is grounded and a high voltage (say 12 V) is applied to the source. The resultant field allows electrons to 'tunnel' through the oxide barrier from the floating gate to the source.

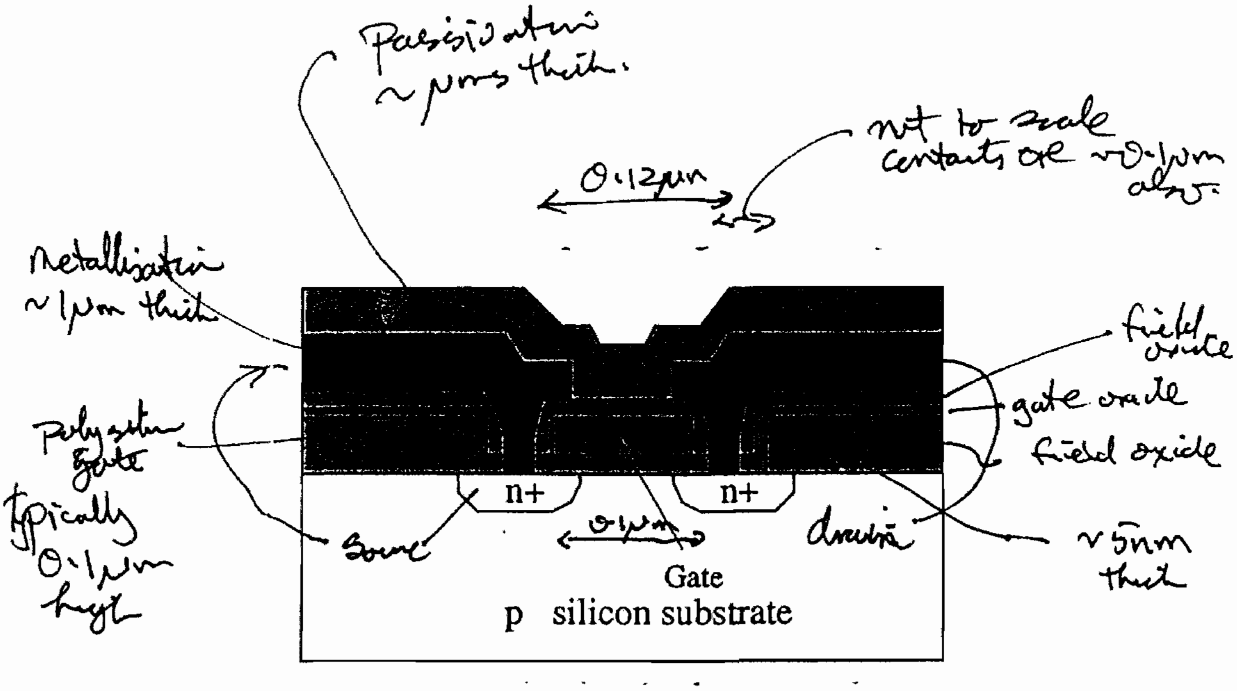
Read operation - is accomplished by applying a moderate voltage (say, 2.5 V) to the drain of the device (bit-line), and a *Read* bias voltage is applied to the control gate.

- If the device is in the **'1'** state, negligible current will flow since the control gate voltage is insufficient to cause a channel with the high V_T.
- If the device is in the **'0'** state, the control gate voltage exceeds the lower V_T, and drain current flows.

The current can be sensed to read out the logic value. Note there is still a delay due to the charge/discharge of the bus capacitance C2, as with the dynamic RAM cell.

More advanced forms of flash memory are now available, in which several different values of V_T may be programmed by injecting different amounts of charge. In this way a single cell can store more than one bit of data.

Q4 (a)

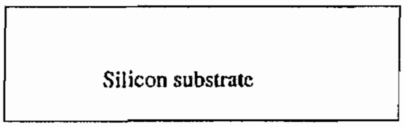


Operation: no bias on gate
no current from source to drain
as either the source or the drain
p-n junction is reverse biased.

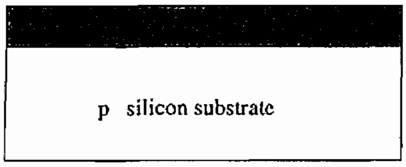
bias on gate : draws electrons
to the interface & form a narrow
n-channel to connect electrically
the source & drain contacts ⇒
current flow.

(b)

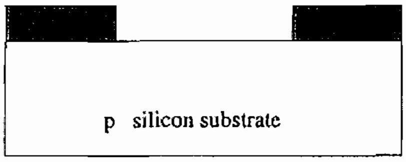
Something reflecting this pattern & sequence.



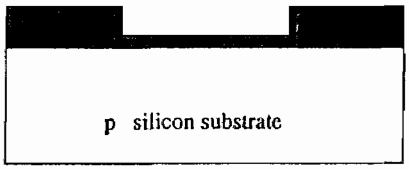
1. start wafer: lowly doped p–type



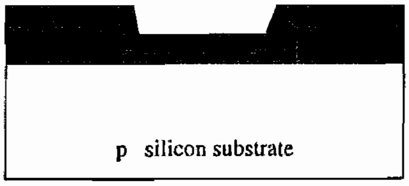
2. oxidation (formation of field oxide)



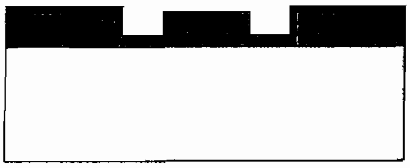
3. field oxide patterning and etching



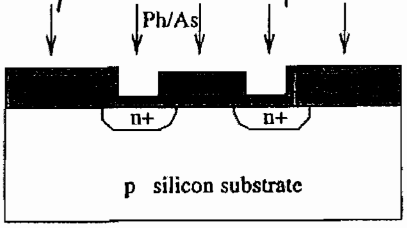
4. gate oxide growth



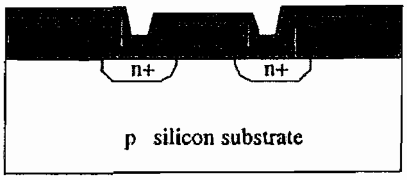
5. polysilicon deposition



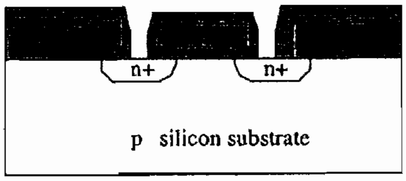
6. polysilicon patterning and etching



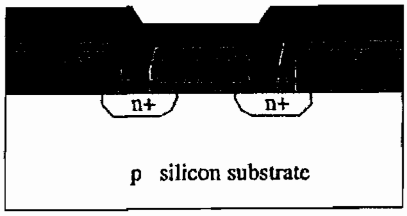
7. implant of + source and drain (As, Ph) followed by diffusion
(note that high energy implants can go through thin oxides)



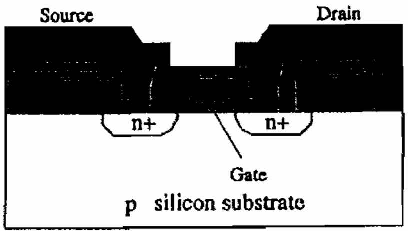
8. depostion of an insulated layer(oxide) or polysilicon oxidation



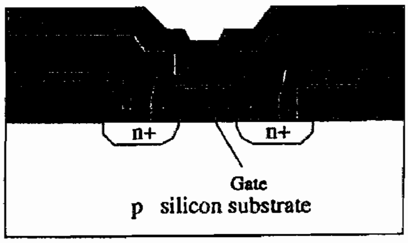
9. oxide etching to form contact windows



10. metalisation



11. metal patterning and etching

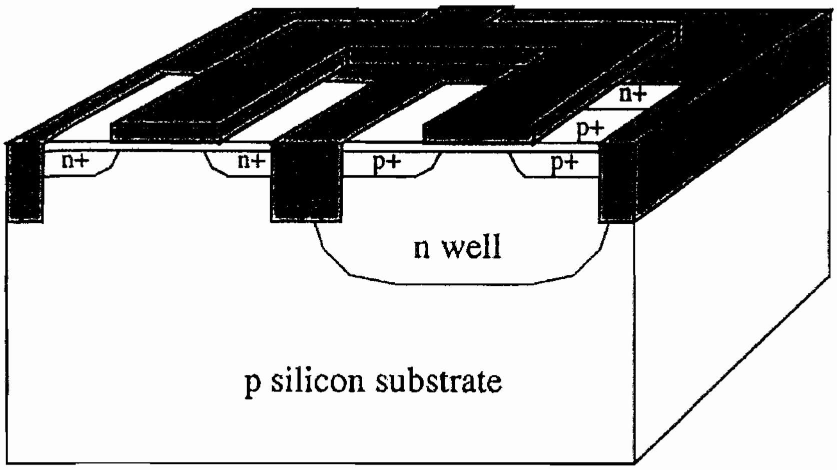


12. passivation (pads are not shown)

but note on lithography, etching, ion-implantation & metallisation

(c)

(c) CMOS.



Extra step: n well
p+ implants
& n+ implant to grnd the
other transistor (p-channel)

(d). Gate capacitance: $C = \frac{\epsilon A}{d}$.
Charge under bias $V = VC$

$$\# \text{ of electrons} = \frac{CV}{e} = \frac{4\times10^{-11}\left(0.05\times10^{-6}\right)^2 \times 3}{2\times10^{-9} \times 1.6\times10^{-19}}$$

$$\approx 1000$$

$\sqrt{1000} = 30$ still a reasonable margin above severe statistical fluctuation:

Note: 0.05 μm → 0.01 μm a only 40 electrons!

Q5 (a) The resistance of a rectangular slab of conducting material is written

$R = \frac{\rho\ \ell}{t\ w}$ (1) where ρ is the resistivity of the material, t its thickness l and w are its length and width. This may be re-written.

$R = R_S\left(\frac{\ell}{w}\right)$ (2) where $R_S = \rho/t$ and incorporates material parameters as well as the thickness.

R_S may be viewed by the circuit designer as the process constant since neither ρ nor t may be controlled by the designer whereas l and w may.

The units of R_S are ohm/square being the resistance of a square of the material of arbitrary side.

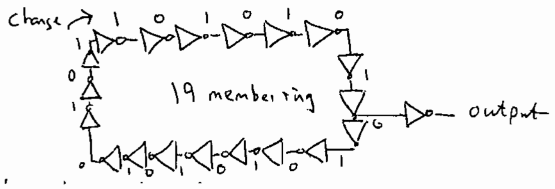
Thus to obtain the resistance of a conductor of rectangular form (2) may be used. for a conductor formed from a series of abutted rectangles an expression like

$R = R_S \sum_i \frac{l_i}{w_i}$ may be used.

Where corners appear the pattern of equipotentials in the conductor is distorted. A finite element analysis shows that the measured resistance is very sensitive to the curvature at acute-angled corners, 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.

(b) Ring Oscillator Circuit



19 member ring

An odd-numbered ring of inverting gates is unstable and oscillates with a period corresponding to 2n gate delays for an n-membered ring because a disturbance propagates round the ring. It is important to have a minimum geometry output gate between the ring devices at the output pad to avoid unnecessary loading of the ring.

In this example 38 gate delays give a periodic signal output waveform of manageable frequency that can be transmitted through the output pads to e.g. a frequency counter.

Many did not mention this gives an average figure

A third order resonance of the ring can also be excited, whereby three consecutive disturbances go round the ring giving the impression of 3x higher performance at the output gate. Higher order resonances are also possible. No second order (or even-order) disturbance can be sustained in an odd-numbered ring.

The simple 19-inverter ring gives an optimistic measurement of circuit performance because the lightly loaded devices switch fast. With a fan-out of 2 or 3 and a long connecting line to the next device, the RC time delay is increased as the switching speed is typically halved.

Q5 (c)

(i) Silicon on sapphire:

The (T011) plane of sapphire has a unit cell that is close to being commensurate with the (100) surface of silicon. Very smooth and clean surfaces of sapphire can be prepared. Growth of high quality thin films of silicon can be deposition by vapour epitaxy. However the pace of the advance in making large area sapphire wafers has not kept pace with that of silicon, so that alternatives have been needed for commercial CMOS production. SOS is particularly radiation hard, so early applications were accelerate for the needs of the space industry.

(ii) SIMOX

A heavy enough dose of high energy (2MV) oxygen can be implanted and then annealed to form a layer of SiO2 buried about 1 micron below the surface. It took some time to perfect the conditions, but is now compatible with handling the largest Si wafers.

(iii) BESOI and Unibond/smart cut

Two wafers with thin oxide layers can be bonded together and the excess silicon on the upper layer can be polished back to a new microns thickness. The thickness uniformity is not sufficient for the most modern IC production. Instead a He layer can be implanted a micron or so below the surface of one wafer to produce a mechanically weak layer at which point heat treatment can be used to lift off the remaining silicon.