ENGINEERING TRIPOS PART IIB 2012 DR D M HOLBURN

4B7 VLSI Design, Technology and CAD - 2012 Solutions

1. (a) Total QA system covers Design (for reliability, verification, packaging, learning from prototyping,...)

Quality control in production (process control, handling of design changes, parts and materials control, environmental control, in process management,)

Total quality management – quality data analysis, periodic reliability testing, warehousing and shipping management, verification of defined product quality)

Customer support – for reassurance, handling complaints, integrity of information held, ...) Quality and reliability improvement – failure analysis, data collection, reliability engineering and failure physics. [40%]

(b) (i) early failure regime, ESD, latchup, thermal runaway

(ii) useful life period with random failures: dielectric breakdown, moisture, contamination, ..

(iii) end-of-life failures from wear-out etc, contamination, moisture, electromigration, hot carrier degradation, mechanical or mechanical failure,.... [30%]

(c) Arrhenius equation: form tf= c exp (Ea/kT) with tf=time to failure, Ea=activation energy, T the absolute temperature and c a constant.

Used in general to try and capture a whole series of processes that require some level of thermal activation to occur.

Thermally stressing a device will shorten the time to failure as T increases.

Same with applied voltage or more general stress S: add in a factor exp ($-\alpha S$)

This analysis forms the basis of accelerated life tests, and failures such as electromigration can be analysed this way. [30%]

Examiner's note: The best candidates had done further reading on the material.

2. (a) (i) SOI technology – straight from notes – improve speed of device, remove weaknesses of free changes induced Si substrate during processing, better isolation between device,...

(ii) Scaling of oxide requires SiO2 to be only 2 monolayers thick – hard to control – used high K dielectric for gate oxide so that a thicker layer can still be deposited with adequate control but still have same capacitance.

(iii) Multilayer metallisation – needed to exploit the full functionality of the chip through relevant interconnections, 7 levels and more. Small local wires at the bottom, longer global and larger cross-section wires at the top. Now in Cu, with near air (gel,...) dielectric. [60%]

(b) Layout: Following the process schedule without too many diversions, concern about gases, pumps, environment, parts, and supply of other relevant materials, ...

Start-up: calibrations (many types), successful prototyping, collection of adequate data base for quality reassurance, ... [40%]

Examiner's note. A popular question, well handled by most candidates, who generally gave a good account of silicon on insulator technology in its various forms.

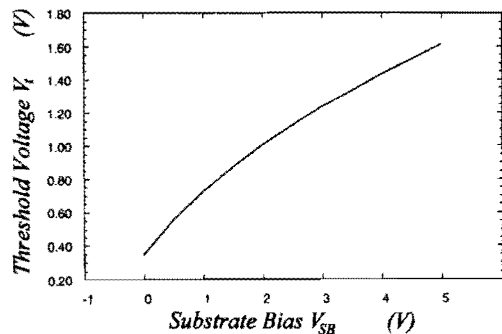
3. (a) The threshold voltage V_T of a MOSFET is that potential which must be applied between gate and source in order to bring about strong inversion within the channel region. There are three main components to this potential:

- ϕ_{GC}, the difference in work functions between the gate material and the Si substrate on the channel side;
- a negative potential arising from the existence of undesired positive charge withinj the gate oxide and at the oxide/substrate interface – referred to as Qox, and assumed to reside entirely at the interface;
- a voltage $-2\phi_F - Q_B/Cox$, needed:
    - to bring the surface potential to the strong inversion condition;
    - to offset the induced depletion layer charge, QB, i.e. to 'unbend' the energy bands that result when the MOS system is first brought together, and to bring the surface potential ϕ_S to be equal to ϕ_F

In essence, the intrinsically p-type semiconductor becomes n-type with this gate potential applied. Further increases in V_{GS} produce only slight changes in surface potential ϕ_S.

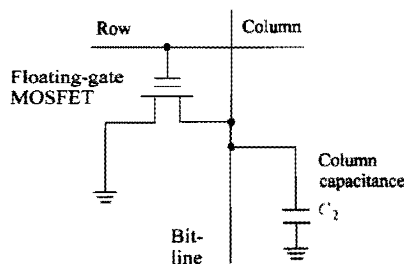
The main factors determining VT are:

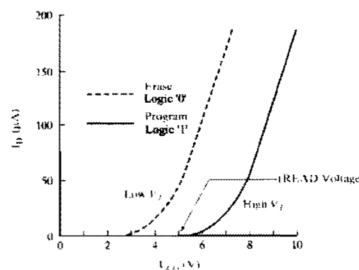
- the materials used for the gate electrode (Al, polySi, ..), determining its work fn
- properties of the dielectric used for the gate insulator, fixing capacitance C_{ox}, and its thickness t_{ox}
- channel dopant density
- impurities, defects, dangling bonds etc at the Si-SiO_2 interface;
- potential between source and substrate – which acts as a second or "back"-gate – see right
- temperature

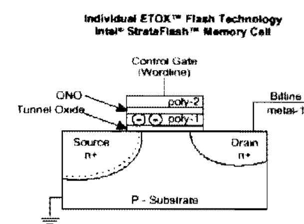


[30%]

(b) **'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. [50%]

Advantages –

- Compact, high density
- Non-volatile
- No capacitor needed
- Less sensitive to charge sharing & noise

Disadvantages –

- More complex fab process
- Slower write operation
- Need for higher voltages

[20%]

Examiner's note: a surprising number of candidates dismissed the discussion of threshold voltage in a couple of sentences, when it was intended to occupy about one-third of the question.

4. (a) CMOS VLSI designs for digital applications are normally implemented with minimum geometry designs wherever possible:
(i) to economise on space

(ii) for faster operation

(iii) for lower power consumption.

Devices like these are unsuitable for direct connection to external circuits since these may impose large capacitive, resistive or inductive loads, and may give rise to transient voltages or currents outside the normal safe operating range, which can induce latchup, and, in extreme cases, can cause permanent damage through static discharge.

To facilitate wired connections it is customary to use bonding pads – squares/rectangles of metallisation typically o(100um) in dimension to which fine wires may be bonded by ultrasonic cold-welding. These wires themselves contribute R, Land C, as do external elements.

Output drivers – pad drivers - must therefore be provided to interface between the min. geometry devices and the pads. They consist of high current-capacity (large W/L) devices which can supply the transient current surges needed to charge/discharge the pad and attached external circuit. They may also have to supply static or dynamic currents to resistive or inductive loads. However, because of their large channel area, such devices have large input devices and themselves need to be driven by transistors or greater size than minimum dimensions. [30%]

(i) To reduce the area occupied:

- use interdigitated/folded structures to economise on space
- successive stages are progressively increased in size and drive capability. The number of stages is minimised subject to constraints of acceptable delay.

(ii) To reduce risk of latchup

- all devices should have multiple well/substrate taps to Vdd/Vss.
- Guard rings may be incorporated to minimise risk of injecting minority carriers into other latchup-sensitive circuits [10%]

(c) To drive a high C load it is necessary in order to minimise delay, to use stages of progressively increasing W/L. Later devices are able to conduct higher current to charge/discharge nodal capacitances, which are themselves bigger because of the use of larger devices.

(i) It can be shown that the optimum number of stages to minimise delay is ln (Cpad/Cgate), where Cpad is the pad capacitance (including external driven capacitance), and Cgate is the capacitance at the input to the pad driver.

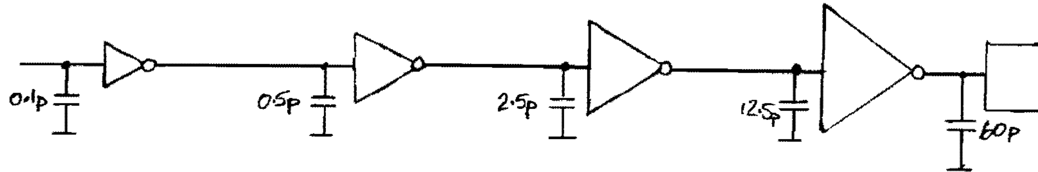
Here 4 stages are to be used. If each stage has its W/L U times that of the previous stage, for the values given: U^4 = Cpad/Cgate = 60/0.1 = 600.

Take logs: $4 \log_{10} U = \log_{10} 600 = 2.778$, so U = 4.95 (say 5 for convenience)

Since the capacitance at the input of the first stage of the driver is 0.1 pF, the successive driven capacitance will be inflated by U~ 5. Hence each device should have W/L 5 times greater than in the previous stage. If channel length is maintained at the minimum dimension, 0.5 μm, successive stages are designed with channel

widths inflated by 5. Hence those stages drive capacitances inflated by that same factor (assuming that gate and pad capacitances dominate e.g. interconnect), so the delay remains constant in each stage.

We also assume the stages are designed for symmetrical rising & falling delays, i.e. p-channel devices are a factor $\mu n/\mu p$ = 2x wider than the n-channel devices.



W/L values should therefore be:

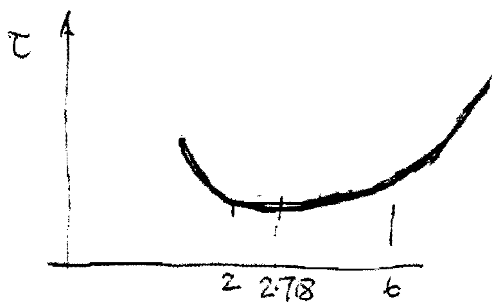
Stage	1	2	3	4
n device W	0.5	0.5	0.5	0.5
n device L	1	5	25	125
p device W	0.5	0.5	0.5	0.5
p device L	2	10	50	250

Note that the p devices are scaled by μ_n/μ_p to achieve equal delays for rising/falling signals.

The delay through each stage is the same, and is equal to:

$$\tau = \frac{3\times 0.5\times 10^{-12}}{2.5\times 10^{-4}\times(1/0.5)\times 3} \quad = \quad 1.0 \text{ ns}$$

Each subsequent stage has to drive 5× the capacitance of its predecessor, but its W/L is scaled ×5 so the delay is unchanged. Hence the total delay is 4 × 1.0 ns = 4 ns. [40%]



(ii) A graph of gate delay versus U has its minimum at U=2.718. The minimum possible delay will therefore be given with number of stages

N = ln 60/0.1 = 6.39. Clearly N must be an integer. N=6 is closest and gives a delay very close to the minimum, and also provides the required non-invert characteristic.

However, note that the area occupied is expected to be considerably greater than for the case of U=6. The total delay curve is fairly flat for values of U between about 2 and 6. A trade-off can be made of area vs. delay, and it can be seen that a big economy of area can be achieved with modest increase in delay

Note that the total of N=6 stages includes the minimum geometry gate generating the signal. Hence the driver itself consists of 5 further inverter stages. For N=6, the total delay τ_6 can be determined:

6 log10 U = log10 (60/0.1) = 2.778 which gives U ~ 2.9 (close to 2.718)

Hence, using proportion, $\tau_6 = \frac{2.9}{5.0}\times 1.0\times 6 = 3.6$ ns, only about 12% faster than the four stage design, but with an area penalty. [20%]

Examiner's note. Not a popular question, but well done by most attempting it.

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

$R = \frac{\rho}{t}\frac{\ell}{w}$ (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

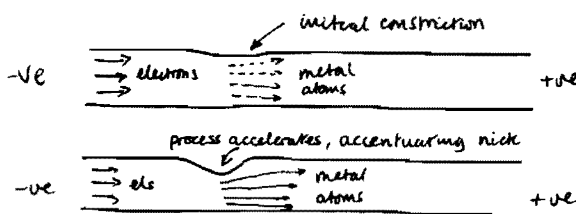
$$R = R_S \sum_i \frac{l_i}{w_i} \text{ may be used.}$$

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 the concave corner 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. [30%]

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

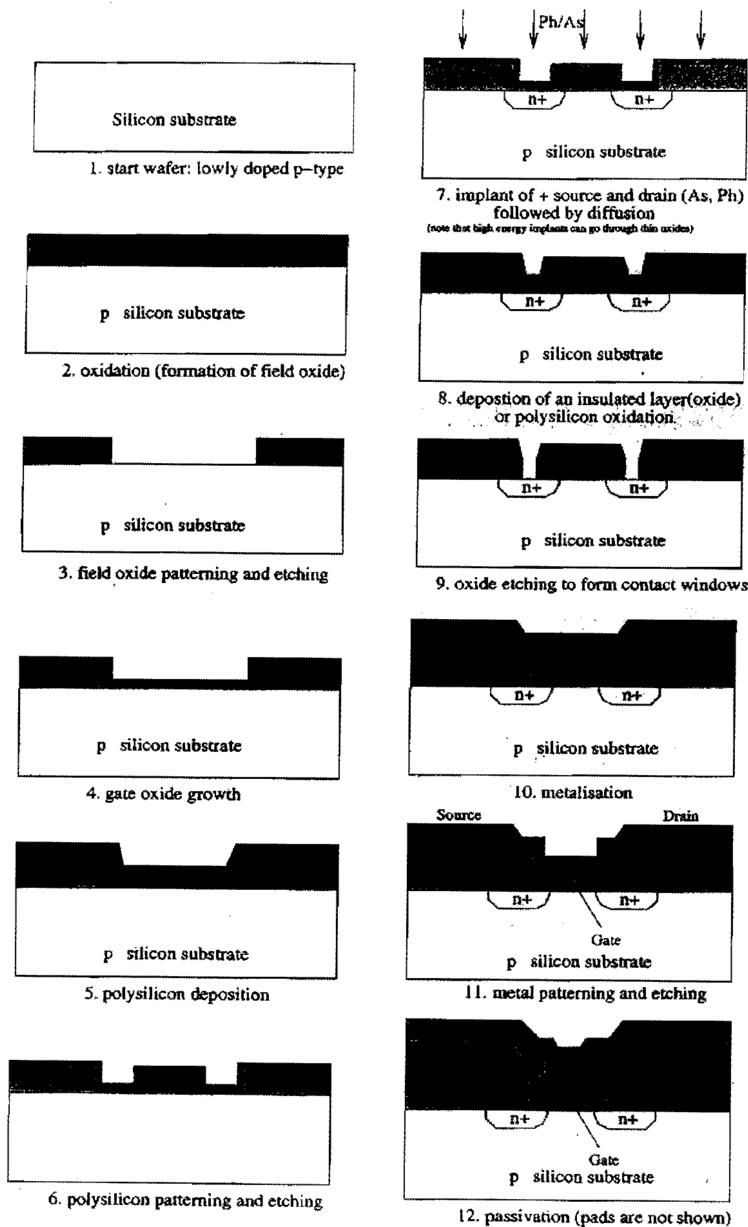
As current flows through a metal line the electrons constantly bombard the metal atoms, transferring momentum to them. Under severest conditions of high current density, the transfer of momentum is sufficient to push the metal ions aside and cause them to drift (i.e. migrate) towards the positive terminal, 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 locally, hence increasing the electron momentum; 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 by keeping current densities low, which demands that all power rails and other current-carrying interconnect be adequately broad in X-section.

Maximum J is typically 10^9 Am^{-2} for Al, translating to a typical 'rule' of thumb' of about 0.5mA per micrometre of width. Other factors may affect the rate of electromigration apart from J: grain size of metal, temperature, duty cycle (AC or DC current flow), metal type, and environmental conditions (e.g. presence of moisture). [20%]

(c) With a set of annotated diagrams, describe the fabrication schedule for making an NMOS transistor. See figure attached below.



[30%]

The annealing process after deep implantation requires very high time-temperature profiles for the wafer, so that its must be performed first, in order not to destroy later processes.
Processes from the bottom up.
Drain engineering before metal and other overlayers are put on.
Local interconnects before some passivation and global interconnects. [20%]

Examiner's note: A popular question, and the early parts were well done. In part (d) some candidates' answers lacked rigour