PART IIB 2010 DR D M HOLBURN

4B7 VLSI Design, Technology and CAD - 2010 Solutions

1. C_{SB}, C_{DB} are source and drain diffusion capacitances to substrate caused by formation of p-n junctions at drain-substrate and source-substrate interfaces
For each of these two components can be distinguished:
(i) an area-dependent component proportional to the plan-view area of the source/drain
(ii) a peripheral component, due to the side-walls of the source/drain, proportional to the perimeter of the 'diffusion'

C_{GS}, C_{GD} are gate-source & gate-drain capacitances due to proximity of these electrodes and to process-dependent overlaps

Cgate is a parallel-plate capacitance between gate and substrate. This depends on floor area, but is strongly dependent on gate potential and whether or not a channel has been formed.

For C_{GD}: in CMOS gates, which are intrinsically inverting structures, as the input swings, the output swings in the opposite direction and the large signal gain is effectively about –1. The opposing swing of V_G and V_D causes an increase in the apparent capacitance being driven at both gate and drain owing to Miller effect. To account for this the static value for C_{GD} is typically doubled.
Total gate capacitance is thus

$$Cg = Cgate + C_{GS} + 2\,C_{GD}$$

The polysilicon gate electrode may also serve as short-range interconnect, where it is not superimposed on the channel, the specific capacitance is much lower, and it is not much affected by potential.

Total drain capacitance or source capacitance is the sum of the area and peripheral components for each. Metal interconnect also contributes capacitance, and other inter-layer capacitances (e.g. between adjacent signal interconnects) may also be identified.

(b) **Numerical**. We consider only those capacitances that are driven with signals. Hence the V_{SS} line is not evaluated. Hence:

Cinput = Cpoly-substr + Cgate-substr + ($C_{GS} + 2 \times C_{GD}$)
Coutput = Cmetal-substr + C_D-substr + ($2 \times C_{DG}$)

The factors of 2 in the brackets arise from Miller effect. For Cmetal-substr and Cpoly-substr there is an *area* and a *peripheral* component.

Input: consider first the poly not over active, then the gates themselves::

$$Apoly = (60-4) \times 2 = 112 \times 10^{-12}\ m^2$$
$$Ppoly = (60-4) \times 2 + 4 \times 2 = 120 \times 10^{-6}\ m$$
$$Agate = (4 \times 2) = 8 \times 10^{-12}\ m^2$$
$$Pgate = 4 + 4 = 8 \times 10^{-6}\ m^2$$

For Pgate we consider only the part over the channel, since this is where the gate-drain and gate-source overlaps occur. Exactly half the length is associated with the drain, half with the source. The part of the gate perimeter at the edge of the channel is accounted for in Cpoly-substr.

$$Cpoly\text{-}substr = 112 \times 10^{-12} \times 4 \times 10^{-5} + 120 \times 10^{-6} \times 5 \times 10^{-11} = 5.68\ fF$$
$$Cgate\text{-}substr = 8 \times 10^{-12} \times 7 \times 10^{-4} = 5.6\ fF$$
$$C_{GS} = 4 \times 10^{-6} \times 3 \times 10^{-10} = 1.2\ fF$$
$$C_{GD} = 4 \times 10^{-6} \times 3 \times 10^{-10} = 1.2\ fF$$

Hence Cinput $= 5.68 + 5.6 + (1.2 + 2 \times 1.2) = 14.9$ fF

Output: consider first the metal interconnect, then the drain diffusion, and we assume the gate is centred on the active region. C_{DG} was dealt with. above.

Amet	$= 80 \times 4$	=	320×10^{-12} m^2
Pmet	$= 80 \times 2 + 2 \times 4$	=	168×10^{-6} m
A_D	$= 6 \times 4$	=	24×10^{-12} m^2
P_D	$= (6 + 4) \times 2$	=	20×10^{-6} m

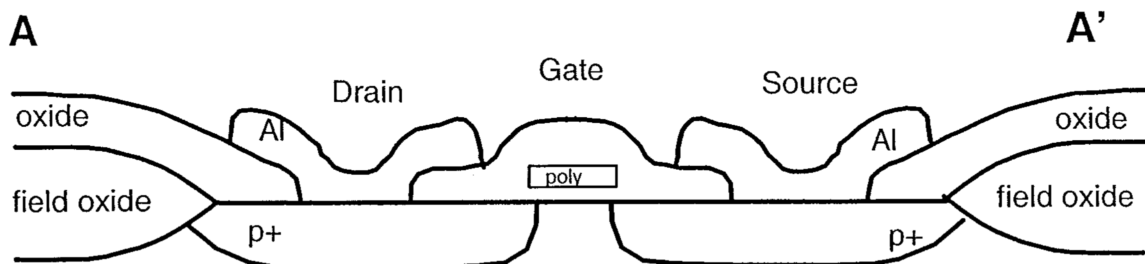
Hence Cmetal-sub $= 320 \times 10^{-12} \times 3 \times 10^{-5} + 168 \times 10^{-6} \times 4 \times 10^{-11} = 16.3$ fF

Cdrain-sub $= 24 \times 10^{-12} \times 10^{-4} + 20 \times 10^{-6} \times 4 \times 10^{-10} = 10.4$ fF

Hence Coutput $= 16.3 + 10.4 + (2 \times 1.2) = 29.1$ fF

(c) C_D-subst is expected to fall as V_D rises and the degree of reverse bias increases. Metal and poly-substr capacitances are substantially constant. Cgate varies as per the discussion above.

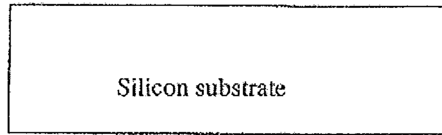
(d) Cross Section



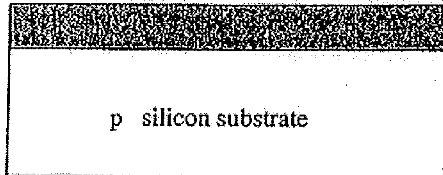
Examiner's comment:

A moderately popular and relatively straightforward question, attempted by over half the candidates. The most common error was an inability to distinguish which contributions to parasitic capacitance were linked to input, output or other circuit nodes.

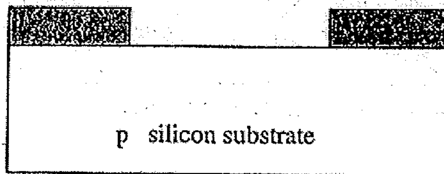
2. (a) With a set of annotated diagrams, describe the fabrication schedule for making an NMOS transistor. See figure attached below.
(b) The dual well and the specific interconnects in the semiconductor (see image)
(c) Comment on the refinements to CMOS technology represented by (i) silicon-on-insulator technology: removing parasitics from proximate silicon, better isolation between devices,
(ii) multiple layer interconnects: single layer interconnects are strictly limited in what the resulting circuit can achieve. Modem circuits have six or more layers of metal separated by dielectrics and graded in size from the very small cross-section and short wires at the bottom next to the wafer, up to large and longer interconnections between different parts of blocks. (iii) hafnium oxide dielectrics: when Si02 becomes to thin, because of scaling, pinholes threaten the integrity - by using a layer of higher dielectric, the charge control in the channel can still be achieved with a thicker layer of the dielectric. (iv) Bi-CMOS: for some applications, one needs the current levels from the output of an HBT, as in driving a laser. If we need CMOS to process the signal for the output, we need a hybrid of CMOS and bipolar to achieve this.



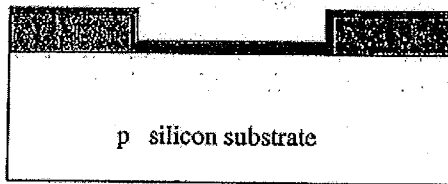
1. start wafer: lowly doped p–type



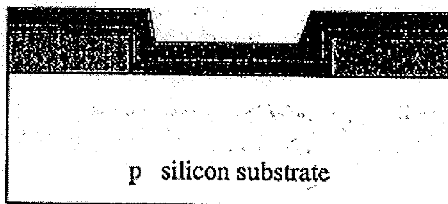
2. oxidation (formation of field oxide)



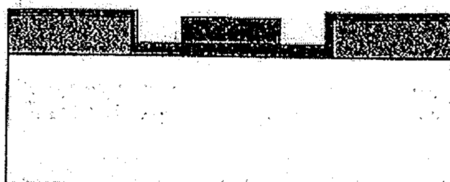
3. field oxide patterning and etching

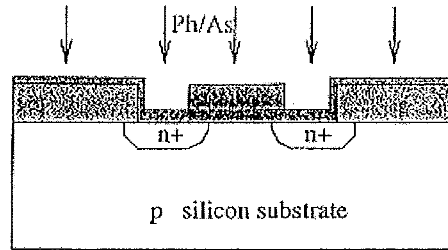


4. gate oxide growth

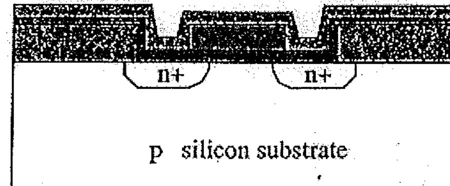


5. polysilicon deposition

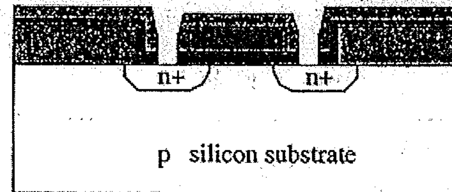

6. polysilicon patterning and etching



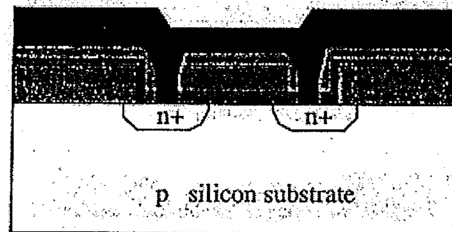
7. implant of + source and drain (As, Ph) followed by diffusion
(note that high energy implants can go through thin oxides)



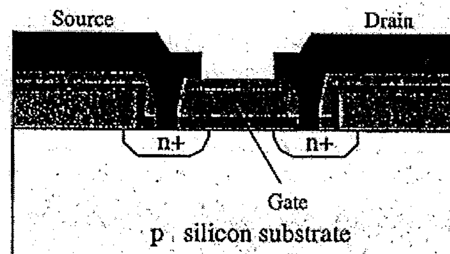
8. depostion of an insulated layer(oxide) or polysilicon oxidation



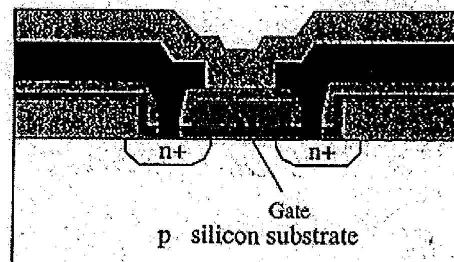
9. oxide etching to form contact windows



10. metalisation



11. metal patterning and etching



12. passivation (pads are not shown)

Examiner's comment:

A popular question where the first 60% of the marks closely followed the contents of one lecture. The last 40% was on refinements to CMOS and the answers were a bit scrappy.

3. (a) Most electronic products are designed to be in service for 10 years. In this time, there is a requirement that failures in service are kept to a minimum. This is achieved by maintaining a quality control process over the manufacture so that errors that might lead to issues of unreliability (as later in the question) are eliminated at the design stage as far as possible, and not allowed to creep in during manufacture.

(b) Annotated standard picture showing early failure rate dropping to a low level of random failure during service life, with an increase in wear-out failure at the end of design life. In the case of VLSI: ESD, latch-up in early failure, as well as any contamination, as well as moisture penetration in the early days after manufacture. During life, dielectric breakdown, contamination and moisture penetration can occur leading to random failure. Hot carrier degradation, electromigration and chemical/mechanical failures can occur at the end of life as wear out or failure.

(c) Describe the failure factors, failure mechanisms, signatures of failure and the implications for a refined original design associated with each of (i) the passivation of integrated circuits, how the overlayer protects the IC underneath: caused by pinholes, cracks, thickness variations, contamination or any engendering for surface inversion: seen by decreased breakdown voltage, short circuits, increase leakage current drift of transistor parameters and noise deterioration.

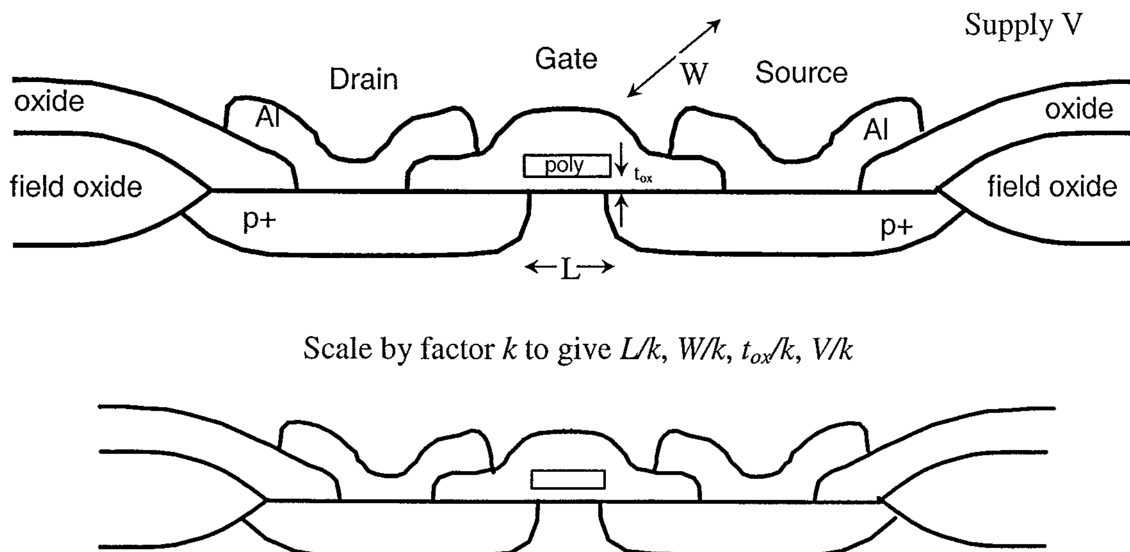
(ii) electromigration in the metal interconnects (or indeed at contact holes or via holes): caused by the divergence of high fields producing a net force on the atoms in the wire, reach a stage where the atoms would move out of the high field regions. This would open up holes or cracks at grain boundaries, and usually leading to a catastrophic breakdown,. see as an open circuit, closed circuit or if arrested, an increased resistance. (iii) hot carrier degradation within or adjacent to the transistors: lucky hot electrons that do not suffer the optic phonon emission, can gain enough energy to move into the oxide layer, getting trapped in defects there and altering the local electrostatic potential, seen as decreased breakdown voltage, short circuits, increased leakage current and drift in threshold voltages or current gain. (iv) electrostatic discharge: caused by a pick up of static electronics, or by a current or voltage surge elsewhere in the system, or an over current, leading to diffusion junction breakdown, oxide film damage, or destruction of metallization, observed as an open or closed circuit, or increased leakage current.

Examiner's comment:

This questions was attempted by half the cohort. The first part on reliability of ICs had answers with a very wide spread of marks. The second part was well answered and the third on failure mechanism was also answered with a wide range of knowledge and understanding

4. In *constant field* scaling, geometric dimensions are modified (typically reduced) and electrode voltages are also scaled in order to maintain electric fields constant. Historically, device dimensions were scaled from about 6 microns to 1 micron without change in Vdd (*constant voltage scaling*). This offered better delay reduction as well as cost reduction; it maintained continuity in supply and logic level standards. However, it has proved impracticable to push dimensions to sub-micron with Vdd unchanged, since the increasing electric fields impact the operation of the MOSFET, requiring other major process alterations (e.g. doping densities, etc). There would also otherwise be greater risk of breakdown. Some adjustments to other process parameters may also be required.

Assume all dimensions & voltages are reduced by a scaling factor k ~ 1, as in the diagram. The MOS transistor has critical dimensions L and W (channel length and width respectively) and gate oxide thickness t_{ox}. The applied voltages are represented by V. We shall consider scaling down all dimensions and voltages by a scaling factor k.



Scale by factor k to give L/k, W/k, t_{ox}/k, V/k

Gate area α LW decreased by k^2

Field at channel α V/t_{ox} unchanged

Since carrier velocity is μE and distance travelled α L

Transit time τ thru channel α L^2/V decreased by k

Hence gate delay is reduced by k

Capacitance at gate etc α LW/t_{ox} decreased by k

Current I consumed I α CV/τ decreased by k

DC Power consumption α IV decreased by k^2

Power density/area α IV/WL constant

For interconnect, scaled in length l, width w, and thickness d, and dielectric thickness h, the key parameters are changed as follows:

Resistance	α	l/dw	increased by k
Capacitance to substr	α	lw/h	decreased by k
Current density J	α	I/dw	increased by k

Note that this assumes device currents are scaled down by k as above.

Time constant RC	α	RC	unchanged

Hence the speed of propagation of signals along interconnect is unchanged.

The industry typically scales process generations with $k \sim \sqrt{2}$, which is roughly the ratio described in the question. The reduction in *Vdd* is consistent with constant-field scaling. This doubles the number of transistors per unit area with each generation and doubles transistor performance every two generations under constant field scaling. Process shrinks of $k \sim 1.05$ are commonly applied as a process becomes mature to boost the speed of components in that process

Major benefits – scaled devices allow:

- higher packing density
- greater speed of operation
- lower current consumption

Size The dependence of area on k^2 gives a clear advantage in terms of packing density, cost, etc and the potential to pack more functionality into an equivalent space. In digital design typically the smallest possible devices should be used to minimise parasitics and provide best speed-power product. This desire has driven the 'push' to smaller geometries in the microprocessor and memory industry.
Speed The reduction in C leads to a valuable increase in intrinsic device speed. However, this is not maintained for interconnect unless the chip size shrinks. There is evidence of such advantage being gained when process shrinks are applied to existing products, e.g. Pentium processor, but in many instances product development incorporates many more scaled devices in a larger chip, so that global interconnect does not follow scaling rules. The delay along such interconnect in terms of (faster) clock ticks is significantly greater.
Power Dissipation Scaling reduces power density for scaled devices by about k, but this may not apply to many key elements like pad drivers, which may be responsible for much of the current draw.

Problem areas

- Charge stored in transistor gate reduced by k^2, hence scaled devices (e.g. memories are more liable to soft errors
- Roff/Ron decreases as dimensions decrease and the role of sub-threshold conduction becomes more important. Hence static power consumption will become a more serious issue.
- Faster clock speeds - these have risen far faster than classical scaling would predict – with Vdd still somewhat higher (viewed historically) than constant field scaling would demand, have led to skyrocketing power dissipation.
- Dynamic power dissipation cannot continue to increase unchecked because it will be uneconomic to cool the chips.

- Increasing clock speeds allied with trend towards larger devices leads to longer interconnect delays as a function of τclock. Clock skews – differential timing changes – may rise as k^3. Manufacturers may attempt to circumvent this by use of Cu interconnect, low-permittivity dielectrics, and multiple interconnect layers scaled less aggressively
- Contact resistances rise as contact structures are scaled
- Some structures e.g. I/O pads, power amplifiers, do not scale
- Reduction in yield at smaller geometries
- Digital cells are typically well characterised in a new process before linear designs (amplifiers, oscillators, mixers) can be adequately verified and reliable models developed. This may delay the introduction of a new process for mixed signal applications.
- Increased fab. costs and lower yields at the beginning of process lifetime mean that for an evolving design that is sensitive to market price, the point at which transition should be made to a smaller process must be carefully judged.

Examiner's comment:

A popular question, and generally quite well done, as the descriptive part is widely documented and was based on the content of a lecture. Discussion of implications and applicability to deep sub-micron processes was less competently done in a number of cases.

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

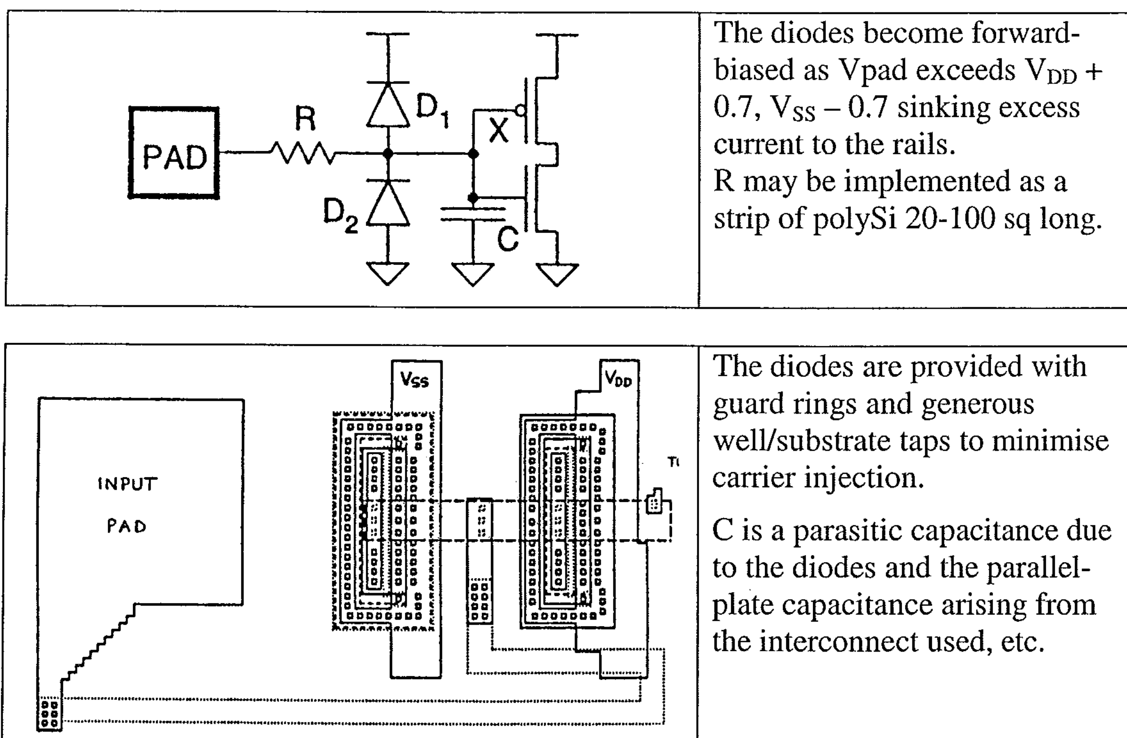
- over and under-voltages
- consequential latchup conditions
- electrostatic discharge

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. 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) Discuss quality, cost, delivery and service as critical success factors for the successful manufacture of integrated circuits. Quality: meeting the customer's requirements in the form of some agreed specification, leading to precision on each of the hundreds of steps made in the manufacture. Cost: lower-price but equal quality win out usually, once the requirement is met, while the manufacture can still make an adequate profit for investment for the future. Delivery: Delays in delivery costs money and opportunity to the customer, and faster delivery can demand premium prices or higher market share. Service: Soft side: how to identify best product, how to troubleshoot, how to adjust to customers evolving requirements, all as an edge when quality/cost/delivery are all of a high standard among competitors.

(c) Describe some of the processes required before a manufacturing line can be approved for the production of quality-assured VLSI products. (i) qualification of all the materials used (ii) layout in of all the equipment to aid throughput (iii) agreed common format for allowed specifications (iv) utilities requirements and control, and for maintaining system stability (v) equipment checks and qualification - cleanliness, integrity. functional operation (vi) characterisation and ' qualification of all processes (vii) full characterisation of products - parameter checking, wafer level reliability and uniformity and reproducibility, by probe testing (viii) agreed characterisation data to go with product release.

Examiner's comment:

An unpopular question, and some evidence that it was tackled as the last attempt with little time available. In the first part on input pads, a satisfactory though brief account that displayed only limited knowledge of the issues. A similarly weak attempt in the latter parts exploring general knowledge of IC manufacture.