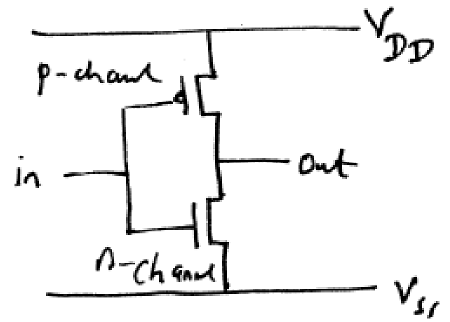
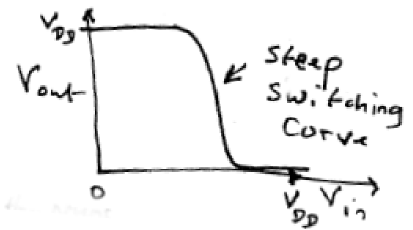
1. (a)

Advantages



For digital circuits, complementary MOS technology has the decisive advantage over NMOS that either the p-channel or the n-channel transistor is in the off state at any time (except during the switching transition) and the current drawn is small, i.e. **low power dissipation**. Hence, CMOS technology has low power consumption at moderate operating frequencies and intermediate geometries. This allows dense circuits to be fabricated without exceeding thermal limitations during operation.



In addition the technology has a wide margin of device operation, and hence CMOS VLSI is a high-yield and successful technology.

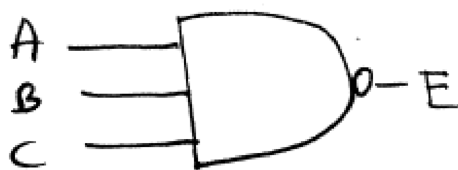
Disadvantages:

One disadvantage is the relatively low hole mobility of p-devices, requiring them to be designed with wider channels to achieve symmetry of operation and high performance. This leads to increased capacitance. GaAs devices are faster but the materials technology is complex and not yet suited to very large scale integration.

However, CMOS fabrication technology is more complex than nMOS, but the performance and economics advantages now heavily outweigh these issues. CMOS is well established and dominant.

[30%]

(b)

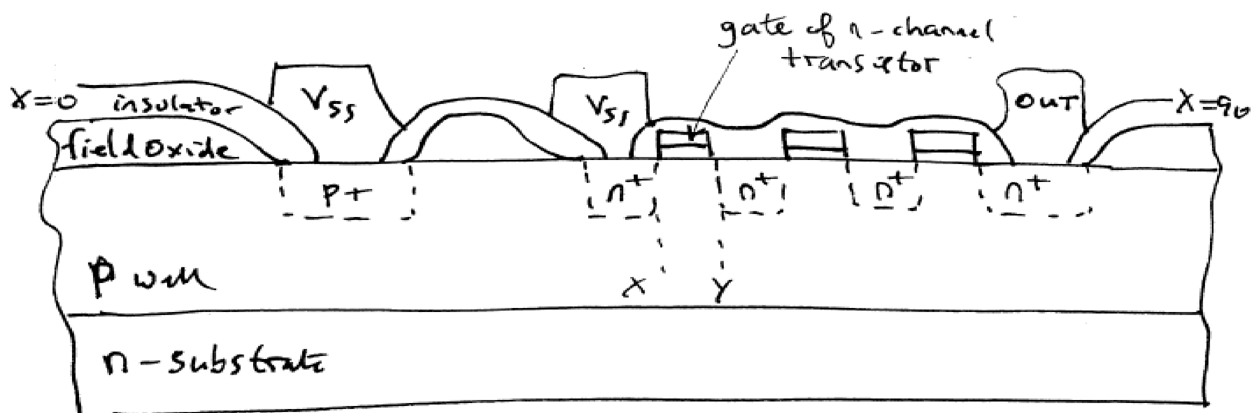


Three-input NAND E = $\overline{A.B.C}$

Upper right	p-type for PMOS
Lower right	n-type for NMOS
Upper left	n-type for ohmic to substrate
Lower left	p-type for ohmic to p-well

[10%]

All three n-channel gates (and hence all inputs) must be high to connect *out* to Vss.



[20%]

(b) (iii) In self-aligned technology, the polysilicon gate is in place prior to performing the implantation of the n+ source and drain regions. This has the result that the undoped channel is solely the region underneath the gate, and it is *self-aligned*. The lithographic step for the definition of the polysilicon gate therefore determines the distance *xy* which governs the switching speed, which is dependent on the carrier transit time. [10%]

(c) (i) The electrical width is determined by the active area. In Fig. 1 the channels are formed in the rectangular regions where active area and polysilicon coincide. The channel width is the dimension perpendicular to the flow of current from D to S; in all six devices it is the longer dimension of the rectangle. [10%]

(ii) In this case, from direct measurement,

$$\frac{width\ of\ p\text{-}channel}{width\ of\ n\text{-}channel} = \frac{2}{3}$$

[10%]

(iii) Worst-case rise time is with 1 p-channel device alone conducting
Worst-case fall time is via the series connection of 3 n-channel devices.

But,

$$\frac{electron\ mobility}{hole\ mobility} = 2$$

Hence,

$$\frac{rise\ time}{fall\ time} = \frac{3}{2} \times \frac{1}{3} \times 2 \quad = \quad 1$$

i.e. they are equal. [10%]

Assessor's Comments

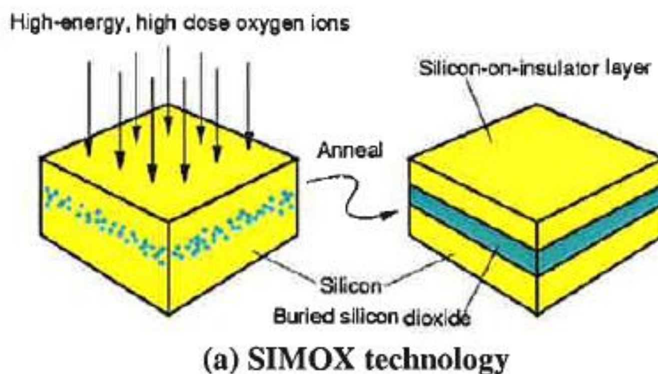
This was a popular question, answered by a large proportion of the candidates. Most solutions demonstrated a good understanding of the principles of CMOS layout, but a number confused the role of n and p type implants. The numerical parts were in general done well. Several had difficulty constructing a cross-section of the device represented by the layout. Most knew the basics of self-aligned structures, but not all understood which lithography steps were the most critical.

2. (a) (i) Scaling of transistors to smaller dimensions generally comes with undesired effects associated with device performance. Provide five adverse effects of scaling. [20%]

Lack of saturation, gate oxide degradation due to hot electrons, threshold voltage shifts, gate-induced drain leakage, and drain-induced barrier lowering, which refers to the lowering of the potential barrier at the source-channel side with increasing drain voltage.

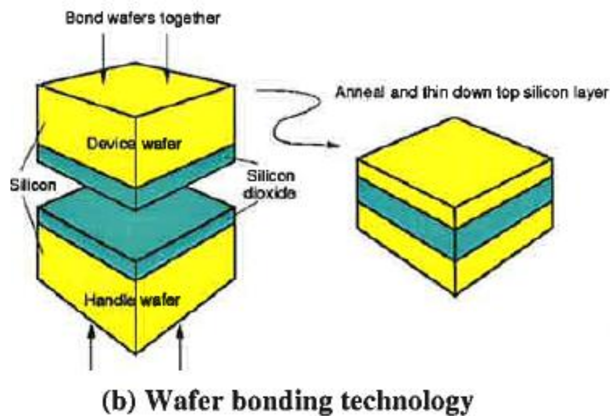
(b) (i) In VLSI circuits, there is a strong quest for running circuit at higher clock speeds. Besides the obvious route of scaling feature sizes to attain higher operating speeds, the silicon on insulator or SOI technology can provide significant speed improvements at reduced power consumption. Describe, with the aid of illustrations, three SOI process technologies that are commonly used in the IC industry. [40%]

The SOI technology allows fabrication of circuits in a thin layer of silicon that is electrically isolated from the semiconductor substrate through use of a buried insulating layer such as SiO2. This eliminates undesirable effects such as high leakage current, parasitic bipolar action leading to latch-up, and crosstalk between devices on the same chip. With SOI, the parasitic capacitance is greatly reduced thus pushing operating speed to higher frequency limits at reduced power consumption. Several processes have been developed for the SOI technology – the most common, as illustrated below, are (a) SIMOX (separation by ion implantation of oxygen), (b) Wafer Bonding (of two silicon wafers one with a top oxide), and (c) UNIBOND (using implantation of He and wafer bonding).



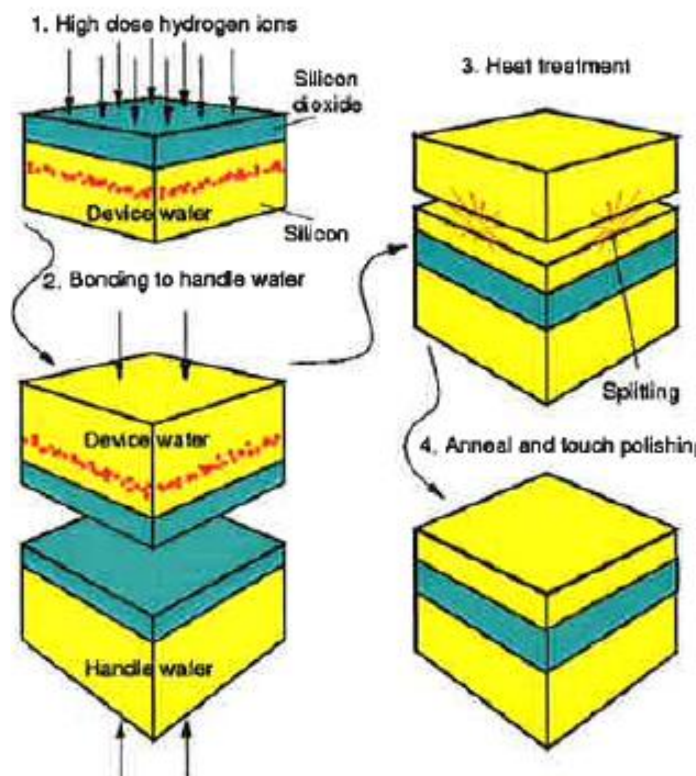
(a) SIMOX technology

To realise a SIMOX SOI wafer, a Si substrate is implanted with a high dose of oxygen ions at high energy giving rise to silicon and oxide layers that are 200nm and 400nm thick, respectively. Newer processes can yield respective 50 nm and 100nm thick layers.



(b) Wafer bonding technology

When flat, polished wafers are brought together the bond due to van der Waals forces. The bond is then strengthened by post-bond annealing. The top Si wafer is then polished to crate a thin SOI layer.



© Unibond-SmartCut SOI technology

This (smart cut) method yields highly uniform layers. The device wafer which has a layer of SiO2 on it, is implanted with a high dose of hydrogen ions and then bonded to a handle wafer. When subject to a heat treatment of 600C, the wafers are divided along the line of implanted hydrogen. This leaves behind a thin and uniform SOI layer.

(ii) Describe SiO_2 based gate dielectrics, the issues related to scaling the gate dielectric, and the requirements new dielectrics must fulfill for compatibility with Si technology. [40%]

Today's mainstream processes use an oxide thickness of 1.5-2 nm. These ultrathin dielectric layers are grown in a standard thermal diffusion furnace using pure oxygen or mixture of oxidizing gases between 900˚C and 1000˚C. Before oxidation, the silicon surface must be cleaned carefully with high purity chemicals. The SiO2 grows controllably in a layer-by-layer mode.

Thinning down the oxide raises severe technological problems: dielectric thickness variation, penetration of impurities (particularly boron) from the highly doped poly-Si gate, reliability and lifetime problems, and the high gate current. The leakage current at a gate bias of 1V can change from 1pA/cm2 at 3.5 nm to 10 A/cm2 at 1.5 nm, which is over 13 orders of magnitude. The practical limit (<1A/cm2) is when

the gate current becomes equal to the off-state source to drain sub-threshold leakage current. Beyond this limit, direct tunneling of electrons from p-Si through the thin oxide into the n+-poly-Si becomes the dominant leakage mechanism – tunneling of holes is much smaller.

New dielectrics must fulfill a number of requirements for compatibility with Si technology: dielectric properties, thermodynamic stability, electronic properties, microstructural stability, deposition tools and chemistry, and process compatibility. Also because of the superior low interface state density at the Si/SiO_2 interface, the first monolayer of the gate dielectric needs to be SiO2 even if high-k oxides are used. But since the capacitances of the SiO_2 and high-k layer are in series, there is a limit to the influence of the high-k capacitance on the total capacitance.

The permittivity of the new dielectric should be considerably larger than that of SiO_2 (ε_r~3.9). Amorphous oxides always show relatively low permittivities while epitaxially grown oxides show really high permittivities.

Thermodynamic stability implies a gate dielectric stack that is chemically unreactive since the dielectric is in direct contact with Si and gate contact material (poly-Si or metal) and is subject to severe temperature treatments (up to 1000˚C). The most likely reactions are the growth of SiO_2, formation of metal oxides, and formation of silicides.

In selecting the appropriate high-k material, the band gap and the (conduction and valence) band offset energies must be taken into account – see Fig. 3.15. The offset must be at least 1 eV for V_G<1 V. This means that the band gap of the material must be at least 3.1 eV since Eg-Si is 1.1 eV and the offsets are each 1 eV. In spite of the large band gap energies of many oxides, the band alignment with Si is often highly asymmetric. Compounds such as $BaZrO_3$, $BaTiO_3$, and Ta_2O_5 have a large band-gap but lack sufficient barrier for the conduction band offset.

In terms of microstructure, there are three possibilities: (perfectly) epitaxial, poly-crystalline, or amorphous. The structure should be stable throughout the full processing. Meeting high temperature stability is a key requirement. The preferred structure is either perfectly epitaxial or amorphous. Poly-crystalline layers give rise to detrimental diffusion because of grain boundaries, enhanced leakage, and high interface state density. Epitaxial layers must be lattice matched with Si.

Dielectric deposition must meet uniformity requirements (non-uniformity < 2% on 200 and 300 nm wafers) and processing should be fast (< 5 mins). Most viable deposition processes are metal organic chemical vapor deposition (MOCVD) and atomic layer deposition (ALD), and typical deposition temperature is 500˚C. ALD is a special CVD application whereby the precursors are not mixed prior to introduction into the reactor. Instead alternating pulses of reactants are introduced. In a first pulse, a (sub) monolayer of the first reactant is absorbed on the wafer surface and the reactor is then purged, and a second reactant is introduce to react with the adsorbed layer.

For most high-k materials, patterning is of prime concern. The (dry) etch chemistry should be Cl/Br-based rather than F-based (because fluorides are non-volatile). Good patterning processes are particularly vital for System on Chip where two or

even three gate oxide processes are employed on a single chip to combine high speed logic (EOT 1 nm), low power logic (EOT 2 nm) and input/output logic (EOT >3.5 nm) applications.

Assessor's Comments

This question on fabrication technology was answered by relatively few candidates. It called for a good understanding of the (less-often explored) disadvantages and challenges of scaling in MOS technology. The answers produced spanned a considerable range, but in general candidates demonstrated good recall of three key SoI processes and of gate-oxide technology.

3. (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 a 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.

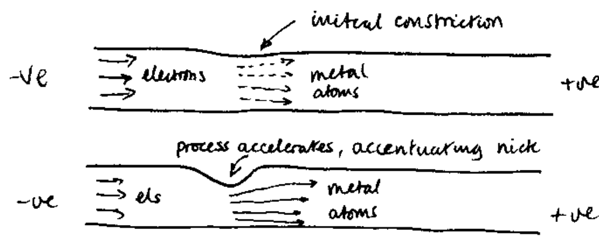
Where corners appear the pattern of equipotentials in the conductor is distorted. A finite element analysis shows that the measured resistance is very sensitive to the curvature at concave corners, which may not be well defined for many cases.

- 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. [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). [30%]

(c) Input pad structures are primarily required to protect MOSFET inputs from:

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

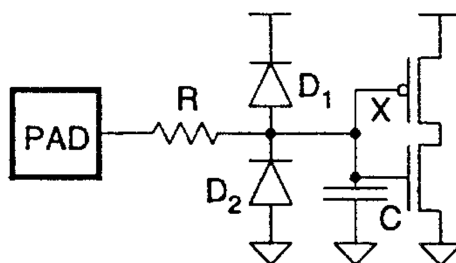
In addition they may contain inverting circuitry, or Schmitt trigger circuitry if the input signals to be fed to the circuit are not known to be proper CMOS level signals.

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.

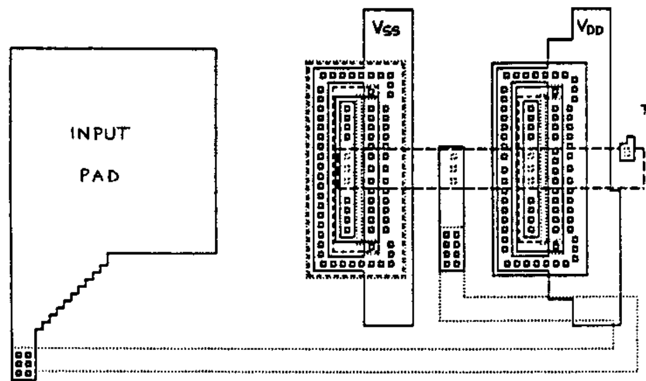
Gate oxide thicknesses in modern processes are o(20 nm) thick with breakdown voltages of 5 V or so. Input resistances may exceed 10^{12} ohms. Since the gate electrode typically has capacitance of a few fF, only a very small packet of charge is required to generate voltages far in excess of $V_{breakdown}$.

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). [40%]

Assessor's Comments

This was a reasonably popular question, attempted by about half the candidates. Nearly all candidates could explain the concept of sheet resistance, but a number gave weak explanations of its limitations with irregular interconnect forms. Electromigration was described satisfactorily by nearly all. There was wide variation in candidates' ability to described the main features of input pads.

4. The account of design rules should include the following major points:
- Design rules allow a ready translation from circuit concepts to actual geometry in silicon.
- They are the effective interface between the circuit/system designer and the fabrication/process engineer. They provide a reliable and workable compromise which is friendly to both sides.
- The designer is concerned to achieve:
  - best possible electrical performance - speed. noise margin, etc
  - minimum area of Si per circuit – lower cost, better yield, reliability
- The process engineer seeks:
  - to maximise tolerances on all parts – easier fabrication, better yield
- There are 3 basic tolerances that set limits to the shapes the designer can use
  - dimensional resolution governed e.g. by λ of light used in photolith, photoresist characteristics
  - alignment errors – registration, temperature changes, bowing/distortion
  - reproducibility of processing – wet etching, layer thickness control

- For practical purposes all 3 effects can be reduced to linear dimensions on a plan view of the mask layout. The permissible dimensions are often highly specific to a manufacturer's process.
- The simplest rules originate from the need for continuity and avoidance of unintended short-circuits. Layers such as polysilicon, metal and diffusion are associated with minimum dimensions and minimum separations. They may also be associated with ohmic resistance (electrical origin). Violation of these rules may lead (as in PCB technology) to open-circuits in conducting traces, or short-circuits, where tracks are too close.
- With metal Al interconnect it is necessary to ensure that the current density does not exceed about 10^9 Am^{-2}, otherwise there is risk of electromigration induced by transfer of momentum from the electronic carriers to metal atoms and causes progressive thinning of interconnect at circuit bottlenecks – e.g. as metal crosses a step. Interconnect width is hence governed by the anticipated peak (rather than mean) current, and not simply by lithographic considerations.
- Since fabrication involves several sequentially masked steps, there is a need to accommodate the possibility of mis-registration between successive masks. For this reason,
    - implant masks overlap the diffusions to which they correspond by a significant margin
    - polySi gates extend beyond the edge of the underlying diffusion
    - metal, diffusion and polySi are required to surround contact cuts by a significant margin
- It is possible to define an 'alignment tree' which summarises the statistical probability of mis-registration between related mask layers.
- The use of metal (Al/Cu) rather than polySi is dictated for
    - power distribution
    - signal transmission over significant distances – for example, clock lines, to avoid skew.
- Where significant currents are transmitted from one metal layer to transistors, or to another metal layer, the contact structure must be capable of carrying the current. Since contact conductance is proportional to cut perimeter (not area), this is achieved through use of many minimum geometry cuts filling the available space.
- Other rules that may be mentioned: contacts and vias of fixed size, antenna rules, well/substrate tap spacings.

[50%]

Numerical Part

The units comprises 120,000 memory cells each driving 20 fF capacitance. Each is clocked at 60 MHz. In the worst case, a pattern of 0-1-0-1 etc on successive clocks will generate maximum dynamic current in these cells.

Whenever a cell output switches 0-1 or 1-0 a packet of energy $1/2\ CVdd^2$ is transferred between the supply rails for that stage. We assume that the dynamic dissipation arising from this dominates other effects. Note that if all inputs remain at 1 (or 0), every stage in the 15,000 bit register is presumed to stay in the corresponding state and no dynamic dissipation would be observed. This assumes no resistive or other losses of charge occur requiring that charge be replenished at each output (e.g. refresh).

In the worst case, each stage of the unit alternates its output state at each successive clock edges, giving rise to maximum dynamic dissipation. This will occur when each unit receives at its input a 0101010 waveform synchronised with the clock, and at half the clock frequency.

Each stage thus dissipates energy at a rate:

$\frac{1}{2} CVdd^2 \times fc$ where fc is the clock frequency

Hence the total power dissipation is thus

$W = \frac{1}{2} \times 8 \times 15000 \times 20 \times 10^{-15} \times 3.3^2 \times 60 \times 10^6 = 0.784$ W

Hence the average current consumption

$I = 0.784/3.3$ A = 238 mA [30%]

Let the interconnect width be W. Then the current density J is:

$J = 238 \times 10^{-3} / (W \times 0.4 \times 10^{-6})$ A m^{-2}

This must be significantly less than the electromigration limit of ~10^{10} Am^{-2}. A factor of 10 is usually considered adequate. To satisfy this we have:

$J < 10^9$, giving:

$$J = \frac{238 \times 10^{-3}}{W \times 0.4 \times 10^{-6}} < 10^9 \quad \text{which gives } W > \frac{238 \times 10^{-3}}{0.4 \times 10^{-6} \times 10^9} = 594\ \mu\text{m}$$

The total capacitance being driven is approximately 120 × 20 pF or 2.4nF. To this needs to be added the capacitance being driven at the 8 output pins, which may each drive a capacitance of order 50pF, making a total of a further 400 pF. Hence the total true worst-case current may be significantly greater than that calculated. Typically, the pads have their own suitably dimensioned power ring, but nonetheless in a conservative design verification of the total current will be necessary.

Note also that the peak current may be many times I, with current transients synchronised to clock edges. [20%]

Assessor's comments

A popular question, attempted by more than half the candidates. Candidates varied in their ability to cite good examples of design rules, but there were some very comprehensive accounts. The numerical part was well done overall, but a number of candidates were unable to interpret the results, or to enumerate the assumptions on which their solutions depended.

5. (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 organo-Si glass or SoI to minimise capacitance to substrate and hence the delay. [30%]

Numerical Part

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

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

Total resistance 10 mm $= \frac{10 \times 10^3}{1} \times 60 = 6 \times 10^5 \Omega$

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

Hence for 10 mm trace, $T = \frac{1}{2} \times 6 \times 10^5 \times 2 \times 10^{-12} = \underline{600\text{ ns}}$ [15%]

(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 $T \propto RC$. [15%]

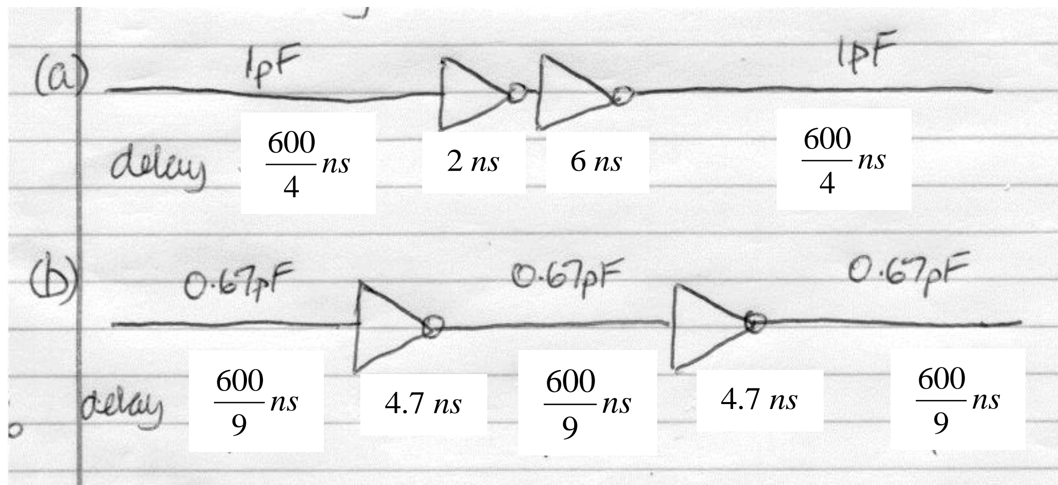
(ii) Using the salicide process reduces the resistance by a factor 60/10 = 6, but leaves the capacitance effectively unchanged since there is no change in geometry or permittivity. Hence *T* is reduced by a factor 6, to 100 ns. [10%]

(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 2 ns. The second drives 1pF so its delay is 6 ns. Each 5 mm segment has 150 ns delay. The total is 150 + 2 + 6 + 150 = 308 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 $(2 + 0.67 \times 4) = 4.7$ ns. Each 3.3 mm segment will have delay 600/9 ns or 67.7 ns.

Hence the total delay will be: 67.7 + 4.7 + 67.7 + 4.7 + 67.7 ns = 209.5 ns.

This comfortably meets the requirement stated, i.e. a factor of two reduction, and is the simplest arrangement to do so. [30%]

Assessor's Comments

A very popular question, attempted by every candidate, though there were indications that some candidates attempted this as their final question with insufficient time. Most were able to come up with a definition of clock skew, but there was wide variation in candidates' ability to describe design approaches to avoid this. The use of buffers was well understood by almost all candidates.

Answers

1. (c) (ii) Ratio of widths p:n = 2:3; (iii) ratio of worst case rise time : fall time = 1

2.

3.

4. (b) (i) 238 mA; (ii) 594 μm

5. (b) 600 ns; (c) 209.5 ns