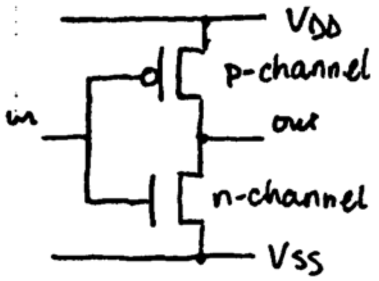
2014 – Part IIB Module 4B7 – VLSI Design, Technology and CAD Dr D M Holburn

1. (a) CMOS inverter

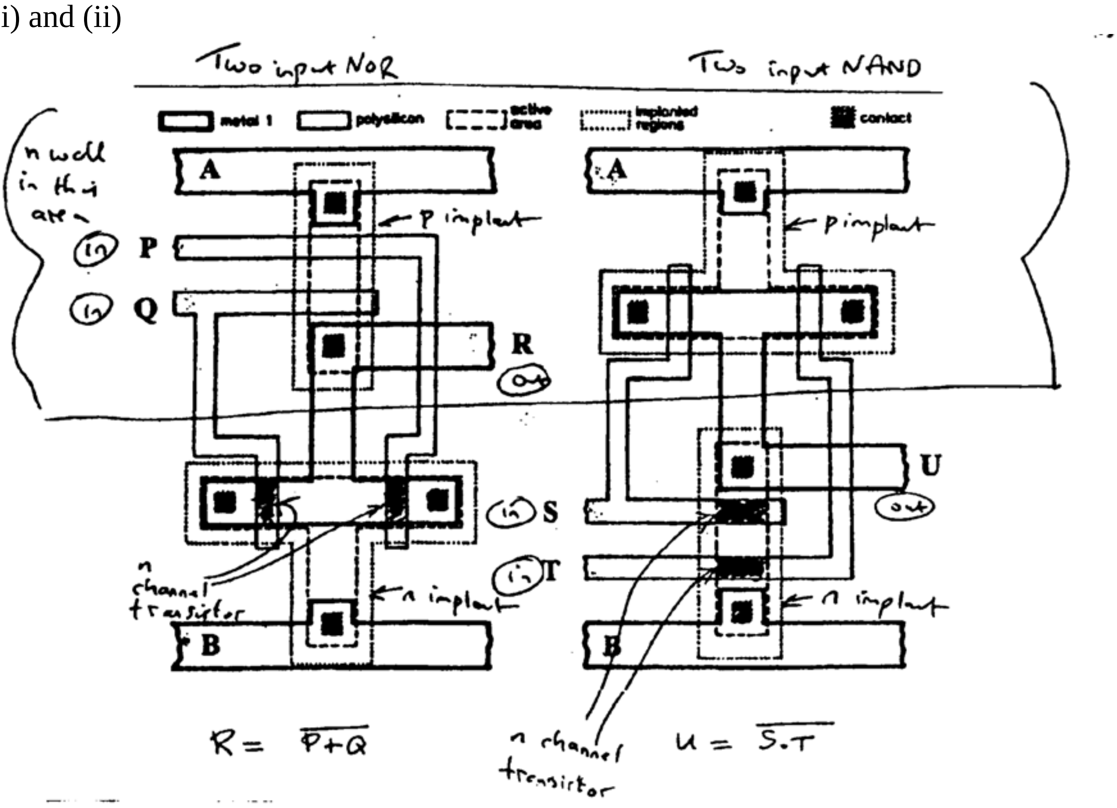


An important advantage of CMOS technology is the low power consumption at moderate operating frequencies and intermediate geometries, since very little current is drawn except during a switching transition. This allows dense circuits to be fabricated without exceeding thermal limitations during operation.

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.

Silicon continues to dominate because of low costs and an exceedingly well-developed technology. [20%]

(i) and (ii)



[10%]

(iii) From inspection of Fig. 1, [10%]

2NOR	*Wn* = 9 mm	*Wp* = 9 mm	r_{NP} = 1
2NAND	*Wn* = 9 mm	*Wp* = 9 mm	r_{NP} = 1

[10%]

c) Since μ_N/μ_P is quoted as 2, the worst-case fall time for the 2NOR device is via one standard n-channel device. The worst-case rise time is determined by 2 series connected p-channel devices, and hence is much slower, about a factor 4 so.
Similar rise and fall times could be obtained by choosing r_{NP} = ¼ for the 2NOR.

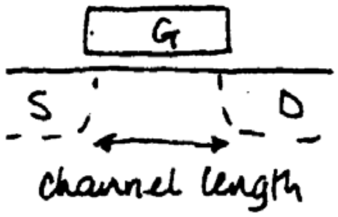
..

For the 2NAND device, the worst-case for rising outputs is via a single p-channel device ; the worst-case for falling output involves two series-connected n-channel devices. For similar delay rising/falling, we require $r_{NP} = 1$ for the 2NAND – as drawn.

In fact, for a general purpose logic system we are interested in minimising the overall delay, so one approach might be to consider how to minimise the sum of rising and falling delays for a number (say, a pair) of gates in cascade. A more detailed analysis shows that in these circumstances for the 2NOR, r_{NP} must be somewhat greater than ¼ to allow for capacitive effects. [20%]

(d) The most important dimension determining switching speed is the channel length (source-drain separation). Reducing this gives faster carrier transit time and improved device performance, limited ultimately by short-channel and punch-through effects in sub-100nm devices.

Cross section



Presently manufactured devices have effective electrical channel lengths of about 40 nm. Research devices have been demonstrated with about 10 nm channel lengths and satisfactory electrical characteristics

If lithographic and manufacturing technologies continue to improve as expected over the next decade, it seems entirely possible that of-order 10 nm channel-length devices will be in large scale manufacture by then. However, there are many other challenges for the industry, including the 'push' to larger wafers (450 mm now being targeted) to reduce unit costs. [30%]

Assessors' Note: *This was a popular question, well answered by a large proportion of the candidates. Most candidates were able to recognise the gates and identify the key regions. Virtually all candidates indicated correctly where the implants, n-well and active regions were to be found. The biggest variation in marks arose from the level of detail provided in sections (a) and (d).*

2. (a) Advantages:

- excellent electrical isolation between devices or blocks of devices (less leakage, no latch-up)
- less area consumed (no buried layers)
- (alternatively one can mention the high junction temperature)

Disadvantages:

- self-heating (the buried oxide layer acts as a thermal barrier)
- expensive (the SOI wafer can be 5-10 times more expensive than a standard bulk silicon wafer). [20%]

(b)

Wafer bonding

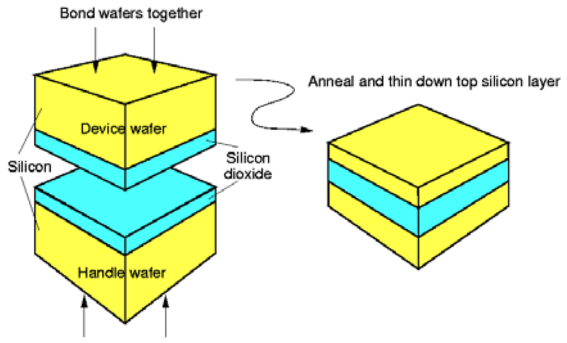
The basic technique relies on the fact that polished and flat wafers, when brought into contact at room temperature, are attracted to each other by van der Waals forces and "bond". To strengthen the bond between the two wafers, a post-bonding anneal at high temperature is usually performed, and the top silicon wafer is then polished to create a thin silicon-on-insulator layer suitable for device manufacturing

Unibond - SmartCut

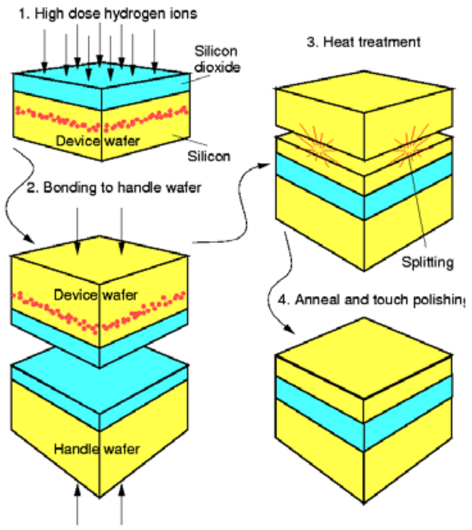
The device wafer, which has a layer of silicon dioxide on top of it, is implanted with a high dose of hydrogen ions (between 3.5 x10^{16} and 1x10^{17} cm^{-2}), after which it is bonded to the handle wafer. A heat treatment at 600 deg C divides the wafers along the line of the implanted hydrogen, leaving behind a thin and uniform silicon-on-insulator layer on the handle wafer. Only a final high-temperature anneal and touch polish are required to the handle wafer to yield the finished wafer.

Wafer bonding is much cheaper than SmartCut. SmartCut, however, is much better in defining accurately the thickness of the SOI layer.

Wafer bonding



Unibond-SmartCut SOI technology



[30%]

..

(c) (i) Time dependent dielectric breakdown (TTDB)

Good quality thermal oxide films have dielectric breakdown strength of 10 MV/cm or more.

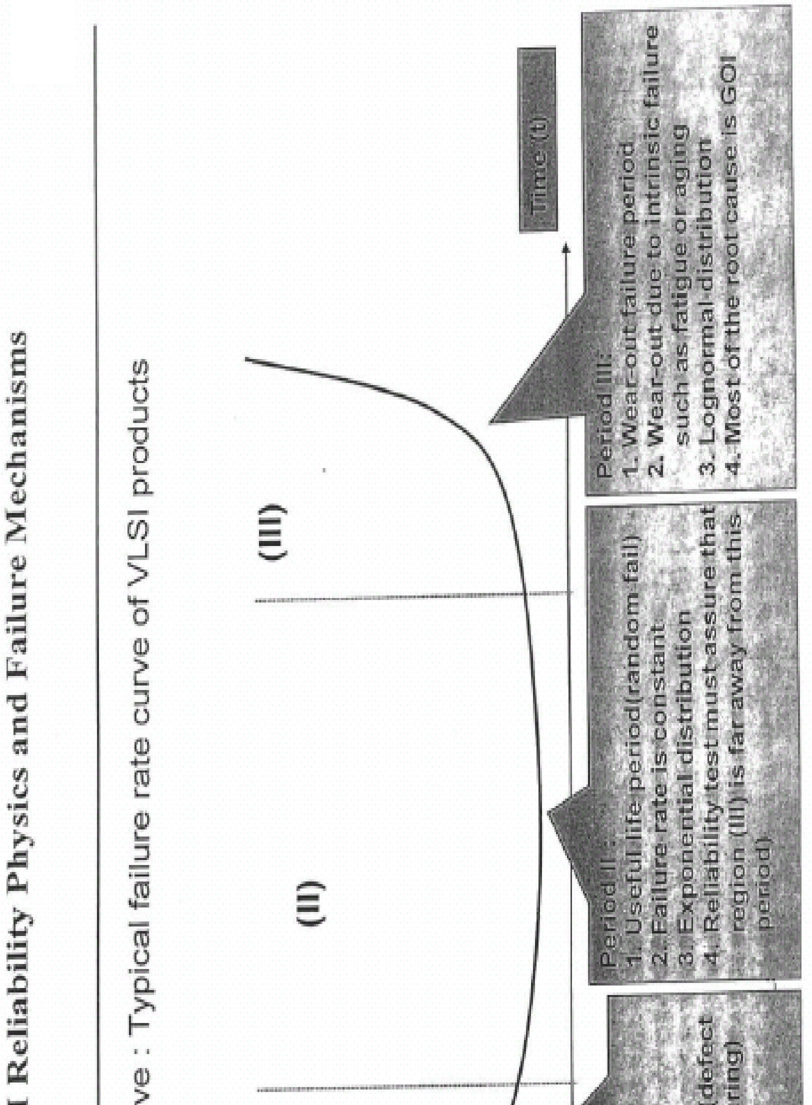
However, oxide film failure over time even in lower electric-field intensity (conditions of practical use) is a major cause of failure. Gate oxides are generally affected by this. When an electric field applied to an oxide film causes the injection of holes into the oxide film to occur and it consequently causes traps to be made in the oxide film. As the number of traps increases, an electric current via the traps is due to hopping or tunnelling. If the number of traps continues to increase and the traps connect between the high voltage and low voltage terminal, the connection carries a high current that causes the gate oxide film to break down.

(ii) Hot carrier Injection

It is generated in MOSFET by the large channel electric fields near the drain region.

Mechanism: Carriers (electrons or holes) that flow into the high electric field area are accelerated by the strong field and gain substantial energy. Some of the carriers have enough energy (that is to say they are hot) to overcome the electric potential barrier existing between the Si substrate and gate oxide film. These hot carriers are injected and subsequently trapped into the gate oxide film. They form a space charge or inversion layer and over a period of time they can affect the threshold voltage (V_T), transconductance (g_m) or breakdown voltage. [20%]

2. (cont)



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

[30%]

Assessors' Note: *This question was answered by over half of the candidates, and the range of marks awarded was wide. Most who attempted it were able to identify two advantages and disadvantages for SoI. However, the level of detail given in the description of process steps was quite variable. A good number of solutions showed that those candidates were aware of the specific failure mechanisms associated with dielectrics, but there was more variation in answers in terms of placing these in the general context of reliability, suggesting that not all candidates had taken the step of reading beyond the lecture notes.*

..

3. Based on lecture notes. A good response should include the following points.

(a) Design rules allow a ready translation from schematic circuit to actual geometry in silicon. They are the effective interface between the circuit/system designer and the fab/process engineer. They provide a workable and reliable compromise which is friendly to both sides.

The designer is concerned to achieve:

- best possible electrical performance – speed, noise margins, linearity, gain;
- minimum area of Si per circuit – lower costs, better yield, reliability.

The process engineer on the other hand seeks:

- to maximise tolerances on all ‘parts’ - easier fabrication, better yield.

There are 3 basic tolerances that set limits to the shapes that the designer can specify:

- dimensional resolution governed by λ of light used in lithography, photoresist characteristics, … ;
- alignment errors – registration, temperature variation, bowing/distortion;
- reproducibility of processing – wet etching, plasma, layer thickness control.

For practical purposes all three 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 of layers and for avoidance of unintended short-circuits. Layers such as polySi, 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 PWB technology – to open-circuits in conducting traces, or short-circuits, where tracks are too close.

Since fab 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 active areas/diffusions to which they correspond, by a significant specified amount;
- polySi gates extend beyond the edge of the underlying diffusion;
- metal, diffusion and polySi are required to surround contact cuts and vias by a significant margin.

For **power pads**, a substantial rectangle of superposed aluminium interconnect layers is required, robust enough and dimensions sufficient to allow a fine gold wire to be ultrasonically bonded to it during packaging. Bond pads have to be around 80-100 microns square, irrespective of the process geometry. Interconnect linked to them has to be dimensioned according to contact/via considerations, and bearing in mind electromigration, and to mimimise series resistance and inductance.

Minimum geometry transistors may be undesirable where high voltage immunity is required – e.g. output or input pads.

..

A number of precautions are required to alleviate the risk of latch-up. As far as design rules are concerned, this calls for use of **well and substrate taps** at a specified maximum pitch and frequency, and in close proximity to power devices (for example, I/O pads). [60%]

(b) The inception of **self-aligned processes** has greatly relaxed the requirements for registration of implant masks. In one example the polySi gate acts itself as a mask for implantation, guaranteeing correct positioning of those implants relative to the gate.

It is possible to define an ‘alignment tree’ which summarises the statistical probability of mis-registration between related mask layers.

The use of metal rather than polySi is dictated for:

- power distribution;
- signal transmission over long distances – e.g. clock lines, to avoid skew,

owing to its lower resistivity/sheet resistance, and (to a lesser degree) lower C.

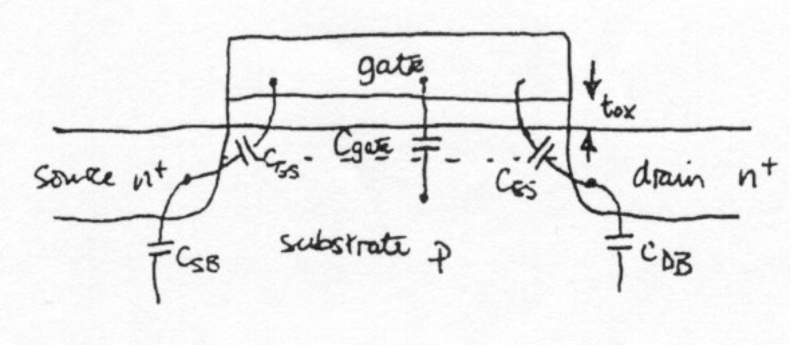
Where significant currents are transmitted from one metal layer to another, or to transistor diffusion, the **contact structure** must be capable of carrying the current. For lithographical reasons, a minimum contact cut lateral dimension is mandated; but since contact conductance is proportional to the cut perimeter (not area) this is achieved through use of many minimum-geometry cuts, filling the available space. A maximum current is often associated with a min-geom cut or via.

Plasma processing involves application of high energy RF fields that can induce high voltages in previously fabricated circuit interconnections, which act as ‘antennas’. These so-called **process-induced gate-oxide damage** instances give rise to the so-called ‘antenna rules’. The fields can be enough to cause breakdown in gate oxide and other fragile structures. The risk is greatest when extensive metal interconnect is coupled to a transistor gate (e.g. long clock line), but is reduced when p-n junctions (e.g. source/drain diffusions) are also connected, since these assist conducting excess energy to ground. Careful design strategies are needed to ensure the area exposed is not too great, taking into account the fragility of the coupled structures.

With metal Al interconnect it is necessary to ensure that the current density does not exceed about 10^9Am^{-2} otherwise there is risk of **electromigration**, which is induced by transfer of momentum from the electronics carriers to metal atoms, and causes progressive thinning of interconnect at current 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. [40%]

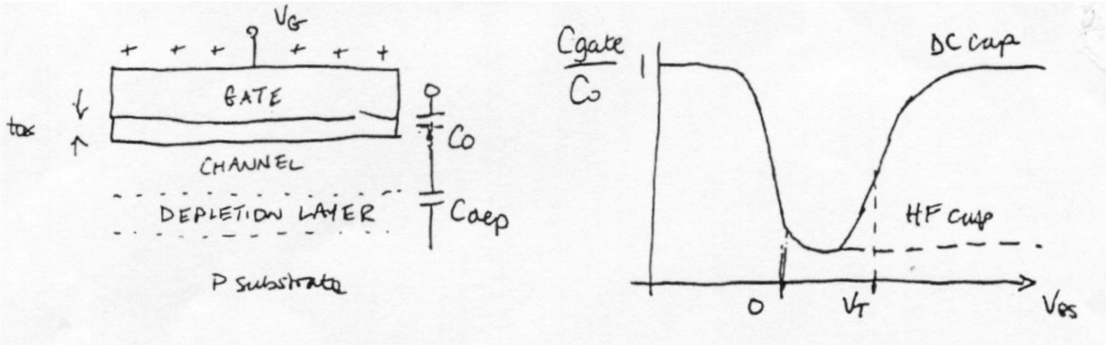
Assessors’ Note : *This question was attempted by over half of the candidates, but was on average answered slightly less well on the whole, possibly because the answer required was entirely descriptive in format. Candidates’ recall of the range of constraints addressed by design rules varied, despite the provision of a number of cases as prompts. There was evidence this was done by some as a ‘last question’.*

4



n-channel enhancement mode MOSFET

(i) 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:
(a) an area-dependent component proportional to the plan-view area of the source/drain
(b) a peripheral component due to the side-walls of the source/drain and proportional to the perimeter of the ‘diffusion’
As device areas shrink, being $\propto L^2$, the peripheral component is of increasing significance ($\propto$ L). Both depend on junction bias, doping profile etc.
(ii) C_{GS}, C_{GD} are gate-source & gate-drain capacitances due to proximity of these electrodes and to process-dependent overlaps
(iii) 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.



When V_G is less than about 0, an accumulation layer is formed as the gate attracts holes. The structure behaves like a capacitor of dielectric thickness t_{ox} and of capacitance Co where Co $\propto (\varepsilon/t_{ox}) \times$ gate area.

As V_G is raised above 0, holes are repelled leaving a region depleted of carriers. The depth of the depletion region, d, increases as V_G increases. This gives rise to a capacitance C_{dep} in series with Co, causing a reduction in measured gate capacitance.

As V_G approaches V_T (threshold voltage) surface inversion gives a relatively high conductivity layer (the channel) which restores the capacitance progressively back towards Co. Only at very high frequencies, or in the absence of a nearby source/drain (which provide carriers), will the channel be unable to form sufficiently rapidly, and a lower capacitance (dashed line) will be observed.

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 the 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.

The 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 overlaid and adjacent signal interconnects) may also be identified. [50%]

Numerical part. We consider only those capacitances that are driven with output signals. Hence the V_{DD}/V_{SS} lines are not evaluated. The two transistor channels have the same dimensions. Hence:

$$Cout = Cmetal\text{-}sub + 2 \times C_D\text{-}sub + 2 \times (2 \times C_{DG}) \qquad (i)$$

The factor of 2 in the brackets arises from the Miller effect. Another factor of 2 comes from the two identical MOSFETs which have drains (output) and gates (input – not under consideration here) connected together. For Cmetal-sub and for C_D-sub there is an *area* and a *peripheral* component.

Output: consider first the metal interconnect. We assume the area over the active region contact should be discounted, for C_D-sub applies here. Then, for each drain diffusion, we assume the gate is centred on the active region, and drain and source (S not involved) each have area half the total. For C_{DG} we consider the length of overlap of gate and drain at the edge of the channel. We look at a single device, noting that the two actually present have identical dimensions; the equation for Cout based on (i) takes into account both devices.

$$Amet = (40 - 2 - 2) \times 2 \times 10^{-12} = 72 \times 10^{-12}\ m^2$$

$$Pmet = ((40 - 2 - 2) + 40 + 2 + 2) \times 10^{-6} = 80 \times 10^{-6}\ m$$

$$A_D = (10 - 1)/2 \times 2 \times 10^{-12} = 9 \times 10^{-12}\ m^2$$

$$P_D = ((10 - 1)/2 + 2) \times 2 \times 10^{-6} = 13 \times 10^{-6}\ m$$

Hence $Cmetal\text{-}sub = 72 \times 10^{-12} \times 3 \times 10^{-5} + 80 \times 10^{-6} \times 4 \times 10^{-11} = 5.36\ fF$

For a single device,

$$C_D\text{-}sub = 9 \times 10^{-12} \times 10^{-4} + 13 \times 10^{-6} \times 4 \times 10^{-10} = 6.1\ fF$$

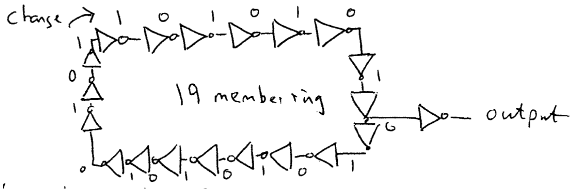
$$C_{DG} = C_{SG} = 2 \times 10^{-6} \times 3 \times 10^{-10} = 0.6\ fF$$

And from (i), $Cout = 5.36 + 2 \times 6.1 + 2 \times (2 \times 0.6) = 19.96\ fF$ [40%]

C_D-sub is expected to fall as V_D rises and the degree of reverse bias increases. Metal and poly-sub capacitances are substantially constant. [10%]

Assessors' note: *A moderately popular and relatively straightforward question, attempted by over half the candidates. There was considerable variability in the descriptive section. In the quantitative section, the most common error was an inability to distinguish which contributions to parasitic capacitance were relevant to the capacitance at the output node. Some contributions were overlooked and others were counted but should not have been. A small number of candidates did not through the quantitative section at the end.*

5 (a) 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 around 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 (for example) a frequency counter.

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. [50%]

(b) Discuss quality, cost, delivery and service as critical 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 a competitive edge when quality/cost/delivery are all of a high standard among competitors. [25%]

(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. [25%]

Assessors' Note: *A less popular question addressing two themes, one related to module coursework, and the other a descriptive review of IC manufacture, which called for some diversity of reading. Those who attempted this had a basic understanding of the ring oscillator's mode of operation, but were unable to explain its value in process evaluation. The descriptive sections needed greater detail.*