Dr D M Holburn

4B7 VLSI Design, Technology and CAD - 2016 Solutions

1. (a) What is yield and why is it so important in Integrated Circuit (IC) manufacturing? [20%]

YIELD = % of dies on a wafer that meet performance specifications.
Yield decreases rapidly with increasing die area - plays major role in economics of IC design, to the extent that estimates about the die area and yield are always considered when making decisions about whether IC development for specific needs will be economically justifiable. Average values of key parameters at the lot level are monitored, and lots are rejected if they are off specs. Statistical circuit parameter variations large enough to cause failure are more common in analogue than in digital circuits.

(b) Briefly describe soft and hard faults in analogue and digital ICs. [20%]

Soft Faults = Failures due to parameter variations or parameter drifts which affect yield in a statistical sense.
Hard Faults = Complete transistor failure or interconnection errors.
Soft faults are accepted in digital ICs, but not hard faults. In analogue circuits, both faults are rejected.

(c) What is meant by the *capability index* (C_P)? State the probability that a fabricated device parameter lies inside the specified design process window? [20%]

Capability Index C_P defines the relationship between the specified process window (which dictates design spec range), actual wafer-level variations and the corresponding yield, i.e. C_P = (Design Spec Width)/(Process Width). The process width is generally defined as $\pm 3\sigma$ about the mean. If the process window is centred around the design spec window, then the probability that a fabricated device parameter lies inside the design spec window is

$$P = \frac{1}{\sqrt{2\pi}} \int_{-3Cp}^{3Cp} \exp\left(-x^2/2\right) dx$$

(d) If 1000 devices on a chip must have a specific parameter within the specified design process window, determine the soft yield if the process has been characterised by a capability index of (i) $C_P = 0.5$ (ii) $C_P = 1.0$ (iii) $C_P = 1.5$ (iv) $C_P = 2.0$ [30%]

Using the expression, $P = \frac{1}{\sqrt{2\pi}} \int_{-3Cp}^{3Cp} \exp\left(-x^2/2\right) dx$

(i) $C_P = 0.5 \rightarrow P \sim 0.8664$. The probability that all 1000 devices have a parameter within the design spec window is $P_{1000} = (0.8664)^{1000} \rightarrow 0$.
(ii) $C_P = 1.0 \rightarrow P \sim 0.9973$, yield = P^{1000} = 6.7%.
(iii) $C_P = 1.5 \rightarrow P \sim 0.999993$, yield = P^{1000} = 99.3%.
(iv) $C_P = 2.0 \rightarrow P \sim 0.999999998$, and yield = 100%.

(e) Comment on the physical implications of your result in part (c) above. [10%]

If C_P is small, the probability that a device parameter is within the process window is small, and the yield is poor. The yield gets worse if the mean of a parameter for a wafer is not centred within device spec window. Setting the design specification width so that C_P =2 will result in reasonable yield, although this may place unrealistic performance demands on the designer.

Models used to predict yield: Probability that a given die is good,

Seed Model $P = e^{-\sqrt{AD}}$

Murphy Model $P = \left(\frac{1-e^{-AD}}{AD}\right)^2$

A = die area and D = average defect density.

$$P = \frac{1}{\sqrt{2\pi}}\int_{-3Cp}^{3Cp} \exp\left(-x^2/2\right)dx$$

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

Assessor's Comments

This was a popular question, well answered by a large proportion of the candidates. Most candidates were able to give a basic explanation of yield but the level of detail given varied. The quantitative part of the question was on the whole done well.

2. (a) In sequential systems operation may need to be synchronised to a master clock. In a large design, different parts of the circuit may receive the clock signal via different routes, so that the timing of the critical edges may differ. Clock skew may arise from the following sources

- Differential delay along different lengths or styles of interconnect
- Passage through different numbers of controlling gates/buffers with different delays
- Need for additional inverters to generate ϕ_{bar} from ϕ

The effects are as for a mistimed discrete circuit. Clock pulses may arrive too late to latch data before it decays to an unknown state.
Several approaches may be used to minimise clock skew, including:

- Use of pipelineing
- Use of buffers to split clock lines into shorter segments
- Complete avoidance of polySi for clock lines

Use of an SoI process can also reduce delay through reducing parasitic capacitance. [10%]

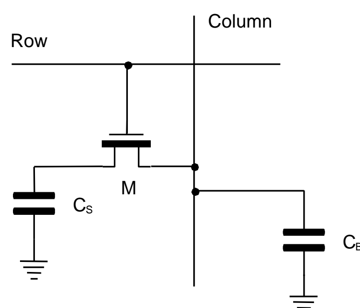
(b) A DRAM cell may be achieved using an n-channel MOS transistor in association with parallel capacitor elements. This makes for an extremely small cell and leads to high memory densities. C_S is a parasitic element typically O(20fF). Much effort has gone towards fabricating capacitors with the highest possible C and minimum area, e.g. trench capacitor.

Reading and *writing* are accomplished by applying logic high to the gate of M via the row/address line in order to select the cell. The cell must periodically be refreshed (O(10ms)) because of charge leakage from C_S.

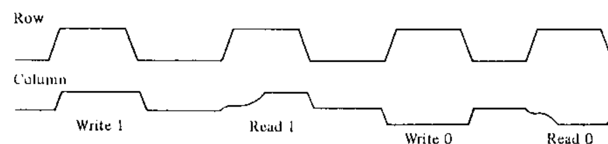
Data can be *written* into the cell by forcing logic 0 or logic 1 on the column/bit-line while the cell is selected. C_S charges to this value, which is retained when the cell is deselected.

When *reading*, the cell is selected by applying logic high to the row line, making it conduct. The column line is connected to a sensitive comparator.

The potential change seen on the column line may be 1 mV or less. Design of suitable sensing comparator in a noisy environment is a great challenge. Normally a regenerative amplifier is used and the column line is precharged to the mean of the logic levels.



DRAM Cell

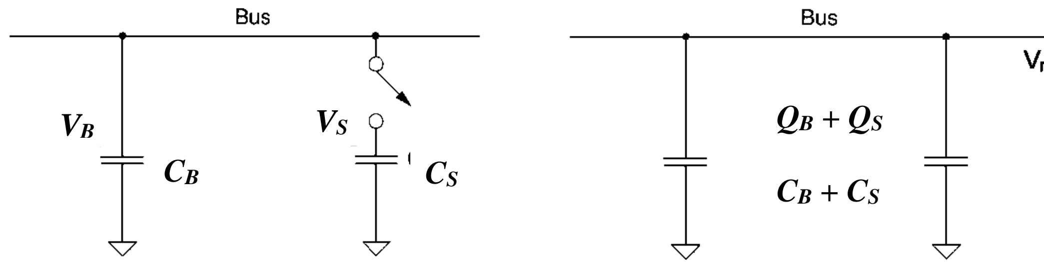


Representative timing diagram

[30%]

(c) A dynamic memory can be modelled as a storage capacitor C_S, in series with a switch S. To store a value, the signal **1** or **0** is brought in from the Bus, capacitance C_B, through closed switch S, which then charges towards the signal potential, i.e. it

'samples' the value on the Bus. S then opens to isolate the storage element. The effectiveness of this depends on the transfer of charge from C_B to C_S.



When reading out the stored value, the switch is again closed. The value stored on C_S is now coupled to the Bus, which transfers the signal read out to where it is needed. The magnitude of charge stored on C_S in comparison with that of C_B determines the potential change that is observed.

The magnitudes of the signals observed in each of these cases depend on how charge is *shared* between the two capacitors, i.e., how it redistributes itself between the two capacitors.

We ask what is the potential on the storage capacitor C_S immediately after the switch closes. This represents the situation that applies when data is being written to the memory cell.

Charge on each capacitor: $Q_B = C_B V_B$ and $Q_S = C_S V_S$

Total charge Q_t: $Q_t = Q_B + Q_S = C_B V_B + C_S V_S$

Total capacitance C_t: $C_t = C_B + C_S$

Immediately after S closes, the resultant voltage V_r is Q_t/C_t or: $V_r = \frac{C_B V_B + C_S V_S}{C_B + C_S}$

If $V_B = V_{DD}$ and V_S is close to 0 i.e. $V_B >> V_S$

This represents the situation where C_S currently stores a logic **0**, and **1** is to be written to it. (The converse situation is easily dealt with in a similar way.)

Then: $V_r = \frac{C_B V_{DD}}{C_B + C_S}$

In other words, in order that the potential V_B be transferred reliably to V_S the following must be satisfied:

$C_B >> C_S$ (say, >10 times as great)

Fortunately, for the case of a DRAM, the value of C_S is invariably much smaller than C_B and the condition is easily met.

However, for the converse case, where data is to be read out of the memory, the opposite requirement holds, that C_S should be much greater than C_B, and both conditions cannot be satisfied simultaneously. The consequence of this is that the readout process generates only a small incremental voltage change proportional to $C_S/(C_S + C_B)$, which may be only a few tens of microvolts. This places great demands on the sense amplifier that converts the signal into a logic level. As RAM sizes increase, C_S is likely to grow. [30%]

(c) Assume the following stage is designed to switch at $V_{sw} = V_{DD}$, and that the capacitor C is charged to $V_{DD} = 5$ V as stated. Hence, if C loses half its charge, having been set to logic 1, it will (incorrectly) indicate logic 0.
Leakage from the capacitor is at a fixed rate of 0.1 nA, assumed independent of potential.

Hence the time take to discharge C to 2.5V is:

$$\tau = \frac{C \times (5-2.5)}{I_{leak}} = \frac{60 \times 10^{-15} \times 2.5}{0.1 \times 10^{-9}} = 1.5 \times 10^{-3} = 1.5 \text{ ms}$$

The cell must be refreshed more often than this.
If the bus line is at 2.5 V and the memory capacitor is charged to 5V, then using charge sharing:

$$V_{sense} = \frac{(0.06 \times 5 + 1.5 \times 2.5) \times 10^{-12}}{(1.5 + 0.06) \times 10^{-12}} = \frac{0.3 + 3.75}{1.56} = 2.596 \text{ V}$$

Hence the change in potential observed is 96 mV [30%]

Assessor's Comments

This was answered by over half of the candidates, and quite a wide range of marks was awarded. Most were able to explain clock skew satisfactorily, but not all were able to develop a generic condition for safe operation of a memory based on charge storage. The quantitative part was on the whole done well.

3. 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. [30%]

(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 = $(60 - 2) \times 1$ = 58×10^{-12} m^2
Ppoly = $(60 - 2) \times 2 + 4 \times 1$ = 120×10^{-6} m
Agate = (2×1) = 2×10^{-12} m^2
Pgate = $3 + 3$ = 6×10^{-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-substr = $58 \times 10^{-12} \times 4 \times 10^{-5} + 120 \times 10^{-6} \times 5 \times 10^{-11}$ = 8.32 fF
Cgate-substr = $2 \times 10^{-12} \times 5 \times 10^{-4}$ = 1.0 fF
C_{GS} = $2 \times 10^{-6} \times 3 \times 10^{-10}$ = 0.6 fF
C_{GD} = $2 \times 10^{-6} \times 3 \times 10^{-10}$ = 0.6 fF

Hence Cinput = $8.32 + 1.0 + (0.6 + 2 \times 0.6)$ = 10.1 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.

$A_{met} = 80 \times 2 \times 10^{-12} = 160 \times 10^{-12}\ m^2$

$P_{met} = (80 \times 2 + 2 \times 2) \times 10^{-6} = 164 \times 10^{-6}\ m$

$A_D = 4.5 \times 2 \times 10^{-12} = 9 \times 10^{-12}\ m^2$

$P_D = (4.5 + 2) \times 2 \times 10^{-6} = 13 \times 10^{-6}\ m$

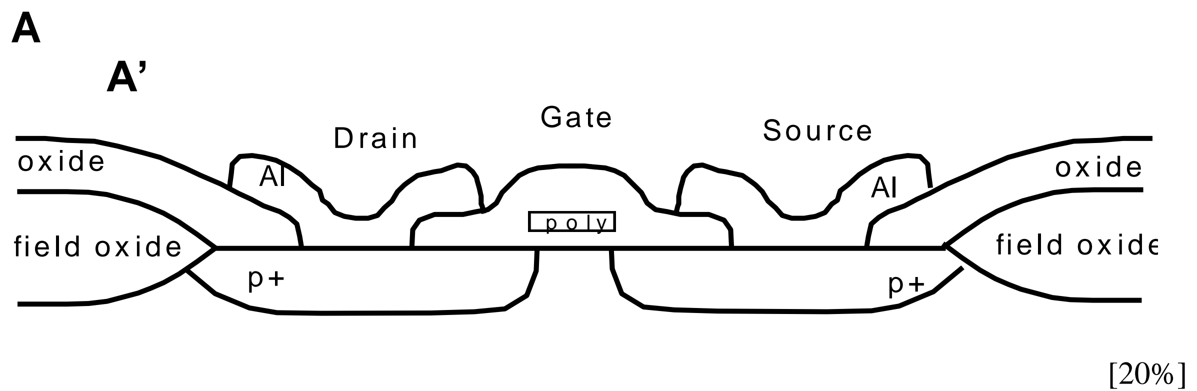
Hence $C_{metal\text{-}sub} = 160 \times 10^{-12} \times 3 \times 10^{-5} + 164 \times 10^{-6} \times 4 \times 10^{-11} = 11.36\ fF$

$C_{drain\text{-}sub} = 9 \times 10^{-12} \times 1 \times 10^{-4} + 13 \times 10^{-6} \times 4 \times 10^{-10} = 6.1\ fF$

Hence $C_{output} = 11.4 + 6.1 + (2 \times 0.6) = 18.7\ fF$ [40%]

(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. [10%]

(d) Cross Section

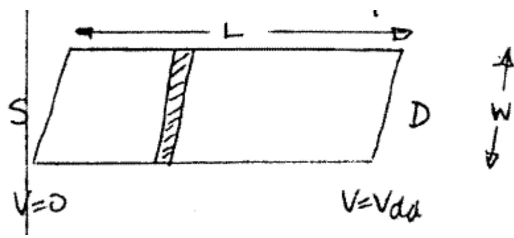


[20%]

Assessor's Comments

This was a popular question, attempted by all candidates. Candidates' recall of the origins of parasitic capacitance varied. Most knew how to go about estimating capacitance, but some overlooked important contributions.

4.



Consider the shaded channel element near the source.

Let Vgate = Vdd to make the device conduct

Let the charge density in the shaded element be Q per unit length

Q	=	C V W	V is strictly the excess of voltage above VT
	~	Cox Vdd W	where Cox is the oxide capacitance pua
Current I	=	Q µn E	where the field E is assumed invariant along
	~	Q µn Vdd/L	

Hence conductance

$$G = I/V = \text{Cox Vdd } \mu n \;\; W/L$$ [20%]

(b) (i) The delay in the first min-geom inverter, if connected directly to the 75 pF pad is:

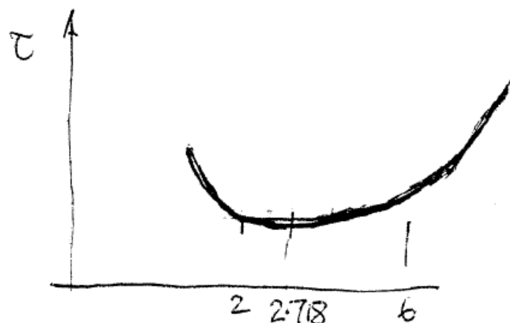
$$\tau_{direct} = \frac{3 \times 75 \times 10^{-12}}{10^{-4} \times 2 \times 3} = 375 \text{ nS}$$ [20%]

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.

(ii) 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.

Hence for minimum delay, for the values given, the number of stages would be:

ln (75/0.2) = ln 375 = 5.9 or 6, rounding up. This suggests 6 stages each a magnification factor U = e = 2.718 larger than the previous one. Other values of U may be advantageous and could result in reduced area, but strictly, such alternatives do not give minimum overall gate delay.



A graph of gate delay versus U has its minimum at U=2.718, but is fairly flat between about 2 and 6.

Hence larger values of U, say 4-6, give only slightly longer delays but may reduce the number of stages and significantly reduce the total pad driver area.

This optimisation is not required here.

Note that the total of 6 stages includes the minimum geometry gate generating the signal. Hence the driver itself consists of 5 further inverter stages.

If channel length is maintained at the minimum dimension, 0.5 μm, successive stages are designed with channel widths inflated by 2.718. 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 shall also assume that the stages are designed for symmetrical rising & falling delays, meaning that the p-channel devices are a factor μn/μp = 2 larger than the n-channel devices.

Using the graded drivers, the delay observed in the min geom. stage – and in each subsequent stage – is

$$\tau_1 = \frac{3\times 2.718\times 0.2\times 10^{-12}}{10^{-4}\times 2\times 3} = 2.7 \text{ ns}$$

Hence the delay in the remaining 5 stages is 5 × 3.7 = 13.5 ns [60%]

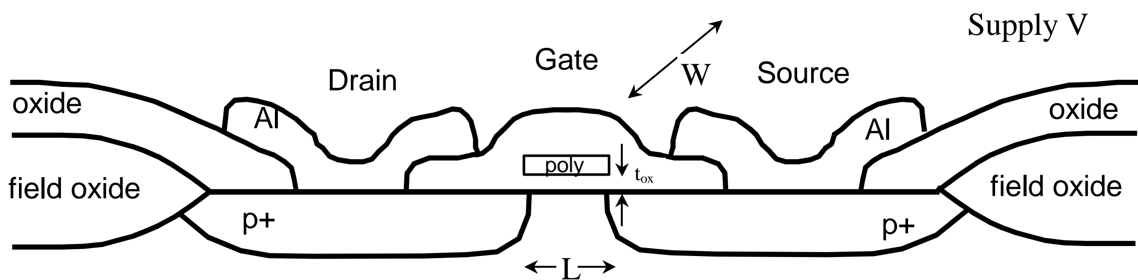
This is the characteristic minimum delay for the driver.

Assessor's comments

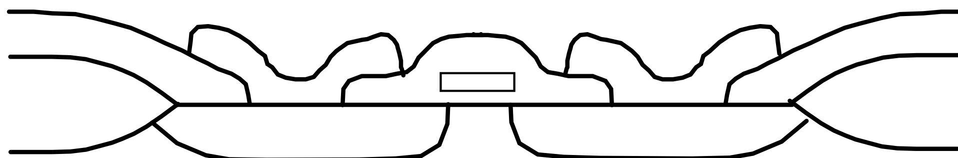
A less popular question. A surprising proportion of candidates could not develop an expression for MOSFET conductance. Although case studies for pad driver design were covered in lectures, some candidates introduced unnecessary complications in the analysis and were unable to complete it.

5. 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 2 micron with limited 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 (between 1.1.and 1.4), as in the diagram. The MOS transistor has critical dimensions L and W (channel length and width respectively) and gate oxide thickness t_{ox}. Applied voltages are represented by V. We consider scaling down all dimensions and voltages by the scaling factor k.



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



[20%]

Packing density α $1/LW$ increased 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 carrier speed is increased 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

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.

For linear circuit elements, there are additional considerations. While 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.

Off-chip loads Devices may have to source or sink off-chip loads, which may be resistive, capacitive, inductive, or any combination. They must therefore be sized to provide these currents. As for digital designs, operation at high frequencies normally dictates the use of small devices, since these have lower parasitics.

Power dissipation Linear circuits may operate at a range of source-drain voltages and currents. This may mean that power dissipation may be significant within devices: Maximum Power Transfer Theorem shows that in a CMOS output stage, power dissipation is maximised when the output terminal lies midway between the supply rails. The heat produced has to be absorbed and conducted away if destructive temperature rises are to be avoided, and calls for devices of larger W and L (though with W/L unchanged) to maintain the energy density at a safe value.

Balanced and matched devices Differential amplifiers and operational amplifiers require balanced and matched devices to minimise offsets and secure acceptable Common Mode Rejection Ration (CMRR). Because of process tolerances and statistical variations, this is less easily achieved with smaller devices.

Electrical noise Electrical noise may be an important factor in linear circuit performance. The principal sources in MOS circuits are: (a) thermal noise, (b) flicker noise. Of these, flicker noise may be most troublesome and it dominates at low frequencies, but may still be significant at frequencies in excess of 1 MHz. It can be reduced by keeping the device gate capacitance large (high $W \times L$); and by using large channel length L, although this reduces the available gain. The front-end stages of a low-noise amplifier need to be of large area, and should operate at low current levels. P-type devices tend to generate less noise than n-type.

Control of channel conductance Where transistors are used as active resistors, control of the resistance can be effected by use of a suitable choice of aspect ratio L/W, as well as by control of V_{GS}. In current sources and current mirrors, device aspect ratio may routinely be optimised to obtain current outputs in the desired ratios.

Major benefits – scaled devices allow:

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

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~1.2 to 1.4 are commonly applied as a process becomes mature to boost the speed of components in that process.

Problem areas

- Some structures e.g. I/O pads, power amplifiers, do not scale
- Loss of compatibility with existing or legacy processes owing to the reduction in supply voltage,
- At the time, the newer process would have offered a lower yield than the existing one; the economic argument would need to be based on a careful review of performance gains, engineering costs, and reduction in yield (leading to possibly poorer sales margins) at the 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.
- 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
- 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. [80%]

Assessor's Comments

A less popular question based on lectured material, but which called for some additional reading in order to score high marks. The attempts seen showed a basic understanding of the scaling concept, but did not provide sufficient detail.

Answers

1. (d)

(i) $C_P = 0.5 \rightarrow P \sim 0.8664$. The probability that all 1000 devices have a parameter within the design spec window is $P^{1000} = (0.8664)^{1000} \rightarrow 0$, and hence the yield => 0%.

(ii) $C_P = 1.0 \rightarrow P \sim 0.9973$, yield = 6.7%.

(iii) $C_P = 1.5 \rightarrow P \sim 0.999993$, yield = 99.3%.

(iv) $C_P = 2.0 \rightarrow P \sim 0.999999998$, yield => 100%.

2. (c)

Minimum refresh time = 1.5 ms

Change in potential observed = 96 mV.

3. (b)

C_{input} = 10.1 fF

C_{output} = 18.7 fF.

4. (b)

(i) τ_{direct} = 375 nS

(ii) Number of stages = 6

Total delay = 13.5 ns.

5.