4B7 VLSI Design, Technology and CAD - 2015 Solutions

Note: Examination Assessor's comments on questions are given at the end of these solutions.

1. (a) How would you define yield in Integrated Circuit (IC) manufacturing? [15%]

YIELD = % of dies on a wafer that meet performance specs.
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 five factors that affect yield. [30%]

1. Dust particles – defined as any unwanted foreign objects in solid, liquid or gaseous state. These type of defects are assumed to be randomly distributed over the surface of wafer and from wafer to wafer. Dust particles on a mask or reticle lead to repeated (predicted) failures.
2. Crystal defects – are present on the wafer prior to fabrication leading to local circuit defects.
3. Mask defects – may be local or global depending on cause.
4. Alignment errors – typically limited to a single wafer.
5. Parameter drifts – introduced during processing and can be attributed to the inherent practical and physical limitations associated with photolithography, deposition and diffusion. These changes may affect all wafers in a lot or may affect individual transistors on a wafer.

(c) Consider a die whose area is 6 mm x 6mm with a defect density of 0.7 cm^{-2}. The die includes an analogue section of 100 transistors whose characteristics must be matched to within 0.5% of the die average value. Assume that the matching characteristics of the transistors are dominated by a single parameter, which follows a "normal" distribution with a $\pm 3\sigma$ window characterized by a variation from the die average of +0.5%.

(i) Determine the yield using Seed's model if the soft faults in the analogue section are neglected. [20%]

The hard yield is the probability that a die is good:

$P_H = \exp(-[AD]^{1/2}) = \exp(-[6mm]^2[7x10^{-3}mm^{-2}])^{1/2} = 60.5\ \%$

The probability that a die has no soft faults = Probability that all 100 meet specs.

The probability that each transistor meets matching requirements is P= 0.9973. Hence probability that all 100 transistors meet specs is $P_S = (0.9973)^{100} = 0.763$.

Overall yield = $P_S\ P_H = 46\%$

NOTE: With a $\pm 3\sigma$ window, characterised by a variation from the die average of $\pm 0.5\%$, corresponds to a window within which 99.73% of the devices will fall.

(ii) Estimate the yield if both hard and soft faults are considered. [20%]

If a 0.25% matching is required, this implies that the probability that a given transistor meets specs = probability that the normal random variable is within $\pm 1.5\sigma$ from the mean. The probability that any one transistor meets specs is 0.8664 (from the probability density table). Probability that all 100 transistors meets specs is:

PS = (0.8664)100 = 5.9 x 10-7

Overall yield = $(0.605)\ (5.9 \times 10^{-7}) = 3.5 \times 10^{-5}\ \%$

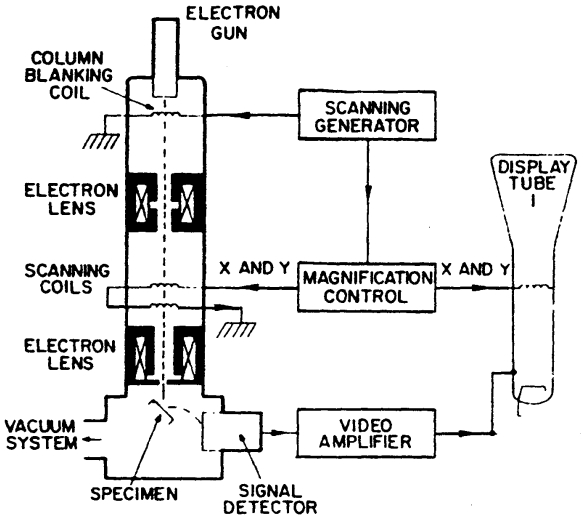
(d) Comment on the physical implications of your result in part (c) above. [15%]

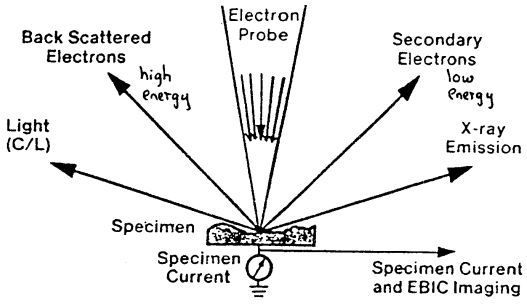
Note that this modest increase in matching requirement 0.5% → 0.25% results in a yield that is so small as to make such a design totally impractical!

Note: Examination Assessor's comments on questions are given at the end of these solutions.

2. The figure shows a schematic electron column, controls and display. The sample is bombarded with electrons in a vacuum. The electron probe is scanned across a portion of the sample in raster fashion to build up a picture. On the right are shown some possible approaches to examining samples with electron beams.

.





Electrons are emitted from a hot filament electron gun at high negative potential with respect to the grounded anode. They travel along the column to the specimen, acquiring energy dependent on the accelerating potential. The specimen is mounted on a micrometer-controlled stage.

Magnetic condenser lenses focus the electron beam to a spot on the surface of the specimen. This spot is scanned across the sample in raster fashion using scanning coils which deflect the beam in x and y.

Secondary electrons emitted from the sample are attracted to the collector screen, which is at a positive potential. Most of the electrons pass through the screen and are further accelerated on to the scintillator (typically biased at around +10 kV relative to the sample) where they produce photons.

In the photomultiplier the photons are absorbed by a photo-emissive surface, which again produces electrons. The electron current is amplified by cascading the electrons down a series of collectors, each of which produces more secondary electrons, i.e. secondary electron multiplication. (It is worth considering why this elaborate approach is used to amplify the small secondary electron current).

The resultant video signal is used to modulate the electron beam on the display cathode ray tube to produce a magnified image of the surface.

The electron energy - which is adjustable by varying the accelerating voltage - determines how far the electron probe penetrates the sample. Samples (particularly biological specimens) are often coated with a metal to make the surface conducting and hence avoid charging effects during observation. Low energy beams (a few kV) may be used to investigate insulating surfaces (e.g. passivated integrated circuits and devices), but with poorer resolution than is obtainable with higher energy beams.

Secondary electrons are emitted at relatively low energy from within the specimen at relatively close to its surface.

Note: Examination Assessor's comments on questions are given at the end of these solutions.

Back-scattered electrons which have high energy comparable with the primary beam (e.g. 40 keV) may also be used to form an image

Using a positively biased secondary electron (SE) detector the electrons scattered from the surface of the specimen are collected efficiently and good surface topography data can be collected

Using a weakly negatively biased detector, only the high energy back-scattered are detected; the signal is noisier but atomic number contrast is obtained. Back-scattered electrons are scattered in single collisions from the atoms in the sample and the signal is not particularly sensitive to the surface condition or topography.

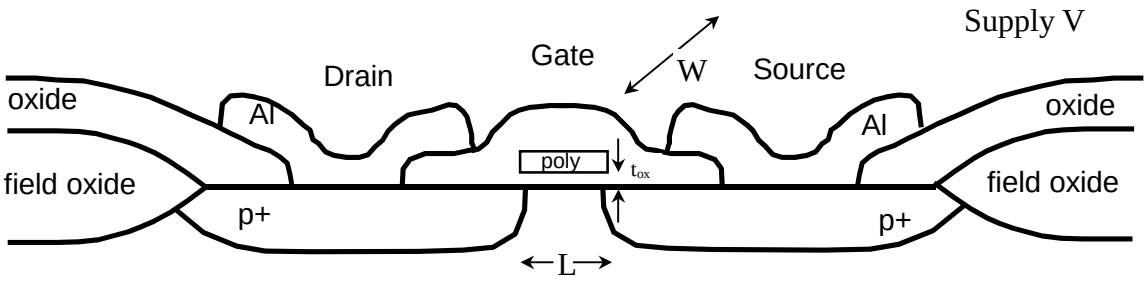
By analysing X-rays generated by the incident electrons, more detail of the chemical composition can be obtained. Light emission (cathodoluminescence) or specimen current are alternative ways of developing an image of the sample.

Under good SE imaging conditions, varying the potential of the specimen surface will modulate the signal obtained at the detector – this is referred to as voltage contrast. This raises the possibility of using the instrument to detect changes of voltage across the specimen, and therefore to obtain information about whether the device is operating correctly. A form of stroboscopic operation may be used in which the beam is chopped at almost the same frequency as that of the device clock, allowing high speed events to be monitored in extended time. [40%]

(b) In *constant field* scaling, geometric dimensions are modified (typically reduced) and electrode voltages are also scaled in order to maintain electric fields constant.

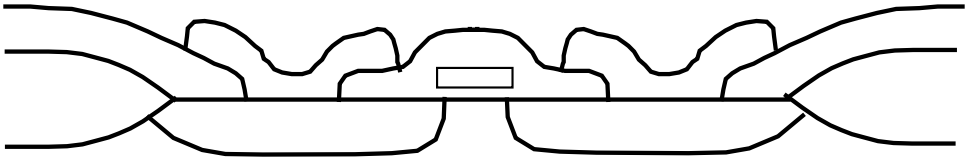
[Historically, device dimensions were scaled from about 6 microns to 1 micron without change in Vdd (*constant voltage scaling*). This offered better delay reduction as well as cost reduction; it maintained continuity in supply and logic level standards. However, it has proved impracticable to push dimensions 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 $k \sim 1$, 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

Note: Examination Assessor's comments on questions are given at the end of these solutions.



Gate area α LW 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	[20%]
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 [10%]

Time constant RC α RC unchanged

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

Capacitance at gate etc	α	LW/t_{ox}	decreased by k
Current consumed I	α	CV/τ	decreased by k
DC Power consumption	α	IV	decreased by k^2
Power density/area	α	IV/WL	constant
Current density J	α	I/dw	increased by k

Note that this assumes device currents are scaled down by k as above. [10%]

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.

Note: Examination Assessor's comments on questions are given at the end of these solutions.

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.

1. Faster clock speeds 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. [20%]

Note: Examination Assessor's comments on questions are given at the end of these solutions.

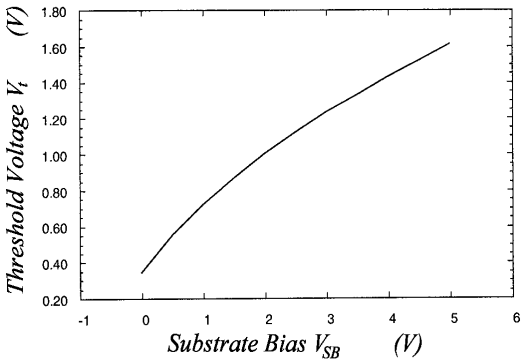
3. (a) The threshold voltage V_T of a MOSFET is that potential which must be applied between gate and source in order to bring about strong inversion within the channel region. There are three main components to this potential:

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

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

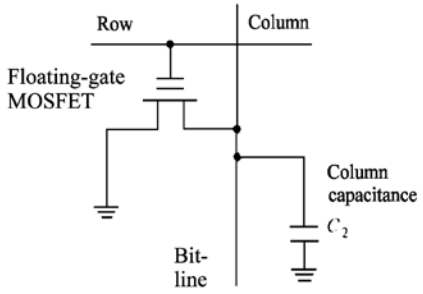
The main factors determining V_T are:

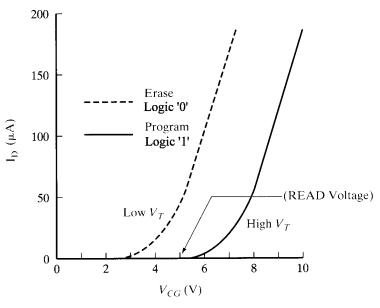
- the materials used for the gate electrode (Al, polySi, ..), determining its work fn
- properties of the dielectric used for the gate insulator, fixing capacitance C_{ox}, and its thickness t_{ox}
- channel dopant density
- impurities, defects, dangling bonds etc at the Si-SiO_2 interface;
- potential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

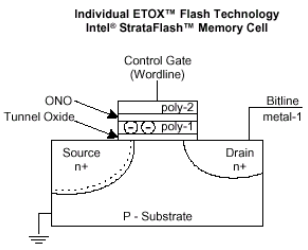


[30%]

(b) **'Flash' memory** - an important type of non-volatile memory, yet has density and speed of operation associated with the DRAM. It has a very simple structure and compact layout - see diagram. The cell closely resembles the one-transistor DRAM cell, except that there is no storage capacitance, and the MOSFET used has an additional *floating gate* between the control gate electrode and the channel. The dielectric separating the floating gate from the control gate is typically a 'sandwich' comprising oxide-nitride-oxide (ONO). The floating gate is electrically isolated, but is capacitively coupled both to the control gate and to the underlying silicon. [20%]







Write operation - consists of placing carefully measured amounts of charge on the floating gate so as to 'program' the MOSFET to have two different values of V_T.

Note: Examination Assessor's comments on questions are given at the end of these solutions.

- If the floating gate contains a large electronic charge, the MOSFET has a higher value of V_T (measured at the control gate) and can be considered to be 'programmed' to the **logic** '**1**' state.
- If the charge is removed from the floating gate, the MOSFET has a lower value of V_T and the cell can be considered to be 'erased' to the **logic** '**0**' state.

Transferring charge in – a high electric field is applied to the drain (bit-line) and to the control gate (row) so that the MOSFET is in saturation. The carriers in the pinch-off region are then highly energetic (hot). If the kinetic energy of the electrons is sufficiently high, a few can become sufficiently hot to be scattered into the floating gate. Once in the floating gate, electrons become trapped in a potential well, and can remain indefinitely without being discharged.

Erase operation - involves removing charge from the floating gate. This is achieved through use of *Fowler-Nordheim* tunneling between the floating gate and source electrode. The control gate is grounded and a high voltage (say 12 V) is applied to the source. The resultant field allows electrons to 'tunnel' through the oxide barrier from the floating gate to the source.

Read operation - is accomplished by applying a moderate voltage (say, 2.5 V) to the drain of the device (bit-line), and a *Read* bias voltage is applied to the control gate.

- If the device is in the '**1**' state, negligible current will flow since the control gate voltage is insufficient to cause a channel with the high V_T.
- If the device is in the '**0**' state, the control gate voltage exceeds the lower V_T, and drain current flows.

The current can be sensed to read out the logic value. Note there is still a delay due to the charge/discharge of the bus capacitance C2, as with the dynamic RAM cell.

More advanced forms of flash memory are now available, in which several different values of V_T may be programmed by injecting different amounts of charge. In this way a single cell can store more than one bit of data. [35%]

Advantages –

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

Disadvantages –

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

[15%]

Note: Examination Assessor's comments on questions are given at the end of these solutions.

4. (a) The key phenomena that lead to dissipation of energy in digital CMOS circuits are:
- charging/discharging of capacitive loads
- crossover currents
- driving resistive loads
- leakage currents

(i) When the logic state of a circuit node of capacitance C changes 1 to 0, the energy stored on the capacitor, $\frac{1}{2}CV^2$, is lost as the capacitor discharges to zero. By symmetry, the same amount of energy is lost as the capacitor charges up from 0 to 1. This is independent of the nature of the charge/discharge paths, and of the signal waveforms. For a collection of nodes, all operating at clock speed *f*, and switching state every clock cycle, the energy loss is $\frac{1}{2}CV^2f$. Since not all nodes will alternate at clock frequency, we introduce an activity factor, $0 < \alpha < 1$ for each node to cover this. α can be determined by statistical observation over many cycles. Hence the total loss across N nodes per second is:

$$V_{DD}^2 \sum_{1}^{N} \frac{\alpha_k}{2} C_k$$

(ii) Both the n and p-channel transistor of a CMOS inverter are partially conducting when the input voltage lies within a range V_{Tn} and $V_{DD} - V_{Tp}$. This means that charge can flow from V_{DD} to ground without ever reaching the load. This is often referred to as cross-over current (dynamic leakage current, short-circuit current, overlap current …). Losses are greater when the input rises/falls slowly, and as the transistor size increases. This leads to a dilemma. To reduce energy loss from this source calls for fast rising/falling edges, but this requires that the previous stage have augmented drive capability and hence greater losses of its own. The best compromise is to have signal rise/fall times about the same and comparable to the propagation delay of the gate. Generally, any step taken to reduce losses from other sources (reducing V_{DD}, α, C , transistor size and node count N) will also reduce crossover losses.

(iii) Resistive loads are encountered in a number of cases in CMOS – for example
- Pseudo nNMOS/pMOS subcircuits (ROM, RAM & PLA structures)
- Amplifiers (e.g. sense amps in RAMs)
- Current sources, current mirrors, voltage dividers
- Oscillators, clock generators, line drivers
- Terminating resistors
- Passive pull-ups/pull-downs on and off-chip
- ESD protection structures.

(iv) Leakage currents are normally minute, but their magnitudes depend on:
- Subthreshold conduction in nominally-off MOSFETs
- Leakage currents in reverse-biased DB and SB junctions
- Leakage currents thru reverse-biased well-well and/or well-substrate junctions
- Electron tunnelling thru the gate oxide (gate leakage)

Such leakage has always been hyper-critical in devices like DRAMS, but not in logic.

The first of these depends critically on $U=V_{DD}/V_T$ and the ratio Ion/Ioff depends on exp(U) . For historic devices U was about 5, but as devices are scaled into the DSM region the ratio has progressively fallen towards around 2. As a result, the significance of leakage currents has risen exponentially.

Leakage in reverse biased junctions is proportional to the number and area of such junctions. It is also heavily temperature dependent. In one top-performing microprocessor operating at $V_{DD} = 0.7V$, it was reported that the power wasted due to leakage grew from 6% to 127% of the dynamic power losses as its temperature rose from 30 to 110°C. [20%]

Note: Examination Assessor's comments on questions are given at the end of these solutions.

(b) Approaches to reduce energy consumption

- Determine loss contributions in functional modules and sub-circuits
- Identify susceptible circuits – e.g. battery operated, circuits idle some of the time
- Consider use of down-scaled CMOS processes to take advantage of:
  - reduced parasitic capacitances
  - reduced V_{DD} needed for given operating speed
- Minimise computational effort for the processing required
  - Evaluate alternative arithmetic/logic designs to reduce activity
  - Simplify activities that don't contribute to processing
  - Consider hard-wired processors vs. software programmed GP processors
  - Avoid DRAMs with mandatory refresh clocks
- Design in *sleep modes* to cut/reduce supply to circuits that are inactive for periods
  - gate clock off or to a lower frequency to reduce activity
  - high and low-side switches (p and n- type MOSFETs) to switch on/off
  - for some sequential logic, can gate clock and reduce Vdd to retain state
- Decide on V_{DD} and V_T according to design requirements:
  - high speed: $V_{DD} >= 4V_T$
  - low power: reduce V_{DD} to minimum necessary for speed required
  - low activity: minimise static currents with high V_T MOSFETs
  - low activity: use channels longer than L_{min} to reduce leakage
- Consider dynamic voltage & frequency scaling where supply voltage is modulated as a function of the time-varying speed requirement
- Minimise the parasitics:
  - Avoid excessive C loads by minimising off-chip connections
  - Trim nodal capacitances by careful attention to layout
  - Avoid R loads
  - Avoid cells with overly strong outputs
  - Downsize MOSFETs wherever possible
  - Sacrifice symmetry of rise/fall to keep p-MOSFETS small and low-C
  - Avoid long runs of parallel buses
- Consider possibility of using sub-threshold operation mode (sometimes called leakage current modulation) for o(10-100x) dynamic energy reduction
- Voltage swing reduction or multi-valued logic, cf. Flash memory
- Adiabatic logic allows recovery of charge from circuit capacitors, e.g. using resonant circuits, at the expense of much greater circuit and operating complexity. [20%]

(c) Numerical Part

The units comprises 40,000 memory cells each driving 10 fF capacitance. Each is clocked at 100 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 $1/2\ CV_{DD}^2$ 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 2,000 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:

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

Note: Examination Assessor's comments on questions are given at the end of these solutions.

Hence the total power dissipation is

$$W = 20 \times 2000 \times ½ \times 100 \times 10^6 \times 10 \times 10^{-15} \times 3.0^2 \quad = \quad 0.18 \text{ W}$$

Hence the average current consumption is

$$I = 0.18/3.0 \quad = \quad 0.06 \text{ A} \quad = \quad 60 \text{ mA}$$ [30%]

(d) The total capacitance being driven here is $40 \times 10^3 \times 10$ fF, or 400 pF. In fact, additional capacitance in other parts of the circuit (e.g. the processor itself) may contribute a comparable amount: hence the total true worst-case current may be twice that calculated. To this must be added any current draw due to pads at input and output taking into account the corresponding driven capacitances (we are not told this). Normally the pads have their own suitably dimensioned power ring, but nonetheless in a conservative design verification of the total current will be necessary. In addition, no account has been taken of losses due to leakage, but in a purely digital design implemented in a 0.5 μm process, these should be small. [15%]

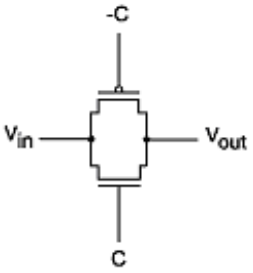
(e) To economise on energy consumption the following should be considered:
Introduce a smaller process geometry (say, 0.25 μm) for this and any other eligible modules, and relax V_{DD} to a value sufficient to accommodate the required clock frequency. Constant-field scaling of CMOS devices is expected to provide a substantially higher switching frequency capability. V_{DD} may correspondingly be reduced since operation at >100MHz is not required, and the dependence of power dissipation on V_{DD}^2 will provide a worthwhile benefit. A similar approach should be applied to other parts of the design.

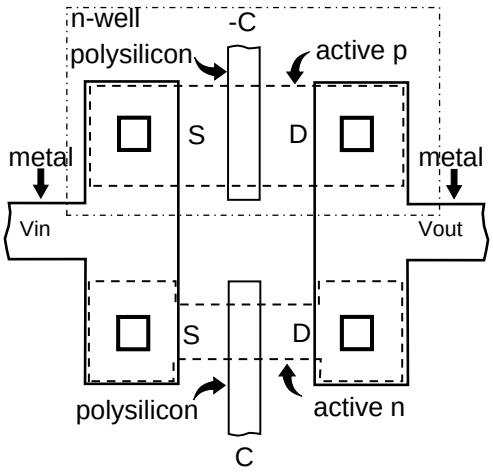
Another approach worth considering is the selective shut-down of parts of the design when not required through the use of a power management policy (are all 18 bits of data required at all times)?

Note also that the peak current may be many times I, with current transients synchronised to clock edges. [15%]

Note: Examination Assessor's comments on questions are given at the end of these solutions.

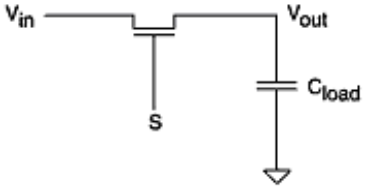
5 (a) A T-gate in CMOS consists of a pair of complementary transistors with source and drain regions strapped together.

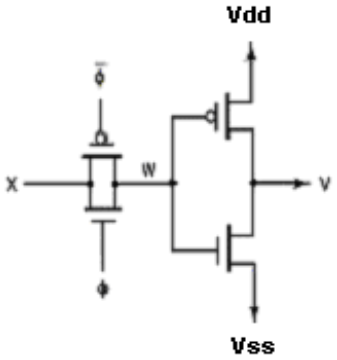




Note: well and substrate taps not shown

The two gate electrodes are driven with complementary control signals C and –C. When C is high and –C is low, both p and n channel devices are conductive. In the opposite situation, both are non-conductive. Note that the device is bilateral when seen from Vin or Vout.





D-type bistable

Consider a single n-channel pass transistor used as a switch. It is conductive when S is high (Vdd), non-conductive when S is low (Vss=0V). To allow a channel to form, V_{GS} must exceed Vt. If C is at Vdd then if Vin is also driven to a high potential around Vdd, Vout cannot rise above Vdd–Vt, typically a 1 volt drop. Thus if Vin were high, Vout would be a so-called weak low. Note that in low state at Vin is transferred reliably to Vout when S is high, since both Vin and Vo << V_{GS}-Vt. As a result it is impracticable to connect single transistor switches in cascade, and they make poor high-side switches. Conversely, a p-type transistor can exert a strong 'high' but only a weak 'low', so poor as a 'low side switch'. By combining the complementary devices in parallel, a switch can be made which suffers from neither of these shortcomings. [25%]

In digital circuits T-gates may be used to realise multiplexers, which may be bilateral. They are commonly used to control feedback paths in sequential (memory) circuits. A major application is in the implementation of a dynamic D-type bistable –see above. Charge is stored on the parasitic capacitance at W.

Advantages

- low device count for multiplexers and bistables

Note: Examination Assessor's comments on questions are given at the end of these solutions.

- bilateral characteristic
- high performance
- can be cascaded

Disadvantages

- effectively a passive device, does not repower logic levels
- requires complementary control signals (extra logic)
- may be sensitive to clock dispersion or skew

Advantages

- efficient switch with low offset voltage
- good frequency response
- good ratio Roff/Ron
- compact structure

Disadvantages

- Insertion loss may be significant and varies with Vin, Vout [15%]

5 (b) An MOS transistor consists electrically of charge stored in the dielectric layers, in the surface/surface states and in the substrate (or well) itself. Switching an enhancement-mode MOST from **off** to **on** consists of applying a gate potential to neutralises these charges and to cause the underlying semiconductor to undergo an inversion due to the E-field from the gate. Hence the threshold gate voltage can be written:

$$Vt = \phi_g + \frac{Q_B - Q_{SS}}{C_O} + 2\phi_{fN}$$

Here, ϕ_g is the WF between gate and Si (typically very small)

ϕ_{fN} is the Fermi potential between the inverted surface and the bulk silicon

C_O is the capacitance per unit gate area

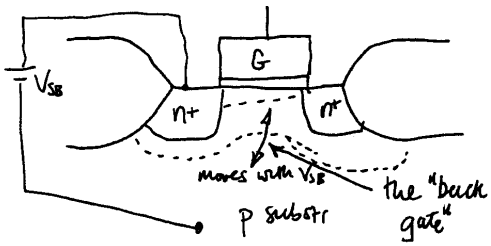
Q_{SS} is the charge density at the $Si:SiO_2$ interface in the channel

Q_B is the charge in the depletion region beneath the gate oxide

With the exception of Q_B, these are dependent only on physical/material parameters and process parameters. However, Q_B depends on ϕ_{fN} and the potential between the transistor source and the substrate, V_{SB}. This is the so-called *body effect*. It is also referred to as *back-gating*, since the substrate in effect acts as a form of gate situated 'behind' the channel.

Increasing V_{SB} causes the channel charge to be depleted; the perceived effect is that Vt is raised for a single transistor, according to the following:

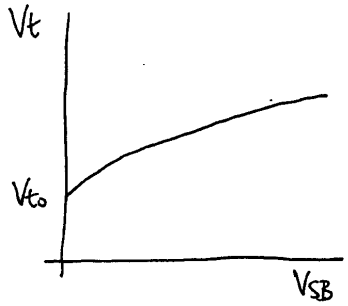
Note: Examination Assessor's comments on questions are given at the end of these solutions.



Change in Vt is given by:

$$V_t = V_{t0} + \gamma\sqrt{V_{SB}}$$

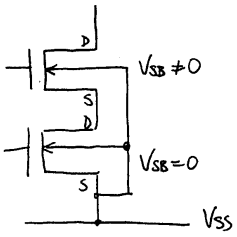
for nMOS devices, where V_{t0} is the threshold voltage for VSB = 0, and γ is typically 0.5 to 1.5, being process-dependent. [30%]



Where transistors are connected in series as in 2/3/4 … input static logic gates, hand computation of the transfer function becomes very difficult, owing to the body effect. The Spice simulator can model this effect accurately

The upper transistor has a higher Vt than the lower owing to body effect. This means that for multi-input gates, the switching level (Vdd/2 for the ideal inverter, is raised for NAND gates and lowered (NOR gates).

This has the subsidiary effect of eroding the noise margins in a corresponding way.



In fact, the switching level and noise margins will change according to which input/s change in a transition.

As far as transient response is concerned, transistors exposed to significant body effect will have a lower apparent conductance in the ON-state, assuming fixed V_{DS}.

As a result, multi-input gates will exhibit slower rise/fall times when parasitic capacitances are charged/discharged through series-connected transistors subject to body effect. [20%]

The designer can compensate for this effect by selecting devices of greater W/L in proportion to the lower conductance in the ON-state of affected devices. Such compensation would have to be done in the light of detailed simulation. [10%]

Note: Examination Assessor's comments on questions are given at the end of these solutions.

Examination Assessor's Comments

Q1 Yield in IC manufacturing; Seed's model

This was a popular question, well answered by a large proportion of the candidates. Most candidates were able to define yield and identify factors affecting it. Two candidates were able to apply Seed's Law as required. Others were unable to decide how to deal with the separate issues of hard and soft errors and lost marks.

Q2 Scanning electron microscope; constant-field scaling

This question was answered quite well. As a part of it was descriptive, a wide range of marks was obtained. Most knew the basics of scanning electron microscopy, but not all could answer the subsidiary questions. The section on scaling was more mixed, but at least one candidate gave a good account of the effect of scaling on all the parameters mentioned. Some gave incomplete answers and lost marks.

Q3 Threshold voltage; flash memory

This question was attempted by all candidates, and was generally well done, despite being mainly of descriptive format. Most knew the definition and determining factors for V_T. The flash memory was quite well described, but candidates' recall of the detail of its structure, operation, advantages and disadvantages varied.

Q4 Power dissipation; power consumption of media processor

A question of limited appeal, attempted by only one candidate, who got the problem part entirely right and scored a first class mark for it, but would have done better if a wider range of measures available to the designer to reduce consumption had been incorporated.

Q5 Transmission gate applications; effect of back-gating on CMOS circuits

A less popular question addressing two themes, one related to characteristics and applications of the transmission gate module, the other to the origin of the body effect / back-gating, both of which called for some diversity of reading. The descriptive sections needed greater detail.