4B7 VLSI Design, Technology and CAD - 2011 Solutions

<table>
<tr><td>1.</td><td></td></tr>
<tr><td>R, I, M1, V_{DD}, V_{REF}, V_{SS}</td><td>It can be seen that as G is shorted to D,
$V_{GS} \equiv V_{DS}$ and M1 is always in saturation.

Hence the second of the S-H equations applies.</td></tr>
</table>

From Kirchhoff, $V_{ref} = V_{DD} - IR$ (1)

From S-H $I = \frac{1}{2}\frac{\mu\varepsilon}{t_{OX}}\frac{W_1}{L_1}\left(V_{ref} - V_{Tn}\right)^2$ (2)

Rearrange (2)

$$V_{ref} = V_{Tn} + \sqrt{\frac{I}{\frac{W_1}{L_1}\frac{\mu\varepsilon}{t_{OX}}}} = V_{DD} - IR$$

Hence

$$V_{ref} = V_{Tn} + \sqrt{\frac{V_{DD} - V_{ref}}{\frac{RW_1}{2L_1}\cdot\frac{\mu\varepsilon}{t_{OX}}}}$$

(b) Substituting the given values:

$$2 = 1 + \sqrt{\frac{6-2}{\frac{R}{2}\cdot\frac{40}{4}\cdot 2\times10^{-5}}} \quad \text{and}$$

$$\frac{6-2}{5R\cdot 2\times10^{-5}} = (2-1)^2 = 1$$

Thus

$$R = \frac{4}{5\times 2\times10^{-5}} = 40\ \text{k}\Omega$$

and

$$I = \frac{6-2}{40\times10^{3}} = 100\ \mu\text{A}$$

Using a standard polySi of typically 50Ω/square, this resistor would require L/W of 800, and would occupy a great deal of space.

(c) An ideal voltage reference is independent of the power supply that services it. Practical references fall short of the ideal and 'PS sensitivity' expresses this in a quantitative way. It is defined:

$$S_X^{V_{REF}} = \frac{\partial V_{REF}}{V_{REF}} \bigg/ \frac{\partial X}{X} = \frac{X}{V_{REF}} \times \frac{\partial V_{REF}}{\partial X}$$

where V_{REF} is the voltage reference, X is the parameter under consideration, V_{DD} in this case; if S is unity, a 10% change in V_{DD} will bring about a 10% change in V_{REF}. The objective is to produce a circuit design in which $S_X^{V_{REF}}$ is as small as possible for relevant parameters X, such as V_{DD}.

An ideal voltage reference will also be independent of ambient temperature. Since most electronic components and materials used in electronic circuits have temperature-dependent properties, which may be of different magnitudes and polarities, this is a non-trivial problem. To quantify the effects of components due to temperature change, we define 'fractional temperature coefficient TC_F, defined as follows for a component of value X:

$$TC_F = \frac{1}{X}\cdot\frac{\partial X}{\partial T} = \frac{1}{T}\cdot S_T^X$$

using the notation above. TC_F is typically expressed as parts per million per deg. C, or ppm/deg C. In circuits comprising several components that determine the output, all may contribute to the TC_F of the output itself, and the various contributing values of TC_F must be taken into consideration with the governing equation that determines the output. TC_F may be positive or negative depending on the materials of which the component is made and on its mode of operation. This opens up the possibility of using devices in combination in circuits such that the temperature-dependent effects cancel to a significant extent.

Achieving minimum $S_{V_{DD}}^{V_{REF}}$ and TC_F simultaneously for a voltage reference is a challenge.

(d) Some of the techniques available are:

(i) use of a Zener diode where the reverse breakdown voltage characteristic is almost independent of current. Most Zener diodes have a significant temperature coefficient, but this can be minimised by use of devices in which the different physical effects (zener & avalanche effect) wholly or partially cancel. Zener diode may not be compatible with commodity CMOS processes and may have to be provided off-chip.

(ii) use a band-gap reference – an arrangement in which the VF of a pair of forward biased diodes operated at different current densities are subtracted in an amplifier to give a result almost independent of temperature.

(iii) use an XFET where the difference in pinch-off voltage of two similar FETs is used to provide a stable reference.

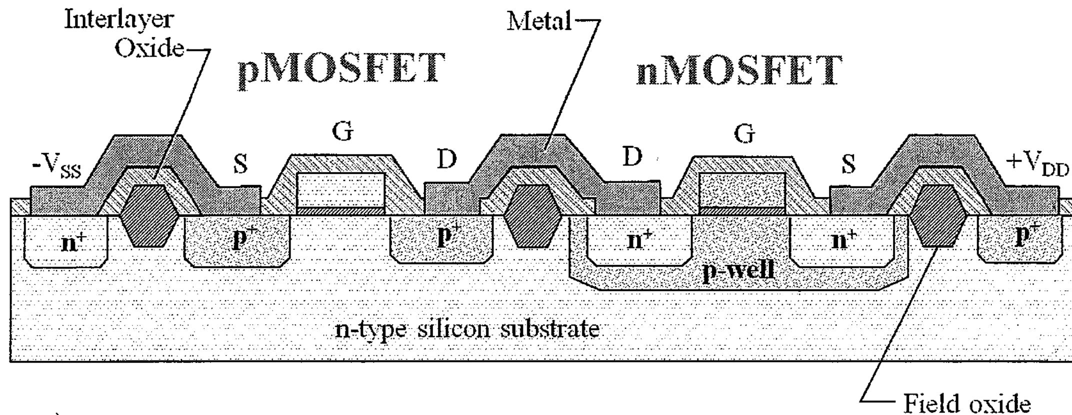
Both (ii) and (iii) call for special processing that may not be compatible with commodity CMOS.

(iv) A reference along the lines proposed by Kwon et al (covered in lectures, section 10) uses only components available in a regular CMOS process. A thermal voltage is developed by use of a pair of MOSFETs operating sub-threshold. This has a positive temperature coefficient. A second voltage is developed from a self-biased β-multiplier circuit, again based on MOS transistors, which has output proportional to V_T: this has a –ve temperature coefficient. When these are scaled appropriately and summed in a suitable amplifier, the result has a near-zero tempco.

Q1 examiner's comment:

This question was attempted by just under half the candidates, and was based on a case study covered in some depth in lectures. Most were able to deal satisfactorily with the numerical part, but not many understood the issues regarding stability, or knew of alternative approaches to voltage reference design.

2. (a) Standard book work – Prof Kelly's lecture notes.



(b)

(i) trench isolation,

(ii) p-well implant into n-type substrate

(iii) gate oxidation

(iv) polysilicon gate deposition

(v) n+ and p+ source and drain contact and p-well body contact

(vi) oxide deposition, patterning and etching to form contact windows

(vii) metallisation and patterning

Order of steps: cannot add anything until what has to be removed is removed.

Field and gate oxide layers have mask functions for various implants.

Ohmic contacts anneal at low temperatures, so these are done last.

(c) SOIT:
Simox, wafer bonding , and unibond-smart-cut examples.
With notes.

Examiner's comment:

A popular question, this was reasonably well answered, although the quality and clarity of the diagrams varied widely, in some cases not doing the students justice. Although part (b) asked to list the processing steps, some chose not to supplement the list with diagrams that make it clear what is going on.

3 (a) Quality assurance
Concern for the life of a product all the way through design, manufacture, mass-production, shipment and field of use.
This includes reliability and qualification (the product does exactly what is says it should).
Achieved reliability is a large part of the success of the IT revolution.
Many applications are in safety critical systems.
Complexity of product means that working demands extra care at each fabrication step, so QA monitoring is essential at each stage.
(b)

Stress:	Limitation on life	Mitigation strategy
temperature,	Materials diffuse and characteristics drift	Don't drive too hard, heat sink
current,	Electromigration, heating, ...	Design to avoid high currents
local electric field,	Electromigration, defect formation, ...	Design and materials to lessen
moisture,	Corrosion affecting materials properties	Encapsulate
stress,	Materials move and electric properties vary	Careful choice of materials
shock,	Contacts break and devices fail	Product design and care in use.

I

3. (continued) (c)
(i) Infant mortality: defect in design or manufacture that fails at or near beginning of life: electrostatic discharge, latch-up, thermal runaway
(ii) Failure in life: Random processes that occur at low but constant rate: e.g. contaminating ions entering key regions of the device, moisture or gas penetration of encapsulant
(iii) Wear and tear end-of-life: hot carrier degradation, electromigration, chemical or mechanical change

Examiner's comment:

This question was attempted by just under half the cohort. It followed straight from parts of different lectures. Some went off on tangents on part (a), but otherwise the answers to the questions showed the range of understanding of the candidates.

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 VDD - V_{Tp}. This means that charge can flow from VDD 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 hitherto in regular 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.

(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 Vdd 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 Vdd and V_T according to design requirements:
    - high speed: Vdd >= $4V_T$
    - low power: reduce Vdd to minimum necessary for speed required
    - low activity: minimise static currents with high V_T MOSFETs
    - low activity: use channels longer than Lmin 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.

(c) Numerical Part

The units comprises 36,000 memory cells each driving 15 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 $CVdd^2$ 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 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:

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

Hence the total power dissipation is

$W = 18 \times 2000 \times \frac{1}{2} \times 100 \times 10^6 \times 15 \times 10^{-15} \times 3.3^2 = 0.29$ W

Hence the average current consumption is

$I = 0.29/3.3$ A = 89.1 mA [20%]

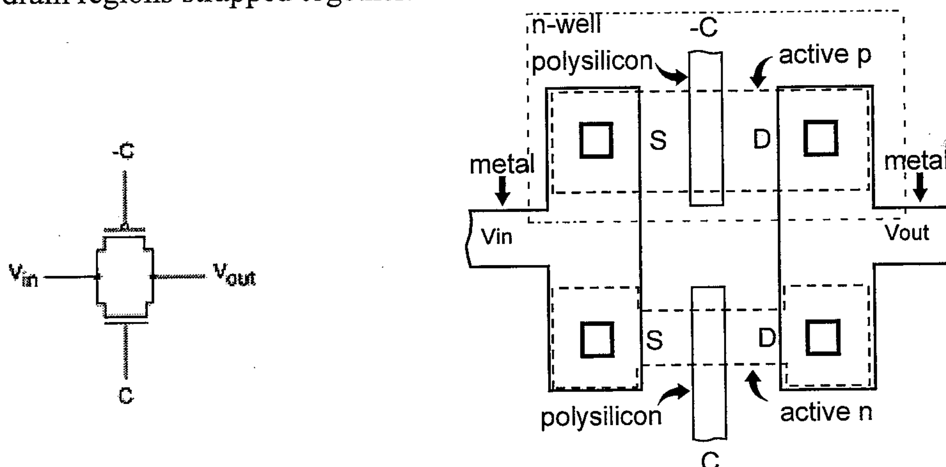
The total capacitance being driven here is $36 \times 10^3 \times 15$ fF, or 540 pF. In fact, additional capacitance in other parts of the circuit (e.g. the processor itself) may contribute a comparable amount: hence the total true worst-case current may be twice that calculated. To this must be added 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 0.7 μm process, these should be small.

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. [20%]

Examiner's comment:

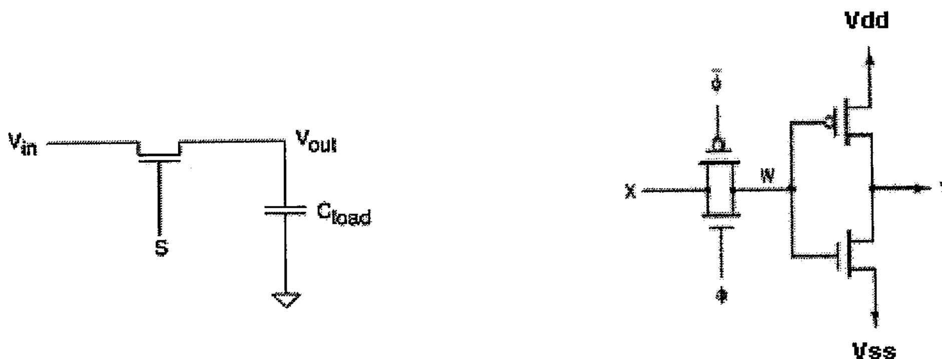
This question was attempted by just over half the candidates. The numerical part was well done by and large, but a number of attempts were let down by relatively weak responses in the descriptive parts which were based on the content of a lecture handout.

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. This allows its use in linear circuits.



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.

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
- 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

In analogue circuits T-gates may also be used in bilateral switches/MUXs for linear signals. A common usage is in sample-hold circuits or electronic exchanges.

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

(c) Critical success factors in VLSI Manufacture
Quality of all aspects of product: continuing business on built-up reputation
Cost of product: for the same performance and quality, lower cost will mean market share
Delivery of product: speed can mean market share and premium prices
Service of product: trouble shooting, accommodating customer schedules, correct product for customer, all adds up to satisfaction and likelihood of repeat business.

(c) Physical layout of production line
(i) Relationship to utility connections
(ii) Wafer moving path for most common products: wafers move constantly and often recycle among equipment, so some optimum can be decided on a least distance path for main product.
(iii) Routing of manufacturers supplies (dummy wafers, quartz-ware, etc)
(iv) Inventory storage
(v) Potential for cross-contamination
(vi) Other: maintenance, parts/tools storage, other house-keeping materials,
(vii) Degree of automation (e.g. cluster tools).

Examiner's Comment:

A fairly popular question, though some evidence that it was tackled as the last attempt with little time available. The first part on transmission gates, which were covered in depth in the lecture notes, required a descriptive answer and was generally well done. For parts (b) and (c): the answers were less impressive in part (c) and some went off track in (c).