Version MWCT/5

EGT3
ENGINEERING TRIPOS PART IIB

Monday 5 May 2025 9.30 to 11.10

Module 4B28

VERY-LARGE SCALE INTEGRATION (VLSI)
(CRIB)

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
Attachment: 4B28 Very-Large Scale Integration (VLSI) data sheet (4 pages).

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 The cross-section of a fabricated PMOS transistor is shown in Fig. 1.

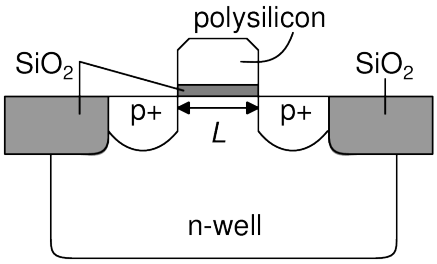


Fig. 1

(a) VLSI fabrication engineers have been trying to reduce the length L of CMOS transistors for years. Explain the most important reason behind the reduction of L. Moreover, discuss an effect of reduction in L on the electrical properties of the fabricated PMOS transistor. [20%]

Solution:

Reason:

The cost of VLSI fabrication is proportional to the power of four of the area (area^4). Therefore, by reducing the physical geometry of the transistor, more transistors can be crammed within the same area. As a result, the cost of fabricating a transistor could be drastically reduced.

Effect:

The current across a short channel saturates quickly, as result of high electric field. A short-channel transistor conducts less current than a long-channel transistor with the same width-to-length ratio.

OR

Technology scaling also reduces the thickness of oxides so there is more sub-threshold current while the transistor is off and more leakage current in short-channel transistors. The static power increases as a result.

(b) The gate-to-channel capacitance C_{GC} is affected by the mode of operations of the transistor. Based on the structure shown in Fig. 1, explain briefly the sources of C_{GC} when the transistor is cut-off and in saturation. [15%]

Solution:

At cut-off, there is no mobile carriers in the channel to form the other conductive plate for the capacitor. Now the n-well substrate acts as the other conductive plate. This capacitance is proportional to the area of the polysilicon gate over the substrate ($C_{GC} = WLC_{ox}$).

At the saturation, the channel pins-off at the source of the PMOS transistor and there the gate capacitance with the source C_{GCS} drops close to 0. And the gate capacitance with the drain (C_{GCD}) rises from $\frac{1}{2}WLC_{ox}$ to $\frac{2}{3}WLC_{ox}$ at full saturation.

(c) If the transistor of length 0.2 μm was fabricated using the generic 0.18 μm technology to provide a 10 μA current from its source to its drain under the following operating conditions: $|V_{GSp}| = 0.8$ V, $|V_{DSp}| = 0.4$ V, $|V_{BSp}| = 0$ V. What should be the width of this transistor, if the transistor is assumed to be a short-channel device? Ignore channel length modulation. [30%]

Solution:

Consider the PMOS as a short-channel device,

$$V_{DSATp} = \frac{(0.8 - 0.5)(4.8)}{(0.8 - 0.5) + 4.8} = 0.283\,\text{V}$$

Since $|V_{DSp}| > V_{DSATp}$, the transistor is in the velocity saturation mode, let w be the width of the transistor, then

$$|I_{DSp}| = (w)(100)(8 \times 10^6)(1 \times 10^{-6}) \left(\frac{(0.8 - 0.5)^2}{(0.8 - 0.5) + 4.8} \right)$$

$$10\mu = 14.1w$$

$$w = 0.7\,\mu\text{m}$$

(d) Explain why a PMOS transistor is not an ideal switch. Discuss a CMOS circuit that could serve better as a switch for digital systems. [20%]

Solution:

PMOS transistor does not pass '0' well when its drain is connected to 0/GND and its source becomes the output of switch. The transistor will be turned off once the source voltage falls below V_{Tp} ($|V_{GSp}| < |V_{Tp}|$).

Transmission gate - a NMOS transistor and a PMOS transistor in parallel with each other - is a better alternative and provides a rail-to-rail signal transfer (V_{DD}-GND) with a stable on-resistance.

(e) An engineer said: "*the resistance of the PMOS transistor in the 0.18 μm generic technology is around 30 kΩ when its width and length are both 200 nm*". Explain whether you agree with this statement or not. [15%]

Solution:

The equivalent resistance model is accurate only when PMOS is charging with a constant channel current in its velocity saturation (or saturation) mode. It is likely between an output voltage from 0 to $0.5V_{DD}$.

Its equivalent resistance increases drastically when the channel current decreases with smaller $|V_{DSp}|$.

Therefore I do not agree with the statement.

Numerical answers: (c) 0.7 μm

Assessor's comments:

Part (a) was well attempted. It is ideal to state the strong electric field as the main reason for velocity saturation in CMOS transistors. Part (b) expects a comparison between the difference in capacitance between cut-off / saturation due to the formation of conducting channels. Most students answered part (c) perfectly. In (d), some were confused between a transmission gate and an inverter, with the former being a proper choice as a switch. While most students agree with the engineering in (e), it is important to recognise that the derivation of the equivalent resistance model assumes a V_{DD}-V_{DD}/2 discharge or 0-V_{DD}/2 charging when the transistor is in the velocity saturation mode.

2 A three-variable Boolean function F is given as:

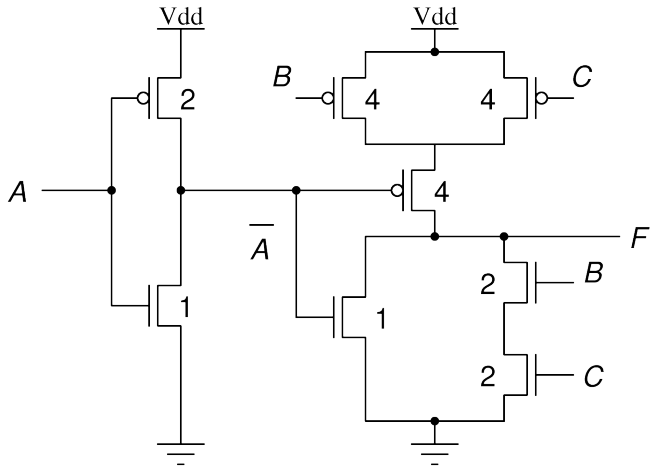
$$F = A(\overline{B} + \overline{C})$$

(a) Design a static CMOS gate for F and provide the size ratios for all transistors such that the *worse-case* pull-up and pull-down equivalent resistances of the gate is equivalent to that of the standard inverter. Assume that only positive phases of the variables (i.e. A, B and C) are available. Use no more than ten transistors and connect as many transistors to V_{DD} or GND as you can. [25%]

Solution:

$$\overline{F} = \overline{A} + BC$$

Therefore an inverter is required to generate $\overline{A}$.



(b) If area is of utmost importance, what change(s) would you make to the static gate in part (a)? Explain one trade-off for such change(s). [15%]

Solution:

Make it pseudo-NMOS: replace the pull-up network (PUN) with a PMOS transistor gated with ground (0) such that it acts as a resistive load.

Trade-off (*any one of the following*):

1. the static power increases since there will be a direct-path current when the output is 0.
2. the voltage-transfer characteristics (VTC) is not symmetric anymore and the low noise margin (NM_L) will be compromised.

(c) Show that the logical effort g for the input B of the static gate in part (a) is 2. Thus, derive the *worse-case* intrinsic delay p. [20%]

Solution:

As shown in the design in part (a), the input capacitance of B is (4 + 2) = 6 units while the input capacitance for the standard inverter is 3. Since the equivalent resistances of the gate is matched with that of the standard inverter, the logical effort g for B is:

$$g = \frac{4+2}{3} = 2$$

The worse-case intrinsic delay happens when A is the slowest signal and it runs through both the inverter and the static gate. So the worse-case intrinsic delay p is:

$$p = 1 + \left(\frac{4+1+2}{3}\right) = \frac{10}{3}$$

(d) Using the logic gate for F as an example, explain the two sources of dynamic power in static CMOS gates. [20%]

Solution:

1. Switching power: energy is drawn from the supply to charge the output capacitance from the diffusions and coupled gates (if any) when F is 1. The charges will then be lost when F turns 0.
2. Short-circuit power: both NMOS and PMOS will be ON during the switch transient so there is a direct current from the supply to the ground for a short period of time. This consumes extra energy as the gate switches.

(e) Suppose the logic gate X is used in a multi-level digital circuit shown in Fig. 2. The

inverter is of the standard size, the 2-input NAND gates are of the same size, and it is given that the input capacitance to the inverter and the final load capacitance are 3 fF and 20 fF respectively. Assume that the ratio between the gate input capacitance and the gate intrinsic capacitance $\gamma = \dfrac{C_g}{C_{int}}$ is 1.

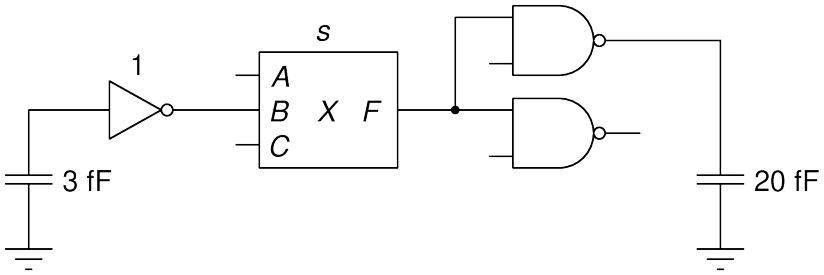


Fig. 2

Based on the logical effort method, determine the size ratio s for X that would minimise the propagation delay to drive the load. [20%]

Solution:

$$F = \frac{20}{3}$$

$$G = 1 \times 2 \times \frac{4}{3} = \frac{8}{3}$$

$$B = 1 \times 2 \times 1 = 2$$

$$H = \frac{20}{3} \times \frac{8}{3} \times 2 = 35.6$$

$$h = \sqrt[3]{35.6} = 3.29$$

$$f_{inv} g_{inv} b_{inv} = 3.29$$

$$f_{inv} = sC_{g,inv}/C_{g,inv}$$

$$s = 3.29$$

Numerical answers: (c) $\frac{10}{3}$ (e) 3.29

Assessor's comments:
A few students missed the stated requirement of connecting as many transistors to VDD or

GND as possible, but otherwise part (a) was well answered. For (b), using the minimum sized transistors are valid solution but the corresponding trade-off of comprised noise margin and worse propagation delay should be discussed. In (c), the worse-case intrinsic delay should include that from the inverter for generating A' (if present in the answer). Parts (d) and (e) are well attempted.

3 A clock distribution network with a clock source S and five sink nodes A, B, C, D and E was routed on metal tracks with a pitch of 1 mm. Fig. 3 shows the network. The width of metal interconnect is 2 μm, and the sheet resistance of the metal is 5 m$\Omega/\square$. It is known that the sink nodes all have a lumped capacitance of 80 fF.

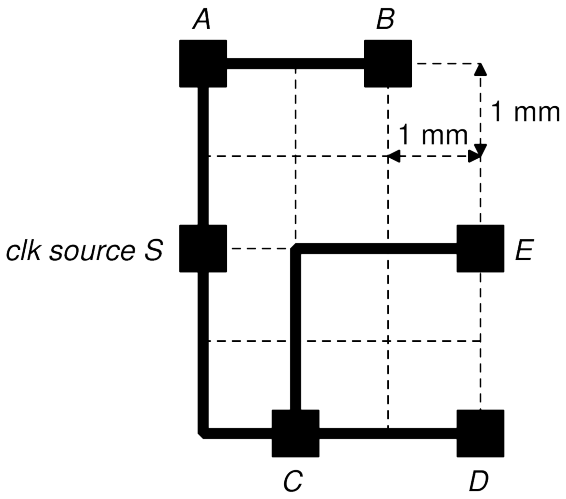
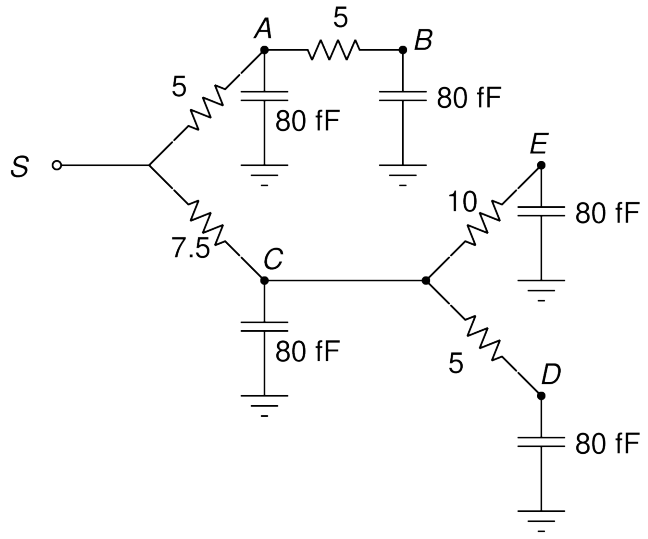


Fig. 3

(a) Calculate the resistance of an interconnect segment of 1 mm length. Using this result, sketch the resistor-capacitor (RC) tree modelled from the clock distribution network in Fig. 3. [30%]

Solution:

$$R = 5\text{m}\Omega \times (1 \times 10^{-3})/(2 \times 10^{-6}) = 2.5\Omega$$



All resistances are in Ω.

(b) Thus, calculate Elmore delays τ_B and τ_E at nodes B and E respectively. [20%]

Solution:

$$\tau_B = 5 \times 80f + (5 + 5) \times 80f = 1.2ps$$
$$\tau_E = 7.5 \times 80f + 7.5 \times 80f + (7.5 + 10) \times 80f = 2.6ps$$

(c) How does the difference in arrival times of the clock signals affect the clock frequency to operate the system? Discuss a measure to mitigate the difference. [15%]

Solution:

Positive clock skew requires a longer clock period such that there is enough time to set up the signal at the next register. This lowers the clock frequency as a result.

Mitigation measures (*any one of the following*):

1. Clock skew may be reduced by routing the clock signal against the direction of logic cells/gates for simpler system with fewer registers.

2. Use wider interconnects to reduce the resistance and therefore the RC delay.
3. A matched RC tree/clock grid/clock mesh should be designed to distribute the clock to minimise the difference in RC delays at each of ther registers.

(d) Fig. 4 illustrates the design of a dynamic, true single-phase clock register. Assume that all transistors had been carefully sized to ensure correct function. Explain the working principle of this positive-edge triggered register. [20%]

Solution:

When CLK = 0, transistors M_2 and M_6 are turned ON, allowing D to drive internal node X ($X = \overline{D}$) and precharge Y to 1. M_4 and M_8 are both turned OFF such that the output Q could not be modified.

Then the positive edge comes and CLK = 1, M_2 and M_6 are now OFF to isolate D from X. At the same time, M_4 and M_8 are now ON to evaluate X. If $X = 0$ ($D = 1$), Y will be remained charged so Q is also 1 (consider two inverters in series); when $X = 1$ ($D = 0$), Y will be discharged through M_4 and M_5 so Q becomes 0 eventually.

(e) Explain whether the register in Fig. 4 is suitable for the clock distribution network in Fig. 3 or not. Assume a perfect clock source at S. [15%]

Solution:

The register is vulunerable when the rise time of the clock signal is long (i.e. the clock edge is not sharp enough) because M_2, M_4, M_6 and M_8 are not supposed to be ON all at the same time.

Luckily, as calculated in (c), the rise time of the clock signal is in an order of ps so the clock edge should be sharp enough for the clock distribution network.

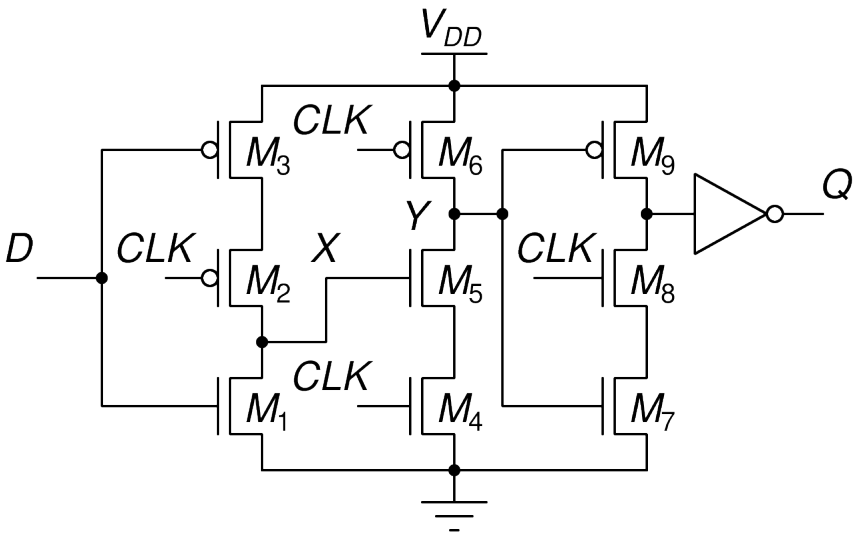


Fig. 4

Numerical answers: (a) 2.5 Ω (b) 1.2 ps; 2.6 ps

Assessor's comments:
Most students answered parts (a) and (b) perfectly. In (c), the impact of more probable, positive clock skew on the clock frequency is crucial to bring forth possible mitigations. Some students forgot to point out the function of transistor pairs of M_2-M_6 and M_4-M_8 in isolating input D from output Q in the flip-flop design. This will lead to the fact that the design is vulnerable to a slow rising clock edge as in (e).

4 Fig. 5 shows a dynamic XNOR gate in the generic 0.13 μm technology with input arriving in both positive and negative phases. The lengths and widths of all NMOS transistors are 100 nm and 300 nm respectively.

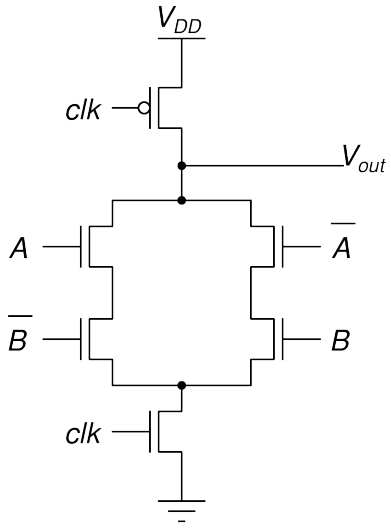


Fig. 5

(a) Calculate the equivalent pull-down resistance of the gate. [15%]

Solution:

$A\bar{B}$ and $\bar{A}B$ are mutually exclusive. So during pull-down, only one side of the network will be conducting.

$R_{eq} = 3 \times R_{eqn} \times \frac{100}{300} = 12.5\ \text{k}\Omega$

(b) Derive the logical effort for this gate when the gate output is 0. Assume that every input (A, $\overline{A}$, B, $\overline{B}$) is provided independently. [15%]

Solution:

Logical effort $g = \frac{R_{gate}C_{gate}}{R_{inv}C_{inv}} = \frac{12.5k \cdot 300/100C_{inv}}{12.5k \cdot 300/100C_{inv}} = 1$

(c) From the measurement of a fabricated gate, the output voltage V_{out} gradually drops during the evaluation phase, even when the output should be 1 and is not connected to any

other gate. Explain this phenomenon. Will the voltage drop increase or decrease as the VLSI technology scales down (i.e. smaller feature size)? [25%]

Solution:

This is because there is charge leakage from reversed-biased *pn* junctions and sub-threshold currents. Charges are lost once the PMOS transistor is turned OFF and the output node is left floating. This can also be due to charge sharing between the output node and the internal nodes. For instance, if A changes from 0 to 1 and B remains 1 during evaluation, the internal node between the two NMOS transistors on the left of Fig. 5 will share some of the charges at the output node.

Leakage, from both reversed-biased junctions and sub-threshold currents, is more serious in smaller feature size as the oxide is thinner between doped silicon areas. So the voltage drop will increase.

(d) Discuss a potential issue when cascading this dynamic gate. Propose a solution for the issue. [25%]

Solution:

Suppose there are two identical XNOR gates that are cascaded by connecting V_{out1} (first level) with A_2 (second level). After precharge, $V_{out1} = V_{out2} = 1$. If B_2 is 0, then the PDN of the second gate is turned ON to discharge V_{out2}.

This is problematic as V_{out1} could also be evaluating down to 0. Since it takes some time for V_{out1} to drop below V_{Tn}, V_{out2} drops unnecessarily and potentially gives an incorrect output depending on the time it takes to complete the evaluation of V_{out1}.

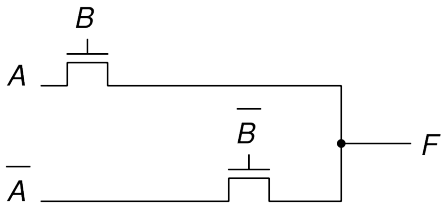
An inverter can be added at the outputs to ensure that the inputs to the next level of dynamic gates make only a 0-to-1 transition.

(e) Design an alternative gate for the same XNOR function but with fewer transistors. State a disadvantage of your alternative design when compared with the dynamic gate. [20%]

Solution:

XNOR function could be implemented by NMOS pass transistors.

(cont.

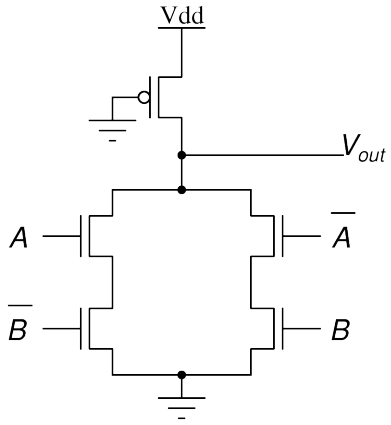


(Pass transistors may be replaced by transmission gates)

For the pass transistor implementation, the robustness is compromised with an imbalanced VTC, and lower $V_{OH} < V_{DD}$ (full rail-to-rail output in dynamic gate).

OR

It could be implemented in the pseudo-NMOS style, as shown below:



For pseudo-NMOS implementation, the static power is high as there is direct-path current from V_{DD} to GND when output is 0. / Also, the pull-up transistor has to be sized carefully.

Numerical answers: (a) 12.5 kΩ (b) 1

Assessor's comments:

Some students did not notice that the pull-down network conducts always with three NMOS transistors in series in (a). Part (b) could be better answered if the equivalent resistances of the gate and the standard inverter are compared (matched in this case) before going on comparing the input capacitances. For (c), many blamed the leakage on backgate coupling despite the lack of output stated in the question. The issue of coupling or cascading should therefore be discussed in (d). Part (e) is generally well answered,

either with pass transistors (better area improvement) or pseudo-NMOS logic.

END OF PAPER