

Digital VLSI Design (ECE 314/514)
Mid-Sem 2023 Rubric

Q.1 a) Why does a FF need to satisfy setup time and hold time constraints? Explain.

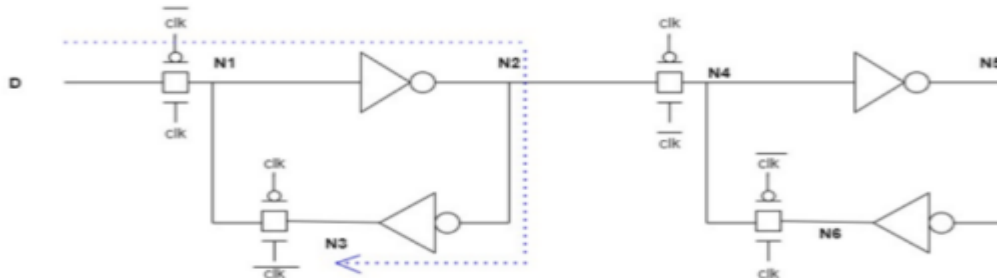
Ans.

Setup and hold time constraints should be satisfied to ensure correct functionality and delays. If setup/hold timings are not met, the FF may enter into a metastable state, and data stored in the FF can be corrupted.

One of the ways to characterise setup/hold times is to ensure that the impact on delays is limited within a defined threshold.

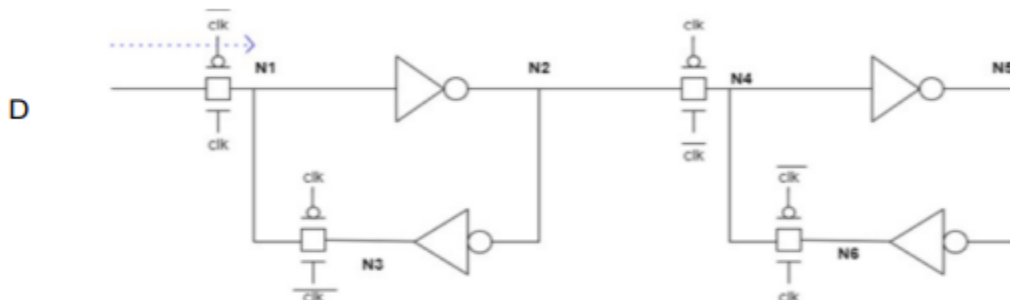
How/ Where is setup time defined?

When the clock (CLK) is low, input D follows the path D→N1→N2→N3 and reaches the input of TG2. This time is known as the setup time of a flop. If data changes within this time, the data won't be able to reach the input of TG2 and when TG2 turns on, it might be two different data reaching node N1, which can cause metastability, glitch at the output and therefore dissipate extra power.



How/ Where is hold time defined?

Hold time arises because of the finite delay between the clock and clock_bar signals which control the switching of transmission gates. The transmission gate takes some time to switch on or off, which is known as the hold time of a flop. So it is necessary to maintain a stable value at D pin to ensure a stable value at node N1 (from the first transmission gate) which is translated to output.



b) Suppose you have designed a chip and after fabrication it fails, how would you determine if it is a setup failure or hold time failure?

Ans:

If the failure is recovered by reducing the operational frequency, then it is a setup failure.

However, if there is no impact of reducing the frequency on failure rate then the failure could be a hold failure.

The hold time requirement is more crucial to meet. After fabrication we cannot add and reduce delay in the data or clock paths to adjust setup or hold requirements. So if there is hold failure, there is no remedy, and the chip has to be rejected as hold requirement is not dependent on clock frequency.

However, we can reduce the clock frequency to meet the setup requirements and we will still be able to sell the chip rated at a lower frequency rating.

Q.2 Consider the circuit of Fig. 2. Assume the inverter switches, ideally at $V_{DD}/2$, neglect body effect, channel length modulation, and all parasitic capacitance throughout this problem.

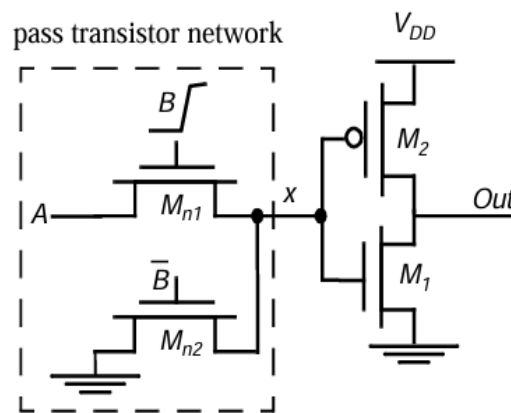


Fig. 2

- a. **What is the logic function performed by this circuit?**

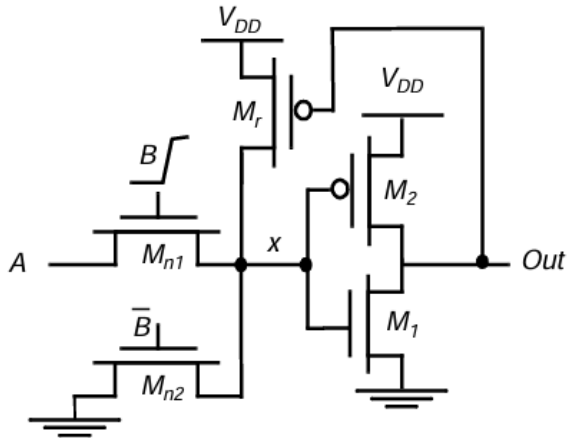
Ans. The circuit is a NAND gate.

- b. **Explain why this circuit has non-zero static dissipation.**

Ans. When $A=B=V_{DD}$, the voltage at node x is $V_x=V_{DD}-V_{tn}$. This causes static power dissipation at the inverter the pass transistor network is driving.

- c. **Using only just 1 transistor, design a fix so that there will not be any static power dissipation. Explain how you chose the size of the transistor.**

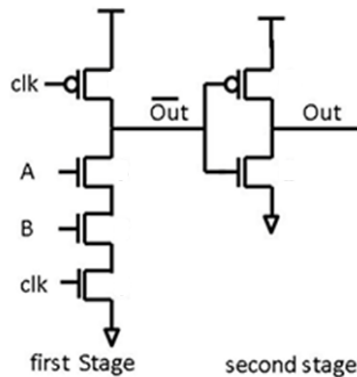
Ans. The modified circuit is shown in the following figure.



The size of M_r should be chosen so that when only one of the inputs, A or B , equals 0, even then one of M_{n1} or M_{n2} can pull down node x to $V_{DD}/2$ or lower.

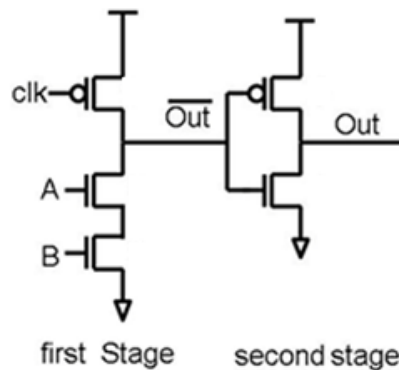
Q3. a) Draw the schematic of a 2-input domino AND gate.

Ans.



b) If one engineer removed the footer nmos from the gate in part (a), how does this change impact the delays compared to the original gate?

Ans.



In this footless 2-input domino gate, the foot NMOS switch connected to clk is eliminated.

Evaluation time: For the same effective pull-down path resistance, we can make NMOS devices smaller. Therefore, we have a smaller logical effort for the first stage. The delay of the evaluation phase will be reduced.

Precharge time: The load of clk pin is also smaller. So, even the precharge delay for a single 2-input domino gate will be reduced.

(c) **Are there any failure modes? Please explain.**

Ans: In the worst-case scenario, if $A=B=1$ and the precharge phase is initiated, then, the precharge current from the PMOS flows through the NMOS pull-down current. To reduce short-circuit current, PMOS size may be reduced. So, this domino gate will become like 'ratioed logic'. If the sizing ratio varies because of PVT variations, then the out_bar node (dynamic node) may not be pulled up high enough to reach the switching threshold and the circuit can fail.

Q.4. Answer the following questions

a) What is sequencing overhead?

Ans. In order to improve the throughput of a system and ensure sequencing, we use pipelining by breaking long combinational logic into smaller paths (T_{pd}). When we do this, sequencing elements like FFs or latches are inserted at appropriate places in the combinational paths. These elements have associated delays (like, propagation delays (T_{pcq}) and setup (T_{setup}) and hold (T_{hold})) that need to be met.

Therefore, to send the data properly to the next stage, the effective clock time period needs to satisfy the following conditions: **$T_{clk} \geq T_{pd} + T_{pcq} + T_{setup}$** . In this constraint, (**$T_{pcq} + T_{setup}$**) delay is called **sequencing overhead** since it adds extra delay to a slow token.

b) What is adaptive sequencing? Explain any 2 adaptive sequencing methods in brief.

Ans. When circuits are operated across a wide range of PVTs, sequencing constraints (like setup/hold) vary significantly. This results in significant performance loss if worst case

constraints are met across all PVTs. To design better circuits, adaptive sequencing methods are used so that we can reduce margins in designs and improve performance.

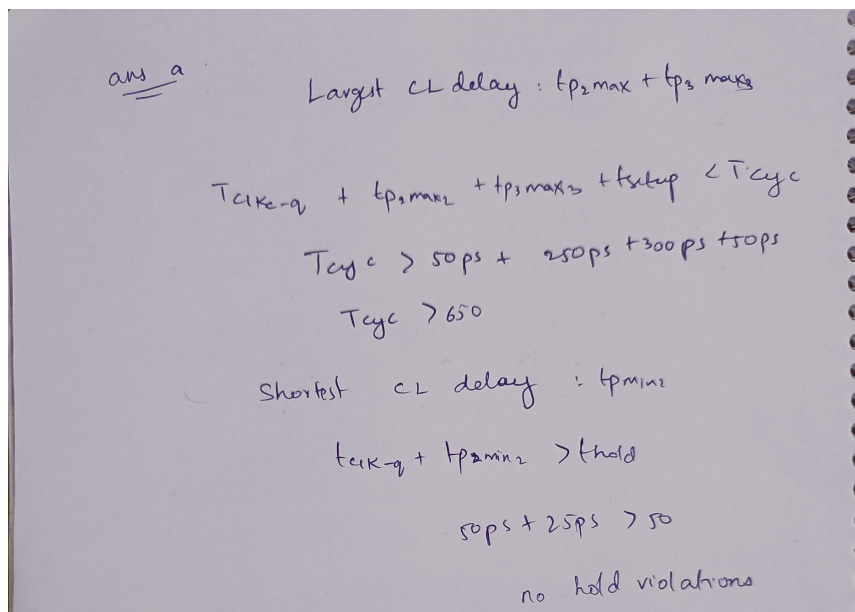
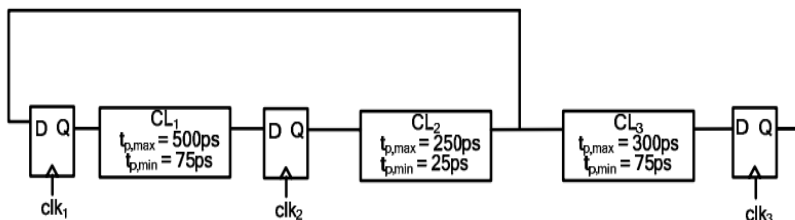
Two common methods that are proposed are:

Razor circuits: Razor circuits typically refer to circuits designed for error detection in processors or other digital systems. These circuits monitor critical signals and detect errors. When a failure is observed, the pipeline is halted, and frequency is reduced (or voltage is raised) to recover from the failure.

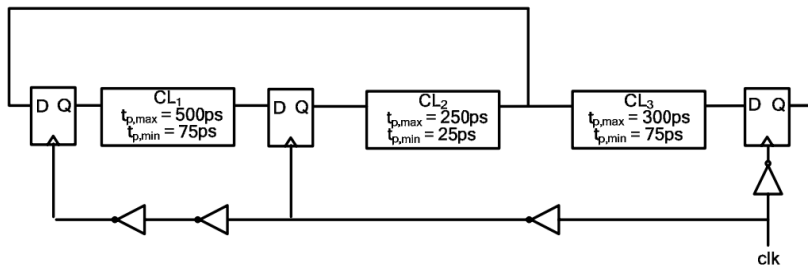
Canary circuits mimic the worst-case path on the chip and use that delay to set the operating frequency. When Canary circuits fail, it indicates that the worst-case path is also susceptible to failure. Voltage is increased, or frequency is lowered to prevent fails.

Q. 5. The minimum and maximum delays through the logic are annotated on the figure, and the flip-flops have the following properties: $t_{\text{clk-q}} = 50\text{ps}$, $t_{\text{setup}} = 50\text{ps}$, and $t_{\text{hold}} = 50\text{ps}$. You can assume that the clock has no jitter.

a) Assuming there is no skew between clk_1 , clk_2 , and clk_3 , what is the minimum clock cycle time for this pipeline? Are there any minimum delay (hold time) violations?



b) Now, we'll include the clock distribution network for this pipeline. Assuming that the delay of each inverter is nominally 50ps, what is the minimum clock cycle time?



In worst case, path through ~~clk~~ must handle an additional delay $\Rightarrow (30+50)ps = 80ps$
 while there is no difference in delay through CL2 and CL3.
 So critical path becomes CL1:
 $T_{cyc} > 50ps + 500ps + 50ps + 80ps.$
 $T_{cyc} > 700ps.$
 $T_{cyc} > T_{era} + T_{comp} + T_{setup} + T_{hold}$

Q6. Answer the following questions [Answer any 4. 5th part would be bonus]

a. Why are non-volatile memories called non-volatile?

Ans. Content is retained even if the power supply is removed.

b. Give two examples of Non-Random Access Memory.

Ans. LIFO, FIFO, SR, CAM [Any 2 should be mentioned]

c. Which type of memory should be preferred if very high speed is the most important requirement?

Ans. SRAM.

d. Why are Flash memories denser than SRAMs? Why are SRAMs better than Flash in terms of endurance?

Ans. In 6T SRAMs, we need six transistors to store one bit of data, while in flash memories, one transistor is used to store a bit of data. In advanced technology nodes, one transistor can even store multiple data bits. This makes the area of a flash cell smaller than that of SRAMs.

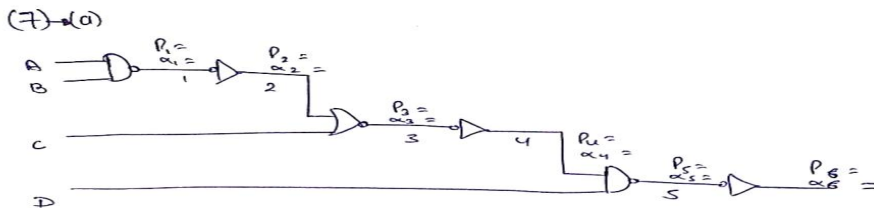
e. High voltages are required to store data in flash memories, which can lead to gate breakdown and other reliability issues, ultimately reducing the number of write cycles for a flash memory cell - thereby limiting its endurance. Since SRAMs don't face any such high voltages, their endurance is better.

f. Why are Design rules for layout allowed to be violated in SRAM array design?

Ans. SRAM occupies nearly 70% of the area in digital SoCs, and any area recovery at the SRAM cell level significantly benefits the chip area and cost overall. Since Memory bit cells are arranged in repetitive patterns, enhanced OPC and RET techniques are applied on arrays to ensure good Yield with denser DRC rules. Hence, many DRCs for example, for contact size, poly extension, minimum metal area, poly-contact spacing, and diffusion spacing, are allowed to be violated in SRAM cells.

Q7. Answer the following questions

- Determine the activity factor at each node in the circuit, assuming the input probabilities $P_a = P_b = P_c = P_d = 0.5$.
- The circuit shown above (part-a) sees a total load capacitance of 10fF and is operated at 0.6V. The time period of the system clock is 2ns. Find the switching power of the circuit.



$$P_a = P_b = P_c = P_d = 1/2$$

- $P_1 = 1 - P_a P_b = 1 - \frac{1}{4} = \frac{3}{4}$
- $\alpha_1 = P_1 \bar{P}_1 = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$
- $P_2 = \bar{P}_1 = \frac{1}{4}$
- $\alpha_2 = P_2 \bar{P}_2 = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$
- $P_3 = \bar{P}_2 P_c = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$
- $\alpha_3 = P_3 \bar{P}_3 = \frac{3}{8} \cdot \frac{5}{8} = \frac{15}{64}$
- $P_4 = \bar{P}_3 = \frac{5}{8}$
- $\alpha_4 = P_4 \bar{P}_4 = \frac{5}{8} \cdot \frac{3}{8} = \frac{15}{64}$
- $P_5 = 1 - P_4 P_d$
 $= 1 - \frac{5}{8} \cdot \frac{1}{2} = \frac{16-5}{16} = \frac{11}{16}$
- $\alpha_5 = \bar{P}_5 P_5 = \frac{11}{16} \cdot \frac{5}{16} = \frac{55}{256}$

$$P_6 = \bar{P}_5 = 5/16$$

$$\alpha_6 = P_6 \bar{P}_6 = \frac{5}{16} \times \frac{11}{16}$$

$$= 55/256$$

$$(b) P_{\text{switching}} = \alpha C V_{dd}^2 f$$

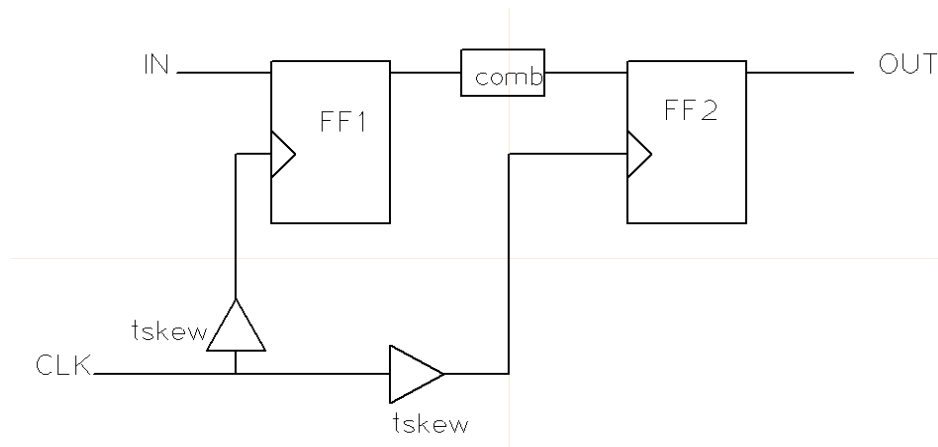
$$= \frac{55}{256} \times 10 \times 10^{-15} \times (0.6)^2 \times \frac{1}{2 \times 10^{-9}}$$

$$= \frac{55}{256} \times 10^{-14} \times \frac{36}{100} \times 2 \times 10^9$$

$$= 15.468 \times 10^{-7} \text{ W}$$

$$[P_{\text{switching}} = 1.546 \mu\text{W}]$$

Q8 .a i) Calculate the maximum operation frequency of the following sequential circuit.



$$T_{pd} = 1\text{ns}$$

$$T_{hold} = 150\text{ps}$$

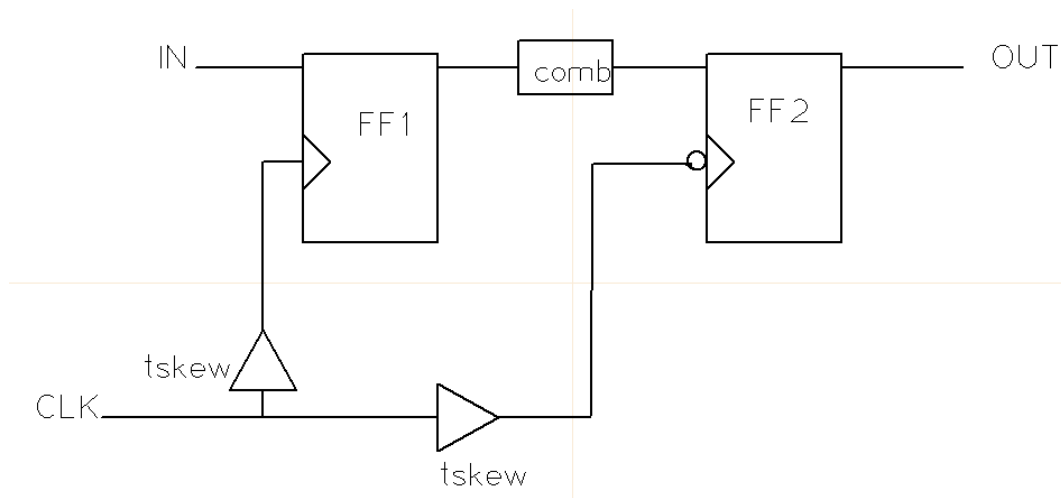
$$T_{setup} = 200\text{ps}$$

$$T_{cq} = 300\text{ps}$$

$$T_{skew} = 100\text{ps}$$

ii) Calculate the setup slack if the clock period is 2ns.

b) If the 2nd FF is changed with a negative edge-triggered FF, Calculate the maximum operating frequency in this case, keeping the timing parameters the same. **[Bonus for B.Tech students]**



Sol:

Solution

a>

(i) $T_c \geq t_{pd} + t_{cq} + t_{setup} + (t_{skew} - t_{skew})$

$T_c \geq 1000ps + 300ps + 200ps + 0$

$T_c \geq 1500ps$

$T_{min} = 1.5ns$

$f_{max} = \frac{1}{1.5ns} = 0.66666 MHz$

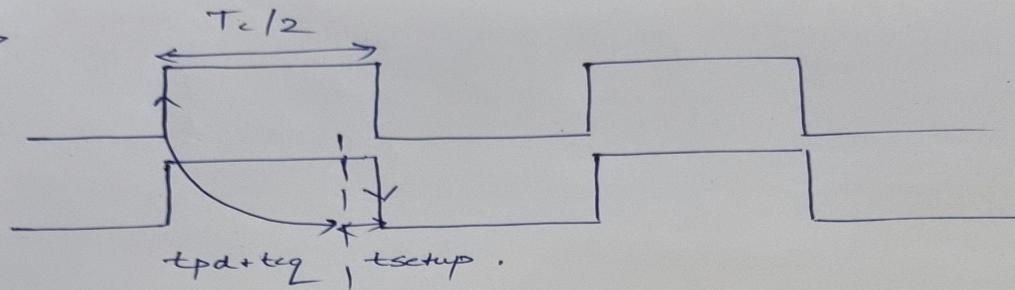
(ii) Setup slack = R.T - AT ; where $T_c = 2ns$

$= (2ns - t_{setup}) - (t_{cq} + t_{pd} + t_{skew})$

$= (2000ps - 200) - (300ps + 1000ps)$

$= 500ps$

b>



$$\frac{T_c}{2} \geq t_{pd} + t_{setup} + t_{eq}$$

$$\frac{T_c}{2} \geq 2000ps + 200ps + 300ps$$

$$\frac{T_c}{2} \geq 1500ps$$

$$T_c \geq 3000ps$$

$$T_c \geq 3ns$$

$$T_{cmin} = 3ns$$

$$f_{max} = \frac{1}{3ns} = 333.33MHz$$