Question 1. What is the maximum frequency you are able to achieve in both the technology nodes? (15 points)

The strategy adopted was:

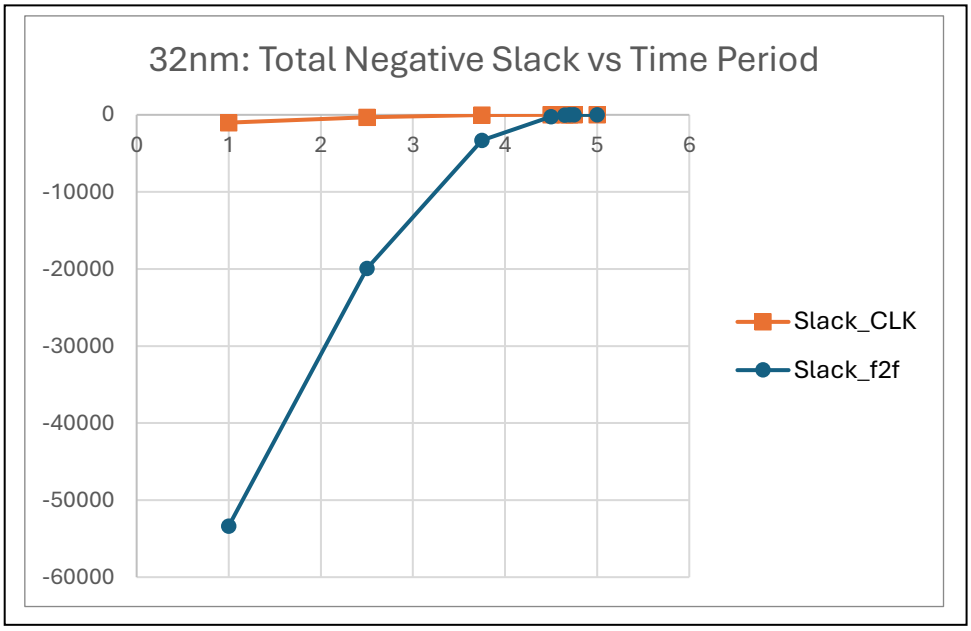
- Binary search – started from an arbitrary search space for time period from 1ns to 10ns for the case of 32nm.
- To determine the max frequency, the TOTAL NEGATIVE SLACK parameter was taken into account:
    - Hold violations is ignored as it can be fixed by adding buffers during the PnR stage.
    - However, the setup constraint is to be satisfied during the synthesis and hence, the parameter concerned at this stage is SLACK.

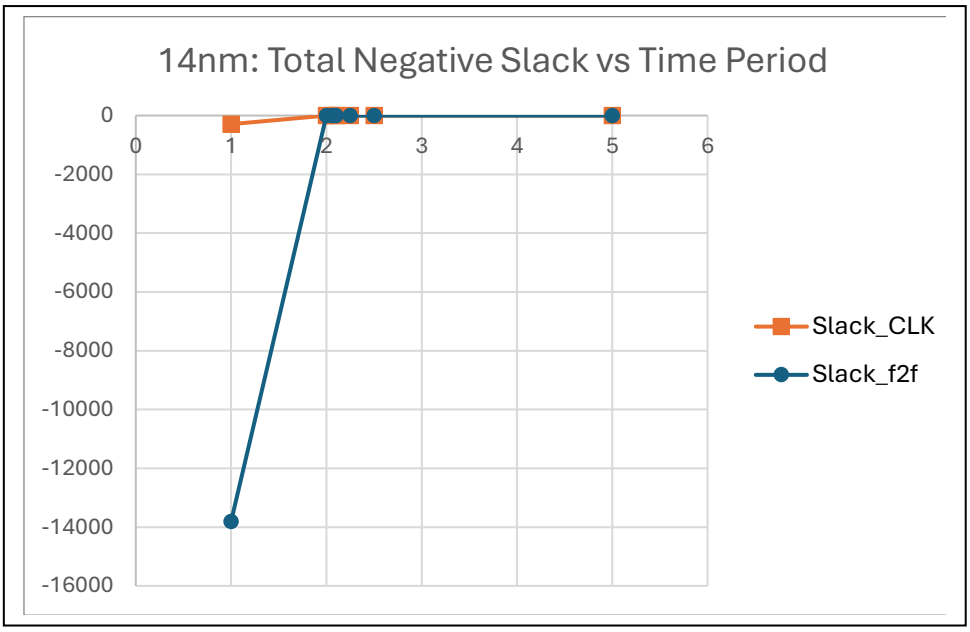
Max frequency Achievable:

Tech Node	Time Period (ns)	Frequency (MHz)
32 nm	4.7	212.77
14 nm	2.06	485.43

As expected from scaling theory, the **time period has nearly halved** upon scaling to a node of ~0.5x;

Total Negative Slack Trends analysed:





Question 2. What is the area of the c-class core in 32 nm and 14 nm? (5 points)

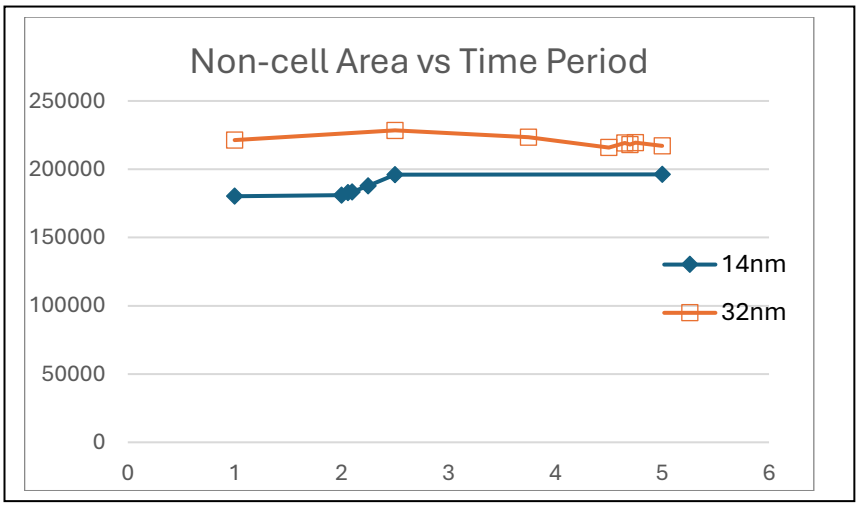
Following observations are noted regarding area of c-class core:

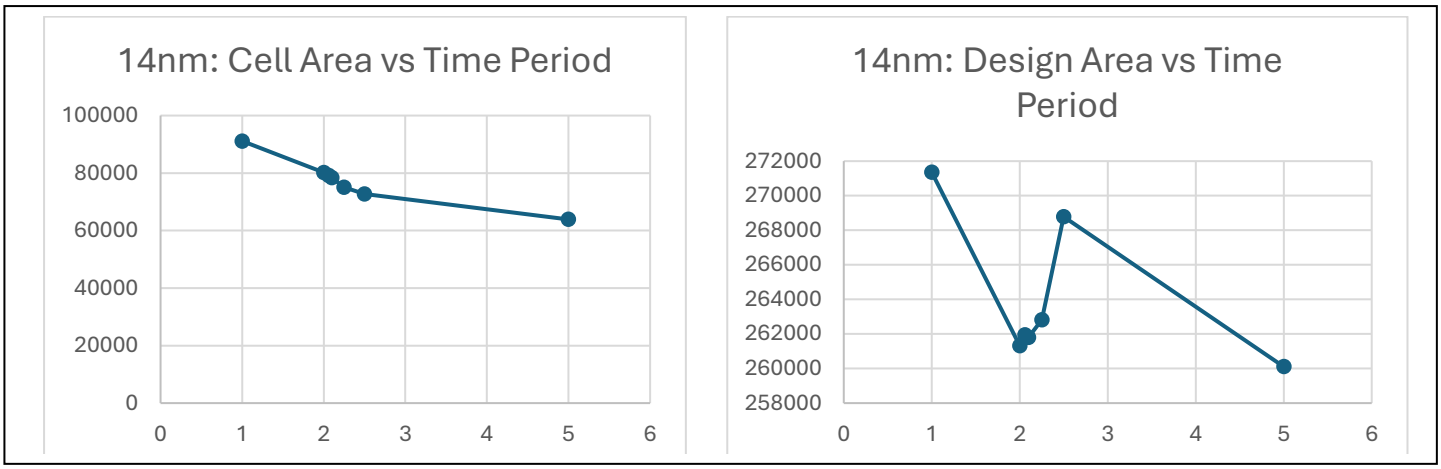
- Within each tech node, the cell area across frequencies in the search space has a standard deviation of 1.5% of the mean – largely independent of the frequency.
- Across all tech nodes, the **non-cell area remains nearly constant.**

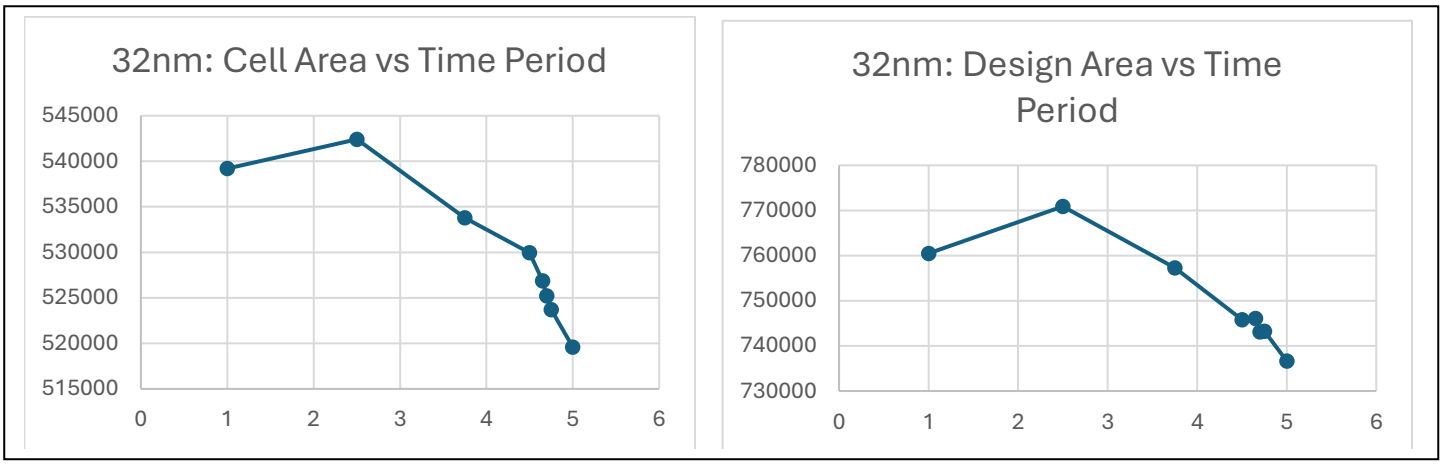
Area for the highest achievable frequency:

Tech Node	Time Period (ns)	Frequency (MHz)	Cell Area (μm^2)	Design Area (μm^2)	Non – Cell Area (μm^2)
32 nm	4.7	212.77	525196.7146	743144.0653	217947.3507
14 nm	2.06	485.43	79139.09227	261940.9745	182801.8823

The distribution across frequencies analysed:







Question 3. What is the relationship between the maximum frequency above to the length of the critical path (the path with the longest delay) in the design? Which units does the critical path pass through? What can you do to improve the maximum frequency? (25 points)

Critical Path length in 14 nm = 1.05 ns

Critical Path length in 32 nm = 3.61 ns

Relationship between the maximum frequency:

- STA engines are required to model:
  - Load dependent slew in setup time. This is usually done using linear extrapolation of an LUT which is defined in the library.
  - Uncertainty in clock. Ideal clock is taken for calculating the propagation delay, hence an uncertainty term is added to the propagation delay that is obtained
- Expression for clock period:

$$T_{clock} = T_{propagation\text{-}delay} + T_{clock\_uncertainty} + T_{lib\_setup\_time}$$

$$T_{propagation\text{-}delay} = T_{clk\text{-}q} + T_{combinational\_delay} + T_{setup\_ideal}$$

- This propagation delay is reported by the tool as “Critical Path Length”.

- Hence, it is the final $T_{propagation\text{-}delay}$ that is being used as RAT for the slack.

To increase the operating frequency, one of the following strategies can be adopted:

- The f2f path synthesized is a very long combinational circuit. The pipeline stages can be increased. More generally, the circuit can be retimed to change the logical depth and of each stage.
- Based on available area budget, the devices can be selectively sized up to have a higher driving strength. However, this cannot be done mindlessly as gate sizing for delay optimization is a convex problem.

Path Traversed by one of the critical paths of 14nm node:

Order	Instance	Cell Type
0	Launch Flop	inst_dpfm_add_sub_ff_stage5_reg_0_
1	U21916	EN2_4
2	U3261	BUF_12
3	U10691	NR2_MM_10
4	U16000	INV_12
5	U28867	INV_S_16
6	U29115	ND2_CDC_2
7	U2859	OAI21_0P5
8	U29116	AOI21_1P5
9	U29117	MUX2_2
10	U2585	OAI21_3
11	U29274	NR2_MM_3
12	U29275	NR2_MM_4
13	U29276	AOI21_3
14	U31079	INV_S_1
15	U3861	ND2_CDC_2
16	U31080	OAI21_0P5

17	U3850	NR2_MM_3
18	U5677	ND2_CDC_4
19	U17786	ND2_4
20	U3845	NR2_MM_6
21	U23193	ND2_5
22	U3864	INV_6
23	U22882	ND2_8
24	U36629	INV_S_4
25	U37924	ND2_CDC_1
26	U37925	NR2_1
27	U37926	EN2_1
28	U37927	INV_S_0P5
29	U23060	INV_S_0P5
30	U16508	ND2_1
31	U16487	OAI21_V1_4
32	U21656	NR2_MM_4
33	U21655	INV_S_4
34	U16402	INV_6
35	U21211	AN4_8
36	U21209	ND2_8
37	U357	BUF_12
38	U613	AO21_4
39	U362	ND2_MM_10
40	U498	INV_S_4
41	U354	OR2_4
42	U323	BUF_10
END	Capture Flop	arr_reg_6__36_

Likewise, all the stages of pipeline are operating at a similar arrival time.

This happens due to the down-sizing of gates performed by the tool to get rid of excess slack.

Critical Path Length in case of 32nm: 3.61ns

Order	Instance	Cell Type
0	LAUNCH FLOP	inst_dpfpu_divider_int_div_rg_state_reg_1_
1	U12808	OR2X1_LVT
2	U12808/Y	OR2X1_LVT
3	U5840	OR2X1_LVT
4	U5840/Y	OR2X1_LVT
5	U2879	INVX2_LVT
6	U2879/Y	INVX2_LVT
7	U12572	OR2X2_LVT
8	U12572/Y	OR2X2_LVT
9	U5773	INVX8_LVT
10	U5773/Y	INVX8_LVT
11	U8380	INVX4_LVT
12	U8380/Y	INVX4_LVT
13	U2770	INVX8_LVT
14	U2770/Y	INVX8_LVT
15	U6293	AO22X1_LVT
16	U6293/Y	AO22X1_LVT

17	U5282	INVX1_LVT
18	U5282/Y	INVX1_LVT
19	U9013	AND2X1_LVT
20	U9013/Y	AND2X1_LVT
21	U8383	INVX0_LVT
22	U8383/Y	INVX0_LVT
23	U17135	AOI21X1_LVT
24	U17135/Y	AOI21X1_LVT
25	U9212	OA21X1_LVT
26	U9212/Y	OA21X1_LVT
27	U13808	OA21X1_LVT
28	U13808/Y	OA21X1_LVT
29	U679	INVX1_LVT
30	U679/Y	INVX1_LVT
31	U7188	NAND2X0_LVT
32	U7188/Y	NAND2X0_LVT
33	U7186	NAND3X0_LVT
34	U7186/Y	NAND3X0_LVT
35	U11695	AO21X1_LVT
36	U11695/Y	AO21X1_LVT
37	U10235	INVX4_LVT
38	U10235/Y	INVX4_LVT
39	U10385	OA21X1_LVT
40	U10385/Y	OA21X1_LVT
41	U6322	XNOR2X2_LVT
42	U6322/Y	XNOR2X2_LVT
43	U6320	OA22X1_LVT

44	U6320/Y	OA22X1_LVT
45	U546	INVX1_LVT
46	U546/Y	INVX1_LVT
47	U13117	AND2X1_LVT
48	U13117/Y	AND2X1_LVT
49	U2157	INVX0_LVT
50	U2157/Y	INVX0_LVT
51	U12911	OA21X1_LVT
52	U12911/Y	OA21X1_LVT
53	U2087	INVX0_LVT
54	U2087/Y	INVX0_LVT
55	U12910	AOI21X1_LVT
56	U12910/Y	AOI21X1_LVT
57	U12909	OA21X1_LVT
58	U12909/Y	OA21X1_LVT
59	U12208	AND2X1_LVT
60	U12208/Y	AND2X1_LVT
61	U14318	NBUFFX2_LVT
62	U14318/Y	NBUFFX2_LVT
63	U1709	OA21X1_LVT
64	U1709/Y	OA21X1_LVT
65	U3428	INVX0_LVT
66	U3428/Y	INVX0_LVT
67	U10334	XNOR2X1_LVT
68	U10334/Y	XNOR2X1_LVT
69	U7065	OAI22X1_LVT
70	U7065/Y	OAI22X1_LVT

71	U1225	NOR2X2_LVT
72	U1225/Y	NOR2X2_LVT
73	U10234	NOR2X0_LVT
74	U10234/Y	NOR2X0_LVT
75	U9812	AND2X1_LVT
76	U9812/Y	AND2X1_LVT
77	U10325	AO21X1_LVT
78	U10325/Y	AO21X1_LVT
79	U14286	AOI21X1_LVT
80	U14286/Y	AOI21X1_LVT
81	U10791	OA21X1_LVT
82	U10791/Y	OA21X1_LVT
83	U10790	XNOR2X2_LVT
84	U10790/Y	XNOR2X2_LVT
85	U10789	OA21X1_LVT
86	U10789/Y	OA21X1_LVT
87	U1790	INVX1_LVT
88	IF_IF_inst...d1396[107] (net)	(net)
89	inst_dpfpu_divider_int_div_rg_inter_stage_reg_107_/D	(capture FF D pin)

- In the case of 14nm, it can also be noticed that on decreasing frequency much less than highest frequency,
    - In the **fetch stage**, a latch to clock gate path, which goes to the **BPU** (Branch Prediction Unit) seems to be the **critical path**.

Question 4. What is the total power consumption of the design operating at the maximum frequency in both nodes? What is the average power density? (5 points)

Node	Time Period	Switch Power	Internal Power	Leakage Power	Total Power	Leakage Power Density (W/mm^2)	Peak Power Density (W/mm^2)
14	2.06ns	26.021mW	4.152mW	6.84E+07pW	30.241mW	2.61E-04	0.115449674
32	4.7ns	497.861µW	3.85E+03µW	1.13E+10pW	1.56E+04µW	1.52E-02	0.02099189205

Peak power is attained only for a fraction of the clock period – Hence it can be modelled as a sum of static power and dynamic power. Leakage power is considered static power since it is present for majority of the on-time of the chip.

Question 5. Using your knowledge of leakage-limited scaling regime, can you estimate the area, maximum frequency, power consumption and the power density of this design at the 2 nm node? (20 points)

Using the post Dennard scaling regime:

If the process technology nodes scales as **s**, the following can be estimated:

		By Extrapolation
Area	S^2	*Note: Core area (cell area) alone scales* 200000 + (50000*2^2/32^2) = 201953 µm^2
Maximum Frequency	1/S	667 MHz
Power Consumption	1	20 – 30 mW
Power Density	$1/S^2$ or 1	*Note: Non-cell area doesn't scale* ~ 0.1 W/mm^2 [peak power – not the static power]

Explanation of the scaling terms:

- Note, while scaling the underlying circuit logic is kept constant. Hence the number of devices is nearly constant.
- Core Area is therefore scaled as the square of node (S^2), as the device length and width will scale linearly

    - As mentioned in one of the previous questions, the non-cell area remains nearly constant. This becomes a major contributor to the 14nm node. Hence, it will continue to be the major contributor to the 2nm node.
- Capacitance scales as S.
- The transistor switching frequency also scales as 1/S
- In the leakage-limited regime, the threshold V cannot be scaled down any further. Hence, the operating voltage also remains constant
- Dynamic Power, the major contributor to peak power:
    - Dynamic power = $\alpha * C_{eff} * V^2 * f$
    - Here, **α,** the switching factor is nearly constant as the underlying circuit is constant
    - C & f compensate for each other, due to the inverse effect of both.
    - V remains constant upon scaling
    - Therefore, the dynamic power remains nearly constant upon scaling
- Dynamic Power scaling can also be analysed as:
    - Power = Net Saturation Current (I_{sat})*V
    - Saturation current remains nearly constant (From the I_{DS} equation)
    - Voltage also remains constant
    - Hence, Power can be concluded to remain constant upon scaling
- Power Density:
    - Power density, while seems to scale as $1/S^2$, also depends on the non-cell area. Hence, between 14nm and 2nm: the power density also remains nearly constant. However, if it is calculated for the core-cell area alone, the power density scales as $1/S^2$
    - For the table, the power density is taken as constant accordingly.

Question 6. Assume that the wafer fabrication costs at 2 nm are 5X that at 14 nm. Let us say you are the architect tasked to improve the performance of the single-core 14 nm design by going to multiple cores at 2 nm. Assume the applications to be run on this chip are embarrassingly parallel (communication overheads are low). Assuming constant total chip price, power and cooling budgets, how many cores will your design have? How many of them will be operating at the max frequency? Why? (30 points)

- I am assuming that the fabrication costs at 2nm is increasing by 5x per unit area, since the question mentions fabrication of an entire wafer.
- Therefore, the chip area can at max be $1/5^{th}$ of the original area at 14nm.

NOTE: As per the description given in the answer to previous questions, the non-cell area remains nearly fixed. Therefore, about $4/5^{th}$ of the area of the chip is required to be preserved inspite of scaling. However, that will not allow to meet the chip's price budget in any way. Hence, it might not be possible to make any core at all in the given price budget. However, for answering the other aspects of the question, I will consider only the CELL AREA as the area affected by the tech node's price.

- Cell area scales as S^2 with the tech node. Hence the cell area per core is 7^2 times smaller in 2nm node.
- Therefore, **to meet the area and price budgets**, it would be possible to accommodate floor(49/5) = **9 cores** in the chip.

This does not consider the power/cooling budgets.

- If each of the cores are operating at the max frequency (677 Mhz), the power budget shoots over the roof.
- Hence, the chip can have only 1 core running at the max frequency, to continue to meet the power budget.

To meet the cooling budget:

- If the cell area alone is taken into account, then the total power density scales as $1/S^2$
    - This allows only for 0.2 core to be running at max frequency
- However, when including the non-cell area as it provides extra surface for dissipation of heat, this keeps a near constant power density upon scaling.
    - Therefore, by meeting the power budget, the cooling budget is also automatically met.

To allow each of the 9 cores to operate simultaneously, the operating frequency of each core must be reduced to $1/9^{th}$ of the original frequency at 2nm = $7/9^{th}$ of the frequency of 14nm chip.

However, this frequency is only for the operation of individual cores which is at 2nm. Further circuitry can be configured to enable UP-FOLDING of the architecture and ensuring the over-all throughput to be identical to that of 14nm.

Data obtained for frequency:

Node	Time Period	Slack_CLK	Slack_f2f
32	5	0	0
32	1	-1010.74	-53357.27
32	2.5	-333.15	-19936.53
32	3.75	-17.3	-3312.5
32	4.5	-0.01	-215.76
32	4.75	0	0
32	4.65	0	-28.32
32	4.7	0	0
14	5	0	0
14	1	-292.67	-13808.66
14	2.5	0	0
14	2	-2.12	0
14	2.25	0	0
14	2.1	0	0
14	2.06	0	0