International Institute of Information Technology Bangalore

A Fully Digital SRAM-Based Four-Layer In-Memory Computing Unit Achieving Multiplication Operations and Results Store

April 11, 2024

1. Divyansh Singhal-IMT2021522
2. Daksh Sharma-IMT2021533
3. Chinmay Sultania-IMT2021540
4. Yash Gupta-IMT2021514

CONTENTS

OBJECTIVE

- The objective of the paper is to address the limitations of the von Neumann computing architecture, which separates memory and arithmetic logic units, leading to inefficiencies in big data and high-performance computing.
- The paper introduces a fully digital static random access memory (SRAM)-based In-Memory Computing (IMC) architecture that simplifies multiplication operations, improves computational efficiency, and provides good scalability.
- This new architecture aims to significantly reduce latency and power consumption in data processing.

INTRODUCTION TO SRAM

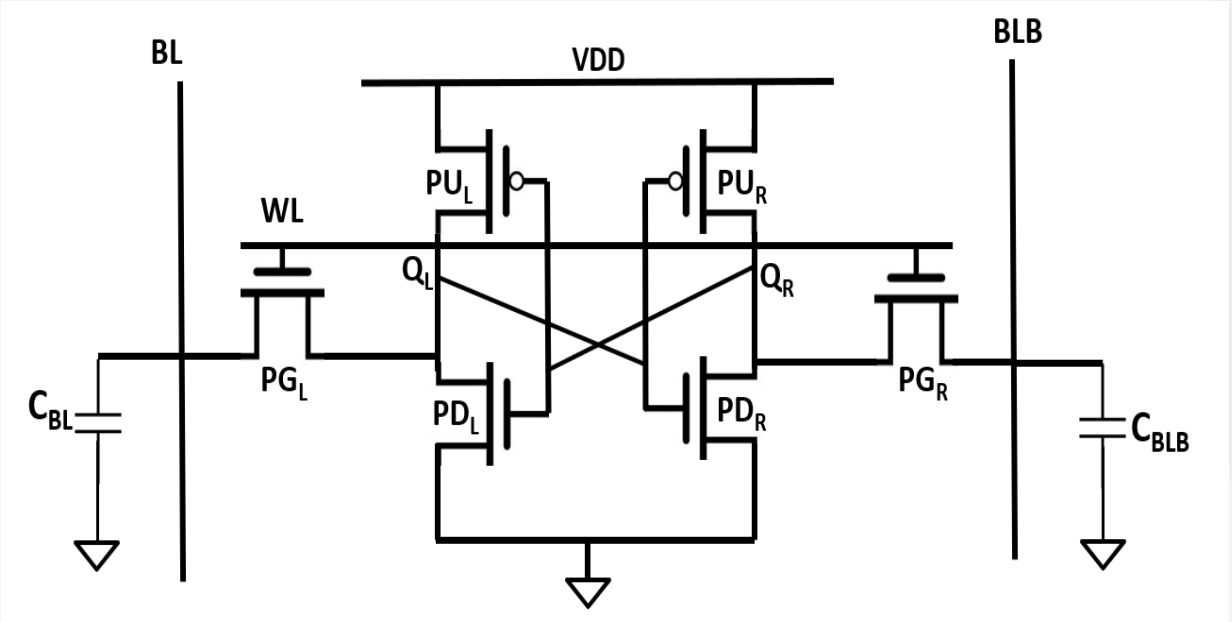
Static random-access memory (static RAM or SRAM) is a type of **random-access memory (RAM)** that uses **latching circuitry (bistable latch)** to store each bit. SRAM will **hold its data permanently in the presence of power**, but data is lost as soon as power is removed.

SRAM is a popular choice for microcontrollers when **fast and efficient operations with low power** consumption are needed. In computer systems, it is commonly used as the **primary working memory,** including the internal registers and cache of a CPU.

6T SRAM CELL – READ OPERATION

1. Pre charge BL and BLB to HIGH.
2. Turn on WL
3. BL or BLB pulled down to 0 based on Q_L and Q_R. So, for eg – if Q_L=0, Q_R=1, BL discharges via PG_L , PD_L, GND; and; BLB stays HIGH but Q_L bumps up slightly.

NOTE – In order to prevent Q from changing its value, **strength of $PD_L > PG_L$**, that is, ON resistance of $PD_L < PG_L$.

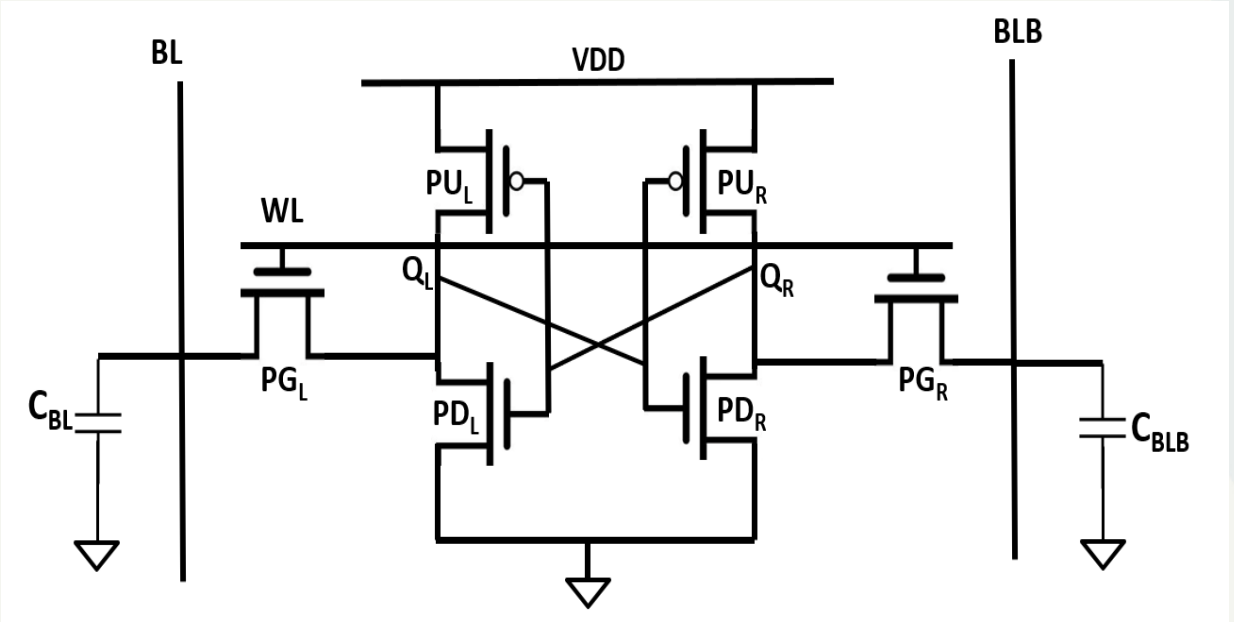


6T SRAM CELL – WRITE OPERATION

1. Drive BL and BLB to necessary values.
2. Turn on WL.
3. BL or BLB will drive Q_L and Q_R with new values. So, for eg – if Q_L=0, Q_R=1, BL =1 and BLB =0, This would force Q_R to low and Q_R to high.

NOTE – Strength of PGL > PUL, that is, ON resistance of $PG_L < PU_L$.

Hence strength of $PD_L > PG_L >> PU_L$.



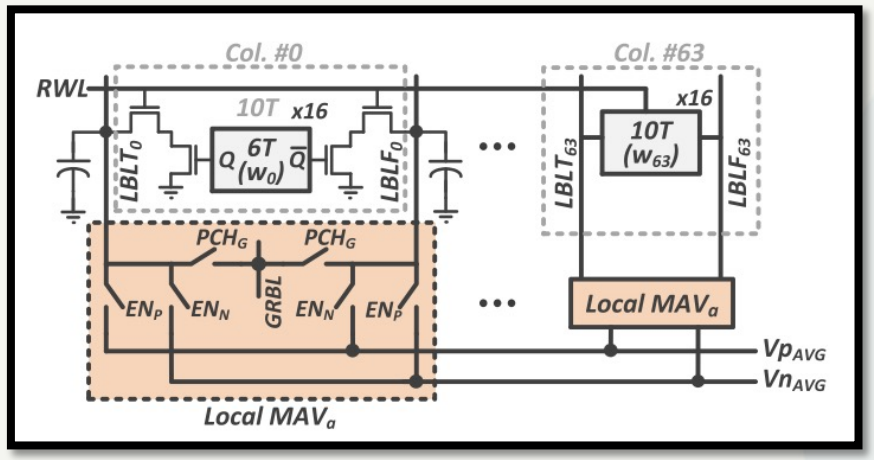
MOTIVATION

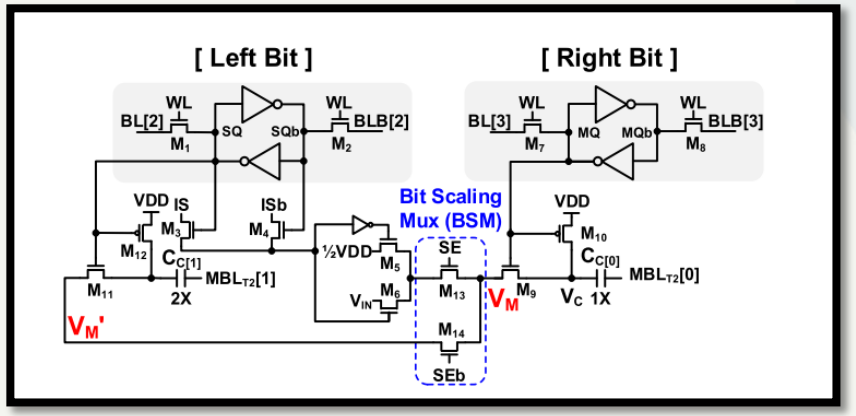
The motivation is to address the limitations of the von Neumann computing architecture, particularly the "memory wall" issue, which hinders the development of big data and high-performance computing. The paper introduces a fully digital SRAM-based In-Memory Computing (IMC) architecture that aims to significantly reduce latency and power consumption during data processing, thereby improving computational efficiency and offering a scalable solution for different bit-width multiplication operations. This innovation has the potential to advance the field of IMC and contribute to the progress of technologies like artificial intelligence, biological systems, and neural networks.

This paper's work aims to provide a design that would improve efficiency of MAC operation using SRAM based IMC (In-Memory Computing).

PREVIOUS WORKS 1 – Analog IMC

Disadvantage - Analog IMC may not be suitable for high-precision applications because it has the disadvantage of low conversion accuracy limited by the low-cost analog-to-digital converters (ADCs), while digital IMC has the advantage of high computational accuracy.

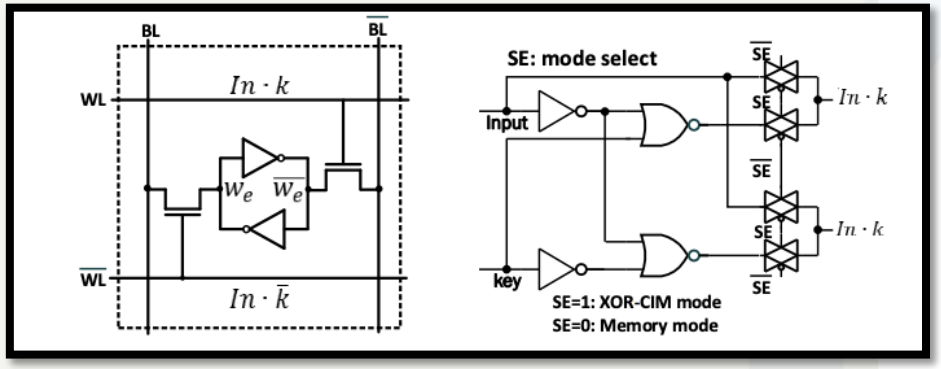


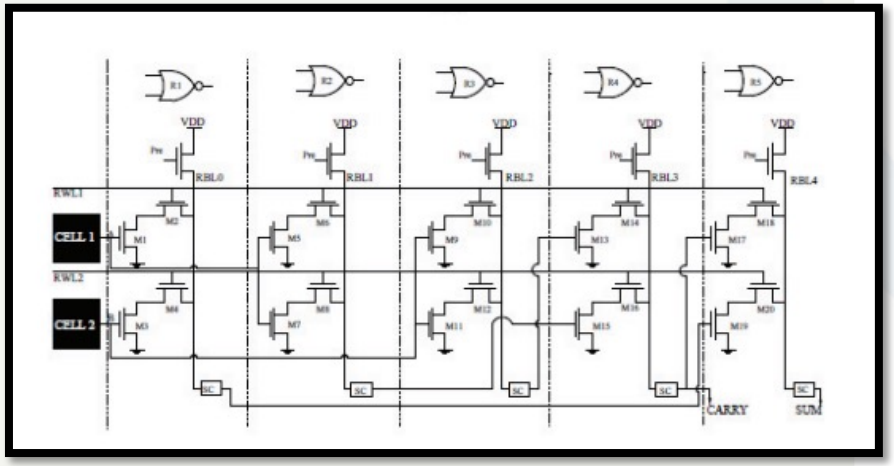


PREVIOUS WORKS 2 – Digital IMC-Approach 1

Integrate basic logic operations such as NAND, AND, OR, and NOR around the memory array .

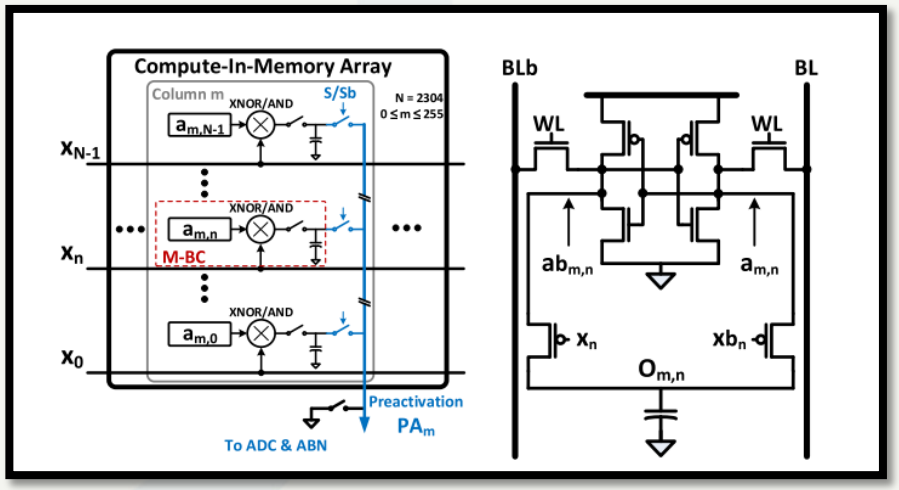
Disadvantage - However, because the bitline discharge is susceptible to process, voltage, and temperature (PVT), the accuracy of logic operations is reduced, limiting the benefits of IMC.

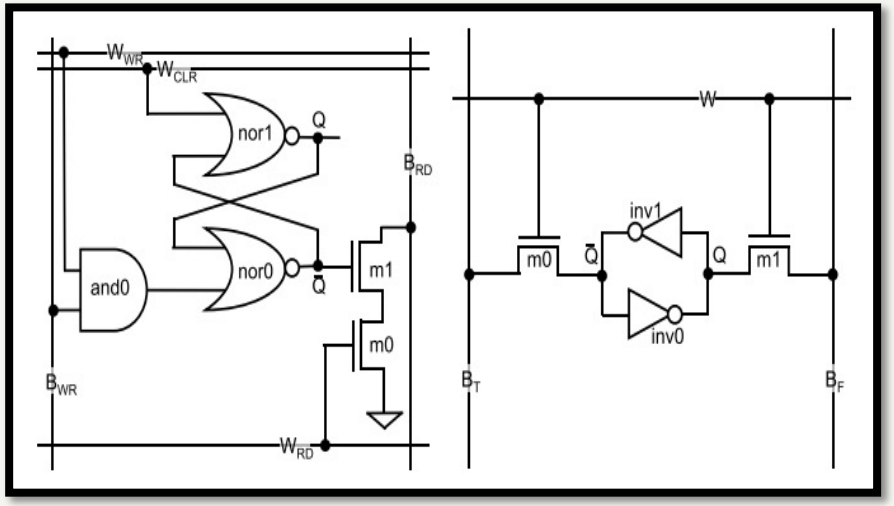




PREVIOUS WORKS 3 – Digital IMC-Approach 2

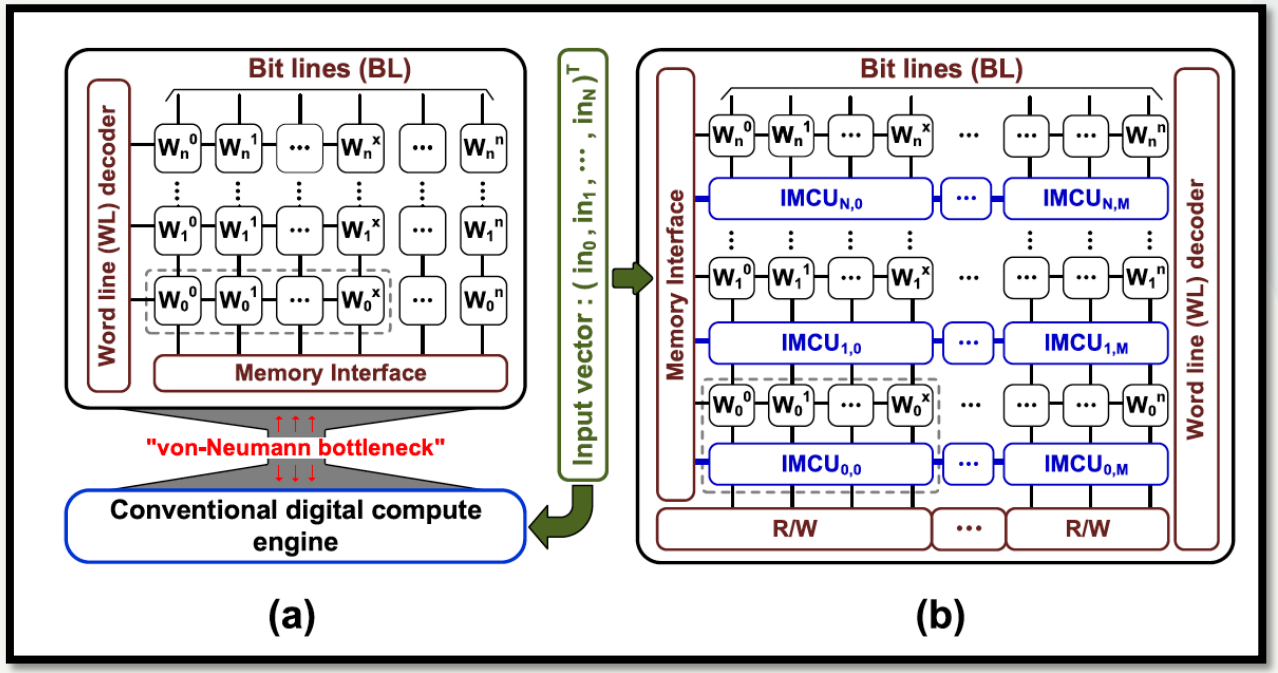
The other approach is to closely combine logic cells with static random access memory (SRAM) cells.



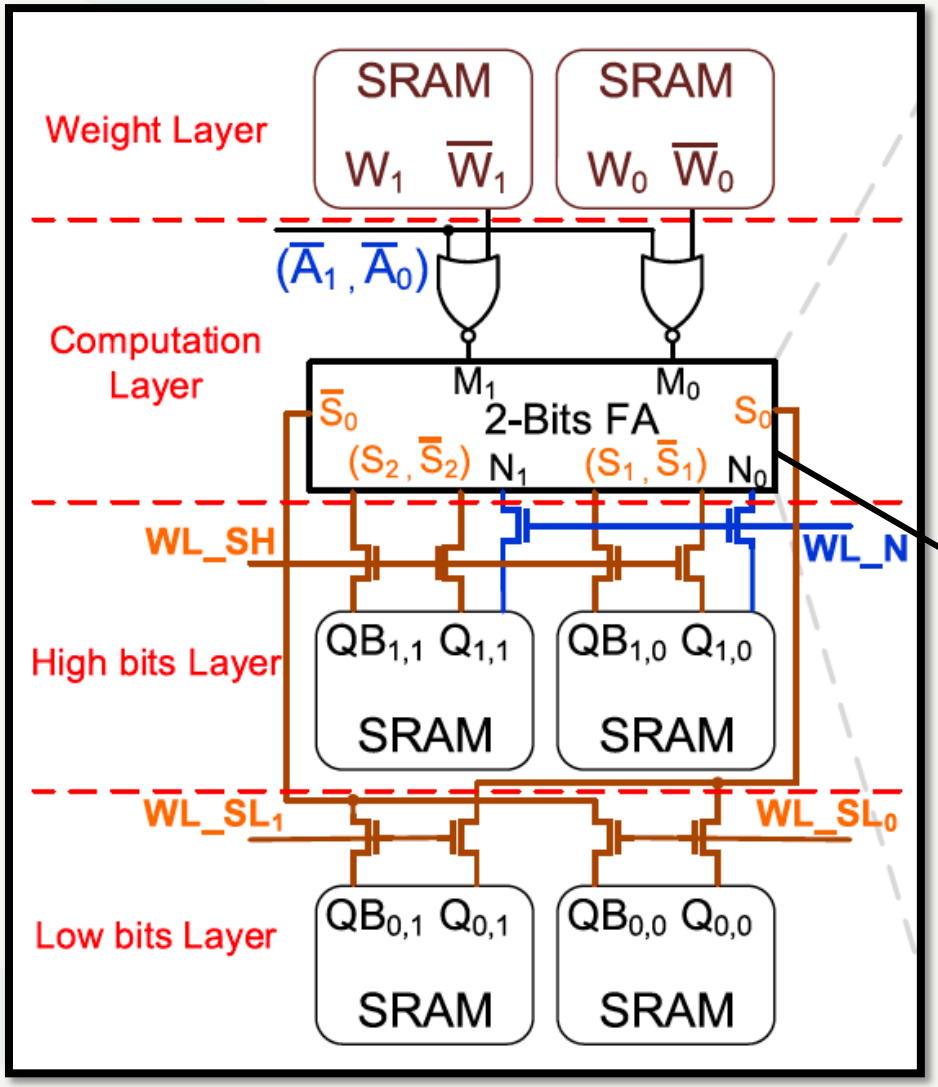


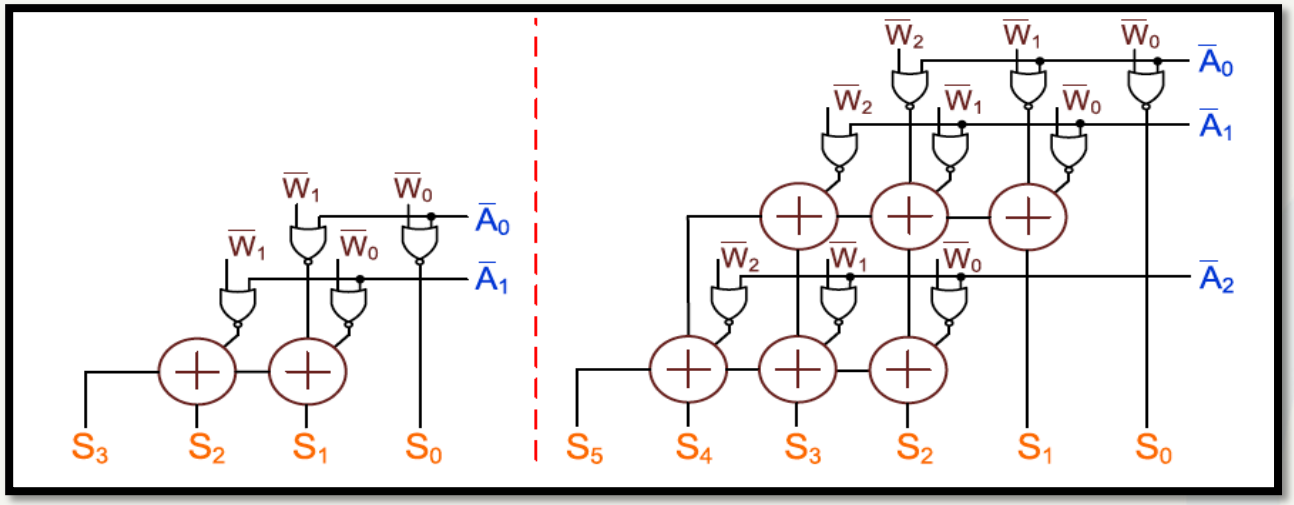
PROPOSED WORK

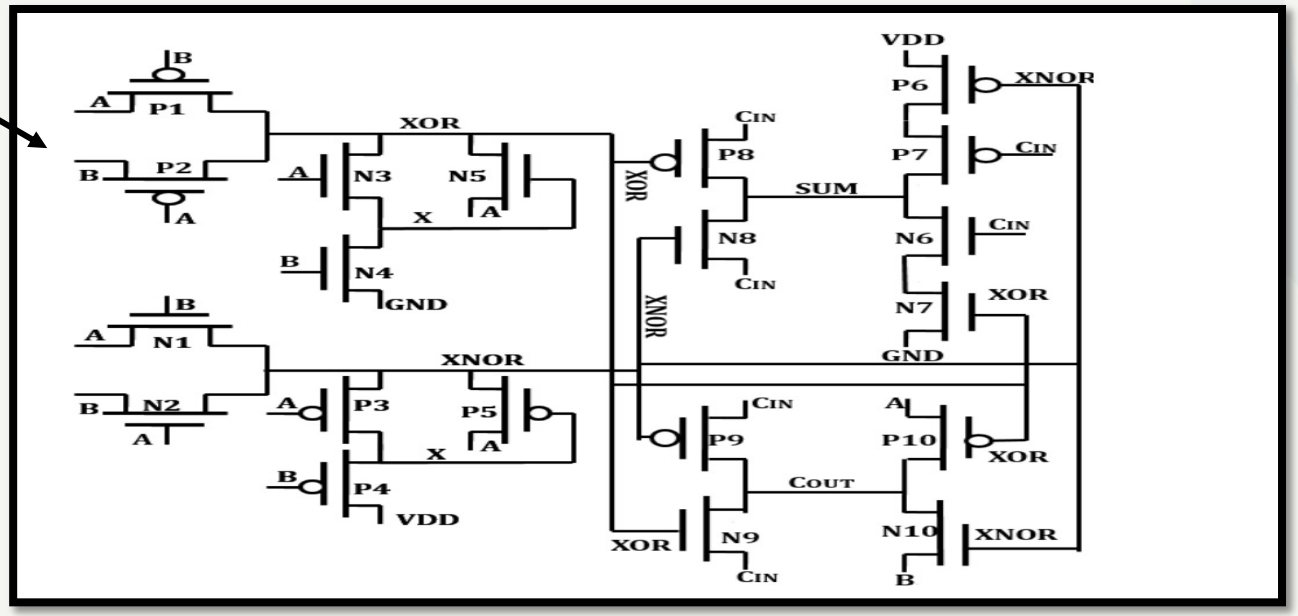
- To reduce computational complexity and optimize the arithmetic operation while improving memory cell use at the same time, the author propose a fully digital IMC that can reuse memory space.



PROPOSED WORK – COMPLETE CIRCUIT







PROPOSED WORK

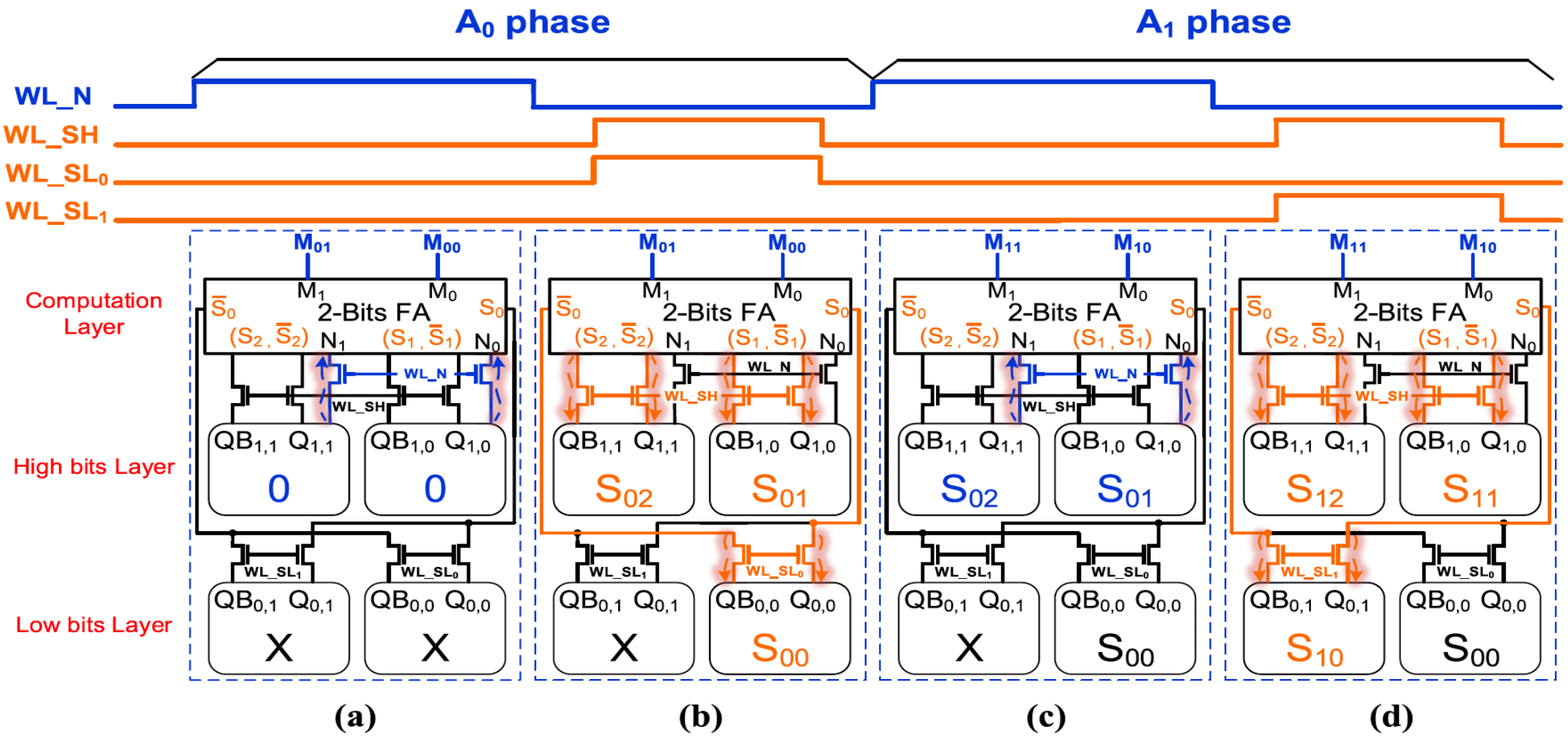
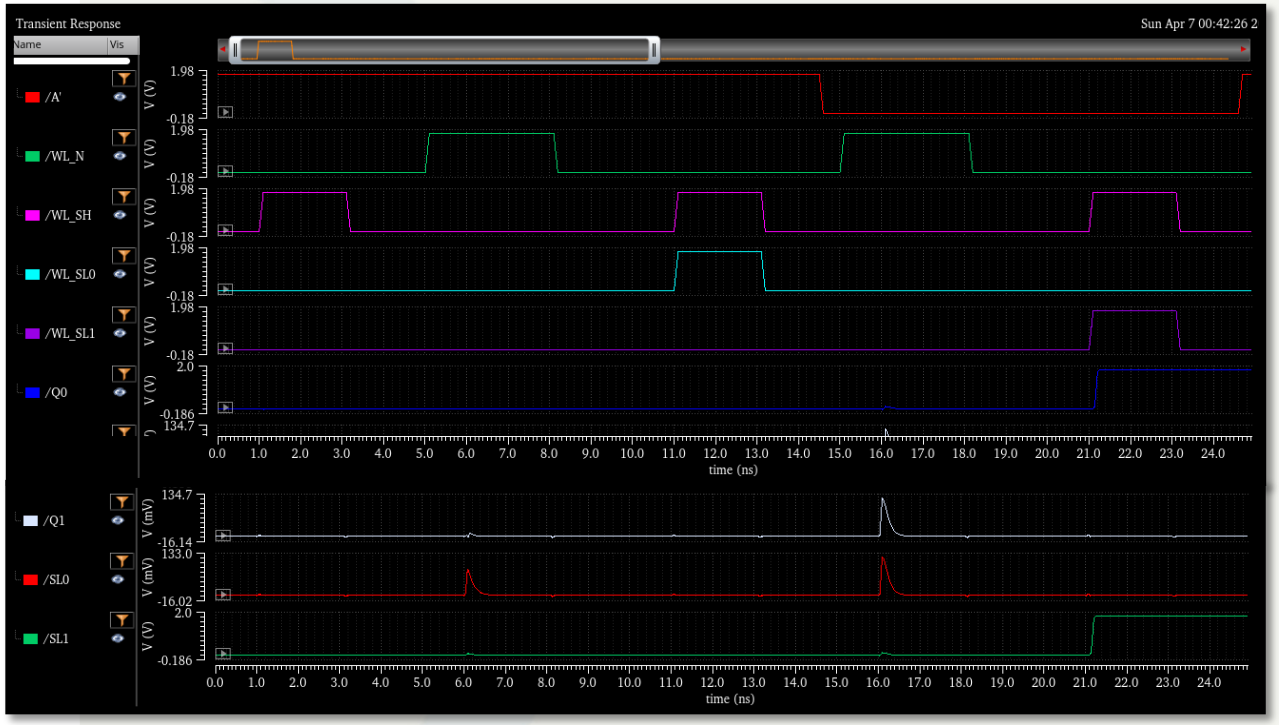
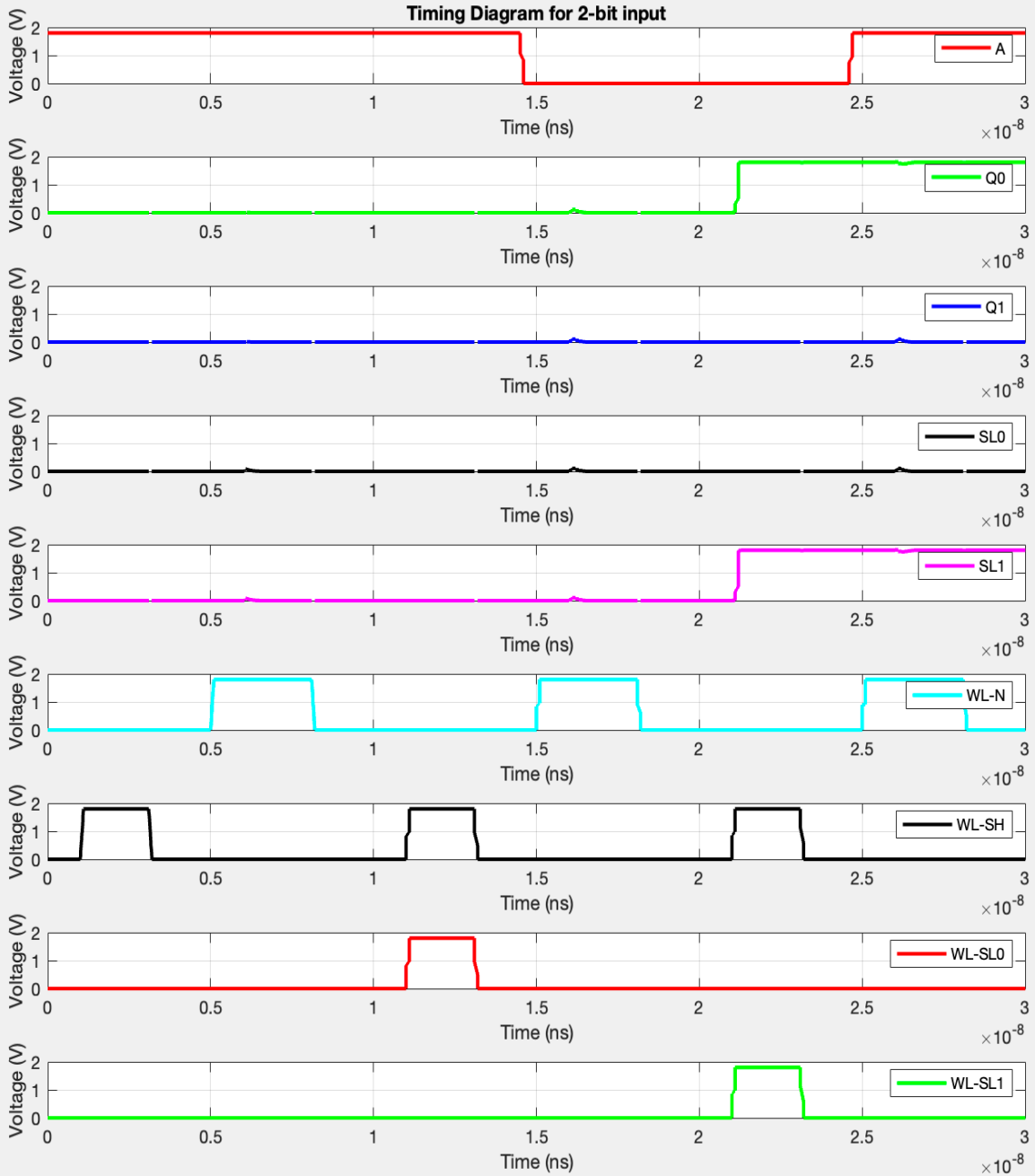


Fig. 5. (a) Calculation in A_0 phase: WL_N = 1, and the input of the adder is provided by the high bits layer. (b) Write-back in A_0 phase: signal WL_SH = 1, $WL\_SL_0 = 1$, and the outputs of adder are written to the high and low bits layers. (c) Calculation in A_1 phase: signal WL_N = 1, data are provided by the high bits layer to the input of the adder. (d) Write-back in A_1 phase: signals WL_SH = 1 and $WL\_SL_1 = 1$, and the outputs of adder are written to the high and low bits layers.

PROPOSED WORK





International Institute of Information Technology Bangalore

PROPOSED WORK – Advantages

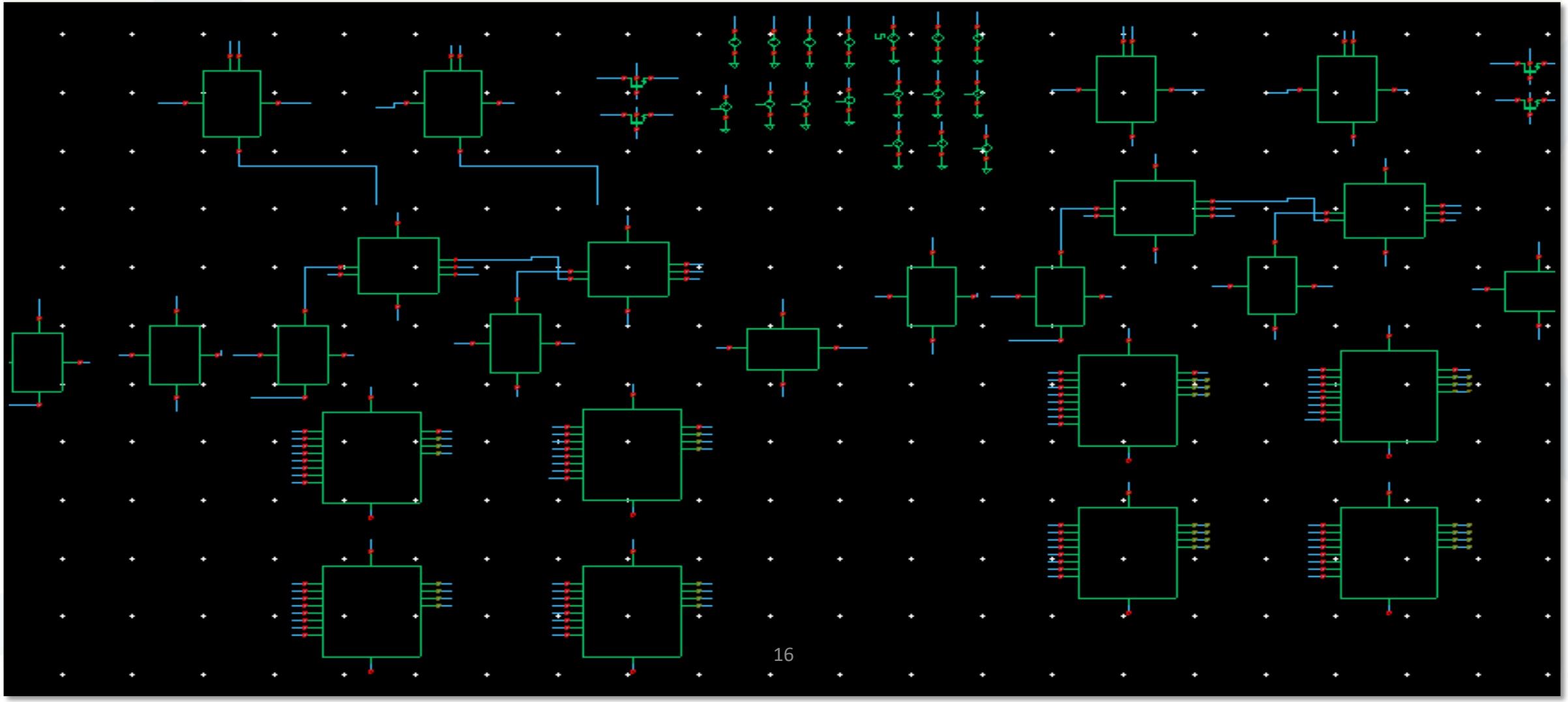
- The advantages of the proposed structure over the existing IMC architecture are given as follows:

1) The multicycle input activation scheme reuses logic units, which reduces the overhead.

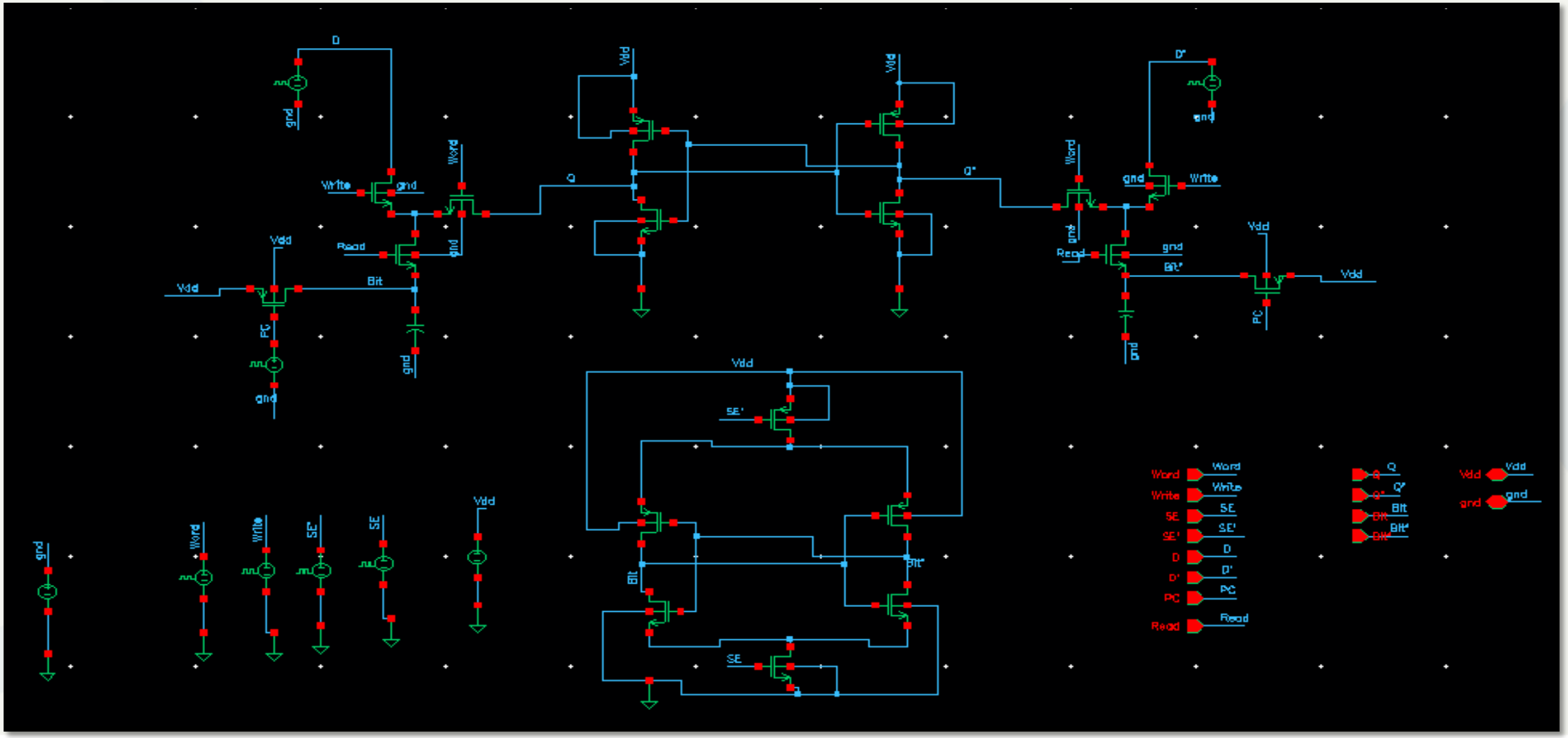
2) Internal write-back is implemented in the IMC unit (IMCU), which improves the efficiency. Multiplication results are stored locally so that there is no need to read the results immediately.

3) The idea provided by this scheme can be easily extended to multiplications with different bit widths, which allows good scalability.

Schematic of 4x4 bit Multiplier



PROPOSED WORK- Modified 6T-SRAM



Novelty

- We have implemented a novel 6T SRAM using Sense Amplifiers along with 3 extra NMOS transistors for Read, Write and Pre-Charge respectively along with Wordline and Bitline transistor to improve the read and write operations by restricting the effect of pre-charged bitline on the stored value Q and vice-versa.

SIMULATION RESULTS AND COMPARISIONS

Cadence Virtuoso circuit Schematics, Graphs and Comparison Tables

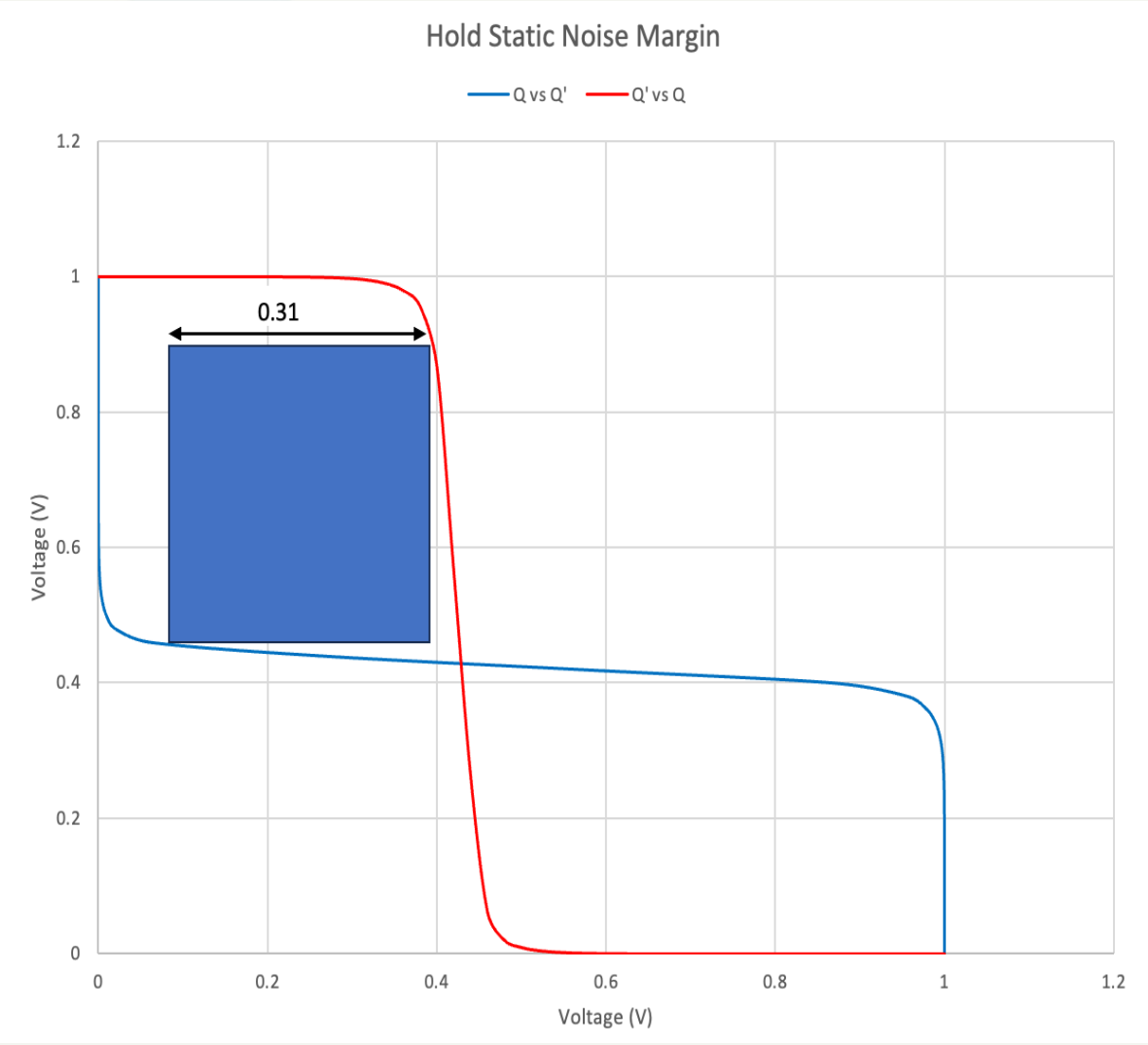
Static Noise Margin in 6T-SRAM

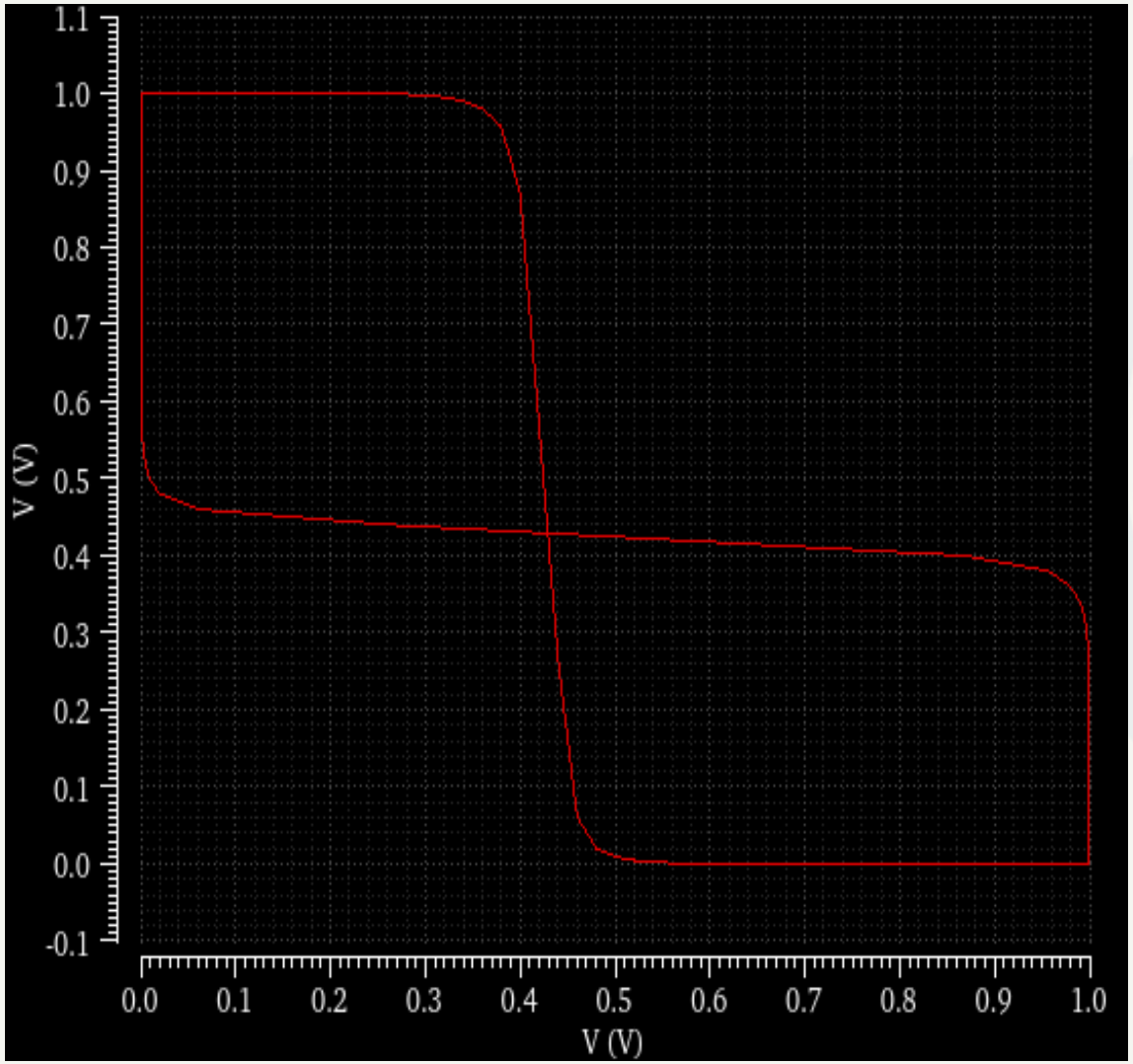
- The Static Noise Margin (SNM) in Static Random Access Memory (SRAM) is a measure of the stability of the stored data. It represents the voltage margin between the '0' and '1' states, ensuring reliable operation despite variations in process, temperature, and voltage. A larger SNM indicates a more robust memory cell design.

- In SRAM, the SNM is typically evaluated through simulations or measurements. It's influenced by various factors including transistor sizing, layout, voltage levels, and manufacturing process variations. Designers aim to maximize SNM to ensure reliable operation under worst-case conditions.

- SNM is crucial for SRAM reliability, especially in high-performance and low-power applications where data integrity is paramount. It's often a key consideration in SRAM design and optimization.

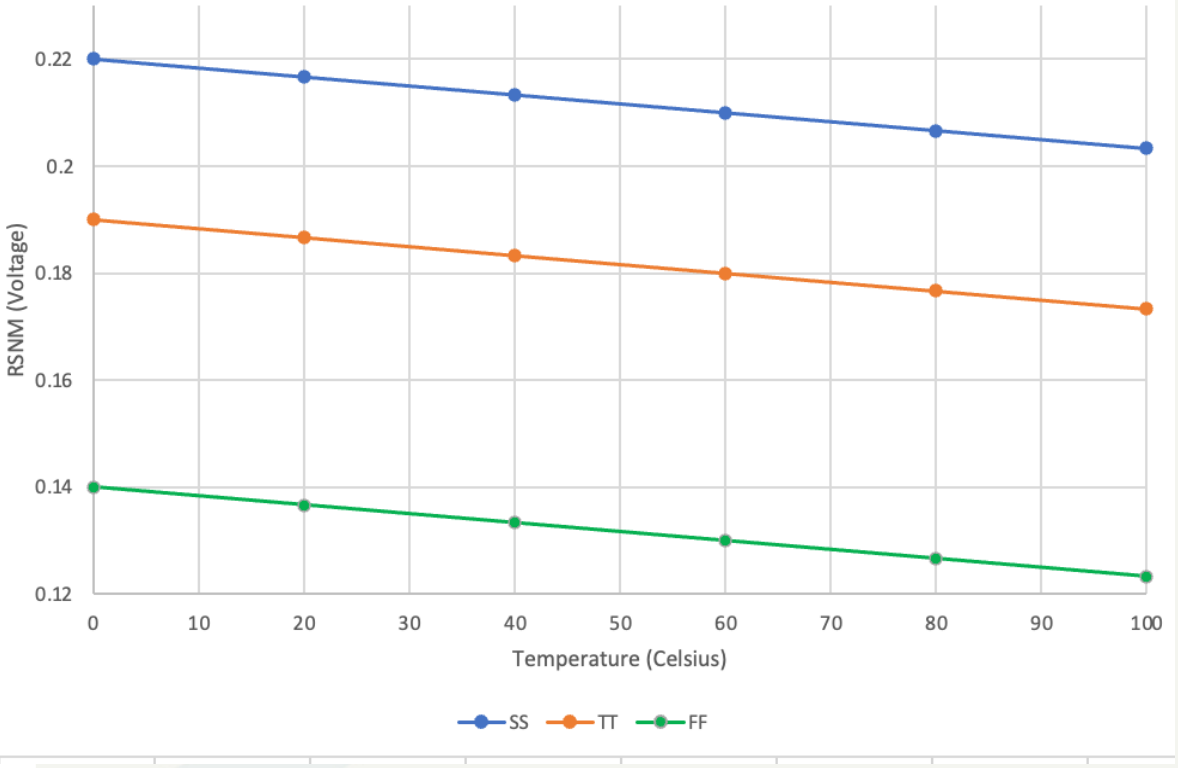
Calculation



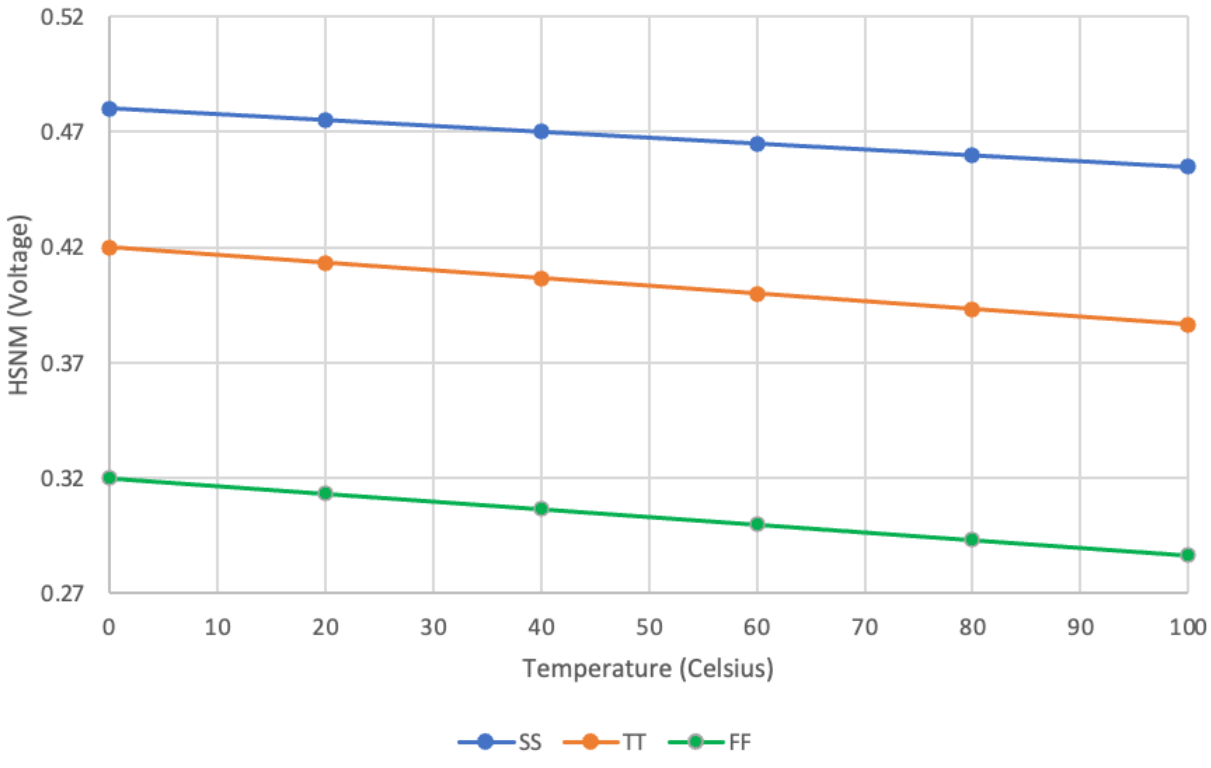


Noise margin Results for different temperature

RSNM



HSNM



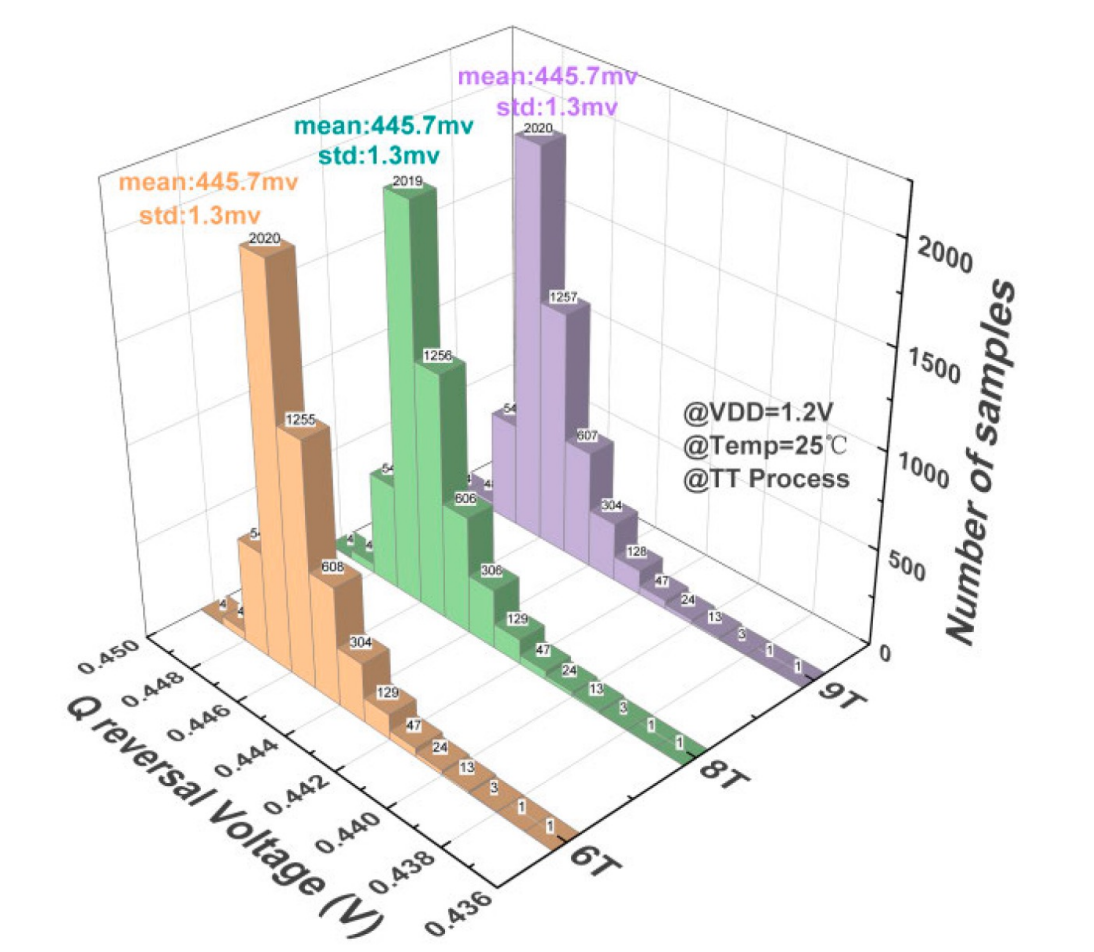
Monte carlo Simulation

1. In the context of Cadence, Monte Carlo simulation refers to a simulation technique used in electronic design automation (EDA) to analyze the performance of analog and mixed-signal circuits. Cadence is a leading provider of EDA software tools commonly used in the design of integrated circuits (ICs) and electronic systems.

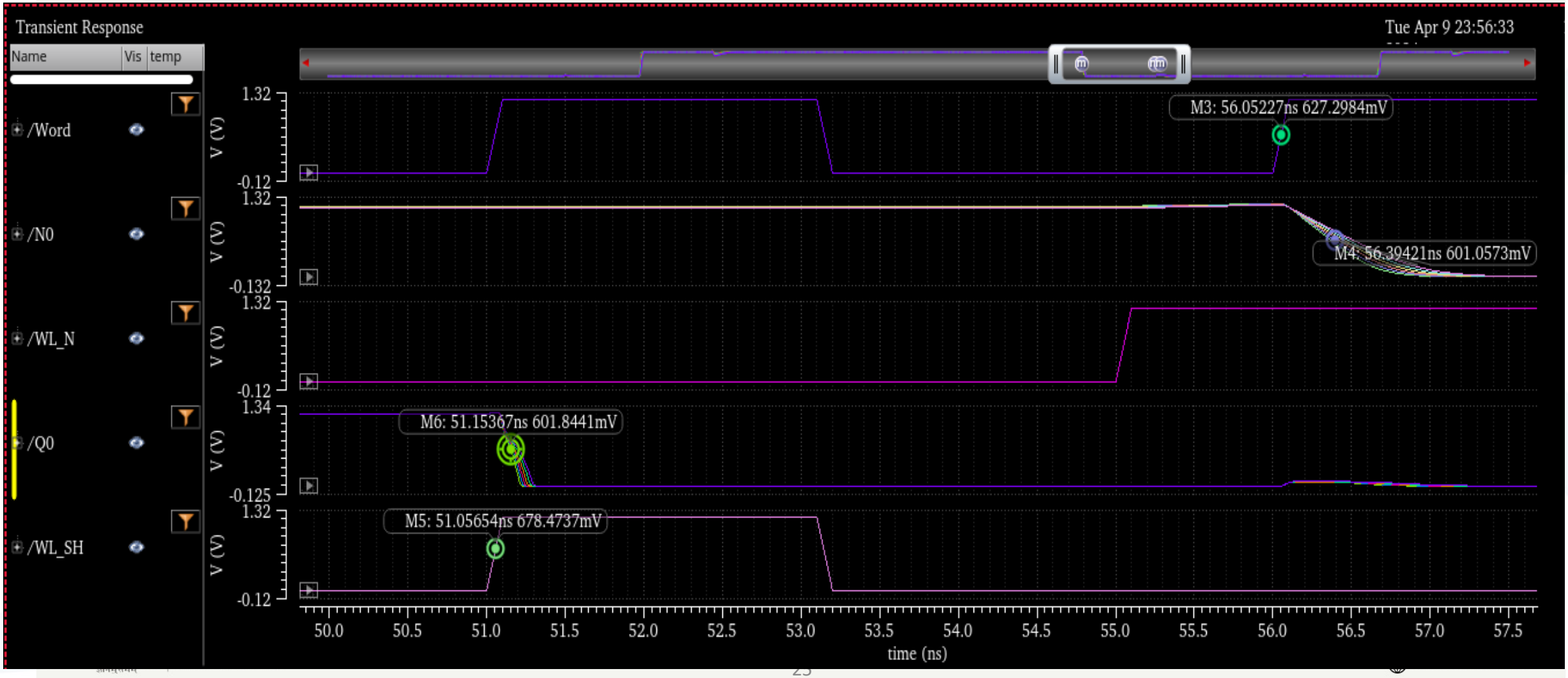
2. Monte Carlo simulation in Cadence involves varying component values, process parameters, and environmental conditions within specified ranges according to their statistical distributions. By repeatedly simulating the circuit with these randomized parameters, Monte Carlo analysis helps designers understand how variations in these factors affect the circuit's behavior and performance metrics such as voltage, current, and timing.

3. This analysis is crucial for assessing the robustness of circuit designs against manufacturing variations, environmental changes, and component tolerances. It helps designers identify potential yield issues, optimize circuit performance, and make informed design decisions to meet specifications under real-world operating conditions.

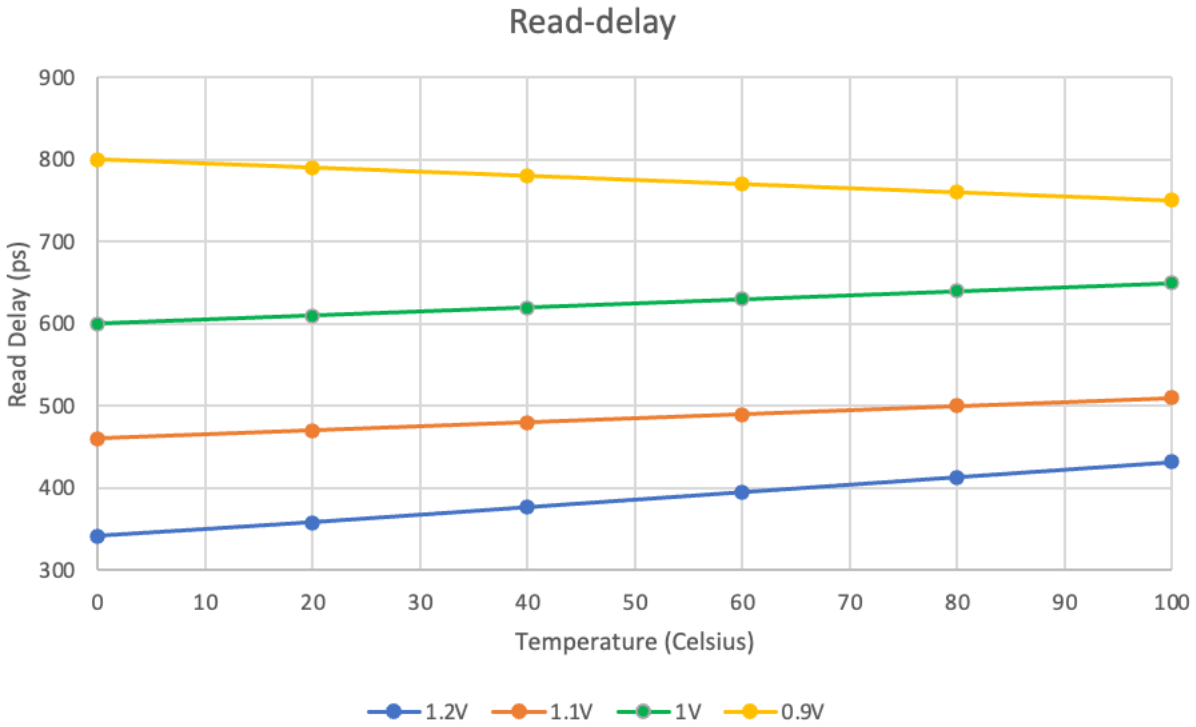
Monte carlo Simulation

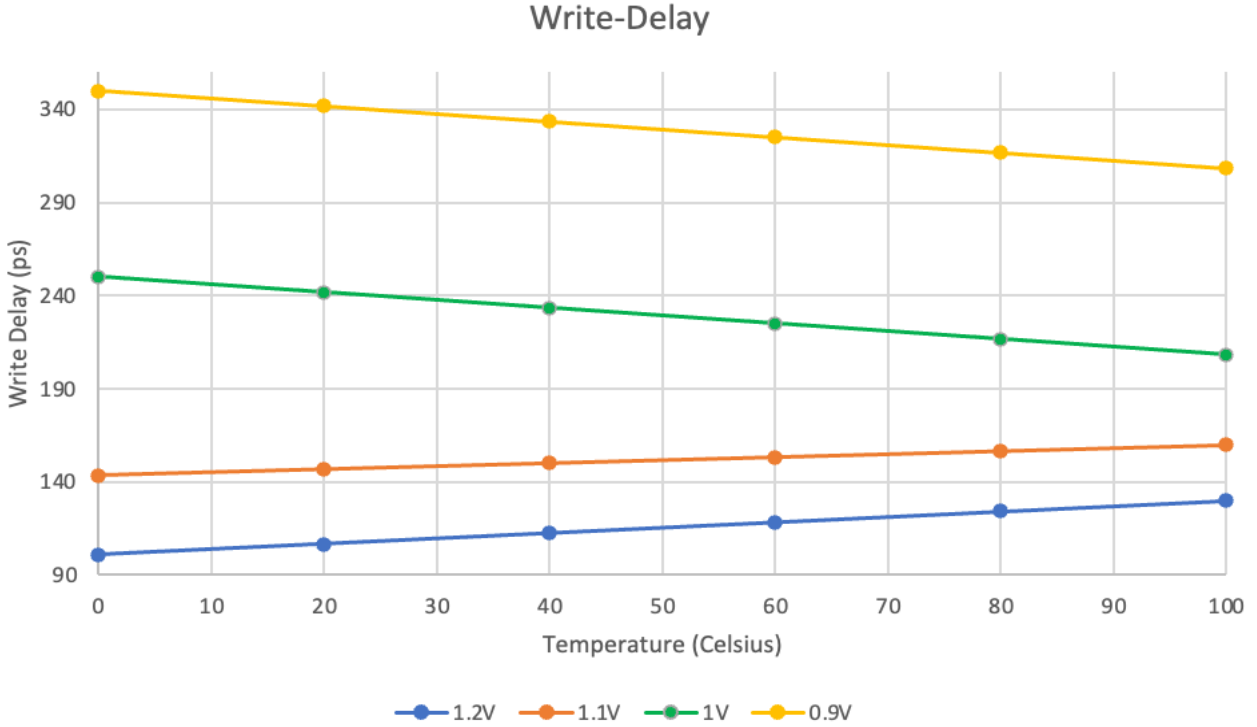


Write and Read delay with varying temperatures

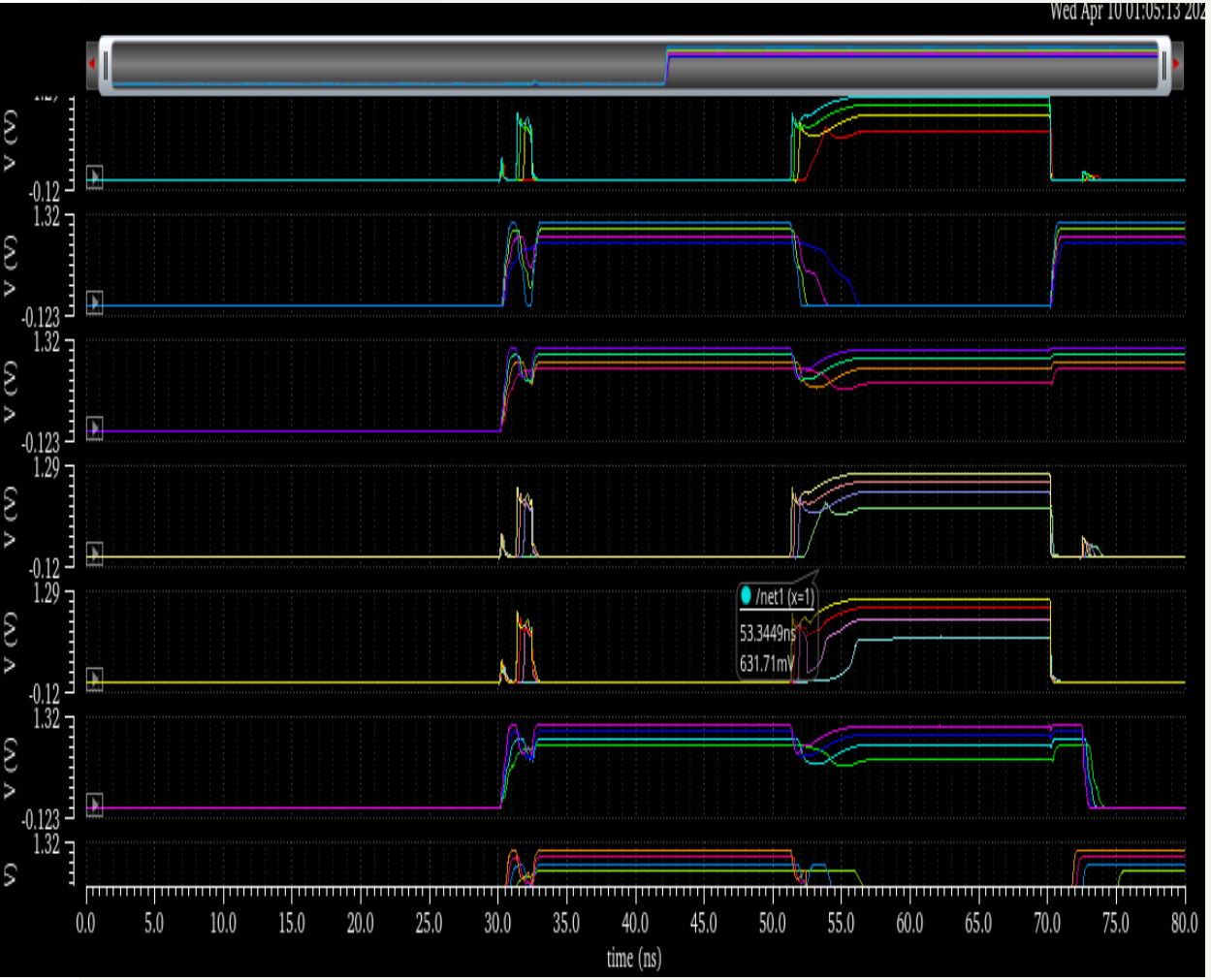


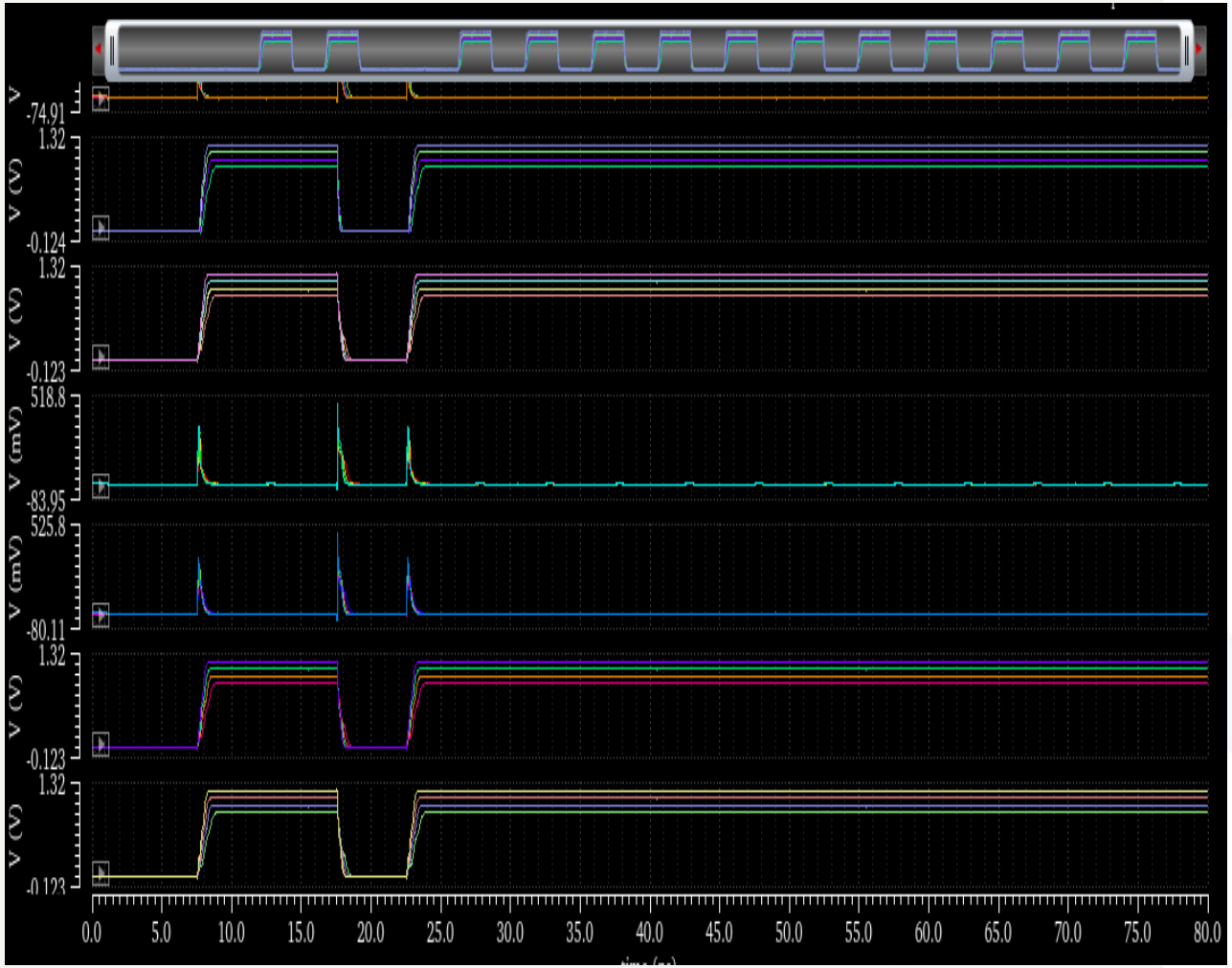
Write and Read Delay



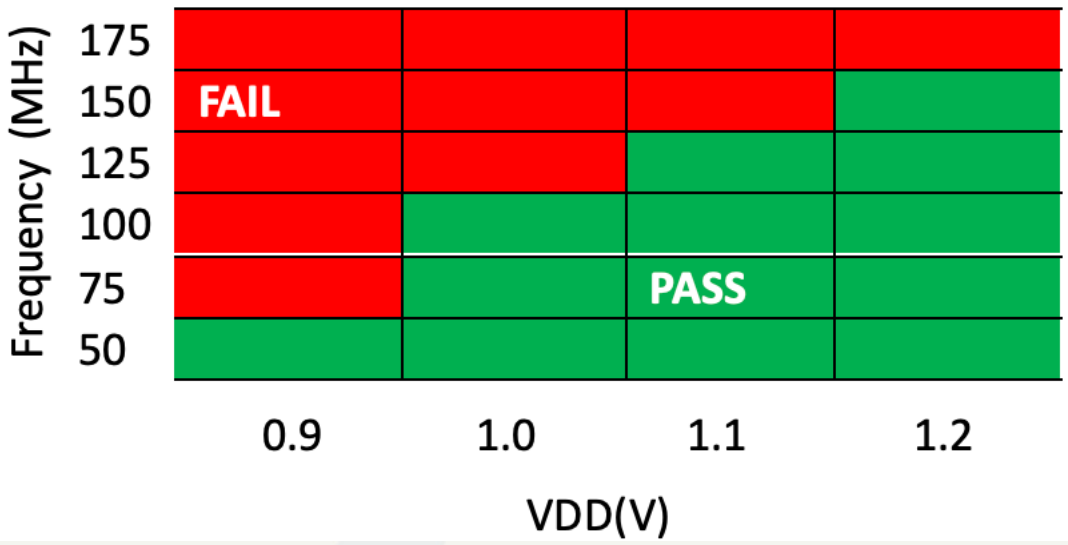


Results of Read/Write operations for different voltages at same frequency

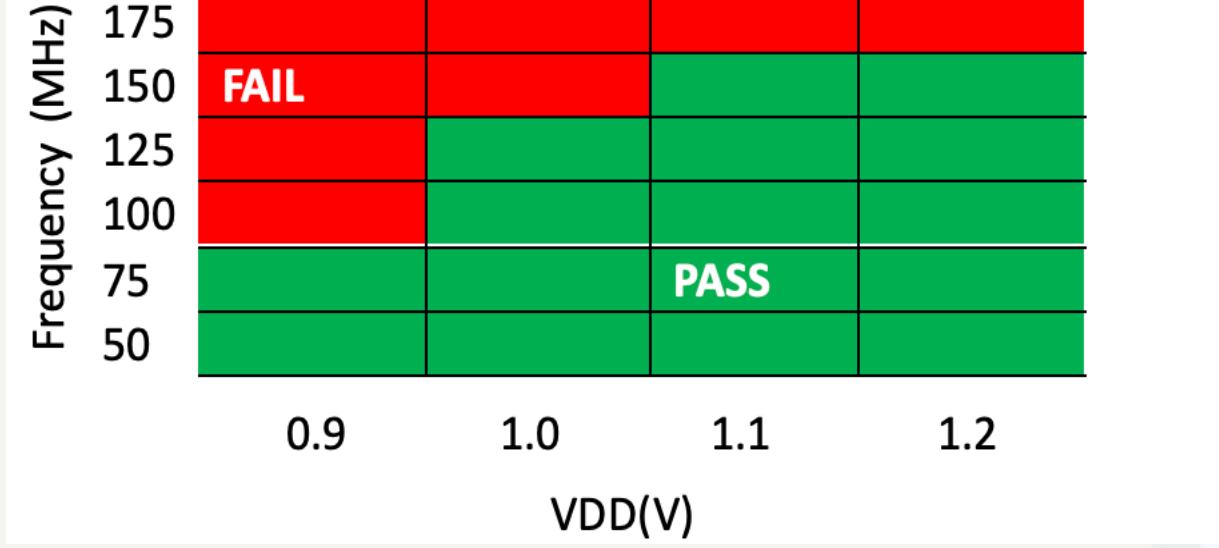




Results of (a) read/write and (b) calculation operations at different supply voltages and frequencies.

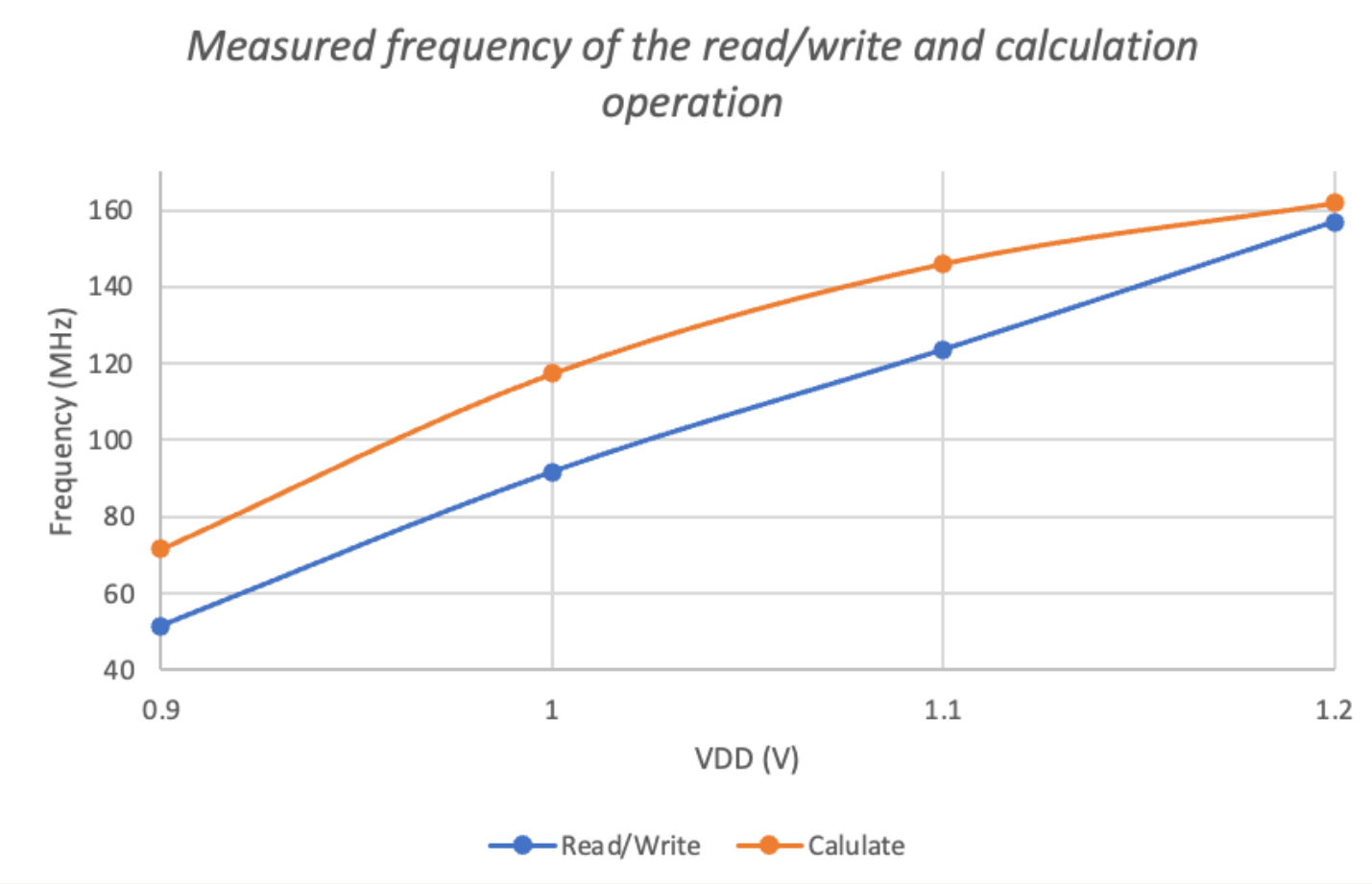


(a)

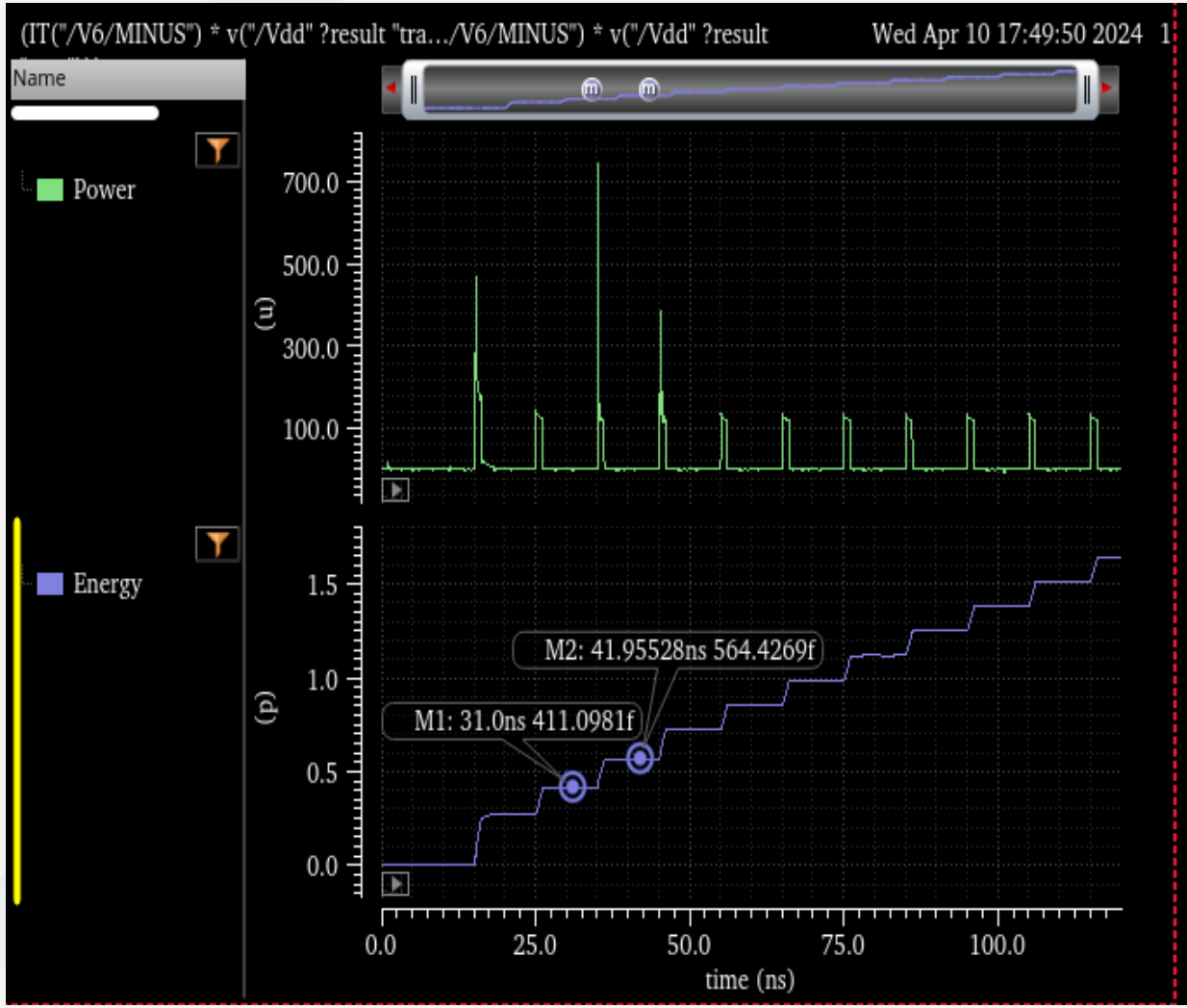


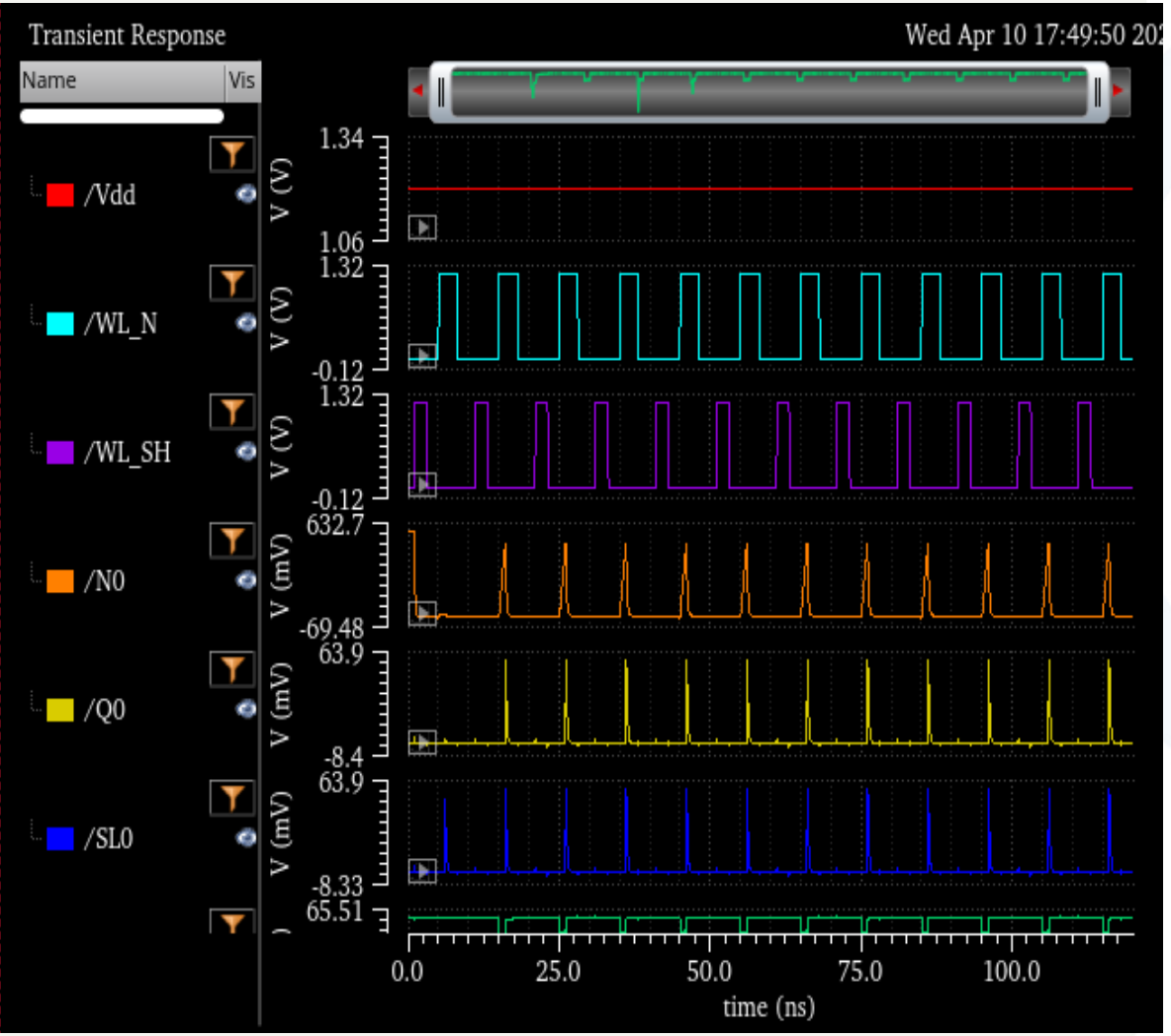
(b)

Measured frequency of the read/write and calculation operations.

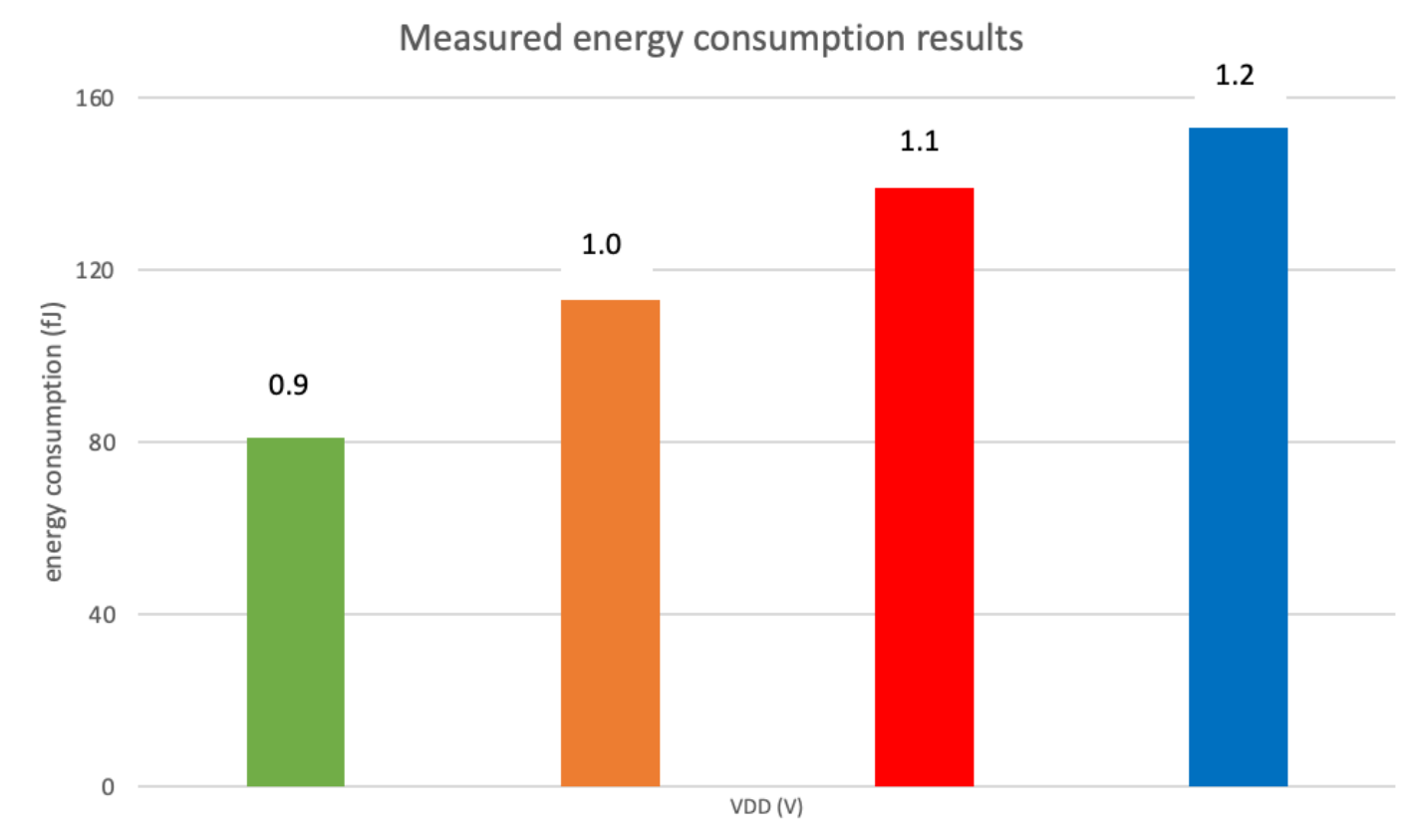


Energy Consumption

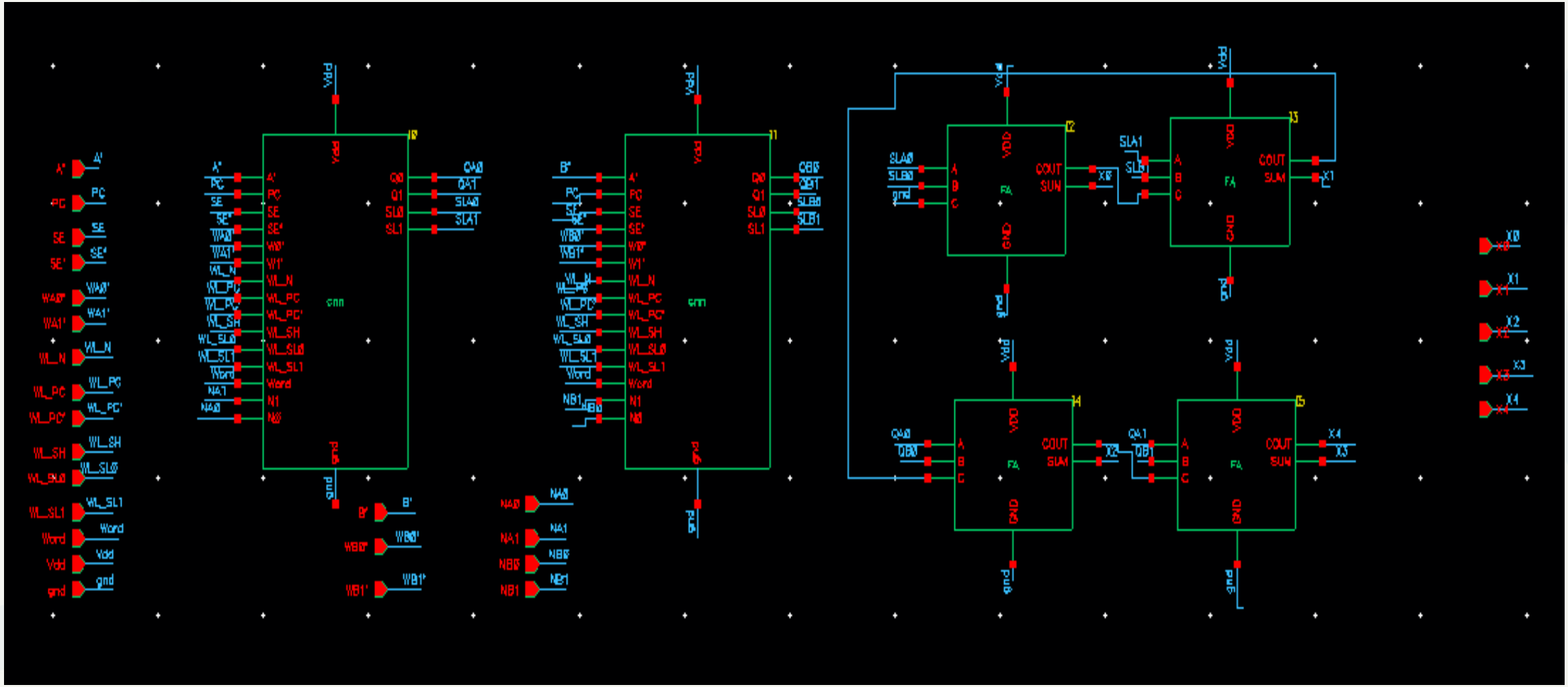




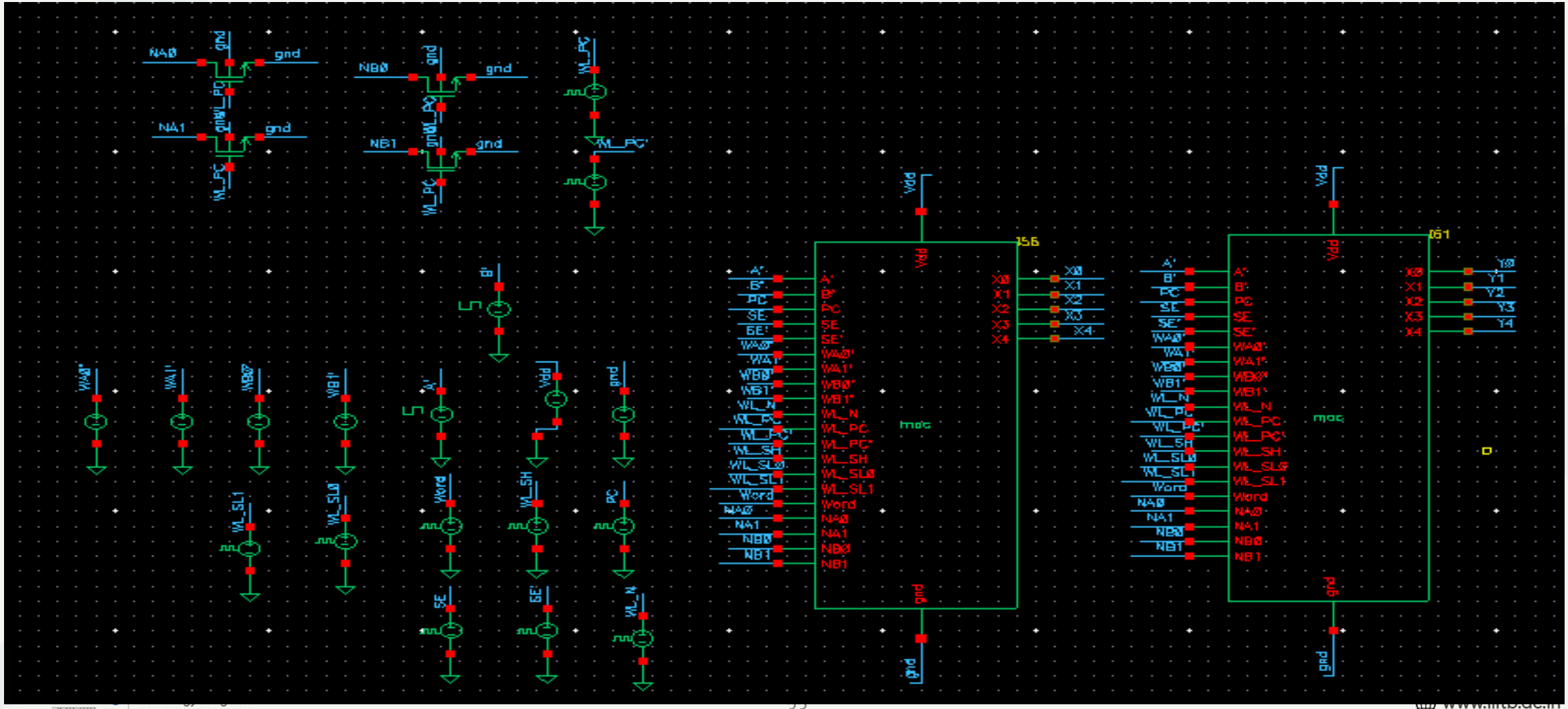
Energy consumption with varying voltage



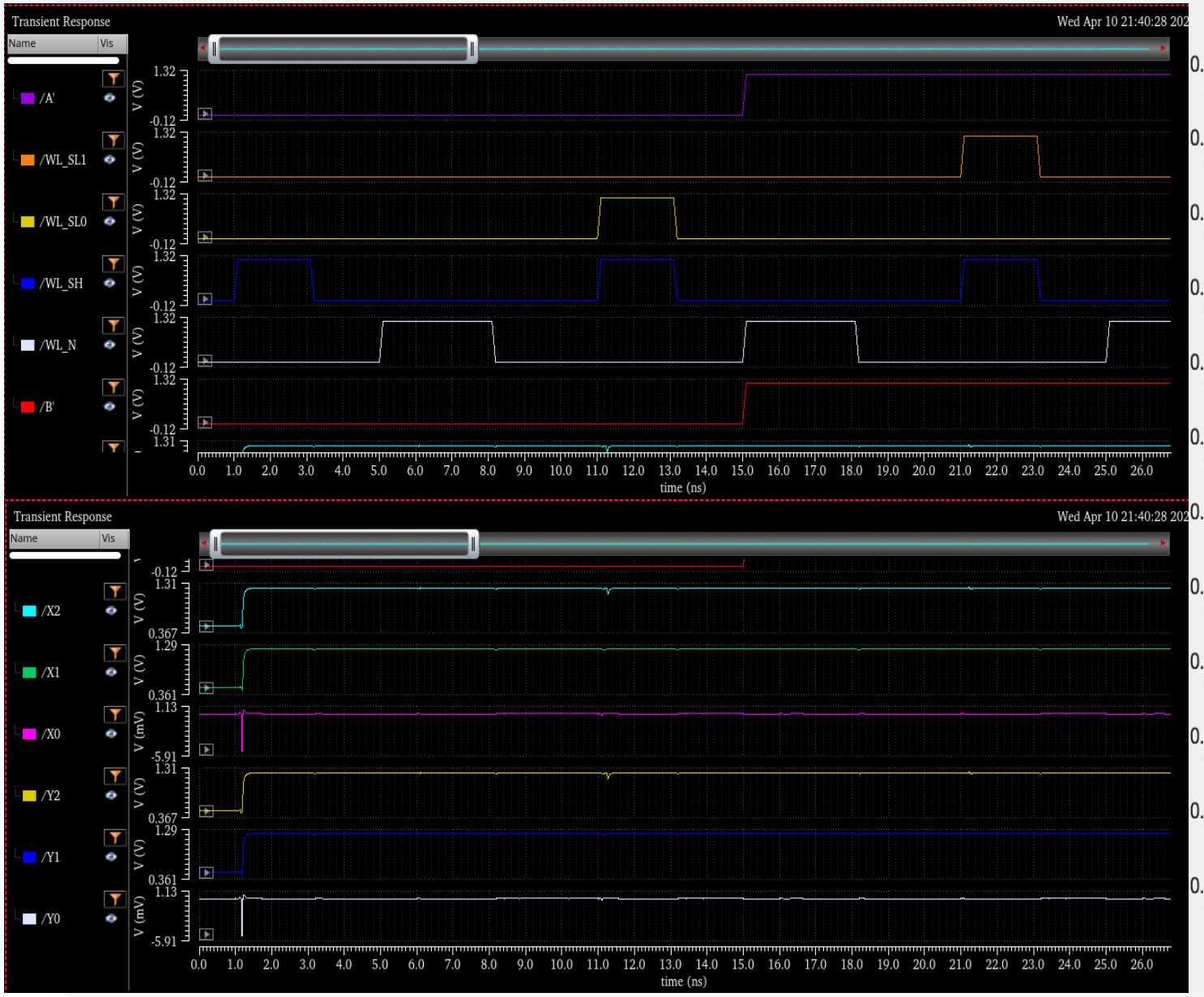
Application of proposed multiplier – 4 bit Single layer Neural Network

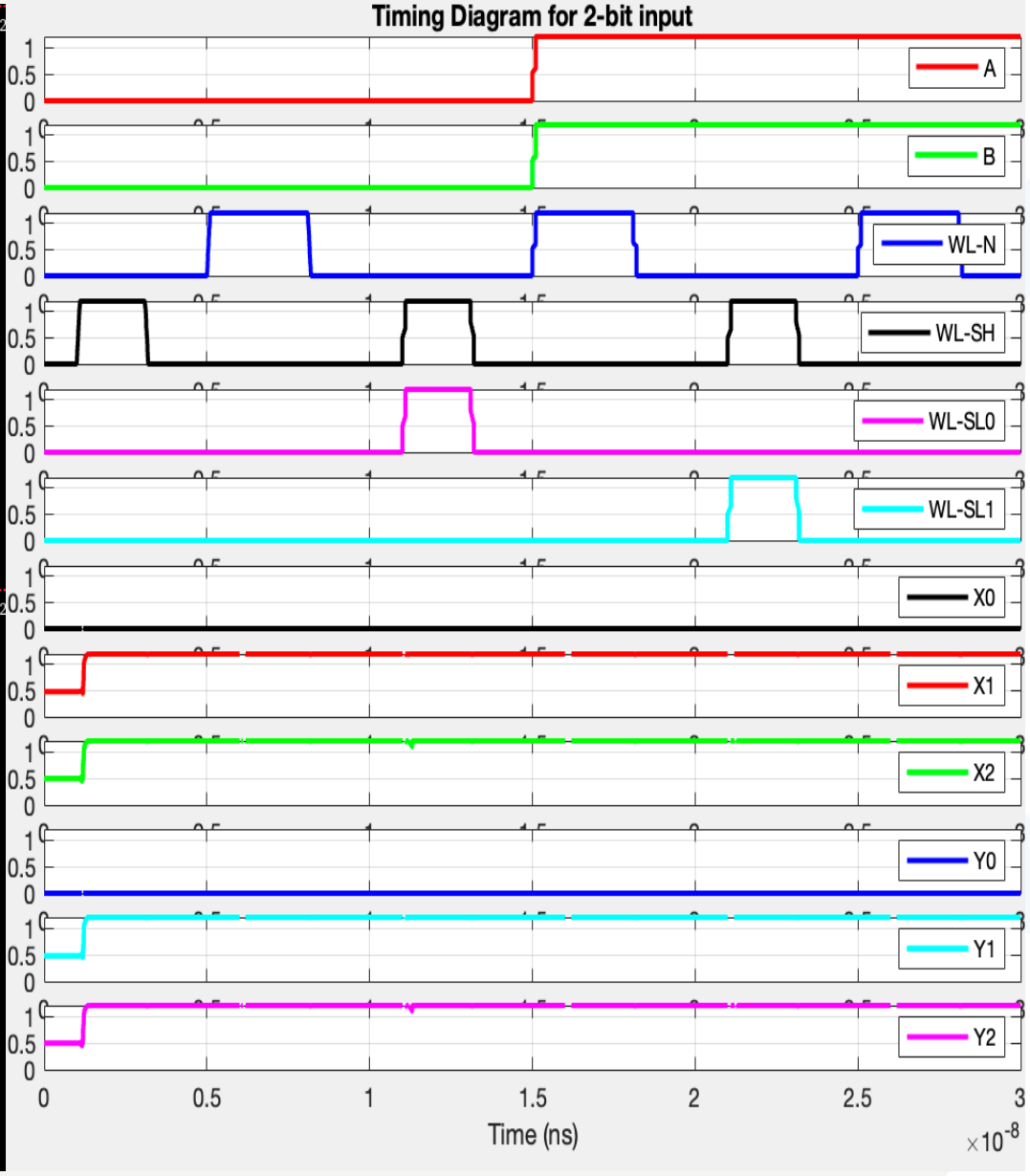


Application of proposed multiplier – 4 bit Single layer Neural Network



4 bit Single layer Neural Network





FUTURE WORK

- Plan to implement In-memory Homomorphic Encryption
- Plan to extend to 4-layer Neural Network for 128 bit input

REFERENCES

1. Static noise margin of 6T and 8T SRAM Cell in 28-nm CMOS Riya Pateliyaa Vishwakarma Government Engineering college, Ahmedabad-Gujarat, India
2. High-Speed Hybrid-Logic Full Adder Using High-Performance 10-T XOR–XNOR Cell
3. K. Kim, H. Jeong, J. Park, and S.-O. Jung, “Transient cell supply voltage collapse write assist using charge redistribution,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 63, no. 10, pp. 964–968, Oct. 2016.
4. J. Chen, W. Zhao, and Y. Ha, “Area-efficient distributed arithmetic optimization via heuristic decomposition and in-memroy computing,”in Proc. IEEE 13th Int. Conf. ASIC (ASICON), Chongqing, China,Oct. 2019, pp. 1–
5. B. Moons and M. Verhelst, “An energy-efficient precision-scalable ConvNet processor in 40-nm CMOS,” IEEE J. Solid-State Circuits, vol. 52, no. 4, pp. 903–914, Apr. 2017.

International Institute of Information Technology Bangalore

26/C, Electronics City, Hosur Road,

Bengaluru – 560 100, Karnataka, India

www.iiitb.ac.in

https://www.facebook.com/IIITBofficial/

https://www.linkedin.com/school/iiit-bangalore/

https://www.instagram.com/iiitb_official/

https://twitter.com/IIITB_official

https://www.youtube.com/user/iiitbmedia